

ON DIFFERENTIAL EVASION GAMES*

JIONGMIN YONG†

Abstract. The notions of evadability and strict evadability are defined for a nonlinear differential game with general closed convex terminal set. Sufficient conditions for strict evadability and evadability and a necessary condition for evadability are proved. The results are strengthened for the linear cases, including the heretofore untreated case in which the terminal set is an $(n-1)$ -dimensional subspace.

Key words. differential evasion games, evadability, strict evadability

AMS(MOS) subject classifications. 90D25, 90D26

1. Introduction. The problem of differential evasion games was first formulated and discussed by Pontryagin and Miščenko [17]. Since then, this problem has been studied extensively by many authors in various investigations [2]–[6], [8]–[18], [20] and [22]–[25].

The purpose of this paper is to complement, in some aspects, the known results about differential evasion games. We now state the evasion problem and some known results.

Suppose the differential game is governed by the following system:

$$(1.1) \quad \begin{aligned} \dot{z} &= f(t, z, u, v), \\ z(0) &= z_0 \end{aligned}$$

where $z \in \mathbb{R}^n$, $u \in U \subseteq \mathbb{R}^p$, $v \in V \subseteq \mathbb{R}^q$, $t \in \mathbb{R}^+ \equiv [0, \infty)$, $f: \mathbb{R}^+ \times \mathbb{R}^n \times U \times V \rightarrow \mathbb{R}^n$ is a given map and z_0 is the initial state. In the game, u is the pursuit control and v is the evasion control. The sets U and V are compact. Suppose that we are also given a subset M of \mathbb{R}^n . The game is terminated if, at some time $t_0 < \infty$, the trajectory $z(\cdot)$ of (1.1), corresponding to a pair of controls $u(\cdot)$ and $v(\cdot)$, satisfies

$$(1.2) \quad z(t_0) \in M.$$

Naturally, the set M is called the terminal set of the game. The goal of the pursuer is to terminate the game by choosing a suitable control $u(\cdot)$, while the goal of the evader is to prevent the game from terminating by choosing a proper control $v(\cdot)$.

We denote

$$\begin{aligned} \mathcal{U} &= \{u: [0, \infty) \rightarrow U \mid u(\cdot) \text{ is measurable}\}, \\ \mathcal{V} &= \{v: [0, \infty) \rightarrow V \mid v(\cdot) \text{ is measurable}\}. \end{aligned}$$

By admissible controls $u(\cdot)$ and $v(\cdot)$ we always mean that $u(\cdot) \in \mathcal{U}$ and $v(\cdot) \in \mathcal{V}$. Any $u(\cdot) \in \mathcal{U}$ is called a pursuit control and $v(\cdot) \in \mathcal{V}$ an evasion control.

In an evasion game, the evader is regarded as the “primary” player, i.e., the evader has some advantages in choosing his controls. The pursuer chooses his control $u(\cdot) \in \mathcal{U}$ at the start of the game. The evader, on the other hand, chooses the values of his control as the game evolves and can use $\{z(s), u(s) \mid 0 \leq s \leq t\}$ when he chooses the value $v(t)$ of the evasion control $v(\cdot) \in \mathcal{V}$ at time t . We denote $\hat{\mathcal{V}}$ to be the set of all such evasion controls.

We now give some definitions.

* Received by the editors August 26, 1985; accepted for publication (in revised form) December 23, 1986.

† Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

DEFINITION 1.1. Game (1.1) with terminal set M is said to be evadable if, for any $z_0 \notin M$ and any given pursuit control $u(\cdot) \in \mathcal{U}$, there exists an evasion control $v(\cdot) \in \mathcal{V}$ such that

$$(1.3) \quad d(M, z(t)) > 0 \quad \forall t \geq 0.$$

Here, $z(\cdot)$ is the corresponding trajectory of (1.1) resulting from $u(\cdot)$ and $v(\cdot)$, and $d(\cdot, \cdot)$ is the Euclidean distance in \mathbb{R}^n .

DEFINITION 1.2. Game (1.1) with terminal set M is said to be strictly evadable if, for any $z_0 \notin M$, there exists a $\delta(z_0) > 0$ such that for any $u(\cdot) \in \mathcal{U}$ we can find $v(\cdot) \in \mathcal{V}$ so that the corresponding trajectory $z(\cdot)$ of (1.1) satisfies

$$d(M, z(t)) \geq \delta(z_0) > 0 \quad \forall t \geq 0.$$

In [2]–[6], [9]–[18], [20] and [22]–[25], the terminal set M was assumed to be a linear subspace of \mathbb{R}^n with $\dim M \leq n-2$. For various forms of $f(t, z, u, v)$ in (1.1), some sufficient conditions for evadability of the game were established in [2], [3], [9]–[18], [20] and [22]–[25]. Some sufficient conditions for strict evadability of the game were given in [4]–[6], where the strict evadability was called the possibility of l -escape. In [8], a strict evadability result was given for the game governed by (1.1) with $f(t, z, u, v) \equiv f(z, u, v)$ and having the terminal set M not a linear subspace, but with some unpleasant restrictions.

In this paper, we generalize the results of [8] by relaxing the restrictions on M . We get some sufficient conditions for the strict evadability and evadability of game (1.1) with general closed terminal set M by using a Lyapunov type method. It turns out that in some cases, our sufficient condition for evadability is very close to a necessary condition for the evadability of the game. As consequences of our main results, we get some conditions for the (strict) evadability of the game (1.1) with f linear in z and with the terminal set M an $(n-1)$ -dimensional subspace of \mathbb{R}^n . This case has not been discussed heretofore.

For $0 \leq a < b \leq +\infty$, we denote

$$\mathcal{U}[a, b] = \{u \text{ restricted to } [a, b] | u \in \mathcal{U}\}.$$

By $u(\cdot) \in \mathcal{U}[a, b]$, we emphasize the control is restricted to the specified interval $[a, b]$. Similarly, we can define $\mathcal{V}[a, b]$, $\hat{\mathcal{V}}[a, b]$, etc. Also, in the following, when $u(\cdot)$, $v(\cdot)$ are given, $z(\cdot)$ always means the trajectory corresponding to $u(\cdot)$ and $v(\cdot)$.

2. Sufficient conditions for strict evadability and evadability. In this section we are going to prove the main results of this paper. Let us start with some general assumptions.

(A1) $f(t, z, u, v)$ is continuous in $(t, z, u, v) \in \mathbb{R}^+ \times \mathbb{R}^n \times U \times V$. There exists a $K > 0$ such that

$$(2.1) \quad \|f(t, z, u, v) - f(t, \hat{z}, u, v)\| \leq K \|z - \hat{z}\|$$

for all $z, \hat{z} \in \mathbb{R}^n$ and $(t, u, v) \in \mathbb{R}^+ \times U \times V$. $U \subseteq \mathbb{R}^p, V \subseteq \mathbb{R}^q$ are compact.

(A2) $M \subset \mathbb{R}^n$ is closed.

(A2') $M \subset \mathbb{R}^n$ is convex and closed.

Remark 2.1. From (A1) we have

$$(2.2) \quad (z, f(t, z, u, v)) \leq K_1(t)(1 + \|z\|^2)$$

for all $(t, z, u, v) \in \mathbb{R}^+ \times \mathbb{R}^n \times U \times V$, where

$$(2.3) \quad K_1(t) = K + \max_{(u, v) \in U \times V} \|f(t, 0, u, v)\| \in L^1_{\text{loc}}[0, \infty)$$

and (\cdot, \cdot) is the usual inner product in \mathbb{R}^n .

We denote, for $0 \leq s < \delta \leq +\infty$, that

$$(2.4) \quad T_s^\delta(M) \triangleq \{z \in \mathbb{R}^n \mid s < d(M, z) < \delta\}.$$

THEOREM 2.2. *Suppose game (1.1) and terminal set M are given and (A1) and (A2) hold. Let $V(\cdot): \mathbb{R}^n \rightarrow \mathbb{R}^+$ be a continuous function with the following properties:*

(1) *There exist strictly increasing functions $\alpha(\cdot), \beta(\cdot): \mathbb{R}^+ \rightarrow \mathbb{R}^+$, with $\alpha(0) = \beta(0) = 0$ such that*

$$(2.5) \quad \alpha(d(M, z)) \leq V(z) \leq \beta(d(M, z)) \quad \forall z \in \mathbb{R}^n.$$

(2) *There exists a $\delta > 0$ such that $(\partial/\partial z)V(z)$ exists in $T_0^\delta(M)$, and*

$$(2.6) \quad \inf_{t \geq 0} \inf_{u \in U} \max_{v \in V} \inf_{z \in T_0^\delta(M)} \left(\frac{\partial}{\partial z} V(z), f(t, z, u, v) \right) \geq 0.$$

Then the game is strictly evadable.

Proof. Suppose $z_0 \notin M$ and $u(\cdot) \in \mathcal{U}[0, \infty)$. By Filippov's lemma and (2.6), we can find $\hat{v}(\cdot) \in \hat{\mathcal{V}}[0, \infty)$ such that

$$(2.7) \quad \inf_{t \geq 0} \inf_{z \in T_0^\delta(M)} \left(\frac{\partial}{\partial z} V(z), f(t, z, u(t), \hat{v}(t)) \right) \geq 0.$$

We claim that by using such an evasion control we have

$$(2.8) \quad d(M, z(t)) \geq \beta^{-1} \circ \alpha(\delta \wedge d(M, z_0)) \triangleq \eta_1 > 0 \quad \forall t \geq 0$$

where β^{-1} is the inverse function of β and $\delta \wedge d(M, z_0) = \min\{\delta, d(M, z_0)\}$. In fact, if (2.8) is not true, then there exist $0 \leq t_1 < t_2$ such that

$$(2.9) \quad d(M, z(t_2)) < \eta_1,$$

$$(2.10) \quad 0 < d(M, z(t)) < \delta, \quad t \in (t_1, t_2],$$

$$(2.11) \quad d(M, z(t_1)) = \delta \wedge d(M, z_0).$$

Here we should know that since $\alpha(s) \leq \beta(s)$, for any $s \in \mathbb{R}^+$,

$$(2.12) \quad \eta_1 = \beta^{-1} \circ \alpha(\delta \wedge d(M, z_0)) \leq \delta \wedge d(M, z_0) \leq \delta.$$

Thus, for $t \in (t_1, t_2]$, we have

$$(2.13) \quad \frac{d}{dt} V(z(t)) = \left(\frac{\partial}{\partial z} V(z(t)), f(t, z(t), u(t), \hat{v}(t)) \right) \geq 0.$$

Integrating on $(t_1, t_2]$, we get

$$(2.14) \quad \beta(d(M, z(t_2))) \geq V(z(t_2)) \geq V(z(t_1)) = \alpha(\delta \wedge d(M, z_0)).$$

This contradicts (2.9). Hence (2.8) holds and our theorem follows. \square

Roughly speaking, the geometric meaning of (2.6) is that whenever the state z of the game gets close to the terminal set M , the evader has enough power to force the velocity of the state towards the direction of leaving M , so that he can prevent the state from getting too close to the terminal set. In other words, (2.6) means that the evader is more (at least not less) powerful than the pursuer when the state is near M .

We note that (2.6) allows us to find, via (2.7), a successful evasion control $\hat{v}(\cdot)$ which only depends on $u(\cdot)$. However, (2.6) seems very restrictive. Thus, we now try to relax it, but in order to do so, we need a little bit more about $V(\cdot)$.

THEOREM 2.3. *Suppose game (1.1) and terminal set M are given satisfying (A1) and (A2). Let $V(\cdot): \mathbb{R}^n \rightarrow \mathbb{R}^+$ be a continuous function satisfying (1) given in Theorem 2.2 and the following:*

(2') There exist constants $\delta, \delta_0, c > 0$, and a continuous function $w(\cdot, \cdot) : (0, \delta] \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$ with the property that for all $s \in (0, \delta]$, $w(s, \cdot)$ is strictly increasing with $w(s, 0) = 0$ such that $(\partial/\partial z)V(z)$ exists for all $z \in T_0^\delta(M)$, and

$$(2.15) \quad \left\| \frac{\partial}{\partial z} V(z) \right\| \leq c \quad \forall z \in T_0^\delta(M),$$

$$(2.16) \quad \left\| \frac{\partial}{\partial z} V(z) - \frac{\partial}{\partial z} V(\hat{z}) \right\| \leq w(s, \|z - \hat{z}\|) \quad \forall z, \hat{z} \in T_s^\delta(M),$$

$$(2.17) \quad \inf_{t \geq 0} \inf_{z \in T_0^\delta(M)} \min_{u \in U} \max_{v \in V} \left(\frac{\partial}{\partial z} V(z), f(t, z, u, v) \right) \geq \delta_0.$$

Then the game is strictly evadable.

Remark 2.4. We note that, in general, the left-hand side of (2.17) is greater than or equal to that of (2.6). Therefore, in general, Theorems 2.2 and 2.3 do not contain each other. However, if we have

$$(2.18) \quad f(t, z, u, v) = h(t, z) + g(t, u, v),$$

then the left-hand sides of (2.6) and (2.17) are equal and then Theorem 2.3 is a trivial consequence of Theorem 2.2. Now, in the general case, the difficulty of proving Theorem 2.3 is that we cannot define an evasion control similar to that in the proof of Theorem 2.2, since if we did so, we would get a control of form $v(t, z)$ which is only measurable in z . Then we would meet some trouble in getting the existence of the trajectory $z(\cdot)$ of (1.1). That is why we need more conditions on $V(\cdot)$ and we need “ $\geq \delta_0$ ” in (2.17).

Proof of Theorem 2.3. Suppose $u(\cdot) \in \mathcal{U}[0, \infty)$ is given and $0 < d(M, z_0) < \delta$. We claim that there exist $T^* \geq 1$ and $\hat{v}(\cdot) \in \hat{\mathcal{V}}[0, T^*]$ such that

$$(2.19) \quad d(M, z(t)) \geq \eta_0 \triangleq \beta^{-1} \circ \alpha \left(\frac{\delta}{2} \wedge d(M, z_0) \right), \quad t \in [0, T^*],$$

$$(2.20) \quad d(M, z(T^*)) \geq \delta.$$

To prove our claim, we first prove the following.

LEMMA 2.5. There exist

$$0 < t^* \leq T \triangleq \frac{2(\beta(\delta) - \alpha(\delta/2 \wedge d(M, z_0)))}{\delta_0}$$

and $\hat{v}(\cdot) \in \hat{\mathcal{V}}[0, t^*]$ such that

$$(2.21) \quad d(M, z(t)) \geq \eta_0, \quad t \in [0, t^*],$$

$$(2.22) \quad d(M, z(t^*)) \geq \delta.$$

Note here that, in general, we cannot assure $t^* \geq 1$.

To prove this lemma, by (2.2), we can let $F \geq \delta/2$ such that

$$(2.23) \quad \|f(t, z(t), u(t), v(t))\| \leq F, \quad t \in [0, T+2]$$

for all possible $u(\cdot) \in \mathcal{U}$, $v(\cdot) \in \mathcal{V}$, and corresponding trajectory $z(\cdot)$ of (1.1) starting from z_0 . Also, we let \hat{t} be such that

$$(2.24) \quad w(s, Ft)F + cKFt \leq \frac{\delta_0}{2} \quad \forall t \in [0, \hat{t}], \quad s \in \left[\frac{\eta_0}{2}, \delta \right].$$

We then define $T_0 = \hat{t} \wedge (\eta_0/2F)$, which only depends on T and z_0 . Now, since $0 < d(M, z_0) < \delta$, we can choose $\hat{v}(\cdot) \in \hat{\mathcal{V}}$ such that

$$(2.25) \quad \left(\frac{\partial}{\partial z} V(z_0), f(t, z_0, u(t), \hat{v}(t)) \right) \geq \delta_0, \quad t \geq 0.$$

Then for the resulting $z(\cdot)$ we set

$$(2.26) \quad \theta_0 = \inf \{t \in [0, T] \mid d(M, z(t)) \geq \delta\}$$

and let $t_1 = T_0 \wedge \theta_0$. Thus, we have

$$(2.27) \quad \delta > d(M, z(t)) > d(M, z_0) - F \frac{\eta_0}{2F} \geq \frac{\eta_0}{2}, \quad t \in [0, t_1].$$

Hence, for $t \in [0, t_1]$, by using (1.1), (2.1), (2.15), (2.16), and (2.23)–(2.25), we have

$$(2.28) \quad \begin{aligned} \frac{d}{dt} V(z(t)) &= \left(\frac{\partial}{\partial z} V(z(t)), f(t, z(t), u(t), \hat{v}(t)) \right) \\ &\geq \delta_0 - w \left(\frac{\eta_0}{2}, \|z(t) - z_0\| \right) F - cK \|z(t) - z_0\| \\ &\geq \delta_0 - w \left(\frac{\eta_0}{2}, Ft \right) F - cKFt \geq \frac{\delta_0}{2}. \end{aligned}$$

Integrating over $[0, t]$, $t \leq t_1$, we get

$$(2.29) \quad \begin{aligned} d(M, z(t)) &\geq \beta^{-1}(V(z(t))) \geq \beta^{-1} \left(V(z_0) + \frac{\delta_0}{2} t \right) \\ &\geq \beta^{-1} \left(\alpha(d(M, z_0)) + \frac{\delta_0}{2} t \right) > \eta_0. \end{aligned}$$

If $d(M, z(t_1)) \geq \delta$, by taking $t^* = t_1$, we are done. Otherwise, we have $t_1 = T_0$. Then, repeating the above argument replacing z_0 by $z(T_0)$, we can get $\hat{v}(\cdot) \in \hat{\mathcal{V}}[T_0, \infty)$ such that

$$(2.30) \quad \left(\frac{\partial}{\partial z} V(z(T_0)), f(t, z(T_0), u(t), \hat{v}(t)) \right) \geq \delta_0, \quad t \geq T_0.$$

Then we define

$$(2.31) \quad \theta_1 = \inf \{t \in [T_0, T] \mid d(M, z(t)) \geq \delta\},$$

$$(2.32) \quad t_2 = T_0 + T_0 \wedge (\theta_1 - T_0).$$

For $t \in [T_0, t_2]$ we have that

$$(2.33) \quad \begin{aligned} \delta &> d(M, z(t)) > d(M, z(T_0)) - F(t_2 - T_0) \\ &\geq \eta_0 - F \frac{\eta_0}{2F} = \frac{\eta_0}{2}. \end{aligned}$$

Thus, as in (2.28) and (2.29) we have that

$$(2.34) \quad \frac{d}{dt} V(z(t)) \geq \frac{\delta_0}{2}, \quad t \in [T_0, t_2],$$

$$(2.35) \quad d(M, z(t)) \geq \eta_0, \quad t \in [T_0, t_2].$$

By induction, we have for some $i \geq 1$, $0 < t_i < T$, either

$$(2.36) \quad \delta > d(M, z(t)) \geq \eta_0, \quad t \in [0, t_{i-1}),$$

$$(2.37) \quad d(M, z(t)) \geq \eta_0, \quad t \in [t_{i-1}, t_i),$$

$$(2.38) \quad d(M, z(t_i)) \geq \delta$$

(in this case, we take $t^* = t_i$ and the proof is finished) or for all $i \geq 1$, $0 < t_i < T$, (2.36), (2.37) hold and

$$(2.39) \quad \delta > d(M, z(t_i)) \geq \eta_0.$$

In this case, we know that $t_i = iT_0$. Thus, we have

$$(2.40) \quad \delta > d(M, z(t)) \geq \eta_0, \quad t \in [0, T),$$

$$(2.41) \quad \frac{d}{dt} V(z(t)) \geq \frac{\delta_0}{2}, \quad t \in [0, T).$$

Then we have

$$\begin{aligned} d(M, z(T)) &\geq \beta^{-1}(V(z(T))) \geq \beta^{-1}\left(V(z_0) + \frac{\delta_0}{2} T\right) \\ &\geq \beta^{-1}\left(\alpha(d(M, z_0)) + \beta(\delta) - \alpha\left(\frac{\delta}{2} \wedge d(M, z_0)\right)\right) \geq \delta. \end{aligned}$$

Thus, we only need to take $t^* = T$. The lemma is proved.

Now, we go back to the proof of the theorem. It is clear that if $t_0 \in [0, T + 2 - \delta/2F]$, with $d(M, z(t_0)) \geq \delta$, then by taking $\hat{v}(\cdot) \equiv v_0 \in V$, we have

$$(2.42) \quad d(M, z(t)) \geq \frac{\delta}{2}, \quad t \in \left[t_0, t_0 + \frac{\delta}{2F}\right].$$

Now, since $0 < d(M, z_0) < \delta$, by Lemma 2.5 we have $t_0^* \in [0, T]$ and $\hat{v}(\cdot) \in \hat{\mathcal{V}}[0, t_0^*]$ such that

$$(2.43) \quad d(M, z(t)) \geq \eta_0, \quad t \in [0, t_0^*],$$

$$(2.44) \quad d(M, z(t_0^*)) \geq \delta.$$

If $t_0^* \geq 1$, we are done. Otherwise, we let $\hat{v}(t) \equiv v_0$, for $t \in [t_0^*, \tau_1]$, where (noting (2.42))

$$(2.45) \quad \tau_1 = \inf \left\{ t \in [t_0^*, 1] \mid d(M, z(t)) \leq \frac{\delta}{2} \right\} \geq t_0^* + \frac{\delta}{2F}.$$

By Lemma 2.5 again, we have $t_1^* \in [\tau_1, \tau_1 + T]$ and $\hat{v}(\cdot) \in \hat{\mathcal{V}}[\tau_1, t_1^*]$ such that

$$(2.46) \quad d(M, z(t)) \geq \eta_0, \quad t \in [\tau_1, t_1^*],$$

$$(2.47) \quad d(M, z(t_1^*)) \geq \delta.$$

By induction, we can define τ_i if $t_{i-1}^* < 1$. We say that for some $k \leq 1 + [2F/\delta]$, $t_k^* \geq 1$. If not, then for $k = 1 + [2F/\delta]$ we have

$$(2.48) \quad 1 > t_k^* \geq \tau_k \geq t_{k-1}^* + \frac{\delta}{2F} \geq \cdots \geq t_0^* + \frac{k\delta}{2F} \geq 1,$$

which is a contradiction. Hence we have proved our claim. For the case $d(M, z_0) \geq \delta$, we can prove the same claim. Then our theorem easily follows. \square

If we relax (2.17), then we get the following result for evadability.

THEOREM 2.6. *Suppose all the assumptions of Theorem 2.3 hold except that (2.17) is replaced by the following:*

$$(2.49) \quad \inf_{z \in T_0^\delta(M)} \min_{u \in U} \max_{v \in V} \left(\frac{\partial}{\partial z} V(z), f(t, z, u, v) \right) > 0 \quad \forall t \geq 0.$$

Then the game is evadable.

The proof is contained in the proof of Theorem 2.3. However, we should note that in the present case we do not have Lemma 2.5 (actually, we do not have (2.22)). Thus, we can only get the evadability.

The author does not know whether “ >0 ” in (2.49) can be replaced by “ ≥ 0 ”. In our arguments, condition “ >0 ” is crucial.

Now, let us consider the case that the terminal set M is convex and closed. Then, for any $x \in \partial M \equiv$ the boundary of M ,

$$(2.50) \quad N(x) \triangleq \{\eta \in \mathbb{R}^n \mid \|\eta\| = 1, \eta \text{ is an outer normal of } M \text{ at } x\} \neq \emptyset.$$

Also, we have that for any $z \in \mathbb{R}^n \setminus M$, there exists a unique $x \in \partial M$ such that

$$(2.51) \quad d(M, z) = \|z - x\|.$$

Thus, we get a map $x: \mathbb{R}^n \setminus M \rightarrow \partial M$. Now, let us define

$$(2.52) \quad V(z) = d(M, z) \equiv \|z - x(z)\|.$$

Then we can prove the following (see [19]).

LEMMA 2.7. *$(\partial/\partial z)V(z)$ exists in $\mathbb{R}^n \setminus M$ and*

$$(2.53) \quad \frac{\partial}{\partial z} V(z) = \frac{z - x(z)}{\|z - x(z)\|} \in N(x(z)).$$

Thus, (2.15) holds with $c = 1$ and (2.16) holds with $w(s, \gamma) = 3\gamma/s$. Also, since f satisfies (2.1), we can replace $f(t, z, u, v)$, $z \in T_0^\delta(M)$ in (2.17) and (2.49) by $f(t, x, u, v)$, $x \in \partial M$. Thus, we get the following.

COROLLARY 2.8. *Suppose (A1) and (A2') hold.*

(i) *The game is evadable if*

$$(2.54) \quad \inf_{x \in \partial M} \inf_{\eta \in N(x)} \min_{u \in U} \max_{v \in V} (\eta, f(t, x, u, v)) > 0 \quad \forall t \geq 0.$$

(ii) *The game is strictly evadable if there exists a $\delta_0 > 0$ such that*

$$(2.55) \quad \inf_{t \geq 0} \inf_{x \in \partial M} \inf_{\eta \in N(x)} \min_{u \in U} \max_{v \in V} (\eta, f(t, x, u, v)) \geq \delta_0.$$

Remark 2.9. The method we used in this section is very close to that used in general stability theory, namely the Lyapunov method (see [1], [21], for example). The difference is that in evasion games we want to keep the state away from the terminal set M , while in stabilizing a system to a closed set M we hope to obtain the result that the state approaches to M as $t \rightarrow \infty$. This shows that the Lyapunov method should give a closer relation between the stability theory and the differential pursuit games in which we want to bring the state to the terminal set M in finite time. Also, it is clear that the reachability of control systems should be closely related to pursuit games. We will discuss these in a forthcoming paper.

3. Necessary conditions for evadability. In this section we will give some necessary conditions for the evadability of our differential game.

For a given convex and closed subset M of \mathbb{R}^n , we know that for any $x \in \partial M$, $N(x)$ is a compact set. Thus, N can be regarded as a map from ∂M to the space

$$(3.1) \quad H \triangleq \{C \subset \mathbb{R}^n \mid C \text{ is compact}\}.$$

Let us set

$$(3.2) \quad \rho_H(A, B) = \frac{1}{2} \left\{ \max_{a \in A} d(a, B) + \max_{b \in B} d(A, b) \right\},$$

which is a standard Hausdorff metric. Then (H, ρ_H) is known to be a complete metric space. We say that N is continuous at some point $x_0 \in \partial M$, if $N: \partial M \rightarrow (H, \rho_H)$ is continuous at x_0 .

We now state our theorem.

THEOREM 3.1. *Suppose game (1.1) and terminal set M are given and (A1) and (A2') hold. Suppose that the game is evadable and N is continuous at $x_0 \in \partial M$. Then*

$$(3.3) \quad \sup_{\eta \in N(x_0)} \min_u \max_v (\eta, f(0, x_0, u, v)) \geq 0.$$

Proof. We prove our result by contradiction. Suppose (3.3) does not hold. By the continuity of N at x_0 , and f in (t, z, u, v) , we have the existence of an $\varepsilon > 0$ and a $\delta > 0$ such that

$$(3.4) \quad (\eta, f(t, z, u_0, v)) < -\varepsilon$$

for all $t \in [0, \delta]$, $z \in \mathcal{O}(x_0, \delta) \triangleq \{z \in \mathbb{R}^n \mid \|z - x_0\| < \delta\}$, $\eta \in N(x(z))$, $v \in V$ and for some element $u_0 \in U$. Now, let $F_0 > 1$ be such that

$$(3.5) \quad \|f(t, z, u_0, v)\| \leq F_0,$$

for all $z \in \mathcal{O}(x_0, \delta)$, $v \in V$ and $t \in [0, \delta]$. Set

$$(3.6) \quad t_0 = \frac{\delta}{2F_0},$$

$$(3.7) \quad \alpha = \min \left\{ \frac{\delta}{4}, \varepsilon t_0 \right\}.$$

Then we claim that for $z_0 = x_0 + \alpha \eta_0$, with η_0 being a fixed element in $N(x_0)$ and any $v(\cdot) \in \mathcal{V}$, there exists $t^* \in [0, t_0]$ such that the trajectory $z(\cdot)$ of (1.1) corresponding to $u(t) \equiv u_0$ and $v(\cdot)$ satisfies

$$(3.8) \quad d(M, z(t^*)) = 0.$$

If the claim were not true, i.e., if there exists a $v(\cdot) \in \mathcal{V}$ such that

$$(3.9) \quad d(M, z(t)) > 0 \quad \forall t \in [0, t_0],$$

then

$$(3.10) \quad \theta(z(t)) \equiv \frac{z(t) - x(z(t))}{\|z(t) - x(z(t))\|} \in N(x(z(t)))$$

is well defined for $t \in [0, t_0]$. By (3.5), (3.6) and (3.7) we have

$$(3.11) \quad \begin{aligned} \|z(t) - x_0\| &\leq \|z(t) - z_0\| + \|z_0 - x_0\| \\ &\leq F_0 t_0 + \alpha \leq \frac{3\delta}{4} < \delta \end{aligned}$$

for all $t \in [0, t_0]$. Then, since $t_0 < \delta$, it follows from (3.4) that

$$(3.12) \quad \frac{d}{dt}(d(M, z(t))) = (\theta(z(t)), f(t, z(t), u_0, v(t))) < -\varepsilon.$$

Integrating (3.12), we get

$$(3.13) \quad d(M, z(t)) \leq d(M, z_0) - \varepsilon t, \quad t \in [0, t_0].$$

Thus, at $t = t_0$, we have

$$(3.14) \quad \begin{aligned} d(M, z(t_0)) &\leq d(M, z_0) - \varepsilon t_0 \\ &= \alpha - \varepsilon t_0 \leq 0, \end{aligned}$$

which contradicts (3.9). Thus our claim is true. The claim, however, contradicts the evadability of the game; hence (3.3) holds. \square

We now let M be a convex and closed set of \mathbb{R}^n with $\dim M = n - 1$, i.e., M is contained in some $(n - 1)$ -dimensional hyperplane and has at least one interior point with respect to that hyperplane. Let $\eta \in \mathbb{R}^n$, $\|\eta\| = 1$. We set

$$(3.15) \quad p(\eta) = \{\eta\}^\perp \triangleq \{z \in \mathbb{R}^n \mid (\eta, z) = 0\}.$$

Then we can assume that

$$(3.16) \quad M \subseteq y_0 + p(\eta)$$

for some $y_0 \in M$ and $\eta \in \mathbb{R}^n$, $\|\eta\| = 1$. Note that in this case

$$(3.17) \quad M = \partial M.$$

We denote all the interior points of M with respect to $y_0 + p(\eta)$ by $\text{Int}_{n-1} M$. Then for any $x_0 \in \text{Int}_{n-1} M \subseteq \partial M$,

$$(3.18) \quad N(x_0) = \{\pm \eta\}.$$

COROLLARY 3.2. *Suppose that game (1.1) is given, (A1) holds and the terminal set is an $(n - 1)$ -dimensional convex and closed subset in \mathbb{R}^n satisfying (3.16). Suppose that the game is evadable. Then for any $x_0 \in \text{Int}_{n-1} M$,*

$$(3.19) \quad \min_{u \in U} \max_{v \in V} (\pm \eta, f(0, x_0, u, v)) \geq 0.$$

The proof is obvious.

Remark 3.3. It is clear that in both Theorem 3.1 and Corollary 3.2 we only need the continuity of $f(t, z, u, v)$ at $\{0\} \times \partial M \times U \times V$.

The following result gives a necessary condition for the evadability of the game where f is linear in z and the terminal set M is an $(n - 1)$ -dimensional convex and closed set.

COROLLARY 3.4. *Suppose that in (1.1)*

$$(3.20) \quad f(t, z, u, v) = A(t)z + g(t, u, v)$$

with $A(\cdot)$, $g(\cdot, \cdot, \cdot)$ continuous. The terminal set M is an $(n - 1)$ -dimensional convex and closed set in \mathbb{R}^n satisfying (3.16) and containing an $(n - 1)$ -dimensional cone. Suppose that the game is evadable; then

$$(3.21) \quad A(0)^* \eta = \lambda \eta$$

for some $\lambda \in \mathbb{R}$.

Proof. Without loss of generality, we assume that

$$(3.22) \quad y_0 + K(\eta) \subseteq M \subseteq y_0 + p(\eta)$$

where $K(\eta)$ is a cone in $p(\eta)$ and $y_0 \in M$.

We suppose that the game is evadable and that for all $\lambda \in \mathbb{R}$

$$(3.23) \quad A(0)^* \eta \neq \lambda \eta.$$

We show that this leads to a contradiction. We claim that if (3.23) holds, then there exists a $z \in \text{Int}_{n-1} K(\eta)$ such that

$$(3.24) \quad (\eta, A(0)z) \neq 0.$$

If (3.24) were not true, then for all $z \in \text{Int}_{n-1} K(\eta)$

$$(3.25) \quad (A(0)^* \eta, z) = (\eta, A(0)z) = 0.$$

Since $\dim K(\eta) = n - 1$, (3.25) implies that for all $z \in p(\eta)$

$$(3.26) \quad (A(0)^* \eta, z) = 0.$$

Thus, we get

$$(3.27) \quad A(0)^* \eta \in p(\eta)^\perp = \text{span} \{ \eta \},$$

which contradicts (3.23). Hence there is a $z \in K(\eta)$, satisfying (3.24). Since $K(\eta)$ is a cone, $\beta z \in K(\eta)$ for all $\beta > 0$, by (3.22)

$$(3.28) \quad y_0 + \beta z \in M, \quad \beta > 0.$$

Without loss of generality, we assume that

$$(3.29) \quad (\eta, A(0)z) > 0.$$

Then we set

$$(3.30) \quad \beta = \{ |(\eta, A(0)y_0)| + \max_{(u,v) \in U \times V} \|g(0, u, v)\| + 1 \} / (\eta, A(0)z).$$

We get

$$(3.31) \quad \begin{aligned} & \min_{u \in U} \max_{v \in V} (-\eta, f(0, y_0 + \beta z, u, v)) \\ & \leq (-\eta, A(0)\beta z) + |(\eta, A(0)y_0)| + \max_{(u,v) \in U \times V} \|g(0, u, v)\| = -1, \end{aligned}$$

which contradicts (3.19) and hence the corollary is proved. \square

Notice that when $A(t) \equiv A$, then the above corollary says that the evadability of the game implies that η is an eigenvector of A^* . Also, we can see that (3.3) is very close to (2.54), especially in the case that $f(t, z, u, v) \equiv f(z, u, v)$.

4. Sharper necessary conditions for evadability. In this section we will give a much stronger result than Corollary 3.4 when the terminal set M is of form $y_0 + p(\eta)$.

We consider the following:

$$(4.1) \quad \begin{aligned} \dot{z} &= A(t)z + g(t, u, v), \\ z(0) &= z_0. \end{aligned}$$

We assume that $A(t)$ is an $(n \times n)$ matrix-valued locally integrable function, and that $g(\cdot, \cdot, \cdot) \in L^\infty([0, t_1] \times U \times V)$ for all $t_1 > 0$. The terminal set is

$$(4.2) \quad M = y_0 + p(\eta).$$

Our main theorem of this section is the following.

THEOREM 4.1. *If game (4.1) with terminal set (4.2) is evadable, then there exists a real-valued locally integrable function $\lambda(t)$ such that*

$$(4.3) \quad A(t)^* \eta = \lambda(t) \eta \quad \text{a.e.}$$

To prove this theorem we need a lemma. We let $\Phi(t)$ be the fundamental matrix of

$$(4.4) \quad \dot{z} = A(t)z,$$

satisfying $\Phi(0) = I$.

LEMMA 4.2. *Let $\eta \in \mathbb{R}^n$, $\|\eta\| = 1$ have the property that for all $z_1 \in \mathbb{R}^n$ such that*

$$(4.5) \quad (\eta, z_1) > 0,$$

we have

$$(4.6) \quad (\eta, \Phi(t)z_1) \geq 0, \quad t \geq 0.$$

Then there exists a real-valued locally integrable function $\lambda(t)$ such that

$$(4.7) \quad \Phi(t)^* \eta = \exp\left(\int_0^t \lambda(\tau) d\tau\right) \eta, \quad t \geq 0,$$

$$(4.8) \quad A(t)^* \eta = \lambda(t) \eta \quad \text{a.e.}$$

Proof. We claim that under our hypothesis we have that for all $x \in p(\eta) \equiv \{\eta\}^\perp$

$$(4.9) \quad (\eta, \Phi(t)x) = 0, \quad t \geq 0.$$

If not, then there exist an $x_1 \in \{\eta\}^\perp$ and a $t_0 > 0$ such that

$$(4.10) \quad (\eta, \Phi(t_0)x_1) \neq 0.$$

Without loss of generality, we assume that

$$(4.11) \quad (\eta, \Phi(t_0)x_1) = -\delta < 0.$$

If we take

$$(4.12) \quad z_1 = x_1 + \varepsilon \eta$$

where

$$(4.13) \quad 0 < \varepsilon \leq \frac{\delta}{1 + (\eta, \Phi(t_0)\eta)},$$

then we have

$$(4.14) \quad (\eta, z_1) = \varepsilon > 0,$$

$$(4.15) \quad (\eta, \Phi(t_0)z_1) = \varepsilon(\eta, \Phi(t_0)\eta) - \delta \leq -\varepsilon < 0,$$

which contradicts (4.6). Thus our claim is true.

It follows from (4.9) that

$$(4.16) \quad \Phi(t)^* \eta \in p(\eta)^\perp = \text{span}\{\eta\}, \quad t \geq 0.$$

Thus, there exists a real-valued function $a(t)$ such that

$$(4.17) \quad \Phi(t)^* \eta = a(t) \eta, \quad t \geq 0.$$

Since $\Phi(t)^*$ is invertible for all t , we get that $a(t) \neq 0$ for all $t \geq 0$. Hence, noting that $(\eta, \eta) = 1 > 0$, we have by (4.6)

$$(4.18) \quad \begin{aligned} a(t) &= (\eta, \Phi(t)\eta) > 0, \quad t \geq 0, \\ a(0) &= 1. \end{aligned}$$

From (4.18), it is clear that $a(t)$ is absolutely continuous. Now, for any $t_1 > 0$, let

$$(4.19) \quad \delta_0 = \min_{t \in [0, t_1]} a(t) > 0.$$

On $[\delta_0, \infty)$ we have

$$(4.20) \quad |(\log s)'| = \frac{1}{s} \leq \frac{1}{\delta_0} < \infty,$$

and so, $\log s$ satisfies a Lipschitz condition on $[\delta_0, \infty)$. Hence, $\log a(t)$ is absolutely continuous on $[0, t_1]$ for all $t_1 > 0$. Thus, there exists a real-valued, locally integrable function $\lambda(t)$ such that (note $a(0) = 1$)

$$(4.21) \quad \log a(t) = \int_0^t \lambda(\tau) d\tau, \quad t \geq 0.$$

Hence, we get (4.7) by combining (4.17) and (4.21).

To obtain (4.8), we consider the following:

$$(4.22) \quad \begin{aligned} \Phi(t)^*[\lambda(t)\eta] &= \lambda(t)\Phi(t)^*\eta \\ &= \lambda(t) \exp\left(\int_0^t \lambda(\tau) d\tau\right) \eta \\ &= \left[\exp\left(\int_0^t \lambda(\tau) d\tau\right) \eta\right]' \quad \text{a.e.} \\ &= [\Phi(t)^*\eta]' \\ &= \Phi(t)^*A(t)^*\eta \quad \text{a.e.} \end{aligned}$$

But $\Phi(t)^*$ is invertible, so (4.8) holds. \square

Remark 4.3. In fact, (4.7) and (4.8) are equivalent. We have seen that (4.7) implies (4.8). Now, if (4.8) holds, we consider the following initial value problem:

$$(4.23) \quad \begin{aligned} \dot{z} &= \lambda(t)z \quad \text{a.e.}, \\ z(0) &= \eta. \end{aligned}$$

Then, $\exp(\int_0^t \lambda(\tau) d\tau)\eta$ and $\Phi(t)^*\eta$ are solutions of (4.23). By uniqueness we get (4.7).

Now, we can prove Theorem 4.1.

Proof of Theorem 4.1. By Lemma 4.2, it is enough to show that for any $z_1 \in \mathbb{R}^n$, with $(\eta, z_1) > 0$, we have (4.6).

We prove this by contradiction. Suppose there exist a $z_1 \in \mathbb{R}^n$ and a $t_0 > 0$ such that $(\eta, z_1) > 0$ and

$$(4.24) \quad (\eta, \Phi(t_0)z_1) < 0.$$

Then we take

$$(4.25) \quad z_0 = y_0 + \alpha z_1$$

where $\alpha > 0$ is as yet undetermined. Then we have

$$(4.26) \quad d(M, z_0) = |(\eta, z_0 - y_0)| = \alpha(\eta, z_1) > 0,$$

i.e., $z_0 \notin M$. Also, for any $u(\cdot)$ and $v(\cdot)$, admissible,

$$(4.27) \quad \begin{aligned} (\eta, z(t_0) - y_0) &= (\eta, \Phi(t_0)z_0 + \int_0^{t_0} \Phi(t_0)\Phi(\tau)^{-1}g(\tau, u(\tau), v(\tau)) d\tau - y_0) \\ &= \alpha(\eta, \Phi(t_0)z_1) + (\eta, \Phi(t_0)y_0 - y_0 \\ &\quad + \int_0^{t_0} \Phi(t_0)\Phi(\tau)^{-1}g(\tau, u(\tau), v(\tau)) d\tau). \end{aligned}$$

Note that the second term is bounded. Hence, if we take α large enough and notice (4.24), we will have

$$(4.28) \quad (\eta, z(t_0) - y_0) \leq 0.$$

When we compare this with (4.26) and note that $z(t)$ is continuous, we see that there exists a $t^* \in (0, t_0]$ such that

$$(4.29) \quad d(M, z(t^*)) = |(\eta, z(t^*) - y_0)| = 0,$$

which contradicts the evadability of the game. \square

COROLLARY 4.4. *If $A(t) \equiv A$, and the game is evadable, then*

$$(4.30) \quad A^*\eta = \lambda\eta$$

for some $\lambda \in \mathbb{R}$.

5. Sufficient conditions for systems linear in the state. In this section we consider a differential game which is governed by (4.1). We assume that $A(t)$ is an $(n \times n)$ matrix-valued locally integrable function, and that $g(t, u, v)$ is in $L^\infty([0, t_1] \times U \times V)$ for all $t_1 > 0$ and Borel measurable in u , continuous in v . The terminal set M is given by

$$(5.1) \quad M = y_0 + Q$$

where Q is a linear subspace of \mathbb{R}^n with $\dim Q < n$. It is clear from (5.1) that

$$(5.2) \quad M = \partial M.$$

Also, for all $x \in M$

$$(5.3) \quad N(x) = \{\eta \in Q^\perp \mid \|\eta\| = 1\}.$$

Let

$$(5.4) \quad B_1^n(0) = \{\eta \in \mathbb{R}^n \mid \|\eta\| \leq 1\}$$

and let $\pi: \mathbb{R}^n \rightarrow Q^\perp$ be the orthogonal projection. Then we have the following theorem.

THEOREM 5.1. *Suppose that game (4.1) is given, $A(t)$, $g(t, u, v)$ are continuous, and*

$$(5.5) \quad \|A(t)\| \leq K, \quad t \geq 0.$$

The terminal set M is given by (5.1). We assume that

$$(5.6) \quad A(t)Q \subseteq Q, \quad t \geq 0,$$

and that there exists a $\delta_0 > 0$ such that

$$(5.7) \quad \delta_0 \pi B_1^n(0) \subseteq \bigcap_{u \in U} [\pi A(t)y_0 + \pi g(t, u, V)], \quad t \geq 0.$$

Then the game is strictly evadable.

Proof. It is clear that under our assumptions, (A1) and (A2') hold. Thus, it suffices to prove that (2.55) holds.

For any $t \geq 0$, $x \in \partial M = M$, and any $\eta \in N(x)$, it follows from (5.7) that

$$(5.8) \quad \delta_0 \eta \in \delta_0 \pi B_1^n(0) \subseteq \bigcap_{u \in U} [\pi A(t)y_0 + \pi g(t, u, V)], \quad t \geq 0.$$

Thus, for each $u \in U$, there exists a $v \in V$ such that (t fixed)

$$(5.9) \quad \delta_0 \eta = \pi A(t)y_0 + \pi g(t, u, v).$$

From the definition of π , we have $\pi^* \eta = \eta$. Hence, we get that

$$(5.10) \quad \begin{aligned} \delta_0 &= (\eta, A(t)y_0 + g(t, u, v)) \\ &\leq (\eta, A(t)y_0) + \max_{v \in V} (\eta, g(t, u, v)). \end{aligned}$$

Since $x - y_0 \in Q$ (see (5.1)), it follows from (5.6) that

$$(5.11) \quad (\eta, A(t)y_0) = (\eta, A(t)x), \quad t \geq 0, \quad \eta \in N(x).$$

Thus, (5.10) implies

$$(5.12) \quad \delta_0 \leq \max_{v \in V} (\eta, A(t)x + g(t, u, v)),$$

for all $t \geq 0$, $x \in \partial M$, $u \in U$ and $\eta \in N(x)$. Thus (2.55) holds and our theorem is proved. \square

We note that (5.6) means that Q is an invariant subspace of $A(t)$ for all $t \geq 0$. In the case $\dim Q = n - 1$, we have seen in the last section that this condition is necessary for the evadability of the game.

COROLLARY 5.2. *Suppose $A(t) \equiv A$, $g(t, u, v) \equiv g(u, v)$, and the terminal set M is an invariant subspace of A . Suppose further that there exists a $\delta_0 > 0$, such that*

$$(5.13) \quad \delta_0 \pi B_1^n(0) \subseteq \bigcap_{u \in U} \pi g(u, V)$$

where $\pi: \mathbb{R}^n \rightarrow M^\perp$ is the orthogonal projection. Then the game is strictly evadable.

The proof is immediate.

Now, if $\dim Q = n - 1$, our strict evadability results can be strengthened. We assume in the following that $A(t)$ and $g(t, u, v)$ are not necessarily continuous but satisfy the assumptions that were made at the beginning of this section. By Theorem 4.1, in our case, (4.3) is a necessary condition for evadability and thus it is necessary for strict evadability. We have the following.

THEOREM 5.3. *Suppose game (4.1) is given, the terminal set M is given by (4.2), and (4.3) holds. Let*

$$(5.14) \quad \mu_\pm(t) = \min_{u \in U} \max_{v \in V} (\pm \eta, g(t, u, v)) - |\lambda(t)|d(M, 0), \quad t \geq 0.$$

Suppose one of the following holds:

(i) If $\lim_{t \rightarrow \infty} \int_0^t \lambda(\tau) d\tau > -\infty$, then

$$(5.15) \quad \mu_\pm(t) \geq 0 \quad \text{a.e.}$$

(ii) If $\lim_{t \rightarrow \infty} \int_0^t \lambda(\tau) d\tau = -\infty$, then (5.15) holds and

$$(5.16) \quad \lim_{t \rightarrow \infty} \mu_\pm(t) > 0,$$

$$(5.17) \quad \left| \int_t^{t+\tau} \lambda(s) ds \right| \leq w(\tau), \quad t \geq 0, \quad w(\cdot) \in L_{\text{loc}}^1[0, \infty).$$

Then the game is strictly evadable.

Remark 5.4. In the case of Theorem 5.3, if we denote $f(t, z, u, v) = A(t)z + g(t, u, v)$, then (5.15) is the same as the following:

$$(5.18) \quad \inf_{x \in \partial M} \inf_{\eta \in N(x)} \min_{u \in U} \max_{v \in V} (\eta, f(t, x, u, v)) \geq 0 \quad \text{a.e.}$$

It is clear that (5.18) is weaker than (2.54) and (2.55). Also, (5.18) is closer to the necessary condition (3.3) than are (2.54) and (2.55).

Proof of Theorem 5.3. Let $z_0 \notin M$, $u(\cdot) \in \mathcal{U}[0, \infty)$ be given. We note that

$$(5.19) \quad d(M, z) = |(\eta, z - y_0)| \quad \forall z \in \mathbb{R}^n.$$

Without loss of generality, we assume that

$$(5.20) \quad d(M, z_0) = (\eta, z_0 - y_0).$$

Then, by Filippov's lemma, we can find $\hat{v}(\cdot) \in \hat{\mathcal{V}}[0, \infty)$ such that

$$(5.21) \quad (\eta, g(t, u(t), \hat{v}(t))) - |\lambda(t)|d(M, 0) = \mu_+(t) \quad \text{a.e.}$$

Under this evasion control, we have (note (5.15))

$$\begin{aligned} d(M, z(t)) &= |(\eta, z(t) - y_0)| \\ &= \left| (\eta, \Phi(t)z_0 + \int_0^t \Phi(t)\Phi(\tau)^{-1}g(\tau, u(\tau), \hat{v}(\tau)) d\tau - y_0) \right| \\ &= \left| (\eta, \Phi(t)(z_0 - y_0)) \right. \\ &\quad \left. + \int_0^t (\eta, \Phi(t)\Phi(\tau)^{-1}[A(\tau)y_0 + g(\tau, u(\tau), \hat{v}(\tau))]) d\tau \right| \\ &= \left| \exp\left(\int_0^t \lambda(s) ds\right) d(M, z_0) \right. \\ &\quad \left. + \int_0^t \exp\left(\int_\tau^t \lambda(s) ds\right) [(\eta, g(\tau, u(\tau), \hat{v}(\tau)) + \lambda(\tau)(\eta, y_0))] d\tau \right| \\ &\geq \exp\left(\int_0^t \lambda(s) ds\right) d(M, z_0) + \int_0^t \exp\left(\int_\tau^t \lambda(s) ds\right) \mu_+(\tau) d\tau. \end{aligned} \quad (5.22)$$

CASE (i). Since $\lim_{t \rightarrow \infty} \int_0^t \lambda(s) ds > -\infty$, there exists $\delta_0 > 0$ such that

$$(5.23) \quad \exp\left(\int_0^t \lambda(s) ds\right) \geq \delta_0, \quad t \geq 0.$$

Thus, from (5.22), we get

$$(5.24) \quad d(M, z(t)) \geq \exp\left(\int_0^t \lambda(s) ds\right) d(M, z_0) \geq \delta_0 d(M, z_0) \quad \forall t \geq 0.$$

This gives the strict evadability.

CASE (ii). By (5.16) and (5.17) we have $t_0 > 0$ and $\delta_0 > 0$ such that

$$(5.25) \quad \mu_+(t) \geq \delta_0 \quad \text{a.e. } t \geq t_0.$$

Then we set

$$(5.26) \quad \hat{\delta} = \min_{0 \leq t \leq 2t_0} \exp\left(\int_0^t \lambda(s) ds\right) > 0.$$

Then, for $t \in [0, 2t_0]$, we have by (5.15), (5.22) and (5.26)

$$(5.27) \quad d(M, z(t)) \geq \exp\left(\int_0^t \lambda(s) ds\right) d(M, z_0) \geq \hat{\delta} d(M, z_0) > 0,$$

and for $t \in [2t_0, \infty)$ we have by (5.15), (5.17), (5.22) and (5.25)

$$(5.28) \quad \begin{aligned} d(M, z(t)) &\geq \exp\left(\int_0^t \lambda(s) ds\right) d(M, z_0) + \int_{t_0}^t \exp\left(\int_\tau^t \lambda(s) ds\right) \delta_0 d\tau \\ &\geq \delta_0 \int_0^{t-t_0} \exp\left(\int_{s-\tau}^t \lambda(s) ds\right) d\tau \\ &\geq \delta_0 \int_0^{t-t_0} e^{-w(\tau)} d\tau \\ &\geq \delta_0 \int_0^{t_0} e^{-w(\tau)} d\tau > 0. \end{aligned}$$

Hence, we complete the proof. \square

Remark 5.5. If $A(t) \equiv A$ as in Theorem 5.3, then $\lambda(t) \equiv \lambda$ is an eigenvalue of A^* (hence of A). In this situation, Case (i) corresponds to $\lambda \geq 0$ and Case (ii) corresponds to $\lambda < 0$.

6. A linear autonomous case. In this section we discuss a differential game of the following form:

$$(6.1) \quad \begin{aligned} \dot{z} &= Az + Bu + Cv, \\ z(0) &= z_0, \end{aligned}$$

where A , B and C are matrices of suitable sizes, $z \in \mathbb{R}^n$, $u \in U \subseteq \mathbb{R}^p$, $v \in V \subseteq \mathbb{R}^q$. The terminal set M is a subspace of \mathbb{R}^n with $\dim M < n$ and $AM \subseteq M$. We will give a strict evadability result which is stronger, in some sense, than Corollary 5.2.

Let us first give the following.

DEFINITION 6.1. Let $\delta > 0$, $U \subseteq \mathbb{R}^p$, $V \subseteq \mathbb{R}^q$ be given. (C, V) is said to be δ -superior to (B, U) if there exists $V_0 \subseteq \mathbb{R}^q$ such that

$$(6.2) \quad V_0 + \delta B_1^q(0) \subseteq V,$$

$$(6.3) \quad BU \subseteq -CV_0,$$

where

$$B_1^q(0) = \{v \in \mathbb{R}^q \mid \|v\| \leq 1\}.$$

The following proposition gives us some intuitive geometric meaning of the above definition.

PROPOSITION 6.2. Let $U = B_1^p(0)$, $V = B_1^q(0)$, and $\mu > 1$, be such that

$$(6.4) \quad \mu B(B_1^p(0)) \subseteq C(B_1^q(0)).$$

Then, $(C, B_1^q(0))$ is δ -superior to $(B, B_1^p(0))$ with $\delta = 1 - (1/\mu)$.

Proof. Take $V_0 = (1 - \delta)B_1^q(0) = -V_0$, where $\delta = 1 - (1/\mu)$. Then it is clear that (6.2) holds and

$$B(B_1^p(0)) \subseteq \frac{1}{\mu} C(B_1^q(0)) = C((1 - \delta)B_1^q(0)) = -CV_0. \quad \square$$

Roughly speaking, (C, V) is δ -superior to (B, U) for some $\delta > 0$ if the evader is more powerful than the pursuer. We will see that more clearly from the following.

LEMMA 6.3. *Suppose $\delta > 0$ and (C, V) is δ -superior to (B, U) , then for any $u(\cdot) \in \mathcal{U}$ there exists a $v_u(\cdot) \in \hat{\mathcal{V}}$ with values in V_0 such that*

$$(6.5) \quad Bu(t) + Cv_u(t) = 0 \quad \text{a.e.},$$

where V_0 is as in Definition 6.1.

Proof. The result follows from (6.3) and Filippov's lemma. \square

REMARK 6.4. As a consequence of the above lemma, we have that for any $u(\cdot) \in \mathcal{U}$ and measurable $w(\cdot)$ with values in $\delta B_1^q(0)$, there exists a $\hat{v}(\cdot) \in \hat{\mathcal{V}}$ such that

$$(6.6) \quad Bu(t) + c\hat{v}(t) = cw(t) \quad \text{a.e.}$$

This gives the superiority of the evader to the pursuer.

Before stating and proving the main result of this section, let us give another lemma.

LEMMA 6.5. *Suppose $a, b \in \mathbb{R}$, $b \neq 0$ and*

$$(6.7) \quad K = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then for $t_0, t \in \mathbb{R}$

$$\int_{t_0}^t e^{(aI+bK)\tau} d\tau = \frac{aI-bK}{a^2+b^2} e^{(aI+bK)\tau} \Big|_{t_0}^t.$$

The proof is straightforward.

We now give our main theorem of this section.

THEOREM 6.6. *Suppose the terminal set M is invariant under A , and $\dim M < n$. Suppose that $[(A^*|_{M^\perp})^*, (C^*|_{M^\perp})^*]$ is completely controllable and that $((C^*|_{M^\perp})^*, V)$ is δ -superior to $((B^*|_{M^\perp})^*, U)$ for some $\delta > 0$. Then the game with terminal set M is strictly evadable.*

REMARK 6.7. In the statement of the theorem, $[(A^*|_{M^\perp})^*, (C^*|_{M^\perp})^*]$ is regarded as a system in M^\perp not in \mathbb{R}^n .

Proof. By choosing suitable orthonormal basis for $M^\perp \oplus M$, we have the representations of A, B, C as follows:

$$(6.8) \quad A = \begin{pmatrix} A_1 & 0 \\ A_3 & A_4 \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \quad C = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}.$$

Then the conditions stated in Theorem 6.6 are the following: $[A_1, C_1]$ is completely controllable; (C_1, V) is δ -superior to (B_1, U) . Let us denote, according to the decomposition of \mathbb{R}^n ,

$$(6.9) \quad z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \quad z_0 = \begin{pmatrix} z_{10} \\ z_{20} \end{pmatrix}.$$

By applying $\pi: \mathbb{R}^n \rightarrow M^\perp$, the orthogonal projection to (6.1), we get

$$(6.10) \quad \begin{aligned} \dot{z}_1 &= A_1 z_1 + B_1 u + C_1 v, \\ z_1(0) &= z_{10}, \end{aligned}$$

and we know that

$$(6.11) \quad d(M, z(t)) = \|\pi z(t)\| = \|z_1(t)\|.$$

Now, we assume that the initial state $z_0 \notin M$, i.e., $z_{10} \neq 0$, and also that the pursuer's control $u(\cdot) \in \mathcal{U}$ is given. Then by Lemma 6.3 and Remark 6.4, we may choose the evasion control as in (6.6), and so

$$(6.12) \quad B_1 u(t) + C_1 \hat{v}(t) = C_1 w(t)$$

with $w(t) \in \delta B_1^q(0)$ being arbitrary. Then (6.10) becomes

$$(6.13) \quad \begin{aligned} \dot{z}_1(t) &= A_1 z_1(t) + C_1 w(t), \\ z_1(0) &= z_{10}. \end{aligned}$$

Then we have from (6.11) and (6.13) that

$$(6.14) \quad \begin{aligned} d(M, z(t)) &= \|z_1(t)\| \\ &= \left\| e^{A_1 t} z_{10} + \int_0^t e^{A_1(t-\tau)} C_1 w(\tau) d\tau \right\|. \end{aligned}$$

Now, let us discuss the different cases.

CASE 1. $\sigma(A_1) \cap \mathbb{R} \neq \emptyset$, where $\sigma(A_1)$ is the spectrum of A_1 .

Since $\sigma(A_1) = \sigma(A_1^*)$, we can assume that there exists an η_1 with $\|\eta_1\| = 1$ such that for some $\lambda_1 \in \mathbb{R}$

$$(6.15) \quad A_1^* \eta_1 = \lambda_1 \eta_1.$$

Since $[A_1, C_1]$ is completely controllable, we have

$$(6.16) \quad C_1^* \eta_1 \neq 0 \quad \text{or} \quad \eta_1^T C_1 \neq 0$$

where η_1^T stands for the transpose of the column vector η_1 .

Now, we set

$$(6.17) \quad w(t) = \begin{cases} \operatorname{sgn}(\eta_1^T z_{10}) \frac{\delta C_1^* \eta_1}{\|C_1^* \eta_1\|} & \text{if } \eta_1^T z_{10} \neq 0, \\ \delta \frac{C_1^* \eta_1}{\|C_1^* \eta_1\|} & \text{if } \eta_1^T z_{10} = 0. \end{cases}$$

Then from (6.14), (6.15) and (6.17), we have that

$$(6.18) \quad \begin{aligned} |\eta_1^T z_1(t)| &= \left| e^{\lambda_1 t} \eta_1^T z_{10} + \int_0^t e^{\lambda_1(t-\tau)} \eta_1^T C_1 w(\tau) d\tau \right| \\ &= e^{\lambda_1 t} |\eta_1^T z_{10}| + \left(\int_0^t e^{\lambda_1 \tau} d\tau \right) \delta \|C_1^* \eta_1\|. \end{aligned}$$

Since $z_{10} \neq 0$, there exists a $\theta > 0$ such that for all admissible controls $u(\cdot)$ and $v(\cdot)$

$$(6.19) \quad d(M, z(t)) = \|z_1(t)\| \geq \frac{\|z_{10}\|}{2}, \quad t \in [0, \theta].$$

For $t \geq \theta$, from (6.18), we have

$$(6.20) \quad \begin{aligned} d(M, z(t)) &\geq |\eta_1^T z_1(t)| \geq \left(\int_0^t e^{\lambda_1 \tau} d\tau \right) \delta \|C_1^* \eta_1\| \\ &\geq \left(\int_0^\theta e^{\lambda_1 \tau} d\tau \right) \delta \|C_1^* \eta_1\| > 0. \end{aligned}$$

Hence, we have the strict evadability of the game in this case.

CASE 2. $\sigma(A_1) \cap \mathbb{R} = \emptyset$.

Note that in this case, $\dim M^\perp \geq 2$.

Let $a \pm ib \in \sigma(A_1^*) = \sigma(A_1)$, where $a, b \in \mathbb{R}$ and $b > 0$, then there exist real vectors α and β in M^\perp such that

$$(6.21) \quad A_1^*(\alpha + i\beta) = (a - bi)(\alpha + i\beta).$$

From this we can see that α and β are linearly independent and

$$(6.22) \quad \begin{aligned} A_1^*\alpha &= a\alpha + b\beta, \\ A_1^*\beta &= -b\alpha + a\beta, \end{aligned}$$

i.e.,

$$(6.23) \quad A_1^*(\alpha, \beta) = (\alpha, \beta) \begin{pmatrix} a & -b \\ b & a \end{pmatrix}.$$

We define $L_0 = \text{span}\{\alpha, \beta\} \subseteq M^\perp$. It is clear that $A_1^*L_0 \subseteq L_0$; hence, we have

$$(6.24) \quad A_1L_0^\perp \subseteq L_0^\perp.$$

Now, by choosing a suitable basis for $M^\perp = L_0 \oplus L_0^\perp$, we have

$$(6.25) \quad \begin{aligned} A_1 &= \begin{pmatrix} A_0 & 0 \\ * & * \end{pmatrix}, & C_1 &= \begin{pmatrix} C_0 \\ * \end{pmatrix}, \\ z_1 &= \begin{pmatrix} z_1^1 \\ z_1^2 \\ z_1^3 \end{pmatrix}, & z_{10} &= \begin{pmatrix} z_{10}^1 \\ z_{10}^2 \\ z_{10}^3 \end{pmatrix}, \end{aligned}$$

where $*$ stands for entries which will not concern us and

$$(6.26) \quad A_0 = \begin{pmatrix} a & b \\ -b & a \end{pmatrix} = aI + bK.$$

Here, K is given by (6.7).

Since $[A_1, C_1]$ is completely controllable, and $A_1L_0^\perp \subseteq L_0^\perp$, we know that $[A_0, C_0]$ is also completely controllable. Hence, in particular, we have $C_0 \neq 0$ and then there exists $w_0 \in \delta B_1^2(0)$ such that

$$(6.27) \quad \varphi \triangleq C_0w_0 \neq 0.$$

By applying the orthogonal projection P_{L_0} to (6.13) we get

$$(6.28) \quad \begin{aligned} \dot{z}_1^1(t) &= A_0z_1^1(t) + C_0w(t), \\ z_1^1(0) &= z_{10}^1. \end{aligned}$$

Thus,

$$(6.29) \quad \begin{aligned} d(M, z(t)) &= \|z_1(t)\| \geq \|P_{L_0}z_1(t)\| = \|z_1^1(t)\| \\ &= \left\| e^{A_0t}z_{10}^1 + \int_0^t e^{A_0(t-\tau)}C_0w(\tau) d\tau \right\|. \end{aligned}$$

Now, we consider two subcases.

(1) $z_{10}^1 \neq 0$.

(i) For $a \geq 0$, we take

$$(6.30) \quad w(t) \equiv 0, \quad t \geq 0.$$

Then for all $t \geq 0$ we have that

$$(6.31) \quad \begin{aligned} d(M, z(t)) &\geq \|z_1^1(t)\| = \|e^{at} e^{bKt} z_{10}^1\| \\ &= e^{at} \|e^{bKt} z_{10}^1\| = e^{at} \|z_{10}^1\| \geq \|z_{10}^1\| > 0. \end{aligned}$$

(ii) For $a < 0$, then, we know that

$$(6.32) \quad \lim_{t \rightarrow \infty} e^{at} e^{bKt} z_{10}^1 = 0,$$

and e^{bKt} is a rotation on the plane L_0 which sweeps all possible angles as t varies from 0 to ∞ . Thus, there exists $t_0 > 0$ such that

$$(6.33) \quad e^{at_0} e^{bKt_0} z_{10}^1 = \delta_0 \psi \equiv \delta_0 \frac{-aI + bK}{a + b} \varphi$$

for some $\delta_0 \in (0, 1)$. Since $\det(-aI + bK) = a^2 + b^2 > 0$, we have $\psi \neq 0$. Now, we take

$$(6.34) \quad w(t) = \begin{cases} 0, & 0 \leq t < t_0, \\ \delta_0 w_0, & t_0 \leq t < \infty \end{cases}$$

where w_0 is as in (6.27). Then, for $0 \leq t < t_0$, we have

$$(6.35) \quad \begin{aligned} d(M, z(t)) &\geq \|z_1^1(t)\| = \|e^{at} e^{bKt} z_{10}^1\| \\ &\geq e^{at_0} \|z_{10}^1\| > 0, \end{aligned}$$

and for $t \geq t_0$, by Lemma 6.5 and (6.33), we have

$$(6.36) \quad \begin{aligned} d(M, z(t)) &\geq \left\| e^{(aI+bK)t} z_{10}^1 + \int_{t_0}^t e^{(aI+bK)(t-\tau)} \delta_0 \varphi d\tau \right\| \\ &= \left\| e^{(aI+bK)t} z_{10}^1 + \int_0^{t-t_0} e^{(aI+bK)\tau} \delta_0 \varphi d\tau \right\| \\ &= \left\| e^{(aI+bK)t} z_{10}^1 + [e^{(aI+bK)(t-t_0)} - I] \frac{aI - bK}{a^2 + b^2} \delta_0 \varphi \right\| \\ &= \|e^{(aI+bK)(t-t_0)} [e^{at_0} e^{bKt_0} z_{10}^1 - \delta_0 \psi] + \delta_0 \psi\| \\ &= \delta_0 \|\psi\| \equiv e^{at_0} \|z_{10}^1\| > 0. \end{aligned}$$

Hence, the game is strictly evadable in this subcase.

(2) $z_{10}^1 = 0$.

Since $d(M, z) = \|z_{10}^1\| > 0$, there exists a $\theta \in (0, 2\pi/b)$ such that (6.19) holds for $t \in [0, \theta]$ if we take $w(t) \equiv w_0$ where w_0 is as above. On the other hand, by choosing such an evasion control, we have

$$(6.37) \quad \begin{aligned} z_1^1(\theta) &= \int_0^\theta e^{(aI+bK)(\theta-\tau)} C_0 w_0 d\tau \\ &= \int_0^\theta e^{(aI+bK)\tau} d\tau \varphi \\ &= [e^{(aI+bK)\theta} - I] \frac{aI - bK}{a^2 + b^2} \varphi \neq 0. \end{aligned}$$

Then, subcase (1) applies. \square

Remark 6.8. From the proof of Theorem 6.6, we see that we only need the condition that there exists a subspace $M \subseteq \mathbb{R}^n$, with $\dim M_0 < n$, $M \subset M_0$, $AM_0 \subset M_0$, and $[A^*|M_0^\perp]^*$, $(C^*|M_0^\perp)^*$ is completely controllable and $((C^*|M_0^\perp)^*, V)$ is δ -superior to $((B^*|M_0^\perp)^*, U)$ for some $\delta > 0$.

Remark 6.9. The condition that $((C^*|M^\perp)^*, V)$ is δ -superior to $((B^*|M^\perp)^*, U)$, which we imposed in Theorem 6.6, does not imply condition (5.13), which we imposed in Corollary 5.2. We will illustrate this point by showing an example in the next section.

7. An example. We present an example in this section.

Consider two objects in \mathbb{R}^3 , one pursuing the other. Suppose the pursuer is x and the evader is y and that they are subject to the following systems:

$$(7.1) \quad \begin{aligned} \dot{x} &= p, \\ \dot{p} &= A_0x + B_0p + C_1u, \end{aligned}$$

$$(7.2) \quad \begin{aligned} \dot{y} &= q, \\ \dot{q} &= A_0y + B_0q + C_2v. \end{aligned}$$

The pursuer wants to “softly” catch the evader; namely, the terminal set M is

$$(7.3) \quad M = \{x = y, p = q\}.$$

The problem models an aerial dogfight in which the evader does not want the pursuer to “get on his tail.” The evader is willing to risk a collision, presumably secure in the belief that the pursuer would not be suicidal.

Now, let us formulate the problem into a standard one. We set

$$(7.4) \quad \begin{aligned} z_1 &= x - y, & z_2 &= p - q, \\ z_3 &= x + y, & z_4 &= p + q. \end{aligned}$$

Then, (7.1) and (7.2) become

$$(7.5) \quad z = \begin{pmatrix} 0 & I & 0 & 0 \\ A_0 & B_0 & 0 & 0 \\ 0 & 0 & 0 & I \\ 0 & 0 & A_0 & B_0 \end{pmatrix} z + \begin{pmatrix} 0 \\ C_1 \\ 0 \\ C_1 \end{pmatrix} u + \begin{pmatrix} 0 \\ -C_2 \\ 0 \\ C_2 \end{pmatrix} v \equiv Az + Bu + Cv$$

and the terminal set is

$$(7.6) \quad M = \{(0, 0, z_3, z_4) \mid z_3, z_4 \in \mathbb{R}^3\}.$$

It is clear that

$$(7.7) \quad AM \subseteq M$$

and (5.13) does not hold for any δ_0 . However, we have the following proposition.

PROPOSITION 7.1. *Suppose*

$$\left[\begin{pmatrix} 0 & I \\ A_0 & B_0 \end{pmatrix}, \begin{pmatrix} 0 \\ -C_2 \end{pmatrix} \right]$$

is completely controllable and (C_2, V) is δ -superior to $(-C_1, U)$. Then “soft” capture can be avoided.

The proof is immediate from Theorem 6.6.

Acknowledgments. The author thanks Professor L. D. Berkovitz for posing the problem, for many suggestive discussions and for encouragement in writing this paper. The author would also like to thank the referee for the suggestions and criticism which led to this much nicer version of the paper.

REFERENCES

- [1] N. P. BHATIA AND G. P. SZEGŐ, *Stability Theory of Dynamical Systems*, Springer-Verlag, New York, Heidelberg, Berlin, 1970.
- [2] A. A. ČIKRII, *The evasion problem in nonlinear differential games*, Kibernetika (Kiev), 11 (1975), pp. 65–68. (In Russian.) Cybernetics, 11 (1975), pp. 412–415. (In English.)
- [3] R. V. GAMKRELIDZE AND G. L. KHARALISHVILI, *A differential game of evasion with nonlinear control*, this Journal, 12 (1974), pp. 332–349.
- [4] P. B. GUSIATINKOV, *On the l-evasion of contact in a linear differential game*, Prikl. Mat. Mikh., 38 (1974), pp. 417–421. (In Russian.) J. Appl. Math. Mech., 38 (1974), pp. 387–392. (In English.)
- [5] ———, *On a problem of l-escape*, Prikl. Mat. Mikh., 40 (1976), pp. 25–37. (In Russian.) J. Appl. Math. Mech., 40 (1976), pp. 20–31. (In English.)
- [6] ———, *l-evasion in linear differential games*, Differentsial 'nye Uravneniya, 12 (1976), pp. 446–455. (In Russian.) Differential Equations, 12 (1976), pp. 311–318. (In English.)
- [7] R. B. HOLMES, *A Course on Optimization and Best Approximation*, Lecture Notes in Mathematics 257, Springer-Verlag, Berlin, Heidelberg, New York, 1972.
- [8] V. N. LAGUNOV, *A nonlinear differential game of evasion*, Dokl. Akad. Nauk. SSSR, 202 (1972), pp. 522–525. (In Russian.) Soviet Math. Dokl., 13 (1972), pp. 131–135. (In English.)
- [9] E. F. MISHCHENKO, *On the problem of evading the encounter in differential games*, this Journal, 12 (1974), pp. 300–310.
- [10] E. F. MISHCHENKO AND N. SATIMOV, *On the possibility of evading encounter in differential games in a critical case*, Trudy Mat. Inst. Steklov, 158 (1981), pp. 121–124. (In Russian.) Proc. Steklov Inst. Math., 158 (1981), pp. 131–134. (In English.)
- [11] ———, *The contact avoidance problem in differential games with nonlinear controls*, Differentsial 'nye Uravneniya, 9 (1973), pp. 1792–1797. (In Russian.) Differential Equations, 9 (1973), pp. 1377–1381. (In English.)
- [12] M. S. NIKOL'SKII, *On a linear escape problem*, Dokl. Akad. Nauk SSSR, 218 (1974), pp. 1024–1027. (In Russian.) Soviet Math. Dokl., 15 (1974), pp. 1462–1466. (In English.)
- [13] V. V. OSTAPENKO, *A nonlinear escape problem*, Kibernetika (Kiev), 14 (1978), pp. 106–112. (In Russian.) Cybernetics, 14 (1978), pp. 594–601. (In English.)
- [14] ———, *A nonautonomous evasion problem*, Avtomat. i Telemekh., 43 (1982), pp. 81–86. (In Russian.) Automat. Remote Control, 43 (1982), pp. 768–773. (In English.)
- [15] L. S. PONTRYAGIN, *On the evasion process in differential games*, Appl. Math. Optim., 1 (1974), pp. 5–19.
- [16] ———, *A linear differential escape game*, Trudy Mat. Inst. Steklov, 112 (1971), pp. 30–63. (In Russian.) Proc. Steklov Inst. Math., 112 (1971), pp. 27–60. (In English.)
- [17] L. S. PONTRYAGIN AND E. F. MISHCHENKO, *A problem on the escape of one controlled object from another*, Dokl. Akad. Nauk SSSR, 189 (1969), pp. 721–723. (In Russian.) Soviet Math. Dokl., 10 (1969), pp. 1488–1490. (In English.)
- [18] ———, *The contact avoidance problem in linear differential games*, Differentsial 'nye Uravneniya, 7 (1971), pp. 436–445. (In Russian.) Differential Equations, 7 (1971), pp. 335–352. (In English.)
- [19] B. N. PSHENICHNYI, *Convex programming in normed spaces*, Kibernetika (Kiev), 1 (1965), pp. 46–54. (In Russian.) Cybernetics, 1 (1965), pp. 46–57. (In English.)
- [20] ———, *The flight problem*, Kibernetika (Kiev), 11 (1975), pp. 120–127. (In Russian.) Cybernetics, 11 (1975), pp. 642–651. (In English.)
- [21] N. ROUCHE, P. HABETS AND M. LALOY, *Stability Theory by Liapunov's Direct Method*, Springer-Verlag, New York, Heidelberg, Berlin, 1977.
- [22] N. SATIMOV, *On the escape problem in differential games with nonlinear controls*, Dokl. Akad. Nauk SSSR, 216 (1974), pp. 744–747. (In Russian.) Soviet Math. Dokl., 15 (1974), pp. 886–890. (In English.)
- [23] ———, *The escape problem for a class of nonlinear differential games*, Differentsial 'nye Uravneniya, 11 (1975), pp. 672–677. (In Russian.) Differential Equations, 11 (1975), pp. 506–510. (In English.)
- [24] ———, *On a way to avoid contact in differential games*, Mat. Sb., 99 (1976), pp. 380–393. (In Russian.) Math. USSR-Sb., 28 (1976), pp. 339–352. (In English.)
- [25] ———, *On the theory of differential games of escape*, Mat. Sb., 103 (1977), pp. 432–444. (In Russian.) Math. USSR-Sb., 32 (1977), pp. 371–383. (In English.)

OPTIMAL PERIODIC CONTROL FOR THE TWO-PHASE STEFAN PROBLEM*

AVNER FRIEDMAN[†], SHAOYUN HUANG[‡] AND JIONGMIN YONG[†]

Abstract. Consider the two-phase Stefan problem in a domain $\{(x, t); x \in D, 0 < t < \infty\}$ with the heat flux across a part Γ_1 of ∂D being a control function $k(x, t)$ periodic in t of period σ , and $N_1 \leq k \leq N_2$, $\int_0^\sigma \int_{\Gamma_1} k(x, t) = M$, where N_1, N_2, M are given positive constants. The solution $u(x, t)$ behaves asymptotically as a periodic function $\hat{u}(x, t)$. We wish to maximize $\int_0^\sigma \int_D p(x) \hat{u}(x, t)$ where p is a given positive function. It is proved that any optimal control k_0 has the form

$$k_0(x, t) = \begin{cases} N_2 & \text{if } 0 < t < \varphi(x), \\ N_1 & \text{if } \varphi(x) < t < \sigma \end{cases}$$

where $\varphi(x)$ is a smooth function.

Key words. two-phase Stefan problems, optimal control, periodic solutions, bang-bang principle

AMS(MOS) subject classifications. 35K85, 35R35, 49A22, 49A29, 49A36

Introduction. Consider the two-phase Stefan problem for the temperature u :

$$(0.1) \quad \frac{\partial}{\partial t} (u + H(u)) - \Delta u = 0 \quad \text{in } D \times \{0 < t < \infty\}$$

with

$$(0.2) \quad \frac{\partial u}{\partial \nu} = k(x, t) \quad \text{for } x \in \Gamma_1, \quad 0 < t < \infty \quad (k < 0),$$

$$(0.3) \quad \frac{\partial u}{\partial \nu} = -g(x, t) \quad \text{for } x \in \Gamma_2, \quad 0 < t < \infty \quad (g > 0)$$

or

$$(0.2') \quad u = k(x, t) \quad \text{for } x \in \Gamma_1, \quad 0 < t < \infty,$$

$$(0.3') \quad u = -g(x, t) \quad \text{for } x \in \Gamma_2, \quad 0 < t < \infty$$

where $\Gamma_1 \cap \Gamma_2 = \emptyset$, $\Gamma_1 \cup \Gamma_2 = \partial D$ and

$$(0.4) \quad u(x, 0) = u_0(x) \quad \text{for } x \in D;$$

here $H(u)$ is the Heaviside function. It is well known (for the Dirichlet boundary conditions (0.2'), (0.3')) that the above problem has a unique global solution [2], [4], [5], [9]. It was recently proved by DiBenedetto and Friedman [3] that if k and g are periodic in t with period σ , then there exists a σ -periodic solution $\hat{u}(x, t)$, independent of $u_0(x)$, such that

$$(0.5) \quad \sup_{x \in D} |u(x, t) - \hat{u}(x, t)| \rightarrow 0 \quad \text{if } t \rightarrow \infty.$$

We now consider k as a control function in a class

$$(0.6) \quad \mathcal{A}_\sigma = \left\{ k; N_1 \leq k \leq N_2, \int_0^\sigma \int_D k(x, t) = M, k(x, t + \sigma) = k(x, t) \right\},$$

* Received by the editors April 14, 1986; accepted for publication December 31, 1986. This work was partially supported by National Science Foundation grants DMS-8501397 and DMS-8420896.

[†] Purdue University, Center for Applied Mathematics, West Lafayette, Indiana 47907.

[‡] Beijing University, Beijing, People's Republic of China. Present address: Department of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

and introduce a functional

$$(0.7) \quad J(k) = \int_0^\sigma \int_D p(x) \hat{u}(x, t) \, dx \, dt$$

and the optimization problem

$$(0.8) \quad J(k_0) = \max_{k \in \mathcal{A}_\sigma} J(k), \quad k_0 \in \mathcal{A}_\sigma.$$

In this paper we analyze the structure of the optimal controls k_0 . We prove that

$$(0.9) \quad k_0(x, t) = \begin{cases} N_2 & \text{if } m\sigma < t < m\sigma + \varphi(x), \\ N_1 & \text{if } m\sigma + \varphi(x) < t < (m+1)\sigma, \end{cases}$$

for $m = 0, 1, 2, \dots$, where $\varphi(x)$ is a smooth function. This result is established not only for the Dirichlet data (0.2'), (0.3') but also for other data, such as (0.2), (0.3') (in which case one must impose restrictions on N_1, N_2 in order to ensure that the free boundary does not meet the fixed boundary ∂D at any time).

The control problem is motivated by the following situation which occurs in a frozen body of sea water: An object D_1 (a pipeline, for instance) is placed in the frozen water and its boundary Γ_1 is heated, causing the ice to melt in the vicinity of D_1 . The temperature of the ice varies with the climate and is thus a periodic function (either on a scale of one year or on a scale of one day). It is therefore natural to choose for the flux k on Γ_1 as a periodic control. The functional (0.7), to be maximized, represents the amount of heat in both the water and ice regions with a weight function p . Since one is usually interested in raising the amount of heat in the melted ice rather than in raising it in the ice, one may choose $p(x)$ to decrease away from D_1 .

In § 1 we briefly establish the asymptotic behavior for the system (0.1), (0.2), (0.3'), (0.4), thereby extending the results of [3] to the case when (0.2') is replaced by (0.2). In § 2 we consider an intermediate control problem for (0.1), (0.2), (0.3'), (0.4) with

$$J(k) = \int_0^T \int_D p(x) u(x, t);$$

here k and g are not assumed to be periodic in t . We establish a bang-bang principle similar to (0.9) for any maximizer k_0 .

For the one-phase Stefan problem a bang-bang principle was established (in [6], [8]) for the functional $J(k)$ representing the volume of ice that has melted by time T . At the end of § 2 we show that such a principle is false for the two-phase Stefan problem.

The analysis of § 2 is used in § 3 where the assertion (0.9) is established. Finally, in § 4 we extend the results of § 3 to the Stefan problem where the boundary condition on Γ_1 is either of the form (0.2) or (0.2') and the boundary condition on Γ_2 is either of the form (0.3) or (0.3'). Here, the fact that the Dirichlet data in (0.2) are in L^∞ , but not continuous in general, requires some approximating procedure.

1. Existence and uniqueness of periodic solutions (under the condition (0.2), (0.3')). Let Γ_1 and Γ_2 be $C^{2+\alpha}$ closed hypersurfaces in \mathbb{R}^N such that Γ_1 lies in the interior of the set enclosed by Γ_2 . Denote by D the domain bounded by Γ_1, Γ_2 , and set $D_T = D \times \{0 < t < T\}$, $\Gamma_{i,T} = \Gamma_i \times \{0 < t < T\}$ for any $T \leq \infty$.

We are given a function $g(x, t)$ satisfying

$$(1.1) \quad \begin{aligned} &g \in L^\infty(D_\infty), D_x g \text{ and } D_t g \text{ belong to } L^\infty(D_\infty), \\ &0 < c_1 \leq g \leq c_2 < \infty \text{ (} c_1, c_2 \text{ are constants);} \end{aligned}$$

a function $h(x)$ satisfying

$$(1.2) \quad h \in C^0(\bar{D}), \quad h = -g \quad \text{on } \Gamma_2 \times \{0\};$$

and a function $k(x, t)$ satisfying

$$(1.3) \quad 0 < N_1 \leq k(x, t) \leq N_2 < \infty \quad \text{on } \Gamma_{1,\infty} \quad (N_1, N_2 \text{ are constants}).$$

The two-phase Stefan problem for the temperature u is formally written in the form

$$(1.4) \quad \begin{aligned} \frac{\partial}{\partial t} (u + H(u)) - \Delta u &= 0 \quad \text{in } \mathcal{D}'(D_\infty), \\ \frac{\partial u}{\partial \nu} &= k \quad \text{on } \Gamma_{1,\infty}, \\ u &= -g \quad \text{on } \Gamma_{2,\infty}, \\ u(x, 0) &= h(x) \quad \text{if } x \in D \end{aligned}$$

where ν is the outward normal to D and H is the Heaviside function. More precisely, by a solution of (1.4) we mean a *weak solution* in the following sense.

For any $0 < T < \infty$ there is a pair (u, ξ) with $\xi \subset H(u)$ such that

$$(1.5) \quad u \in L^\infty((0, T); L^2(D)) \cap L^2((0, T); H^1(D)),$$

$$(1.6) \quad u = -g \quad \text{on } \Gamma_{2,T},$$

$$(1.7) \quad \int_0^T \int_D [-(u + \xi)\varphi_t + \nabla u \cdot \nabla \varphi] dx dt = \int_0^T \int_{\Gamma_1} k\varphi dS dt + \int_D (h + \xi_0)\varphi(x, 0) dx$$

for every $\varphi \in V$. Here ξ_0 is a given initial data with $\xi_0 \subset H(h)$ and

$$V = \{\varphi \in H^1(D_T), \varphi = 0 \text{ on } \Gamma_2 \times (0, T), \varphi = 0 \text{ on } t = T\}.$$

In order to construct a solution of (1.5)–(1.7) we follow the procedure in [3] and introduce, for any $\varepsilon > 0$,

$$H_\varepsilon(s) = \begin{cases} 1 & \text{if } s > \varepsilon, \\ s/\varepsilon & \text{if } 0 < s \leq \varepsilon, \\ 0 & \text{if } s \leq 0. \end{cases}$$

Define

$$(1.8) \quad h_\varepsilon(x) = \begin{cases} h(x) & \text{if } h(x) > \varepsilon \text{ or } h(x) < 0, \\ \varepsilon \xi_0 & \text{if } 0 \leq h(x) \leq \varepsilon. \end{cases}$$

Then

$$(1.9) \quad H_\varepsilon(h_\varepsilon) = \xi_0.$$

Consider the ε -approximate problems

$$(1.10) \quad \begin{aligned} \frac{\partial}{\partial t} (u_\varepsilon + H_\varepsilon(u_\varepsilon)) - \Delta u_\varepsilon &= 0 \quad \text{in } D_T, \\ \frac{\partial u_\varepsilon}{\partial \nu} &= k \quad \text{on } \Gamma_{1,T}, \\ u_\varepsilon &= -g \quad \text{on } \Gamma_{2,T}, \\ u_\varepsilon &= h_\varepsilon \quad \text{on } D \times \{0\}. \end{aligned}$$

Existence for (1.10) can be established by a fixed point argument using standard parabolic estimates. Uniqueness follows from the following more general stability result.

LEMMA 1.1. *Suppose \hat{u}_ε is a solution of (1.10) corresponding to data $\hat{k}, \hat{g}, \hat{h}_\varepsilon$. If $\hat{g} \cong g$ on $\Gamma_{2,T}$, then*

$$(1.11) \quad \begin{aligned} & \int_{D \times \{t\}} [\hat{u}_\varepsilon - u_\varepsilon + H_\varepsilon(\hat{u}_\varepsilon) - H_\varepsilon(u_\varepsilon)]^+ \\ & \leq \int_D [\hat{h}_\varepsilon - h_\varepsilon + H_\varepsilon(\hat{h}_\varepsilon) - H_\varepsilon(h_\varepsilon)]^+ + \int_0^t \int_{\Gamma_1} [\hat{k} - k]^+. \end{aligned}$$

Proof. We take the difference of the parabolic equations for \hat{u}_ε and u_ε , multiply by $H_\lambda(\hat{u}_\varepsilon - u_\varepsilon)$ and integrate over D_t . After an integration by parts, we let $\lambda \rightarrow 0$, and (1.11) follows.

Take any small positive constant δ_0 and denote by $V_0(x)$ the solution of

$$(1.12) \quad \begin{aligned} \Delta V_0 &= 0 \quad \text{in } D, \\ V_0 &= \delta_0 \quad \text{on } \Gamma_1, \\ V_0 &= -c_2 \quad \text{on } \Gamma_2; \end{aligned}$$

c_2 is as in (1.1). By the maximum principle,

$$(1.13) \quad \frac{\partial V_0}{\partial \nu} \geq N_1 > 0 \quad \text{on } \Gamma_1 \quad (N_1 \text{ constant}).$$

We shall henceforth assume that $k(x, t)$ satisfies (1.3) with N_1 as in (1.13); N_2 is an arbitrary constant.

Denote by $V_1(x)$ the solution of

$$(1.14) \quad \begin{aligned} \Delta V_1 &= 0 \quad \text{in } D, \\ \frac{\partial V_1}{\partial \nu} &= N_2 \quad \text{on } \Gamma_1, \\ V_1 &= -c_1 \quad \text{on } \Gamma_2. \end{aligned}$$

By the maximum principle

$$(1.15) \quad V_0(x) < V_1(x) \quad \text{in } \bar{D}.$$

From Lemma 1.1 we obtain the following.

LEMMA 1.2. *There holds*

$$\begin{aligned} \int_{D \times \{t\}} [V_0 - u_\varepsilon + H_\varepsilon(V_0) - H_\varepsilon(u_\varepsilon)]^+ &\leq \int_D [V_0 - h_\varepsilon + H_\varepsilon(V_0) - H_\varepsilon(h_\varepsilon)]^+, \\ \int_{D \times \{t\}} [u_\varepsilon - V_1 + H_\varepsilon(u_\varepsilon) - H_\varepsilon(V_1)]^+ &\leq \int_D [h_\varepsilon - V_1 + H_\varepsilon(h_\varepsilon) - H_\varepsilon(V_1)]^+. \end{aligned}$$

Using the same method as in [3] we can deduce the following estimates for the solution of (1.10).

LEMMA 1.3. *There holds*

$$(1.16) \quad \begin{aligned} \|u_\varepsilon\|_{L^\infty(D_T)} &\leq C, \\ \|\nabla u_\varepsilon\|_{L^2(D_T)} &\leq C(T), \\ \left\| \frac{\partial}{\partial t} u_\varepsilon \right\|_{L^2(E)} &\leq C(T, \text{dist}(E, \partial D_T)) \end{aligned}$$

for any compact set $E \subset D \times (0, T)$; the constants C are independent of ε .

It follows that the family $\{u_\varepsilon\}$ is precompact in $L^2(D_T)$ and that, for a subsequence $\varepsilon \rightarrow 0$,

$$(1.17) \quad \begin{aligned} u_\varepsilon &\rightarrow u \quad \text{in } L^2(D_T) \text{ and a.e. in } D_T, \\ \nabla u_\varepsilon &\rightarrow \nabla u \quad \text{weakly in } L^2(D_T), \\ H_\varepsilon(u_\varepsilon) &\rightarrow \xi \quad \text{weakly in } L^2(D_T). \end{aligned}$$

Using Lemma 1.2 and arguing as in [3] we can establish the following theorem.

THEOREM 1.4. *If*

$$(1.18) \quad V_0(x) \leq h(x) \leq V_1(x),$$

$$(1.19) \quad H(V_0) \leq \xi_0 \leq H(V_1)$$

in D , then there exists a solution (u, ξ) of (1.5)–(1.7) and

$$(1.20) \quad V_0(x) \leq u(x, t) \leq V_1(x),$$

$$(1.21) \quad H(V_0) \leq \xi(x, t) \leq H(V_1)$$

in D_T .

Uniqueness of the solution follows from the more general stability result.

THEOREM 1.5. *Let $(\tilde{u}, \tilde{\xi})$ be a solution of (1.5)–(1.7) for data $\tilde{k}, \tilde{g}, \tilde{h}, \tilde{\xi}_0$ satisfying the same conditions as k, g, h, ξ_0 above. Then*

$$(1.22) \quad \begin{aligned} &\iint_{D_T} [|u - \tilde{u}|^2 + (\xi - \tilde{\xi})(u - \tilde{u})] \\ &\leq C \int_0^T \int_{\Gamma_1} |k - \tilde{k}| + C \int_0^T \int_{\Gamma_2} |g - \tilde{g}| + C \int_D |h - \tilde{h} + \xi_0 - \tilde{\xi}_0| \end{aligned}$$

where C is a positive constant independent of the data and of T ; C depends on N_1, N_2, c_1, c_2 .

The proof is similar to the proof given in [4] (see also [3], [5]) for Dirichlet data. The fact that C is independent of T will be of crucial importance in the sequel.

We next consider the periodic case

$$(1.23) \quad \begin{aligned} k(x, t + \sigma) &= k(x, t) \quad \text{on } \Gamma_1 \times (0, \infty), \\ g(x, t + \sigma) &= g(x, t) \quad \text{on } \Gamma_2 \times (0, \infty) \end{aligned}$$

for some $\sigma > 0$.

LEMMA 1.6. *If (1.23) holds then there exists a function \hat{h}_ε with $V_0(x) \leq \hat{h}_\varepsilon(x) \leq V_1(x)$ in D such that the corresponding solution \hat{u}_ε of the ε -approximating problems (1.10) is periodic in t of period σ .*

The proof which is similar to the proof of [3, Theorem 7.1] uses the Schauder fixed point theorem.

Taking a convergent subsequence of the \hat{u}_ε 's, we obtain the following theorem.

THEOREM 1.7. *If (1.23) holds then there exists a function \hat{h} and a selection $\hat{\xi}_0 \in H(\hat{h})$ with $V_0(x) \leq \hat{h}(x) \leq V_1(x)$, $H(V_0) \leq \hat{\xi}_0 \leq H(V_1)$ in D such that there exists a solution $(\hat{u}, \hat{\xi})$ of (1.5)–(1.7), with $h = \hat{h}$, $\xi_0 = \hat{\xi}_0$, $u = \hat{u}$, which is periodic in t of period σ , and*

$$V_0(x) \leq \hat{u}(x, t) \leq V_1(x),$$

$$H(V_0) \leq \hat{\xi}(x, t) \leq H(V_1)$$

in D_∞ ; the function \hat{u} is uniquely determined by k, g .

The last assertion (about uniqueness) follows from Theorem 1.5. Indeed, if (\tilde{h}, \tilde{u}) is another periodic solution then (1.22) implies that

$$\int_0^\infty \int_D |\hat{u}(x, t) - \tilde{u}(x, t)|^2 dx dt < \infty.$$

Since, however, $\hat{u} - \tilde{u}$ is periodic in t of period σ , it follows that $\hat{u} - \tilde{u} \equiv 0$.

2. Optimal control in finite time. We now consider k to be a control variable belonging to the set

$$\mathcal{A} = \left\{ k \in L^\infty(\Gamma_1 \times (0, T)), N_1 \leq k \leq N_2, \int_0^T \int_{\Gamma_1} k = M \right\}$$

with N_1, N_2 as in § 1; we assume that

$$(2.1) \quad N_1 T \text{ meas}(\Gamma_1) < M < N_2 T \text{ meas}(\Gamma_1),$$

for otherwise the set \mathcal{A} is empty or reduces to one element. Consider the Stefan problem (1.5)–(1.7) written in the form

$$(2.2) \quad \begin{aligned} \frac{\partial}{\partial t} \beta(u) - \Delta u &= 0 \quad \text{in } \mathcal{D}'(D_T), \\ u_\nu &= k \quad \text{on } \Gamma_{1,T}, \\ u &= -g \quad \text{on } \Gamma_{2,T}, \\ \beta(u) &= \beta(h) \quad \text{on } t=0 \end{aligned}$$

where $\beta(u) = u + H(u)$, and set

$$K_T = \{(k, u); k \in \mathcal{A}, u \text{ is the unique solution of (2.2) corresponding to } k\}.$$

Let $p(x)$ be a given function satisfying

$$(2.3) \quad \begin{aligned} p &\text{ is piecewise continuous in } \bar{D}, \\ p &\geq 0 \quad \text{in } \bar{D}, \quad p > 0 \quad \text{on } \Gamma_1. \end{aligned}$$

We introduce the functional

$$(2.4) \quad J(k) = \int_0^T \int_D p(x) u(x, t) dx dt$$

and consider the following maximization problem.

Problem (F). Find k_0 such that

$$J(k_0) = \max_{k \in \mathcal{A}} J(k), \quad k_0 \in \mathcal{A}.$$

THEOREM 2.1. *There exists a solution of Problem (F).*

Proof. Let $(k_m, u_m) \in K_T$ be a maximizing sequence. From the estimates in § 1 (cf. Lemma 1.3) it follows that, for a subsequence,

$$\begin{aligned} k_m &\rightarrow k_0 \quad \text{weakly,} \\ u_m &\rightarrow u_0 \quad \text{weakly in } H^1(D_T) \text{ and a.e. in } D_T. \end{aligned}$$

It is now easily seen that $(k_0, u_0) \in K_T$ and that k_0 is a maximizer.

We shall henceforth consider a particular maximizer k_0 and proceed to study its structure. For this purpose we set $\beta_\varepsilon(u) = u + H_\varepsilon(u)$ and again consider the ε -approximating problems

$$(2.5) \quad \begin{aligned} \frac{\partial}{\partial t} \beta_\varepsilon(u) - \Delta u &= 0 \quad \text{in } D_T, \\ u_\nu &= k \quad \text{on } \Gamma_{1,T}, \\ u &= -g \quad \text{on } \Gamma_{2,T}, \\ u &= h \quad \text{on } t=0. \end{aligned}$$

We define

$K_T^\varepsilon = \{(k, u); k \in \mathcal{A}, u \text{ is the unique solution of (2.5) corresponding to } k\}$, and (as in [1], [6]) introduce the functional

$$J_\varepsilon(k) = \int_0^T \int_D p(x)u(x, t) dx dt - \frac{1}{2} \int_0^T \int_{\Gamma_1} (k - k_0)^2 dS dt$$

for any $(k, u) \in K_T^\varepsilon$.

Problem (F_ε) . Find $(k_\varepsilon, u_\varepsilon) \in K_T^\varepsilon$ such that

$$J_\varepsilon(k_\varepsilon) = \max_{(k, u) \in K_T^\varepsilon} J_\varepsilon(k).$$

Similarly to Theorem 2.1 one can prove that there exists a solution $(k_\varepsilon, u_\varepsilon)$ of problem (F_ε) .

LEMMA 2.2. *If $(k_\varepsilon, u_\varepsilon) \in K_T^\varepsilon$ is a maximizer of problem (F_ε) then, as $\varepsilon \rightarrow 0$,*

$$(2.6) \quad k_\varepsilon \rightarrow k_0 \quad \text{in } L^2(\Gamma_1 \times (0, T)),$$

$$(2.7) \quad u_\varepsilon \rightarrow u_0 \quad \text{weakly in } L^2((0, T); H^1(D)) \text{ and a.e. in } D_T.$$

Proof. For a subsequence,

$$(2.8) \quad \begin{aligned} k_\varepsilon &\rightarrow \tilde{k} \quad \text{weakly in } L^2(\Gamma_1 \times (0, T)), \\ u_\varepsilon &\rightarrow \tilde{u} \quad \text{weakly in } L^2((0, T); H^1(D)) \text{ and a.e. in } D_T \end{aligned}$$

and $(\tilde{k}, \tilde{u}) \in K_T^\varepsilon$. It suffices to show that $\tilde{k} = k_0$, for then, by uniqueness, also $\tilde{u} = u_0$. Let $(k_0, u_\varepsilon^*) \in K_T^\varepsilon$. By the estimates (1.16) of § 1,

$$(2.9) \quad u_\varepsilon^* \rightarrow \tilde{u}_0 \quad \text{weakly in } L^2((0, T); H^1(D)) \text{ and a.e. in } D_T$$

where $(k_0, \tilde{u}_0) \in K_T$ and $\tilde{u}_0 = u_0$ by uniqueness. We have

$$\begin{aligned} J(k_0) &\cong J(\tilde{k}) = \int_{D_T} p\tilde{u} \quad (\text{by maximality of } k_0) \\ &= \lim_{\varepsilon \rightarrow 0} \left[\int_{D_T} p u_\varepsilon - \frac{1}{2} \int_{\Gamma_{1,T}} (k_\varepsilon - k_0)^2 + \frac{1}{2} \int_{\Gamma_{1,T}} (k_\varepsilon - k_0)^2 \right] \quad (\text{by (2.8)}) \\ &\cong \liminf_{\varepsilon \rightarrow 0} J_\varepsilon(k_\varepsilon) + \frac{1}{2} \limsup_{\varepsilon \rightarrow 0} \int_{\Gamma_{1,T}} (k_\varepsilon - k_0)^2 \\ &\cong \liminf_{\varepsilon \rightarrow 0} J_\varepsilon(k_0) + \frac{1}{2} \limsup_{\varepsilon \rightarrow 0} \int_{\Gamma_{1,T}} (k_\varepsilon - k_0)^2 \quad (\text{by maximality of } k_\varepsilon) \\ &= \lim_{\varepsilon \rightarrow 0} \int_{D_T} p u_\varepsilon^* + \frac{1}{2} \limsup_{\varepsilon \rightarrow 0} \int_{\Gamma_{1,T}} (k_\varepsilon - k_0)^2 \\ &= J(k_0) + \frac{1}{2} \limsup_{\varepsilon \rightarrow 0} \int_{\Gamma_{1,T}} (k_\varepsilon - k_0)^2 \end{aligned}$$

by (2.9) and $\tilde{u}_0 = u_0$. It follows that (2.6) holds.

We now proceed to analyze any solution $(k_\varepsilon, u_\varepsilon) \in K_T^\varepsilon$ of Problem (F_ε) . Let l be a function such that $k_\varepsilon + \delta l \in \mathcal{A}$ for all small enough $\delta > 0$, and let $(k_\varepsilon + \delta l, u_{\varepsilon,\delta}) \in K_T^\varepsilon$. Then we have

$$(2.10) \quad \begin{aligned} 0 &\leq J_\varepsilon(k_\varepsilon + \delta l) - J_\varepsilon(k_\varepsilon) \\ &= \int_{D_T} p(u_{\varepsilon,\delta} - u_\varepsilon) - \delta \int_0^T \int_{\Gamma_1} (k_\varepsilon - k_0)l - \frac{\delta^2}{2} \int_0^T \int_{\Gamma_1} l^2. \end{aligned}$$

For fixed ε one can establish by standard parabolic estimates that

$$\left\| \frac{u_{\varepsilon,\delta} - u_\varepsilon}{\delta} \right\|_{L^2((0,T); H^1(D))} \leq C \quad \text{as } \delta \rightarrow 0,$$

and, in fact,

$$(2.11) \quad \frac{u_{\varepsilon,\delta} - u_\varepsilon}{\delta} \rightharpoonup z_\varepsilon \quad \text{weakly in } L^2((0, T); H^1(D)) \text{ and a.e. in } D_T$$

where z_ε is the solution of

$$(2.12) \quad \begin{aligned} \frac{\partial}{\partial t} (\beta'_\varepsilon(u_\varepsilon) z_\varepsilon) - \Delta z_\varepsilon &= 0 \quad \text{in } D_T, \\ \frac{\partial}{\partial \nu} z_\varepsilon &= l \quad \text{on } \Gamma_{1,T}, \\ z_\varepsilon &= 0 \quad \text{on } \Gamma_{2,T}, \\ z_\varepsilon &= 0 \quad \text{on } t = 0. \end{aligned}$$

Denote by Q_ε the solution of the parabolic problem

$$(2.13) \quad \begin{aligned} \beta'_\varepsilon(u_\varepsilon) \frac{\partial Q_\varepsilon}{\partial t} + \Delta Q_\varepsilon &= p \quad \text{in } D_T, \\ \frac{\partial}{\partial \nu} Q_\varepsilon &= 0 \quad \text{on } \Gamma_{1,T}, \\ Q_\varepsilon &= 0 \quad \text{on } \Gamma_{2,T}, \\ Q_\varepsilon &= 0 \quad \text{on } t = T. \end{aligned}$$

Then

$$\begin{aligned} 0 &= \int_{D_T} \left[\frac{\partial}{\partial t} (\beta'_\varepsilon(u_\varepsilon) z_\varepsilon) - \Delta z_\varepsilon \right] Q_\varepsilon \\ &= \int_D \beta'_\varepsilon(u_\varepsilon) z_\varepsilon Q_\varepsilon \Big|_0^T - \int_{D_T} \beta'_\varepsilon(u_\varepsilon) z_\varepsilon Q_{\varepsilon,t} - \int_{D_T} z_\varepsilon \Delta Q_\varepsilon \\ &\quad - \int_0^T \int_{\partial D} [z_{\varepsilon,\nu} Q_\varepsilon - z_\varepsilon Q_{\varepsilon,\nu}] \\ &= - \int_{D_T} p z_\varepsilon - \int_0^T \int_{\Gamma_1} l Q_\varepsilon \quad \text{by (2.12), (2.13)}. \end{aligned}$$

Using this and (2.10), (2.11), we find that

$$\begin{aligned} 0 &\geq \limsup_{\delta \rightarrow 0} \frac{1}{\delta} [J_\varepsilon(k_\varepsilon + \delta l) - J_\varepsilon(k_\varepsilon)] \\ &= \int_{D_T} p z_\varepsilon - \int_{\Gamma_{1,T}} (k_\varepsilon - k_0) l = \int_{\Gamma_{1,T}} l (-Q_\varepsilon - k_\varepsilon + k_0), \end{aligned}$$

i.e.,

$$(2.14) \quad \int_{\Gamma_{1,T}} (Q_\varepsilon + k_\varepsilon - k_0) l \geq 0.$$

From this relation and from (2.6) we can deduce, precisely as in [6], that for any small $\eta > 0$ there is a set Σ_η in $\Gamma_{1,T}$ of measure smaller than η such that

$$(2.15) \quad k_\varepsilon = \begin{cases} N_2 & \text{on } (Q_\varepsilon < \lambda_\varepsilon - \eta) \cap \{\Gamma_{1,T} \setminus \Sigma_\eta\}, \\ N_1 & \text{on } \{Q_\varepsilon > \lambda_\varepsilon + \eta\} \cap \{\Gamma_{1,T} \setminus \Sigma_\eta\}, \end{cases}$$

if ε is small enough; λ_ε is some constant and Σ_η is independent of ε .

We shall analyze Q_ε as $\varepsilon \rightarrow 0$. Observe first that, by comparison,

$$(2.16) \quad \bar{Q} \leq Q_\varepsilon \leq 0$$

where \bar{Q} is the solution of

$$\begin{aligned} \Delta \bar{Q} &= p \quad \text{in } D, \\ \bar{Q}_\nu &= 0 \quad \text{on } \Gamma_1, \quad \bar{Q} = 0 \quad \text{on } \Gamma_2. \end{aligned}$$

Denote by Γ_1^δ the intersection of D with δ -neighborhood of Γ_1 . By comparison, (1.20) holds also for $u = u_\varepsilon$. Hence, by (1.12), (1.14) there holds

$$\beta'_\varepsilon(u_\varepsilon) = 1 \quad \text{in } V_\delta = \Gamma_1^\delta \times (0, T)$$

for some $\delta > 0$ independent of k, ε . From this and from (2.16), (2.13) it follows that, for a sequence $\varepsilon \rightarrow 0$,

$$(2.17) \quad Q_\varepsilon \rightarrow Q \quad \text{uniformly in } V_\delta,$$

and

$$\begin{aligned} Q_t + \Delta Q &= p \quad \text{in } V_\delta, \\ Q_\nu &= 0 \quad \text{on } \Gamma_{1,T}, \\ Q &\leq 0 \quad \text{in } V_\delta, \\ Q &= 0 \quad \text{on } t = T. \end{aligned} \quad (2.18)$$

Next, by differentiation with respect to t of (2.13) we see that the function $\zeta = Q_{\varepsilon,t}$ satisfies

$$\begin{aligned} \beta'_\varepsilon(u_\varepsilon) \zeta_t + \Delta \zeta + \beta''_\varepsilon(u_\varepsilon) u_{\varepsilon,t} \zeta &= 0 \quad \text{in } D_T, \\ \zeta_\nu &= 0 \quad \text{on } \Gamma_{1,T}, \\ \zeta &= 0 \quad \text{on } \Gamma_{2,T}, \\ \zeta(x, T) &\geq 0 \quad \text{if } x \in D \quad (\text{since } Q_\varepsilon \leq 0, Q_\varepsilon(x, T) = 0). \end{aligned}$$

By the maximum principle it follows that $\zeta \geq 0$ in D_T , i.e.,

$$(2.19) \quad Q_{\varepsilon,t} \geq 0 \quad \text{in } D_T.$$

(Actually, since $\beta''_\varepsilon(u)$ is not bounded, one should define $\beta_\varepsilon(u)$ from the outset slightly differently, so that it is in C^2 .) From (2.19), (2.17) it follows that

$$(2.20) \quad Q_t \geq 0 \quad \text{in } V_\delta.$$

We claim that

$$(2.21) \quad Q_t > 0 \quad \text{on } \Gamma_1 \times (0, T).$$

Indeed, otherwise we conclude from (2.20) that Q_t takes local minimum zero at a point (x_0, t_0) on $\Gamma_1 \times (0, T)$. Since (by (2.18))

$$(Q_t)_t + \Delta Q_t = 0 \quad \text{in } V_\delta$$

and

$$Q_{t,\nu}(x_0, t_0) = 0,$$

this is a contradiction to the strong maximum principle for Q_t .

We can now state the main result of this section.

THEOREM 2.3. *For any maximizer k_0 of Problem (F) there exists a function $\varphi(x)$ in $C^{1,\alpha/2}(\Gamma_1)$ such that*

$$(2.22) \quad k_0(x, t) = \begin{cases} N_2 & \text{if } 0 < t < \varphi(x), \\ N_1 & \text{if } \varphi(x) < t < T. \end{cases}$$

Proof. From (2.15) and (2.17) it follows that there exists a $\lambda \leq 0$ such that

$$k_0 = \begin{cases} N_2 & \text{if } Q(x, t) < \lambda, \\ N_1 & \text{if } Q(x, t) > \lambda. \end{cases}$$

From (2.21) we see that the surface

$$\{(x, t) \in \Gamma_1 \times (0, T); Q(x, t) = \lambda\}$$

is given by $t = \varphi(x)$ with φ in $C^{1,\alpha/2}$ (since Q is in $C^{1,\alpha/2}$).

Notice that φ is in $C^{m,\alpha/2}$ if Γ_1 is in $C^{2m,\alpha}$; in particular, $\varphi \in C^\infty$ if $\Gamma_1 \in C^\infty$.

Remark 2.1. Denote by $V(T)$ the measure of the ice that has melted by time T and set $J_0(k) = V(T)$. Consider the problem of maximizing $J_0(k)$ over $k \in \mathcal{A}$. For the one-phase Stefan problems it was proved in [6], [8], that any maximizer k_0 has the form (2.22). We wish to show that this bang-bang principle is false for the two-phase Stefan problem. Consider the one-dimensional Stefan problem

$$\begin{aligned} \frac{\partial \beta(u)}{\partial t} - u_{xx} &= 0 & (0 < x < 1, 0 < t < T), \\ u_x(0, t) &= -k(t) & (0 < t < T), \\ u(1, t) &= -\frac{1}{2} & (0 < t < T), \\ u(x, 0) &= V_0(x) & (0 < x < 1) \end{aligned}$$

where

$$\begin{aligned} V_0(x) &= -\frac{1}{2} + (1-x), \\ \mathcal{A} &= \left\{ k \in L^\infty(0, T), 1 \leq k(t) \leq 2, \int_0^T k \, dt = T+1 \right\}. \end{aligned}$$

If

$$V_1(x) = \frac{1}{2} + (1-2x)$$

then, by comparison,

$$V_0(x) \leq u(x, t) \leq V_1(x) \quad (0 < x < 1, 0 < t < T)$$

and thus the global solution u exists for any k . Set

$$k_1(t) = \begin{cases} 1 & \text{if } 0 \leq t \leq T-1, \\ 2 & \text{if } T-1 < t < \infty, \end{cases} \quad k_2(t) = \begin{cases} 2 & \text{if } 0 \leq t \leq 1, \\ 1 & \text{if } 1 \leq t < \infty, \end{cases}$$

and denote by u_i the solution corresponding to k_i . Then $u_1(x, t) = V_0(x)$ if $0 < t < T-1$ and, consequently, by comparison,

$$u_1(\tfrac{1}{2}, T) > V_1(\tfrac{1}{2}) + \varepsilon = \varepsilon > 0$$

for some ε independent of T . It follows that

$$u_1(\tfrac{1}{2} + \delta, T) \geq 0 \quad \text{for some } \delta > 0,$$

where δ is independent of T . On the other hand, by asymptotic stability (cf. [4])

$$u_1(x, t) \rightarrow V_0(x) \quad \text{if } t \rightarrow \infty$$

and, in particular,

$$u_1(\tfrac{1}{2} + \delta, T) < 0$$

if T is large enough. It follows that

$$J_0(k_2) < J_0(k_1),$$

and thus the bang-bang principle (2.22) does not hold for the maximizer of the functional $J_0(k)$.

3. Periodic optimal control. In this section we consider the case of periodic boundary conditions with period σ . Thus we assume that

$$(3.1) \quad g(x, t + \sigma) = g(x, t)$$

and take the control set

$$(3.2) \quad \mathcal{A}_0 = \left\{ N_1 \leq k \leq N_2, k(x, t + \sigma) = k(x, t), \int_0^\sigma \int_{\Gamma_1} k = M \right\}$$

where N_1, N_2 are as before, and

$$(3.3) \quad N_1 \sigma \text{ meas}(\Gamma_1) < M < N_2 \sigma \text{ meas}(\Gamma_1).$$

For any $k \in \mathcal{A}_\sigma$ consider the Stefan problem

$$(3.4) \quad \begin{aligned} \frac{\partial}{\partial t} \beta(u) - \Delta u &= 0 \quad \text{in } \mathcal{D}'(D_\infty), \\ u_\nu &= k \quad \text{on } \Gamma_{1,\infty}, \\ u &= -g \quad \text{on } \Gamma_{2,\infty}, \\ \beta(u) &= \beta(h) \quad \text{on } t=0 \end{aligned}$$

and set

$$P = \{(k, u); k \in \mathcal{A}_\sigma, u \text{ is the unique solution of (3.4) corresponding to } k\}.$$

From § 1 we have, for any $(k, u) \in P$,

$$(3.5) \quad \int_0^\infty \int_D |u(x, t) - \hat{u}(x, t)|^2 \leq C < \infty, \quad C \text{ independent of } k,$$

where \hat{u} is the unique periodic solution of

$$(3.6) \quad \begin{aligned} \frac{\partial}{\partial t} \beta(u) - \Delta u &= 0 \quad \text{in } \mathcal{D}'(D_\infty) \\ u_\nu &= k \quad \text{on } \Gamma_{1,\infty}, \\ u &= -g \quad \text{on } \Gamma_{2,\infty}, \\ u(x, t + \sigma) &= u(x, t). \end{aligned}$$

In this section we work with the functional

$$(3.7) \quad J(k) = \int_0^\sigma \int_D p(x) \hat{u}(x, t) \, dx \, dt.$$

From (3.5) we get

$$\int_{m\sigma}^{(m+1)\sigma} \int_D |u - \hat{u}| \leq (|D|\sigma)^{1/2} \left\{ \int_{m\sigma}^{(m+1)\sigma} \int_D |u - \hat{u}|^2 \right\}^{1/2} \rightarrow 0 \quad \text{if } m \rightarrow \infty$$

and, consequently,

$$\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{j=0}^{m-1} \int_{j\sigma}^{(j+1)\sigma} p(u - \hat{u}) = 0,$$

i.e.,

$$\lim_{m \rightarrow \infty} \frac{1}{m} \int_0^{m\sigma} \int_D pu = \lim_{m \rightarrow \infty} \frac{1}{m} \int_0^{m\sigma} \int_D p\hat{u} = \int_0^\sigma \int_D p\hat{u}.$$

It follows that for any $(k, u) \in \mathcal{A}_\sigma$

$$(3.8) \quad J(k) = \int_0^\sigma \int_D p\hat{u} = \lim_{m \rightarrow \infty} \frac{1}{m} \int_0^{m\sigma} \int_D pu.$$

Problem (P). Maximize $J(k)$ over $k \in \mathcal{A}_\sigma$.

The existence of a maximizer can be proved as in § 2, using the form (3.7) of $J(k)$ and the fact that

$$(3.9) \quad \|\hat{u}\|_{H^1(D_\sigma)} \leq C$$

where C is a constant independent of k .

LEMMA 3.1. Suppose (k_m, u_m) and (k, u) belong to P , and

$$\begin{aligned} k_m &\rightarrow k \quad \text{weak star in } L^\infty(\Gamma_{1,\sigma}), \\ u_m &\rightarrow u \quad \text{a.e. in } D_\sigma. \end{aligned}$$

Then

$$(3.10) \quad \lim_{m \rightarrow \infty} \frac{1}{m} \int_0^{m\sigma} \int_D p(u_m - u) = 0.$$

Proof. Let \hat{u}_m, \hat{u} be the periodic solutions corresponding to k_m and k , respectively. Using the a priori estimate (3.9) we deduce that for any subsequence of $\{u_m\}$ there is a sub-subsequence, which we still denote by u_m , for which

$$\hat{u}_m \rightarrow \tilde{u} \quad \text{in } L^2(D_\sigma)$$

and \tilde{u} is a periodic solution of (3.6) corresponding to k . By uniqueness $\tilde{u} = \hat{u}$. Using (3.5) we get

$$\begin{aligned}
 & \lim_{m \rightarrow \infty} \frac{1}{m} \left| \int_0^{m\sigma} \int_D p(u_m - u) \right| \\
 & \leq C \lim_{m \rightarrow \infty} \left\{ \frac{1}{m} \int_{D_{m\sigma}} |u_m - \hat{u}_m| + \frac{1}{m} \int_{D_{m\sigma}} |\hat{u}_m - \hat{u}| + \frac{1}{m} \int_{D_{m\sigma}} |\hat{u} - u| \right\} \\
 & \leq C \lim_{m \rightarrow \infty} \left\{ \frac{1}{m} \left(\int_{D_{m\sigma}} |u_m - \hat{u}_m|^2 \right)^{1/2} m^{1/2} + \int_{D_\sigma} |\hat{u}_m - \hat{u}| \right. \\
 & \quad \left. + \frac{1}{m} \left(\int_{D_{m\sigma}} |\hat{u} - u|^2 \right)^{1/2} m^{1/2} \right\} \\
 & \leq C \lim_{m \rightarrow \infty} \int_{D_\sigma} |\hat{u}_m - \hat{u}| = 0,
 \end{aligned}$$

which proves (3.10).

Consider now a pair $(k_0, u_0) \in P$ where k_0 is a solution of Problem (P). In order to analyze k_0 , we introduce, for any positive integer m , the penalized functional

$$(3.11) \quad J_m(k) = \frac{1}{m} \int_0^{m\sigma} \int_D pu - \frac{1}{2} \int_0^\sigma \int_{\Gamma_1} (k - k_0)^2$$

for $(k, u) \in P$.

Problem (P_m). Maximize $J_m(k)$ over $k \in \mathcal{A}_\sigma$.

The existence of a solution can be established as before.

LEMMA 3.2. Suppose $(k_m, u_m) \in P$, k_m a solution of Problem (P_m). Then, as $m \rightarrow \infty$,

$$(3.12) \quad k_m \rightarrow k_0 \quad \text{in } L^2(\Gamma_1 \times (0, \sigma)).$$

Proof. For any subsequence of (k_m, u_m) there is a sub-subsequence, which we denote again by (k_m, u_m) , for which

$$(3.13) \quad k_m \rightarrow \tilde{k} \quad \text{weak star in } L^\infty(\Gamma_{1,\infty}),$$

$$(3.14) \quad u_m \rightarrow \tilde{u} \quad \text{a.e.,}$$

and $(\tilde{k}, \tilde{u}) \in P$. Further, by Lemma 3.1,

$$\begin{aligned}
 J(k_0) & \geq J(\tilde{k}) = \lim_{m \rightarrow \infty} \frac{1}{m} \int_0^{m\sigma} \int_D p\tilde{u} \\
 & = \lim_{m \rightarrow \infty} \left[J_m(k_m) + \frac{1}{m} \int_0^{m\sigma} \int_D p(\tilde{u} - u_m) + \frac{1}{2} \int_{\Gamma_{1,\sigma}} (k_m - k_0)^2 \right] \\
 & \geq \liminf_{m \rightarrow \infty} J_m(k_0) + \limsup_{m \rightarrow \infty} \frac{1}{2} \int_{\Gamma_{1,\sigma}} (k_m - k_0)^2 \\
 & = J(k_0) + \limsup_{m \rightarrow \infty} \frac{1}{2} \int_{\Gamma_{1,\sigma}} (k_m - k_0)^2,
 \end{aligned}$$

and (3.12) follows.

$$\begin{aligned}
(3.15) \quad & \frac{\partial}{\partial t} \beta_\varepsilon(u) - \Delta u = 0 \quad \text{in } D_\infty, \\
& u_\nu = 0 \quad \text{on } \Gamma_{1,\infty}, \\
& u = -g \quad \text{on } \Gamma_{2,\infty}, \\
& u = h \quad \text{on } t = 0.
\end{aligned}$$
$$P_m^\varepsilon = \{(k, u); k \in \mathcal{A}_\sigma, u \text{ is the unique solution of (3.15) corresponding to } k\}.$$
$$(3.16) \quad J_m^\varepsilon(k) = \frac{1}{m} \int_0^{m\sigma} \int_D p(x) u(x, t) - \frac{1}{2} \int_0^\sigma \int_{\Gamma_1} [(k - k_m)^2 + (k - k_0)^2].$$

LEMMA 3.3. Suppose $(k_m^\varepsilon, u_m^\varepsilon) \in P_m^\varepsilon$ where k_m^ε is a solution of Problem (P_m^ε) . Then, as $\varepsilon \rightarrow 0$,

Proof. We may suppose that

$$\begin{aligned} k_m^\varepsilon &\rightharpoonup \tilde{k}_m \quad \text{weakly in } L^2(\Gamma_1, \sigma), \\ u_m^\varepsilon &\rightarrow \tilde{u}_m \quad \text{a.e. in } D_{m\sigma}. \end{aligned}$$

$$\bar{u}_m^\varepsilon \rightarrow u_m \quad \text{a.e. as } \varepsilon \rightarrow 0.$$
$$\begin{aligned} J_m(k_m) &\cong J_m(\tilde{k}_m) = \frac{1}{m} \int_{D_{m\sigma}} p \tilde{u}_m - \frac{1}{2} \int_{\Gamma_{1,\sigma}} (\tilde{k}_m - k_0)^2 \\ &\cong \lim_{\varepsilon \rightarrow 0} \frac{1}{m} \int_{D_{m\sigma}} p u_m^\varepsilon - \frac{1}{2} \liminf_{\varepsilon \rightarrow 0} \int_{\Gamma_{1,\sigma}} (k_m^\varepsilon - k_0)^2 \end{aligned}$$

(since $k_m^\varepsilon \rightarrow k_m$ weak star in $L^\infty(\Gamma_{1,\sigma})$)

$$\begin{aligned}
&= \limsup_{\varepsilon \rightarrow 0} \left[\frac{1}{m} \int_{D_{m\sigma}} p u_m^\varepsilon - \frac{1}{2} \int_{\Gamma_{1,\sigma}} (k_m^\varepsilon - k_0)^2 \right] \\
&= \limsup_{\varepsilon \rightarrow 0} \left[J_m^\varepsilon(k_m^\varepsilon) + \frac{1}{2} \int_{\Gamma_{1,\sigma}} (k_m^\varepsilon - k_m)^2 \right] \\
&\geq \liminf_{\varepsilon \rightarrow 0} J_m^\varepsilon(k_m) + \frac{1}{2} \limsup_{\varepsilon \rightarrow 0} \int_{\Gamma_{1,\sigma}} (k_m^\varepsilon - k_m)^2 \text{ (by maximality of } k_m^\varepsilon) \\
&= \liminf_{\varepsilon \rightarrow 0} \left[\frac{1}{m} \int_0^{m\sigma} \int_D p \bar{u}_m^\varepsilon - \frac{1}{2} \int_{\Gamma_{1,\sigma}} (k_m - k_0)^2 + \frac{1}{2} \limsup_{\varepsilon \rightarrow 0} \int_{\Gamma_{1,\sigma}} (k_m^\varepsilon - k_m)^2 \right] \\
&= \frac{1}{m} \int_0^{m\sigma} \int_D p u_m - \frac{1}{2} \int_{\Gamma_{1,\sigma}} (k_m - k_0)^2 + \limsup_{\varepsilon \rightarrow 0} \int_{\Gamma_{1,\sigma}} (k_m^\varepsilon - k_m)^2 \\
&= J_m(k_m) + \frac{1}{2} \limsup_{\varepsilon \rightarrow 0} \int_0^\sigma \int_{\Gamma_1} (k_m - k_m^\varepsilon)^2,
\end{aligned}$$

from which (3.17) follows.

Consider Problem (P_m^ε) and suppose $(k_m^\varepsilon, u_m^\varepsilon) \in P_m^\varepsilon$ where k_m^ε is a maximizer. Let l be any function such that $k_m^\varepsilon + \delta l \in \mathcal{A}_\sigma$ for any small $\delta > 0$, and denote by $u_m^{\varepsilon, \delta}$ the solution of (3.15) corresponding to $k_m^\varepsilon + \delta l$. Then $J_m^\varepsilon(u_m^\varepsilon) - J_m^\varepsilon(u_m^{\varepsilon, \delta}) \geq 0$, and, proceeding as in § 2, we derive the analogue of (2.14):

$$(3.18) \quad \frac{1}{m} \int_0^m \int_{\Gamma_1} (Q_m^\varepsilon + 2k_m^\varepsilon - k_m - k_0) l \geq 0,$$

where Q_m^ε is the solution of the parabolic problem

$$(3.19) \quad \begin{aligned} \beta_\varepsilon'(u_m^\varepsilon) Q_{m,t}^\varepsilon + \Delta Q_m^\varepsilon &= p \quad \text{in } D_{m\sigma}, \\ Q_{m,\nu}^\varepsilon &= 0 \quad \text{on } \Gamma_1 \times (0, m\sigma), \\ Q_m^\varepsilon &= 0 \quad \text{on } \Gamma_2 \times (0, m\sigma), \\ Q_m^\varepsilon &= 0 \quad \text{on } t = m\sigma. \end{aligned}$$

Let

$$(3.20) \quad P_m^\varepsilon(x, t) = \frac{1}{m} \sum_{j=0}^{m-1} Q_m^\varepsilon(x, t + j\sigma).$$

Since k_m^ε and k_m, k_0 are periodic, (3.18) can be written in the form

$$(3.21) \quad \int_0^\sigma \int_{\Gamma_1} (P_m^\varepsilon + 2k_m^\varepsilon - k_m - k_0) l \geq 0.$$

As in § 2 we have

$$\bar{Q} \leq Q_m^\varepsilon \leq 0 \quad \text{in } D_{\sigma m}$$

with the same function \bar{Q} , and $Q_{m,t}^\varepsilon \geq 0$. It follows that

$$(3.22) \quad \bar{Q} \leq P_m^\varepsilon \leq 0 \quad \text{in } D_\sigma,$$

$$(3.23) \quad P_{m,t}^\varepsilon \geq 0 \quad \text{in } D_\sigma.$$

Also,

$$(3.24) \quad \frac{\partial}{\partial t} P_m^\varepsilon + \Delta P_m^\varepsilon = p \quad \text{in } V_\delta \equiv \Gamma_1^\delta \times (0, \sigma),$$

$$(3.25) \quad \frac{\partial}{\partial \nu} P_m^\varepsilon = 0 \quad \text{on } \Gamma_1 \times (0, \sigma).$$

Using Lemmas 3.2, 3.3 and (3.21)–(3.25) we can proceed to establish, as in § 2, that there exist sequences of $\{m\}$ and $\{\varepsilon_m\}$ such that

$$(3.26) \quad P_m^\varepsilon \rightarrow P \quad \text{uniformly in } V_\delta, \quad P \geq 0,$$

$$(3.27) \quad P_t + \Delta P = p \quad \text{in } V_\delta,$$

$$(3.28) \quad P_\nu = 0 \quad \text{on } \Gamma_{1,\sigma},$$

and

$$(3.29) \quad P_t > 0 \quad \text{on } \Gamma_{1,\sigma}.$$

Further,

$$(3.30) \quad k_0 = \begin{cases} N_2 & \text{on } \{P < \lambda\} \cap \Gamma_{1,\sigma}, \\ N_1 & \text{on } \{P > \lambda\} \cap \Gamma_{1,\sigma} \end{cases}$$

for some $\lambda \leq 0$. Finally, from (3.29), (3.30) we conclude the following theorem.

THEOREM 3.4. *If k_0 is a maximizer of Problem (P), then there exists a $C^{1,\alpha/2}$ function $\varphi(x)$ defined on Γ_1 such that*

$$(3.31) \quad k_0(x, t) = \begin{cases} N_2 & \text{if } m\sigma < t < m\sigma + \varphi(x), \\ N_1 & \text{if } m\sigma + \varphi(x) < t < (m+1)\sigma \end{cases}$$

for $m = 0, 1, 2, \dots$, and $x \in \Gamma_1$.

4. Optimal control for Dirichlet data. In this section we extend the results of §§ 1–3 to other boundary conditions. We begin with

$$(4.1) \quad \begin{aligned} u &= k \quad \text{on } \Gamma_1 \times (0, T), \\ u &= -g \quad \text{on } \Gamma_2 \times (0, T) \end{aligned}$$

(instead of (0.2), (0.3')) and assume that $\Gamma_1 \in C^{3+\alpha}$ and

$$(4.2) \quad \begin{aligned} k &\in L^\infty(\Gamma_1 \times (0, \infty)), \quad 0 < \beta_1 \leq k \leq \beta_2 < \infty, \\ g &\in L^\infty(\Gamma_2 \times (0, \infty)), \quad 0 < c_1 \leq g \leq c_2 < \infty, \\ h &\in C^0(\bar{D}). \end{aligned}$$

Since g and k are not assumed to be Lipschitz continuous, one cannot expect ∇u to be in $L^2((0, T); H^1(D))$. We shall therefore employ the following concept of a weak solution.

A pair (u, ξ) is called a *weak solution* of the two-phase Stefan problem corresponding to the boundary data (4.1) and the initial data $u = h$, $\xi_0 \in H(h)$ if $u \in L^\infty(D_T)$ and

$$(4.3) \quad \begin{aligned} \int_0^T \int_D [(u + \xi)\varphi_t + u\Delta\varphi] dx dt &= \int_0^T \int_{\Gamma_1} k \frac{\partial\varphi}{\partial\nu} dS dt - \int_0^T \int_{\Gamma_2} g \frac{\partial\varphi}{\partial\nu} dS dt \\ &\quad - \int_D (h + \xi_0)\varphi(x, 0) dx \quad \forall \varphi \in V \end{aligned}$$

where

$$V = \{\varphi \in C(\bar{D}_T); D_x\varphi, D_x^2\varphi, D_t\varphi \in C(\bar{D}_T), \varphi = 0 \text{ on } t = T \text{ and on } \partial D \times (0, T)\}.$$

Let $V_0(x)$ and $V_1(x)$ be the solutions of

$$(4.4) \quad \begin{aligned} \Delta V_0 &= 0 \quad \text{in } D, \\ V_0 &= \beta_1 \quad \text{on } \Gamma_1, \\ V_0 &= -c_2 \quad \text{on } \Gamma_2, \end{aligned}$$

and

$$(4.5) \quad \begin{aligned} \Delta V_1 &= 0 \quad \text{in } D, \\ V_1 &= \beta_2 \quad \text{on } \Gamma_1, \\ V_1 &= -c_1 \quad \text{on } \Gamma_2. \end{aligned}$$

THEOREM 4.1. *Assume that*

$$\begin{aligned} V_0(x) &\leq h(x) \leq V_1(x) \quad \text{in } D, \\ H(V_0) &\leq \xi_0 \leq H(V_1) \quad \text{in } D. \end{aligned}$$

Then there exists a unique solution u of the two-phase Stefan problem (4.3), and

$$(4.6) \quad \begin{aligned} V_0(x) &\leq u(x, t) \leq V_1(x), \\ H(V_0) &\leq \xi(x, t) \leq H(V_1). \end{aligned}$$

Proof. Let Ψ_m be smooth functions (say in C^1) in \bar{D}_T such that their restrictions

$$k_m = \Psi_m|_{\Gamma_{1,T}}, \quad g_m = -\Psi_m|_{\Gamma_{2,T}},$$

satisfy (4.2) and

$$\begin{aligned} k_m &\rightarrow k, \quad g_m \rightarrow g, \quad \text{weak star in } L^\infty(\partial D \times (0, T)), \\ h_m &\equiv \Psi_m|_{t=0} \rightarrow h \quad \text{uniformly in } \bar{D}. \end{aligned}$$

By [4] there exists a solution (u_m, ξ_m) corresponding to Ψ_m , and it satisfies (4.6). Set

$$\begin{aligned} D^\delta &= \{x \in D, \text{dist}(x, \partial D) > \delta\}, \\ D_T^\delta &= D^\delta \times (0, T). \end{aligned}$$

In view of (4.6), for some $\delta > 0$ there holds

$$\frac{\partial u_m}{\partial t} - \Delta u_m = 0 \quad \text{in } (D \setminus D^{2\delta}) \times (0, T).$$

But then, by parabolic interior estimates,

$$|u_m|_{C^1(\partial D^\delta \times (\delta, T))} \leq C$$

where C is a constant independent of m . Considering u_m as a solution of the two-phase Stefan problem in $D_T^{2\delta}$, we can apply gradient estimates from [3] and diagonalization in order to conclude that, for a subsequence,

$$\begin{aligned} u_m &\rightarrow u \quad \text{in } L^2(D_T), \\ \xi_m &\rightarrow \xi \quad \text{weak star in } L^\infty(D_T). \end{aligned}$$

Taking $m \rightarrow \infty$ in the identities (4.3) for (u_m, ξ_m) we find that (u, ξ) is a weak solution. Finally, uniqueness is proved as in [4], [5].

The stability estimate (1.22) was established in [4] (see also [3]) for smooth data, say Dirichlet-Lipschitz data. Thus it holds, in particular, for (u_m, ξ_m) . The proof shows that the constant C is independent of m . Hence, by approximation, (1.22) remains true for solutions of (4.3) with data in the class (4.2).

Similarly, for smooth periodic Dirichlet data k_m, g_m there exists a periodic solution and, upon taking $m \rightarrow \infty$, we obtain a periodic solution $(\hat{u}, \hat{\xi})$ when the periodic data (k, g) are in the class (4.2). By the stability estimate, the periodic solution is uniquely determined by k and g .

To extend the results of § 2 we introduce

$$\hat{\mathcal{A}} = \left\{ k \in L^\infty(\Gamma_1 \times (0, T)), N_1 \leq k \leq N_2, \int_0^T \int_{\Gamma_1} k = M \right\}$$

with any $N_2 > N_1 > 0$ and M as in (2.1), and the class of solutions

$$\hat{K}_T = \{(k, u); k \in \hat{\mathcal{A}}, u \text{ is the unique solution of (4.3) corresponding to } k\}.$$

Consider the functional

$$J(k) = \int_0^T \int_D p(x)u(x, t) \, dx \, dt \quad (p > 0)$$

for $(k, u) \in \hat{K}_T$.

Problem (\hat{F}). Maximize $\hat{J}(k)$ over $k \in \hat{\mathcal{A}}$.

Let (k_0, u_0) be a solution of Problem (\hat{F}) and introduce the functional

$$\hat{J}_\varepsilon(k) = \int_0^T \int_D pu - \frac{1}{2} \int_0^T \int_{\Gamma_1} (k - k_0)^2,$$

with $(k, u) \in \hat{K}_T^\varepsilon$, i.e., u is the solution of

$$\begin{aligned} \frac{\partial}{\partial t} \beta_\varepsilon(u) - \Delta u &= 0 \quad \text{in } D_T, \\ (4.7) \quad u &= k \quad \text{on } \Gamma_1 \times (0, T), \\ u &= -g \quad \text{on } \Gamma_2 \times (0, T), \\ u &= h \quad \text{on } t = 0. \end{aligned}$$

We also introduce Problem (\hat{F}_ε) of maximizing $\hat{J}_\varepsilon(k)$ over \hat{K}_T^ε .

We proceed as in § 2 to derive the optimality condition for a solution $(k_\varepsilon, u_\varepsilon)$ of Problem (\hat{F}_ε):

$$(4.8) \quad \int_0^T \int_{\Gamma_1} (Q_{\varepsilon, \nu} + k - k_0) l \leq 0.$$

Here

$$\begin{aligned} \beta'_\varepsilon(u_\varepsilon) Q_{\varepsilon, t} + \Delta Q_\varepsilon &= p \quad \text{in } D_T, \\ Q_\varepsilon &= 0 \quad \text{on } \partial D \times (0, T) \text{ and on } t = T. \end{aligned}$$

We easily deduce that $\hat{Q} \leq Q_\varepsilon \leq 0$ where

$$\Delta \hat{Q} = p \quad \text{in } D, \quad \hat{Q} = 0 \quad \text{on } \partial D,$$

and $Q_{\varepsilon, t} \geq 0$. Hence, for a sequence $\varepsilon \rightarrow 0$, $Q_\varepsilon \rightarrow Q$ uniformly in $V^\delta \equiv \Gamma_1^\delta \times (0, T)$ and

$$\begin{aligned} Q_t + \Delta Q &= p \quad \text{in } V^\delta, \\ Q &= 0 \quad \text{on } \Gamma_1 \times (0, T), \\ Q_t &\geq 0. \end{aligned}$$

The strong maximum principle can now be applied to Q_t ; we conclude that $Q_t > 0$ in V^δ and

$$(4.9) \quad \frac{\partial}{\partial \nu} Q_t > 0 \quad \text{on } \Gamma_1 \times (0, T).$$

This inequality, together with (4.8) and the uniform convergence $Q_{\varepsilon, \nu} \rightarrow Q_\nu$ on $\Gamma_1 \times (0, T)$, yields the following.

THEOREM 4.2. *If k_0 is a maximizer of Problem (\hat{F}) then there exists a $C^{1,\alpha/2}$ function $\varphi(x)$ defined on Γ_1 such that on $\Gamma_1 \times (0, T)$:*

$$(4.10) \quad k(x, t) = \begin{cases} N_2 & \text{if } 0 < t < \varphi(x), \\ N_1 & \text{if } \varphi(x) < t < T. \end{cases}$$

The results of § 3 can now also be extended in the obvious way.

Remark. Using the approximation procedure of k, g by smooth functions k_m, g_m as in the present section, we can extend the results of §§ 1–3 to the case where $g \in L^\infty(D_\infty)$ and $0 \leq c_1 \leq g \leq c_2 \leq \infty$ (instead of g satisfying (1.1)). Finally, the results of this paper extend to the case of a Neumann condition on $\Gamma_2 \times (0, T)$.

REFERENCES

- [1] V. BARBU, *Optimal Control of Variational Inequalities*, Pitman, London, 1984.
- [2] A. DAMLAMIAN, *Some results on the multiphase Stefan problem*, Comm. Partial Differential Equations, 2 (1977), pp. 1017–1044.
- [3] E. DIBENEDETTO AND A. FRIEDMAN, *Periodic behavior for the evolutionary dam problem and related free boundary problems*, Comm. Partial Differential Equations, 11 (1986), pp. 1297–1377.
- [4] A. FRIEDMAN, *The Stefan problem in several space variables*, Trans. Amer. Math. Soc., 133 (1968), pp. 51–87.
- [5] ———, *Variational Principles and Free Boundary Problems*, John Wiley, New York, 1982.
- [6] ———, *Optimal control for parabolic variational inequalities*, this Journal, 25 (1987), pp. 482–497.
- [7] A. FRIEDMAN, S. HUANG AND J. YONG, *Bang-bang optimal control for the dam problem*, Appl. Math. Optim., 15 (1987), pp. 68–85.
- [8] A. FRIEDMAN AND L. JIANG, *Nonlinear optimal control problems in heat conduction*, this Journal, 21 (1983), pp. 940–952.
- [9] S. L. KAMENOMOSTKAJA, *On Stefan's problem*, Math. USSR Sb., 53 (1965), pp. 485–514.

CONTROL OF FREE BOUNDARY PROBLEMS WITH HYSTERESIS*

AVNER FRIEDMAN† AND KARL-HEINZ HOFFMANN‡

Abstract. We consider the problem of controlling the free boundary of the two-phase Stefan problem by means of boundary hysteresis control based on the Preisach model. It is proved that for each control μ there is a corresponding solution of the Stefan problem and that there exists an optimal control.

Key words. Stefan problem, hysteresis, optimal control, control variable, Preisach model

AMS(MOS) subject classifications. 35R35, 49B22, 49A36, 80A99

Introduction. In this paper we are concerned with free boundary problems with a control variable whose structure involves a hysteresis law. In fact we choose the Preisach model for hysteresis [19]; thus the control variable is a measure μ defined on the space of pairs (ρ_1, ρ_2) with $\rho_1 < \rho_2$. The fact that the hysteresis law is discontinuous causes some technical difficulties. We establish the existence of a unique solution u to the free boundary problem for any given μ , and then proceed to prove the existence of an optimal control. For definiteness we shall consider in detail only the two-phase Stefan problem. Other problems, some with free boundary and others with fixed boundary, can be treated by the same methods as briefly mentioned in § 5.

We note that the Preisach model is often used by physicists [3], [6], [17]; for other hysteresis models see Hornung [11]. The two-phase Stefan problem with hysteresis was treated by Hoffmann and Sprekels [10] for a simple situation of a thermostat control (cf. also Jäger [12] for a related reaction diffusion problem).

A systematic mathematical treatment of hysteresis has been carried out by Krasnosel'skii, Pokrovskii and co-workers (see [13]–[15] and the references therein). More recently Visintin [22]–[25] has studied several physical models with hysteresis. In [23] he proved that the hysteresis functional is a continuous mapping, a fact which will be very useful in this work.

In § 1 we describe the hysteresis model for the Stefan problem and then briefly outline the structure and main results of this paper.

1. The Stefan problem. Consider the one-dimensional Stefan problem:

$$(1.1) \quad u_t = u_{xx} \quad \text{if } 0 < x < s(t), \quad t > 0 \quad \text{or if } s(t) < x < a, \quad t > 0,$$

$$(1.2) \quad u(0, t) = \gamma(t) \quad \text{if } t > 0, \quad \gamma(t) \text{ continuous, } 0 < \gamma_1 \leq \gamma(t) \leq \gamma_2 < \infty,$$

$$(1.3) \quad u(x, 0) = u_0(x) \quad \text{if } 0 < x < a, \quad \text{where } u_0(x) > 0 \quad \text{for } 0 < x < s_0, \\ u_0(x) < 0 \quad \text{for } s_0 < x < a,$$

$$(1.4) \quad u_x(a, t) + u(a, t) = g(t) \quad \text{for } t > 0, \quad g(t) \leq 0,$$

$$(1.5) \quad u(s(t), t) = 0 \quad \text{for } t > 0, \quad s(0) = s_0, \quad 0 < s_0 < a,$$

$$(1.6) \quad \dot{s}(t) = -u_x^+(s(t), t) + u_x^-(s(t), t) \quad \text{for } t > 0$$

where $u_x^\pm(s(t), t) = u_x(s(t) \mp 0, t)$; $u_0(x)$ is assumed (for simplicity) to be continuously differentiable for $0 \leq x \leq s_0$ and for $s_0 \leq x \leq a$ and $u_0(s_0 \pm 0) = 0$.

* Received by the editors March 3, 1986; accepted for publication (in revised form) November 20, 1986. This work was partially supported by National Science Foundation grants DMS-8420896 and DMS-8501397.

† Purdue University, Center for Applied Mathematics, West Lafayette, Indiana 47907. Present address: Department of Mathematics, University of Minnesota, Minneapolis, Minnesota 55755.

‡ University of Augsburg, Augsburg, West Germany.

This problem has a unique solution (u, s) for any $g \in L_{\text{loc}}^\infty[0, \infty]$, provided $s(t)$ does not intersect $x = a$. Furthermore, the free boundary $x = s(t)$ is in $C^\infty(0, \infty) \cap C^0[0, \infty]$.

For any $0 < \rho_1 < \rho_2 < a$ set $\rho = (\rho_1, \rho_2)$ and define

$$M_\rho^0(s) = \begin{cases} -1 & \text{if } s \geq \rho_2, \\ 0 & \text{if } \rho_1 < s < \rho_2, \\ 0 & \text{if } s < \rho_1, \end{cases}$$

and

$$M_\rho^1(s) = \begin{cases} -1 & \text{if } s \geq \rho_2, \\ -1 & \text{if } \rho_1 < s < \rho_2, \\ 0 & \text{if } s \leq \rho_1. \end{cases}$$

Given any continuous function $s(t)$, $t \geq 0$, set $s_0 = s(0)$. If $\rho_1 < s_0 < \rho_2$ we define the hysteresis law based on M_ρ^0 as follows: Set $t_0 = 0$ and $M_\rho(s(0)) = M_\rho^0(s_0)$. For $k \in \mathbb{N} \cup \{0\}$ define

$$t_{k+1} = \begin{cases} \inf T_{k+1} & \text{if } T_{k+1} \neq \emptyset, \\ +\infty & \text{if } T_{k+1} = \emptyset \end{cases}$$

where

$$T_{k+1} = \left\{ t \in (t_k, \infty); s(t) = \begin{cases} \rho_1 & \text{if } k+1 \text{ even,} \\ \rho_2 & \text{if } k+1 \text{ odd} \end{cases} \right\}$$

and

$$M_\rho(s(t)) = \begin{cases} 0 & \text{if } t \in [t_k, t_{k+1}) \text{ and } k+1 \text{ odd,} \\ -1 & \text{if } t \in [t_k, t_{k+1}) \text{ and } k+1 \text{ even.} \end{cases}$$

The hysteresis law based on M_ρ^1 is defined in a similar way: Set $t_0 = 0$ and $M_\rho(s(0)) = M_\rho^1(s_0)$ and define

$$t_{k+1} = \begin{cases} \inf T_{k+1} & \text{if } T_{k+1} \neq \emptyset, \\ +\infty & \text{if } T_{k+1} = \emptyset \end{cases}$$

where

$$T_{k+1} = \left\{ t \in (t_k, \infty); s(t) = \begin{cases} \rho_1 & \text{if } k+1 \text{ odd,} \\ \rho_2 & \text{if } k+1 \text{ even} \end{cases} \right\}$$

and

$$M_\rho(s(t)) = \begin{cases} 0 & \text{if } t \in [t_k, t_{k+1}) \text{ and } k+1 \text{ even,} \\ -1 & \text{if } t \in [t_k, t_{k+1}) \text{ and } k+1 \text{ odd.} \end{cases}$$

In the case $s_0 \leq \rho_1$ or $s_0 \geq \rho_2$ we have $M_\rho^0(s_0) = M_\rho^1(s_0)$ and $M_\rho(s(t))$ is defined as follows:

Set $t_0 = 0$, $M_\rho(s(0)) = M_\rho^0(s_0) = M_\rho^1(s_0)$. In case $\rho_1 \geq s_0$,

$$T_{k+1} = \left\{ t \in (t_k, \infty); s(t) = \begin{cases} \rho_1 & \text{if } k+1 \text{ odd,} \\ \rho_2 & \text{if } k+1 \text{ even,} \end{cases} \right\}$$

t_{k+1} is defined as before, and

$$M_\rho(s(t)) = \begin{cases} 0 & \text{if } t \in [t_k, t_{k+1}) \text{ and } k+1 \text{ odd,} \\ -1 & \text{if } t \in [t_k, t_{k+1}) \text{ and } k+1 \text{ even.} \end{cases}$$

In case $s_0 \geq \rho_2$,

$$T_{k+1} = \begin{cases} t \in (t_k, \infty); s(t) = \begin{cases} \rho_1 & \text{if } k+1 \text{ odd,} \\ \rho_2 & \text{if } k+1 \text{ even,} \end{cases} \end{cases}$$

t_{k+1} is defined as before, and

$$M_\rho(s(t)) = \begin{cases} 0 & \text{if } t \in [t_k, t_{k+1}) \text{ and } k+1 \text{ even,} \\ -1 & \text{if } t \in [t_k, t_{k+1}) \text{ and } k+1 \text{ odd.} \end{cases}$$

We shall denote by $M_\rho(s(t), M_\rho^i)$ the hysteresis law $M_\rho(s(t))$ based on $M_\rho^i (i = 0, 1)$.

Set

$$(1.7) \quad \Omega = \{(\rho_1, \rho_2); \eta < \rho_1 < \rho_2 < a - \eta\}$$

where η is a fixed small positive number, $0 < \eta < a/2$. Denote by $M(\Omega)$ the set of all nonnegative measures μ on Ω satisfying the following:

- (i) μ is absolutely continuous with respect to the Lebesgue measure $d\rho \equiv d\rho_1 d\rho_2$;
- (ii) $\int_\Omega d\mu(\rho) = 1$.

Denote by $M_1(\Omega)$ the set of all μ in $M(\Omega)$ satisfying the following:

- (iii) $\mu(A) \leq C \int_A d\rho_1 d\rho_2$ for any Lebesgue measurable sets A ; C is a positive constant depending on μ .

The hysteresis functional corresponding to μ is defined by

$$(1.8) \quad M_\mu(s(t)) = \int_\Omega M_\rho(s(t), M_\rho^i) d\mu(\rho) \quad (i = 0, 1);$$

i will be fixed throughout the paper. This functional was first introduced by Preisach [19].

In thermostat control problems it is natural to assume some hysteresis effects. In this paper we shall be interested in controlling the free boundary of the Stefan problem (1.1)–(1.6) when $g(t)$ is a control variable whose form is given by the hysteresis law

$$(1.9) \quad g(t) = M_\mu(s(t)), \quad \mu \in M(\Omega);$$

thus the actual control variable is the measure μ .

We recall the following results due to Visintin [23]: For any $\mu \in M(\Omega)$,

$$(1.10) \quad \text{if } s \in C[0, T] \text{ then } M_\mu(s(t)) \text{ is also in } C[0, T];$$

$$\text{if } s, s_n \in C[0, T] \text{ and } \max_{0 \leq t \leq T} |s_n(t) - s(t)| \rightarrow 0 \text{ as } n \rightarrow \infty,$$

$$(1.11) \quad \text{then } \max_{0 \leq t \leq T} |M_\mu(s_n(t)) - M_\mu(s(t))| \rightarrow 0 \text{ as } n \rightarrow \infty;$$

if further $\mu \in M_1(\Omega)$, then

$$(1.12) \quad \max_{0 \leq t \leq T} |M_\mu(s_n(t)) - M_\mu(s(t))| \leq C \max_{0 \leq t \leq T} |s_n(t) - s(t)|$$

where C is a constant depending only on $\max |s(t)|$.

In § 2 we shall prove that the Stefan problem (1.1)–(1.9) has a solution for any $\mu \in M(\Omega)$ and that the solution is unique if $\mu \in M_1(\Omega)$.

In § 3 we introduce the functional

$$(1.13) \quad J(\mu) = \max_{0 \leq t \leq T} |s(t) - \sigma(t)| + \int_{\Omega} h(\rho) d\mu(\rho)$$

where $\sigma(t)$ is a given continuous function and $h(\rho)$ is a given nonnegative continuous function. We establish the existence of a control $\mu_0 \in \tilde{M}(\Omega)$ satisfying:

$$J(\mu_0) = \inf_{\mu \in \tilde{M}(\Omega)} J(\mu);$$

here $\tilde{M}(\Omega)$ is a suitable admissible class of controls contained in $M(\Omega)$.

In § 4 we study the asymptotic behavior of solutions of (1.1)–(1.9) as $t \rightarrow \infty$, when (1.2) is replaced by $u_x(0, t) = 0$. In § 5 we briefly mention several other problems with hysteresis (both with free boundary and with fixed boundary) for which the methods of the present paper can be applied.

2. Existence and uniqueness for (1.1)–(1.9). In general a global solution of (1.1)–(1.9) may not exist for all $t > 0$ since $x = s(t)$ may cross $x = a$ in finite time; this can be seen by considering already a stationary solution of the form $C - lx$ and comparing with the case where $s(0) = a - \eta$, $g(t) \equiv -1$. In order to avoid this situation we assume that

$$(2.1) \quad \max \{ \gamma_2, \max_x u_0(x) \} < \eta.$$

Then $u(x) = \theta a - \theta' x$ is a stationary solution of (1.1)–(1.6) when $s(0) = \theta a / \theta'$ (if $\theta < \theta'$) and it majorizes the boundary data and the initial conditions of $u(x, t_0)$, if $s(t_0) = a - \eta$, provided $\theta' < 1$ and $1 - \theta$ is sufficiently small.

THEOREM 2.1. *Assume that (2.1) holds. Then for any $\mu \in M(\Omega)$ there exists a solution of (1.1)–(1.9).*

Proof. Take any small $\delta > 0$ and define

$$g_\delta(t) \equiv 0 \quad \text{if } 0 < t < \delta.$$

Let (u, s) denote the corresponding solution of (1.1)–(1.6) in $\{0 < t < \delta\}$. Next define

$$g_\delta(t) = M_\mu(s(t - \delta))$$

for $\delta < t < 2\delta$. Since $g_\delta(t)$ is known in this interval, we can solve (1.1)–(1.6) for $\delta < t < 2\delta$ with this g_δ and with the initial values $(u(x, \delta), s(\delta))$ obtained in the first step.

We continue by defining $g_\delta(t)$ for $2\delta < t < 3\delta$ as before in terms of the function $s(t)$ obtained in the second step, and solve (1.1)–(1.6) in $2\delta < t < 3\delta$. Continuing in this way we obtain a unique solution of (1.1)–(1.6) with $g = g_\delta(t)$; we designate this solution by (u_δ, s_δ) .

We claim that there exist positive constants η_0, η' such that, for any δ sufficiently small, there holds

$$(2.2) \quad \eta_0 \leq s_\delta(t) \leq a - \eta',$$

that is, the free boundary stays uniformly away from $\{x = 0\}$ and $\{x = a\}$ (and, consequently, the solution (u_δ, s_δ) exists for all $t > 0$).

To prove (2.2) we begin by observing that, by the maximum principle, $u_\delta \leq 0$ in $\{s_\delta(t) < x < a, 0 < t < T\}$ for any $T > 0$ for which $0 < s_\delta(t) < a$ if $0 < t < T$, i.e., u_δ cannot take positive maximum on $\{x = a\}$. Also, by (1.4), u_δ cannot take negative minimum smaller than $\min g_\delta$ on $\{x = a\}$, for at a point of minimum, on $\{x = a\}$, $\partial u_\delta / \partial x < 0$. It follows that $-c_0 \leq u_\delta \leq 0$ on $\{x = a\}$, where $c_0 = \max_{s_0 \leq x \leq a} |u_0(x)| + 1$, and by (1.4),

$$(2.3) \quad \left| \frac{\partial}{\partial x} u_\delta(a, t) \right| \leq c_1 \quad \text{if } 0 \leq t \leq T, \quad c_1 = c_0 + 1.$$

Since $u_\delta(0, t) \geq \gamma_1 > 0$ we can use a comparison theorem [9, p. 691] to compare u_δ with a stationary solution of the Stefan problem (i.e., with a linear function $u(x)$ satisfying $u(0) = \gamma$ for some small $\gamma < \gamma_1$, $u(a) = -c_1$). We then deduce that

$$s_\delta(t) \geq \eta_0 > 0 \quad \text{if } 0 < t < T$$

where η_0 depends only on γ_1, a, c_1 . But then we can apply parabolic estimates for the heat equation in $\{0 < x < \eta_0, 0 < t < T\}$ and obtain:

$$(2.4) \quad \left| \frac{\partial}{\partial x} u_\delta \left(\frac{1}{2} \eta_0, t \right) \right| \leq C \quad (0 < t < T)$$

where C is a constant independent of T . We may take $C > c_1$ and $C > \sup |u_0^1(x)|$.

In one of the proofs of global existence for the Stefan problem (due to L. S. Jiang; see [8, p. 118]) we show that $u_{\delta,x}$ in $\{0 < x < s_\delta(t), 0 < t < T\}$ does not take negative minimum smaller than $-C$ on the free boundary and $u_{\delta,x}$ in $\{s_\delta(t) < x < a, 0 < t < T\}$ also does not take negative minimum smaller than $-C$ on the free boundary. Hence

$$u_{\delta,x}(s_\delta(t) \pm 0) \geq -C \quad \text{if } 0 < t < T.$$

Since the functions on the left-hand side are actually negative, we deduce that

$$|u_{\delta,x}(s_\delta(t) \pm 0)| \leq C$$

and, from (1.6),

$$(2.5) \quad \left| \frac{d}{dt} s_\delta(t) \right| \leq C \quad \text{if } 0 < t < T;$$

C is independent of δ, T .

We can now proceed to establish the right-hand inequality in (2.2). Suppose $x = s_\delta(t)$ intersects $x = a - \eta'$ for the first time at $t = T$. Then there exists a t_0 such that $0 < t_0 < T$,

$$a - \eta = s_\delta(t_0) < s_\delta(t) < s_\delta(T) = a - \eta' \quad \text{if } t_0 < t < T.$$

In view of (2.5), $T - t_0 > 2c(\eta - \eta')$ for some $c > 0$ independent of δ . Then $g_\delta(t) = -1$ if $t_0 + \delta < t < T$ and

$$\frac{\partial}{\partial x} u_\delta(a, t) + u_\delta(a, t) = -1 \quad \text{if } t_0 + \delta < t < T.$$

Again using comparison with a stationary solution introduced following (2.1) we deduce that

$$s_\delta(t) \leq a - 2\eta' \quad \text{for } t_0 + \delta < t < T$$

provided η' is sufficiently small (and provided δ is small enough (depending on c, η, η'), which is a contradiction to the assumption that $s_\delta(T) = a - \eta'$).

Having proved (2.2) we deduce that (u_δ, s_δ) exists for all $t > 0$. From (2.5) it follows that for a sequence $\delta = \delta_n \rightarrow 0$

$$(2.6) \quad s_{\delta_n}(t) \rightarrow s(t) \quad \text{uniformly in } 0 \leq t \leq T$$

for any $T < \infty$. Clearly also

$$(2.7) \quad s_{\delta_n}(t - \delta_n) \rightarrow s(t) \quad \text{uniformly in } \delta_n < t < T$$

and further, by (1.11),

$$(2.8) \quad M_\mu(s_{\delta_n}(t)) \rightarrow M_\mu(s(t)) \quad \text{uniformly in } 0 \leq t \leq T.$$

Representing u_{δ_n} by means of the Green and Robin functions (in $\{u_{\delta_n} > 0\}$ and $\{u_{\delta_n} < 0\}$, respectively) in terms of s_{δ_n} and the boundary values $M_\mu(s(t - \delta_n))$ on $x = a$ (cf. [7], [21]), taking $n \rightarrow \infty$ and using (2.6)–(2.8), we deduce that (u, s) forms a solution of (1.1)–(1.6) with $g(t)$ given by (1.8), (1.9).

THEOREM 2.2. *If $\mu \in M_1(\Omega)$ then the solution of (1.1)–(1.9) is unique.*

Proof. We represent u by the Green and Robin functions in $\{u > 0\}$ and $\{u < 0\}$, respectively, differentiate in x and let $x \rightarrow s(t)$ (cf. [7]). We then obtain a pair of Volterra-type integral equations for

$$v_1(t) = u_x(s(t) - 0, t), \quad v_2(t) = u_x(s(t) + 0, t)$$

having the form

$$v_1 = I_1(v_1, s), \quad v_2 = I_2(v_2, s) + I(s)$$

where I_1, I_2 are Volterra-type operators and

$$I(s) = \int_0^t R(s(t), t, \tau) M_\mu(s(\tau)) d\tau.$$

Here

$$\dot{s}(t) = v_2(t) - v_1(t)$$

and R is a smooth kernel (so long as $s(t) < a$).

By (1.12),

$$\max_{0 \leq \tau \leq t} |M_\mu(s(\tau)) - M_\mu(\tilde{s}(\tau))| \leq C \max_{0 \leq \tau \leq t} |s(\tau) - \tilde{s}(\tau)|$$

for any two continuous functions, where C is a constant depending only on $\max |s|$. Using this fact and the standard estimates of [7], we deduce that the mapping $(v_1, v_2) \rightarrow (\bar{v}_1, \bar{v}_2)$, given by

$$\bar{v}_1 = I_1(v_1, s), \quad \bar{v}_2 = I_2(v_2, s) + I(s)$$

is a contraction in $C[0, \sigma]$ provided σ is sufficiently small. This proves uniqueness for $0 < t < \sigma$. Proceeding step by step we derive uniqueness for all $t > 0$.

Remark 2.1. Theorem 2.1 extends to models where the measure $\mu(\rho)$ is supported on a line segment

$$I = \{(\rho_1, \rho_1 + \delta_0), \eta_1 < \rho_1 < a - \delta_0 - \eta_1\}$$

for some $\delta_0 > 0$, $\eta_1 > 0$ and is absolutely continuous with respect to $d\rho_1$; if further

$$\int_\sigma d\mu(\rho_1) \leq C \int_\sigma d\rho_1$$

for any measurable subset σ of I , then Theorem 2.2 is also valid.

Remark 2.2. Theorems 2.1, 2.2 and Remark 2.1 extend to the case where (1.2) is replaced by

$$-u_x(0, t) + \lambda u(0, t) = \gamma(t)$$

where $\lambda \geq 0$, $\gamma(t) > 0$, or when (1.4) is replaced by

$$u(a, t) = g(t) \quad \text{or} \quad u_x(a, t) = g(t).$$

Without any restriction on λ, γ, u_0 , the free boundary may hit the boundary $\{x = 0\}$.

3. Existence of optimal control. We shall now restrict μ to belong to a subset of $M(\Omega)$ given by

$$(3.1) \quad K = \left\{ \mu; d\mu(\rho) = f(\rho) d\rho, \int_{\Omega} f^{1+\delta}(\rho) d\rho \leq C \right\}$$

where δ and C are some positive constants; the condition (2.1) is assumed throughout this section.

Let $\sigma(t)$ be a given continuous function for $0 \leq t \leq \infty$ and let $h(\rho)$ be a prescribed nonnegative continuous function on $\bar{\Omega}$. Consider the cost functional

$$(3.2) \quad J(f, s) = \max_{0 \leq t \leq T} |s(t) - \sigma(t)| + \int_{\Omega} h(\rho) f(\rho) d\rho$$

for any $f \in K$, where T is a fixed positive number. Here (u, s) is any solution of (1.1)–(1.9) corresponding to $d\mu = f d\rho$. Since we may not have uniqueness (unless $\mu \in M_1(\Omega)$), there may be more than one pair (u, s) . We introduce the set of costs

$$(3.3) \quad J[f] = \{J(f, s); (u, s) \text{ solution of (1.1)–(1.9) with } d\mu = f d\rho\}.$$

Using (1.11) it is easy to see that the set of solutions (u, s) corresponding to a given μ is compact in the uniform topology for $0 \leq t \leq T, 0 \leq x \leq a$; hence there exists a solution (u_0, s_0) such that

$$(3.4) \quad J(f, s_0) = \inf \{J(f, s); J(f, s) \in J[f]\}.$$

Set

$$(3.5) \quad J(f) = J(f, s_0)$$

and consider the minimization problem: Find $f_0 \in K$ such that

$$(3.6) \quad J(f_0) = \min_{f \in K} J(f).$$

THEOREM 3.1. *Problem (3.6) has a solution.*

Proof. Let (f_n, s_n, u_n) be a minimizing sequence. By taking a subsequence we may assume that

$$(3.7) \quad \begin{aligned} s_n &\rightarrow s_0 \quad \text{uniformly in } t, \quad 0 \leq t \leq T, \\ u_n &\rightarrow u_0 \quad \text{uniformly in } (x, t), \quad 0 \leq x \leq a, \quad 0 \leq t \leq T, \\ f_n &\rightarrow f_0 \quad \text{weakly in } L^1(\Omega). \end{aligned}$$

Suppose we can prove that, for any continuous function $\phi(t)$,

$$(3.8) \quad \begin{aligned} I_n \equiv \int_0^T \phi(t) dt \left\{ \int_{\Omega} [M_p(s_n(t), M_p^i) f_n(\rho) \right. \\ \left. - M_p(s_0(t), M_p^i) f_0(\rho)] d\rho \right\} \rightarrow 0 \text{ if } n \rightarrow \infty \quad (i = 0, 1). \end{aligned}$$

Then we represent u_n by means of the Green and Robin functions and, using (3.7), (3.8), we find that the same integral representation holds for u_0 and the corresponding s_0, f_0 . This means that (u_0, s_0) forms a solution of (1.1)–(1.6) with g given by $M_{\mu}(s_0(t))$, $d\mu = f_0 d\rho$. Since, as easily seen from (3.2),

$$J(f_n, s_n) \rightarrow J(f_0, s_0),$$

it follows that f_0 and the corresponding solution (u_0, s_0) form a solution of the optimal control problem (3.6). Thus it remains to prove (3.8).

From the proof of (1.11) we see that the convergence of the $M_\mu(s_n(t))$ to $M(s(t))$ is uniform not only in t but also in μ provided μ is in the class K ; in particular,

$$\max_{0 \leq t \leq T} |M_{\mu_n}(s_n(t)) - M_{\mu_n}(s(t))| \rightarrow 0 \quad \text{if } n \rightarrow \infty$$

when $d\mu_n = f_n d\rho$. Therefore the proof of (3.8) is reduced to proving that

$$(3.9) \quad \tilde{I}_n \equiv \int_{\Omega} \left[\int_0^T M_\rho(s_0(t)) \phi(t) dt \right] [f_n(\rho) - f_0(\rho)] d\rho \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

To prove (3.9) set

$$\psi(\rho) = \int_0^T M_\rho(s_0(t)) \phi(t) dt, \quad d\mu_0 = f_0 d\rho.$$

Let

$$B_{\varepsilon_0} = \{(\rho_1, \rho_2) \in \Omega; 0 < \rho_2 - \rho_1 < \varepsilon_0\}.$$

Given any $\varepsilon > 0$ we can choose ε_0 sufficiently small so that

$$(3.10) \quad \mu_n(B_{\varepsilon_0}) + \mu_0(B_{\varepsilon_0}) < \varepsilon;$$

here we used the fact that $\mu_n \in K$.

Now, by Sard's lemma, the set

$$S^* = \{s^*; \text{there exists a } t \in [0, T] \text{ such that } s_0(t) = s^* \text{ and } s'_0(t) = 0\}$$

has Lebesgue measure zero. Set

$$S_k = \{s; \text{there exists a } t \in [0, T] \text{ such that } s_0(t) = s \text{ and } |s'_0(t)| < 1/k\}, \quad k \in \mathbb{N}.$$

It is easily seen that

$$S^* = \bigcap_{k=1}^{\infty} S_k.$$

Since $\mathcal{L}(S^*) = 0$ (where $\mathcal{L}(A)$ = Lebesgue measure of A), it follows that $\mathcal{L}(S_k) \rightarrow 0$ if $k \rightarrow \infty$. Setting

$$\tilde{S}_k = \{(\rho_1, \rho_2); \text{either } \rho_1 \text{ is in } S_k \text{ or } \rho_2 \text{ is in } S_k\},$$

we conclude that also $\mu_0(S_k) \rightarrow 0$ if $k \rightarrow \infty$ and, consequently,

$$(3.11) \quad \mu_0(\tilde{S}_k) < \varepsilon$$

for some sufficiently large k , say $k \geq k_0$.

Similarly, if $k \geq k_0$,

$$(3.12) \quad \mu_n(\tilde{S}_k) < \varepsilon \quad \text{for all } n.$$

Let $\Sigma_\varepsilon = \Omega \setminus (B_{\varepsilon_0} \cup \tilde{S}_k)$ and take any $\rho = (\rho_1, \rho_2) \in \Sigma_\varepsilon$. Consider an interval $t_1 < t < t_2$ such that $s_0(t_1) = \rho_2$, $s_0(t_2) = \rho_1$, $\rho_1 < s_0(t) < \rho_2$ if $t_1 < t < t_2$. Then

$$s'_0(t) < -\frac{1}{2k} \quad \text{if } |t - t_2| < \delta,$$

$$s'_0(t) < -\frac{1}{2k} \quad \text{if } |t - t_1| < \delta$$

for some $\delta > 0$ independent of t_1, t_2 . It follows that if $\rho' = (\rho'_1, \rho'_2)$ where

$$|\rho'_1 - \rho_1| < \frac{\delta}{2k}, \quad |\rho'_2 - \rho_2| < \frac{\delta}{2k}$$

then $\{x = s_0(t)\}$ crosses $\{s_0 = \rho'_i\}$ at time $t = \tilde{t}_i$, where

$$|\tilde{t}_i - t_i| < 2k|\rho'_i - \rho_i|,$$

and $\rho'_1 < s_0(t) < \rho'_2$ for $\tilde{t}_1 < t < \tilde{t}_2$. Similar consideration holds in case $s_0(t_1) = \rho_1, s_0(t_2) = \rho_2, \rho_1 < s_0(t) < \rho_2$ if $t_1 < t < t_2$.

Summing over the finite number ($\leq 2N$) of such pairs (t_1, t_2) we find that

$$\int_0^T |M_\rho(s_0(t), M_\rho^i) - M_{\rho'}(s_0(t), M_{\rho'}^i)| dt \leq 4Nk|\rho' - \rho|.$$

Consequently,

$$(3.13) \quad |\psi(\rho) - \psi(\rho')| \leq 4CNk|\rho' - \rho| \quad \text{if } \rho \in \Sigma_\varepsilon$$

where $C = \max |\phi|$. Recalling (3.10)–(3.12) we see that for any $\varepsilon > 0$

$$(3.14) \quad |\tilde{I}_n| \leq C\varepsilon + \left| \int_{\Sigma_\varepsilon} \psi(\rho) f_n(\rho) d\rho - \int_{\Sigma_\varepsilon} \psi(\rho) f_0(\rho) d\rho \right|.$$

By (3.13), $\psi(\rho)$ is continuous on Σ_ε . Since $f_n \rightarrow f_0$ weakly in L^1 , it follows that the difference of the integrals on the right-hand side of (3.14) converge to zero as $n \rightarrow \infty$. Consequently,

$$\limsup_{n \rightarrow \infty} \tilde{I}_n \leq C\varepsilon.$$

Since ε can be taken arbitrarily small, (3.9) follows.

Remark 3.1. Theorem 3.1 extends to the model described in Remark 2.1 provided $d\mu = f(\rho_1) d\rho_1$ and $K = \{f; \int [f(\rho_1)]^{1+\delta} \leq C\}$ for some $\delta > 0, C > 0$. Theorem 3.1 also extends to the boundary conditions given in Remark 2.2.

Remark 3.2. If we replace K by $K \cap M_1(\Omega)$ then (u, s) is uniquely determined by f and $J[f]$ consists of one number.

4. Asymptotic behavior as $t \rightarrow \infty$. In this section we return to problem (1.1)–(1.8) with (1.2) replaced by

$$(1.2') \quad u_x(0, t) = 0 \quad \text{if } t > 0.$$

Assuming that (u, s) is a solution for all $t > 0$ (with $0 < s(t) < 1$, i.e., the free boundary does not hit the fixed boundary in finite time), we shall study the asymptotic behavior of the free boundary as $t \rightarrow \infty$.

LEMMA 4.1. *There holds*

$$(4.1) \quad \int_0^\infty |M_\mu(s(t))| dt < \infty.$$

Proof. By comparison we have

$$(4.2) \quad u^+(x, t) \leq A \cos \frac{\pi x}{2a} e^{-\pi^2 t / 4a^2},$$

$$(4.3) \quad u^-(x, t) \leq Bx$$

where A and B are positive constants.

Integrating $u_t = u_{xx}$ over $0 < x < s(t)$, $0 < t < T$ and over $s(t) < x < a$, $0 < t < T$ and adding, we easily obtain, after using (1.2'), (1.3), (1.5), (1.6), (4.2), (4.3) and the fact that $0 < s(t) < a$,

$$(4.4) \quad -s(T) + s_0 + \int_0^T u_x(a, t) dt = \int_0^a u(x, T) dx - \int_0^a u_0(x) dx;$$

it follows that

$$(4.5) \quad \int_0^T u_x(a, t) dt = O(1) \quad \text{as } T \rightarrow \infty.$$

Next we integrate

$$xu_t = xu_{xx}$$

over $0 < x < s(t)$, $0 < t < T$ and, after integration by parts, obtain

$$(4.6) \quad \int_0^T s(t) u_x^+(s(t), t) dt = O(1).$$

Similarly

$$(4.7) \quad a \int_0^T u_x(a, t) dt - \int_0^T s u_x^-(s(t), t) dt - \int_0^T u(a, t) dt = O(1).$$

Using (4.5) and

$$\begin{aligned} \int_0^T [s(t) u_x^+(s(t), t) - s(t) u_x^-(s(t), t)] dt &= - \int_0^T s(t) \dot{s}(t) dt \\ &= -\frac{1}{2} (s^2(T) - s^2(0)) \\ &= O(1), \end{aligned}$$

we get, from (4.6), (4.7) and (4.5),

$$(4.8) \quad \int_0^T [u_x(a, t) + u(a, t)] dt = O(1) \quad \text{as } T \rightarrow \infty,$$

which is the assertion (4.1).

THEOREM 4.2. *There holds that*

$$(4.9) \quad \lim_{t \rightarrow \infty} s(t) \quad \text{exists.}$$

Proof. Let v be a solution of the heat equation in $\{0 < x < a, t > 0\}$ satisfying

$$v_x(0, t) = 0 \quad \text{if } t > 0, \quad v(x, 0) = -B \quad \text{if } 0 < x < a,$$

$$v_x(a, t) + v(a, t) = M_\mu(s(t)) \quad \text{if } t > 0.$$

Representing v by means of the fundamental solution associated with the same boundary conditions as for v and using (4.8), we can deduce that

$$\max_{0 \leq x \leq a'} |v(x, t)| \rightarrow 0 \quad \text{if } t \rightarrow \infty \quad \text{for any } 0 < a' < a.$$

Since $u^- \leq v$ if B is large enough, the same is then true for u^- . Recalling also (4.2), we obtain

$$\max_{0 \leq x \leq a'} |u(x, t)| \rightarrow 0 \quad \text{if } t \rightarrow \infty \quad \text{for any } 0 < a' < a.$$

In view of (4.3) we then also have

$$(4.10) \quad \int_0^a |u(x, T)| dx \rightarrow 0 \quad \text{if } T \rightarrow \infty.$$

Next we write

$$(4.11) \quad \int_0^T u_x(a, t) dt = \int_0^T M_\mu(s(t)) dt - \int_0^T u(a, t) dt.$$

By Lemma 4.1,

$$\lim_{T \rightarrow \infty} \int_0^T M_\mu(s(t)) dt \text{ exists.}$$

Since $\int_0^T u(a, t) dt$ is a monotone function of T , it then follows from (4.5), (4.11) that this monotone function has a finite limit as $T \rightarrow \infty$. Consequently, the left-hand side of (4.11) has a limit as $T \rightarrow \infty$. Using this fact and (4.10) with $T \rightarrow \infty$, the assertion (4.9) follows from (4.4).

Remark 4.1. If $u_x(0, t) = 0$ is replaced by $u_x(0, t) = -\gamma$, $\gamma > 0$, then numerical results suggest that the free boundary is asymptotically periodic.

5. Other control problems with hysteresis.

5.1. The Muscat problem. Consider the Muscat problem of two immiscible fluids in a one-dimensional porous medium:

$$(5.1) \quad u_t = \begin{cases} au_{xx} & \text{if } 0 < x < s(t), \\ bu_{xx} & \text{if } s(t) < x < 1, \end{cases} \quad \begin{matrix} 0 < t < T_0, \\ 0 < t < T_0, \end{matrix}$$

$$(5.2) \quad u(0, t) = g_1(t) \quad \text{if } 0 < t < T_0,$$

$$(5.3) \quad u(1, t) = g_2(t) \quad \text{if } 0 < t < T_0,$$

$$(5.4) \quad u(x, 0) = u_0(x) \quad \text{if } 0 < x < 1,$$

$$(5.5) \quad u(s(t) - 0, t) = u(s(t) + 0, t) \quad \text{if } 0 < t < T_0,$$

$$(5.6) \quad au_x(s(t) - 0, t) = bu_x(s(t) + 0, t) = -\dot{s}(t) \quad \text{if } 0 < t < T_0,$$

$$(5.7) \quad s(0) = s_0,$$

where g_1, g_2, s_0 are given as well as the positive constants a, b .

We recall [1], [4], [5] that for any g_1, g_2 in $C^{0,1}[0, T_0]$ and $u_0 \in C^{0,1}[0, 1]$ there exists a unique classical solution for all $t < T$, with $s(t) \in C^\infty(0, T) \cap C^0[0, T]$ and either $T = T_0$ or else $T < T_0$ and $s(t) \rightarrow 0$ or $s(t) \rightarrow 1$ as $t \rightarrow T$. The function u represents the pressure in the fluids.

Since the above model may represent, for instance, water and oil, it is natural to attempt to control the location of the free boundary. Taking

$$(5.8) \quad g_2(t) = \int_0^t \int_{\Omega} M_{\rho}(s(\tau), M_{\rho}^i) d\mu(\rho) d\tau$$

we see that, if a solution exists,

$$-1 \leq g_2'(t) \leq 0,$$

i.e., $g_2 \in C^{0,1}[0, T]$.

We can now proceed analogously to §§ 2 and 3 and establish the existence and uniqueness of a solution of (5.1)–(5.8), as well as the existence of an optimal control for the problem (3.1), (3.2), (3.6) associated with (5.1)–(5.8).

5.2. One-phase Stefan problem in n -dimension. Consider the one-phase Stefan problem in an n -dimensional domain [9] and denote the temperature by θ . As usual, we introduce the associated variational inequality for u , where $u_t = \theta$. Since θ is a continuous function in (x, t) (see [2]), for any fixed point x_0 the function $u(x_0, t)$ is continuously differentiable in t .

Now form the hysteresis function

$$(5.9) \quad (M_{\mu}u)(t) = \sum_{j=1}^N \int_{\Omega} M_{\rho}(u(x_j, t), M_{\rho}^i) d\mu_j(\rho)$$

where $\mu = (\mu_1, \dots, \mu_N)$, $\mu_j \in M(\Omega)$ and x_j are fixed points; each $M_{\rho}(u(x_j, t), M_{\rho}^i)$ is defined precisely as $M_{\rho}(s(t), M_{\rho}^i)$ in § 1.

We can apply the methods of § 2 and thus establish the existence of a solution to the variational inequality of the Stefan problem for u , under the boundary condition

$$\frac{\partial u}{\partial \nu} + u = M_{\mu}u$$

on the fixed boundary of the water region. We can also easily establish the existence of an optimal control $\mu = (\mu_1, \dots, \mu_N)$ with $\mu_i \in K$ for the functional

$$(5.10) \quad \max_{(x, t) \in D} |u(x, t) - U(x, t)| + \sum_{i=1}^N \int \sigma_i(\rho) d\mu_i(\rho)$$

where U, σ_i are given, $\sigma_i \geq 0$, and D is a prescribed set.

Similar results can be established for the hysteresis law whereby $u(x_j, t)$ in (5.9) is replaced by the following function of t :

$$(5.11) \quad \tilde{u}_j(t) \equiv \int_{\Gamma_j} u^2(x, t) dS_x$$

where Γ_j is a surface contained in the set $\{u(x, 0) > 0\}$.

5.3. Controlling air pollution. Consider the time-dependent two-space dimensional air pollution model: Find $c = c(x, z, t)$ such that

$$(5.12) \quad c_t = uc_x + wc_z + (k_H c_x)_x + (k_V c_z)_z - Ec \quad \text{in } (0, a) \times (0, h) \times (0, T)$$

with the boundary conditions

$$(5.13) \quad c(0, z, t) = c(a, z, t) = 0,$$

$$(5.14) \quad k_V \frac{\partial c(x, h, t)}{\partial z} = 0,$$

$$(5.15) \quad k_V \frac{\partial c(x, 0, t)}{\partial z} = Q(x)$$

and initial condition

$$(5.16) \quad c(x, z, 0) = c_0(x, z).$$

Here

$$\begin{aligned} c &= \text{concentration of pollutant,} \\ (u, w) &= \text{wind velocity, } u \text{ and } w \text{ are positive,} \\ k_H &= \text{horizontal diffusivity,} \\ k_V &= \text{vertical diffusivity,} \\ E &= \text{rate of chemical decay constant} \\ Q &= \text{quantity of pollution emitted from the ground.} \end{aligned}$$

Let $G = \{(x, z), 0 < x < a, 0 < z < h\}$ and let G' be a subset of G . Let

$$s(t) = \int_{G'} (c(x, z, t))^2 dx dz.$$

We wish to control the amount of pollution emitted from the ground by replacing the boundary condition (5.15) by

$$(5.17) \quad k_V \frac{\partial c(x, 0, t)}{\partial z} = Q(x) M_\mu(t)$$

where

$$(5.18) \quad M_\mu(t) = \int_{\Omega} M_\rho(s(t), M_\rho^i) d\mu(\rho);$$

here $\Omega = \{\rho = (\rho_1, \rho_2); 0 < \rho_1 < \rho_2 < A\}$ for some large constant A . The existence of an optimal control can be established by the methods of §§ 2 and 3. For references on pollution models see [18], [20].

5.4. Identification problem. Consider the inverse problem associated with

$$(5.19) \quad \frac{dx}{dt} = f(x, y, t), \quad x(0) = x_0$$

where

$$(5.20) \quad y(t) = \int_{\Omega} M_\rho(x(t), M_\rho^i) d\mu(\rho);$$

that is, we wish to discover the measure μ from measurements of $x(t)$. If we denote by $x_\mu(t)$ a solution of (5.19), (5.20), then the methods of §§ 2 and 3 show that

$$(5.21) \quad \mu_m \rightarrow \mu \text{ weakly implies } x_{\mu_m}(t) \rightarrow x_\mu(t) \text{ uniformly;}$$

the μ_m are assumed to belong to a class K , as in (3.1).

In order to identify the hysteresis law μ , we first select a large number of measures μ_i and compute their corresponding solution $x_i(t)$ (not necessarily unique, unless the μ_i belong to $M_1(\Omega)$). Next we make measurements and obtain a function $x(t)$. We compare $x(t)$ with the functions $x_i(t)$ and denote by i_0 the index for which $x_{i_0}(t)$ is nearest to $x(t)$, in the uniform sense, say. We now identify the unknown measure μ with the measure μ_{i_0} . The justification for this method of approximation is based on the result (5.21). For physical background on identification problems in flows in porous media, see Maulem [16].

REFERENCES

- [1] C. BAIocchi, L. C. EVANS, L. FRANK AND A. FRIEDMAN, *Uniqueness for two immiscible fluids in one-dimensional porous medium*, J. Differential Equations, 36 (1980), pp. 249–256.
- [2] L. A. CAFFARELLI AND A. FRIEDMAN, *Continuity of the temperature in the Stefan problem*, Indiana Univ. Math. J., 28 (1979), pp. 53–70.
- [3] J. A. ENDERBY, *The domain model of hysteresis: Part 1, Independent domains*, Trans. Faraday Soc., 51 (1955), pp. 835–848.
- [3a] ———, *The domain model of hysteresis. Part 2, Interacting domains*, Trans. Faraday Soc., 52 (1956), pp. 102–120.
- [4] L. C. EVANS, *A free boundary problem: The flow of two immiscible fluids in a one-dimensional porous medium II*, Indiana Univ. Math. J., 27 (1978), pp. 93–111.
- [5] L. C. EVANS AND A. FRIEDMAN, *Regularity and asymptotic behavior of two immiscible fluids in a one-dimensional porous medium*, J. Differential Equations, 31 (1979), pp. 366–391.
- [6] D. H. EVERETT AND W. J. WHITTON, *A general approach to hysteresis: Part 1*, Trans. Faraday Soc., 48 (1952), pp. 749–757.
- [6a] D. H. EVERETT AND F. H. SMITH, *A general approach to hysteresis: Part 2, Development of the domain theory*, Faraday Soc., 50 (1954), pp. 187–197.
- [6b] D. H. EVERETT, *A general approach to hysteresis: Part 3, A formal treatment of the independent domain model of hysteresis*, Trans. Faraday Soc., 50 (1954), pp. 1077–1096.
- [6c] ———, *A general approach to hysteresis: Part 4, An alternative formulation of the domain model*, Trans. Faraday Soc., 51 (1955), pp. 1551–1557.
- [7] A. FRIEDMAN, *Free boundary problems for parabolic equations: Melting of solids*, J. Math. and Mech., 8 (1959), pp. 499–518.
- [8] ———, *Analyticity of the free boundary for the Stefan problem*, Arch. Rational Mech. Anal., 61(1976), pp. 97–125.
- [9] ———, *Variational Principles and Free-Boundary Problems*, Wiley-Interscience, New York, 1982.
- [10] K.-H. HOFFMANN AND J. SPREKELS, *Real time control in a free boundary problem connected with the continuous casting of steel*, in Optimal Control of Partial Differential Equations, K.-H. Hoffmann and W. Krabs, eds., Birkhäuser, Berlin, 1984, pp. 127–143.
- [11] U. HORNUNG, *The mathematics of hysteresis*, Bull. Austral. Math. Soc., 30 (1984), pp. 271–287.
- [12] W. JÄGER, *A diffusion-reaction system modelling spatial patterns*, Equadiff., Bratislava, Czechoslovakia, 1981; Teubner-Texte zur Math., 47 (1982).
- [13] M. A. KRASNOSEL'SKII, *Equations with non-linearities of hysteresis type*, VII Int. Konf. Nichtlin. Schwing., Berlin (1975) (in Russian.) Abh. Sächs Akad. Wiss. Leipzig Math.-Natur. Kl. DDR, 3 (1977), pp. 437–458 (English abstract in Zentbl. Math. 406, 93032).
- [14] M. A. KRASNOSEL'SKII AND A. V. POKROVSKII, *Operators representing non-linearities of hysteresis type*, in Theory of Operators in Functional Spaces, G. P. Akilov, ed., Nauka, Novosibirsk, 1977. (In Russian.)
- [15] ———, *Systems with Hysteresis*, Nauka, Moscow, 1983. (In Russian.)
- [16] Y. MUALEM, *Theory of universal hysteretical properties of unsaturated porous media*, Proc. Fort Collins Fluid Internat. Hydrol. Symp. 1977, in Water Resour. Res. Public. 14 (1979).
- [17] L. NEEL, *Théorie des lois d'aimantation de Lord Rayleigh*, Cahiers Phys., 12 (1942), pp. 1–20.
- [18] F. T. M. NIEUWSTADT AND A. P. VAN ULDEN, *A numerical study on the vertical dispersion of passive contaminants from a continuous source in the atmospheric surface layer*, Atmospheric Environment, 12 (1978), pp. 2119–2124.
- [19] F. PREISACH, *Über die magnetische Nachwirkung*, Z. Phys., 94 (1935), pp. 277–302.
- [20] D. RANDERSON, *A numerical experiment in simulating the transport of sulphur dioxide through the atmosphere*, Atmospheric Environment, 4 (1970), pp. 615–632.
- [21] L. I. RUBINSTEIN, *The Stefan Problem*, American Mathematical Society Translations, 27, American Mathematical Society, Providence, RI, 1971.
- [22] A. A. VISINTIN, *A model for hysteresis of distributed systems*, Ann. Mat. Pura Appl., 131 (1982), pp. 203–231.
- [23] ———, *On the Preisach model for hysteresis*, Nonlinear Anal., 8 (1984), pp. 977–996.
- [24] ———, *Partial differential equations with hysteresis functionals*, Pavia, Inst. de Annalisi Numerica # 377, Pavia, Italy, 1983.
- [25] ———, *On the evolution of ferromagnetic media*, Pavia, Inst. de Annalisi Numerica # 378, Pavia, Italy, 1983.

IDENTIFICATION OF MINIMAL ORDER STATE SPACE MODELS FROM STOCHASTIC INPUT-OUTPUT DATA*

Y. BARAM† AND B. PORAT†

Abstract. This paper discusses the problem of identifying a minimal order state space representation of a multivariable linear time invariant system from Gaussian stationary input-output measurements. A procedure for identifying the system's order is proposed, based on an approximate probability distribution of the squared singular values of the Hankel matrix built from the sample cross-covariances. The approximate distribution converges to the true one as the number of measurements becomes large. The order determination procedure also identifies sets of linearly independent rows and linearly independent columns of the Hankel correlation matrix which form a basis for a minimal order representation of the system.

Key words. system identification, stochastic realization

AMS(MOS) subject classifications. 93B, 93E

1. Introduction. The minimal order representation of the relationship between the input and the output of a finite dimensional linear system has been of interest since the pioneering work of Gilbert [1] and of Kalman [2]. The construction of such representations from a finite dimensional Hankel matrix of impulse responses was proposed by Ho and Kalman [3], Rissanen [4] and Silverman [5], among others. In the case where the input is a stochastic process, the impulse response function is naturally replaced by the cross-correlation function between the input and the output. Akaike [6] gave probabilistic and geometric interpretations to the role played by the Hankel correlation matrix when the input is a white noise process.

When the cross-correlation function between the input and the output of the system under consideration is not given, one is faced with an identification problem. When the exact, unknown correlations are replaced by their estimates from measurement data, the finite structure of the Hankel correlation matrix is normally lost due to the inaccuracies associated with finite samples. A major consequence is that the minimal representation order cannot be determined exactly from any finite dimensional Hankel matrix of sample correlations. In the past, direct identification of the order from a Hankel matrix of sample correlations has been essentially given up due to the complexity of the probability distributions involved [7] and other, indirect methods have been proposed by Chow [7], [8] and Woodside [9]. An approximation technique, based on canonical variates analysis of truncated data records has been suggested by Larimore [10]. Order estimation criteria have been suggested by Akaike [11] and by Fine and Hwang [12], whose method has shown to produce consistent estimates.

In this paper we propose a solution to the problem of identifying the minimal order and a minimal order representation for the relationship between the input and the output of a linear system from stochastic measurement data. For the sake of completeness, we first derive a generalized representation following, essentially, Akaike's steps [6]. We then obtain explicit terms for the representation, which define the model to be identified. The key to this model is a maximal set of linearly independent rows and linearly independent columns of the Hankel correlation matrix, which defines both the order and a basis for the representation. The crucial problem is then to identify a maximal set of linearly independent rows and linearly independent columns of the

* Received by the editors June 5, 1985; accepted for publication (in revised form) January 12, 1987.

† Department of Electrical Engineering, Technion, Israel Institute of Technology, Haifa 32000, Israel.

Hankel matrix from measurement data. We propose a procedure which performs this task by sequentially testing the ranks of submatrices of the Hankel matrix. We then derive an approximate distribution for the rank test statistics and show how it can be calculated from measurement data.

2. Minimal order system representation. Let $u_n \in \mathbb{R}^q$ be an uncorrelated Gaussian input of a system and let $y_n \in \mathbb{R}^m$ be the system's output. Suppose that u_n and y_n are jointly Gaussian and stationary with a cross-correlation function given by

$$(2.1) \quad R_k = E\{y_n u_{n-k}^T\}$$

where E denotes the expectation operation. We assume that there exist a positive integer P and a set of positive scalars a_1, \dots, a_{P-1} , $a_0 = 1$, such that for any $k \geq P$

$$(2.2) \quad \sum_{i=0}^{P-1} a_i R_{k-i} = 0.$$

It is well known that when u_n is the input and y_n is the output of a finite dimensional linear system, they satisfy a relationship of the type (2.2) where the a_i 's are the system's characteristic polynomial coefficients. In the context of system identification u_n may represent a known, synthetically generated "pseudo-random" sequence.

It is desired to find a minimal order representation for the input-output relationship between u_n and y_n . Let us denote by $U_n^- = (u_{n-1}^T, u_{n-2}^T, \dots, u_0^T)^T$ the vector of past inputs and by $Y_n^+ = (y_n^T, y_{n+1}^T, \dots, y_{2n-1}^T)^T$ the vector of n -step future outputs. The space $Y_n^+|U_n^-$ of mean square projection of Y_n^+ on U_n^- is spanned by $R(n)[E\{U_n^- U_n^{-T}\}]^{-1} U_n^-$, where

$$(2.3) \quad R(n) = \begin{bmatrix} R_1 & R_2 & \cdots & R_n \\ R_2 & R_3 & \cdots & R_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ R_n & R_{n+1} & \cdots & R_{2n-1} \end{bmatrix}.$$

It follows from (2.2) that the rows beyond the P th block row of $R(n)$ are linearly dependent on the previous ones. This implies that $Y_n^+|U_n^-$ is spanned by $R(P, n)[E\{U_n^- (U_n^-)^T\}]^{-1} U_n^-$, where

$$(2.4) \quad R(P, n) = \begin{bmatrix} R_1 & R_2 & \cdots & R_n \\ R_2 & R_3 & \cdots & R_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ R_P & R_{P+1} & \cdots & R_{P+n-1} \end{bmatrix}.$$

Suppose that $R(P, n)$ has p linearly independent rows and let x_n denote the vector whose elements are the Euclidean inner products between some maximal set of linearly independent rows of $R(P, n)$ and $[E\{U_n^- U_n^{-T}\}]^{-1} U_n^-$. Then x_n is a vector of minimal dimension which spans $Y_n^+|U_n^-$. It follows that there exist matrices A_n , B_n , C_n and D_n such that

$$(2.5) \quad x_{n+1} = A_n x_n + B_n u_n, \quad y_n = C_n x_n + D_n u_n.$$

Let $U_n^-(P) = (u_{n-1}^T, u_{n-2}^T, \dots, u_{n-P}^T)^T$ denote the P -step past of u_n and let $Y_n^+(P) = (y_n^T, y_{n+1}^T, \dots, y_{n+P-1}^T)^T$ denote the P -step future of y_n . Let us denote

$$R = R(P) = E\{Y_n^+(P) U_n^-(P)^T\} = \begin{bmatrix} R_1 & R_2 & \cdots & R_P \\ R_2 & R_3 & \cdots & R_{P+1} \\ \vdots & \vdots & \ddots & \vdots \\ R_P & R_{P+1} & \cdots & R_{2P-1} \end{bmatrix}.$$

Since the columns beyond the P th block column of $R(P, n)$ are linearly dependent on the previous ones, R is of rank p . Since we have

$$E\{E\{Y_n^+(P)|U_n^-\}U_n^-(P)^T\} = E\{Y_n^+(P)U_n^-(P)^T\}$$

and since x_n consists of elements of $E\{Y_n^+(P)|U_n^-\}$, there exists a p -dimensional vector ω_n of elements of $U_n^-(P)$, such that the matrix $E\{x_n\omega_n^T\}$ is of full rank. By simple operations on (2.5) one obtains

$$(2.6a) \quad A_n = E\{x_{n+1}\omega_n^T\}[E\{x_n\omega_n^T\}]^{-1},$$

$$(2.6b) \quad B_n = E\{x_{n+1}u_n^T\}[E\{u_nu_n^T\}]^{-1},$$

$$(2.6c) \quad C_n = E\{y_n\omega_n^T\}[E\{x_n\omega_n^T\}]^{-1},$$

$$(2.6d) \quad D_n = E\{y_nu_n^T\}[E\{u_nu_n^T\}]^{-1}.$$

This is, essentially, Akaike's representation [6]. We note that the term D_nu_n is missing in Akaike's representation due to inclusion of u_n in U_n^- . Here we have chosen to obtain the form (2.5) which is often encountered in stochastic system estimation and control.

We proceed to obtain explicit expressions for the matrices A_n , B_n , C_n and D_n . Let W denote the submatrix of R which consists of the intersection between the first maximal set of linearly independent rows and an equal number of the first linearly independent columns of R , i.e., the elements of R which belong to both sets. Let

$$S = \begin{bmatrix} R_2 & R_3 & \cdots & R_{P+1} \\ R_3 & R_4 & \cdots & R_{P+2} \\ \vdots & \vdots & \ddots & \vdots \\ R_{P+1} & R_{P+2} & \cdots & R_{2P} \end{bmatrix}$$

and denote by U the matrix obtained by intersecting the rows and the columns of S having the same indices as the rows and the columns of W in R . Then, by the arguments leading to (2.6), we obtain

$$(2.7a) \quad A_n = A = UW^{-1}.$$

Denote by V the matrix which consists of the intersection between the first q columns of R with its first p linearly independent rows; then

$$(2.7b) \quad B_n = B = V[\text{cov}(u_n)]^{-1}.$$

Let Z denote the matrix which consists of the intersection of the first block row of R with the columns of W . Then

$$(2.7c) \quad C_n = C = ZW^{-1}.$$

Finally,

$$(2.7d) \quad D_n = D = R_0[\text{cov}(u_n)]^{-1}.$$

A minimal order representation for the input-output relationship between u_n and y_n is then given by

$$x_{n+1} = Ax_n + Bu_n, \quad y_n = Cx_n + Du_n$$

where A , B , C and D are given by (2.7a-d). We note that due to the finite initial input time, $n = 0$, imposed in the definition of U_n^- , the state x_n , as defined above is nonstationary. Passing to initialization in the remote past, x_n becomes stationary, with $\Pi = E\{x_nx_n^T\}$ satisfying

$$\Pi = A\Pi A^T + B \text{cov}(u_n) B^T.$$

We have seen that the key to the construction of a minimal order representation for the input-output relationship between the processes u_n and y_n from their cross-correlation function is the construction of the matrix W , i.e., the selection of maximal sets of linearly independent rows and linearly independent columns of the matrix R . The following procedure, based on a technique for partial realization from impulse responses suggested in [13], selects a maximal set of linearly independent columns of R . Modifying the technique so as to select linearly independent rows of R instead of linearly independent columns is straightforward. Define

$$\tilde{R} = R(2P) = \begin{bmatrix} R_1 & R_2 & \cdots & R_{2P} \\ R_2 & R_3 & \cdots & R_{2P+1} \\ \vdots & \vdots & \ddots & \vdots \\ R_{2P} & R_{2P+1} & \cdots & R_{4P-1} \end{bmatrix}.$$

Set $j=0$ and $i=1$. Check the rank of the matrix $H_{j,i}$ composed of the upper left $(2P-j) \times (jm+i)$ submatrix of \tilde{R} . If $H_{j,i}$ is rank deficient, delete the $(km+i)$ th columns, $k=j, j+1, \dots, P$ from \tilde{R} and repeat the rank check for the new $H_{j,i}$. If $H_{j,i}$ has a full rank and $i < m$, set $i = i+1$ and repeat the rank check. If $i = m$, set $j = j+1$ and $i = 1$. This procedure is continued until rank $H_{j,i} = P$ or until $j = Pm$. The rank p of R is the rank of the final $H_{j,i}$. A maximal set of the first linearly independent columns of R is obtained by taking the first p elements of the columns of $H_{j,i}$.

It can be seen that the above procedure guarantees that each tested matrix has, at most, rank deficiency of order one, i.e.,

$$(2.8) \quad \text{rank } H_{j,i} \geq (jm + i - 1)$$

(clearly, $\text{rank } H_{j,i} \leq jm + i$). This property is consistent with the statistical inference method discussed in the following section.

3. Statistical inference. We now turn to the problem of finding a minimal order representation for a vector valued process, when the correlations R_i are not given. Instead, we assume that estimates \hat{R}_i are available. Specifically, we take \hat{R}_i to be the sample covariances. A natural approach to this problem is substituting the values of \hat{R}_i instead of R_i in the representation derived in § 2. The problem is, however, that the minimal order is unknown and \hat{R} is likely to be of full rank for any P due to the stochastic nature of the problem. The decision on the order and on the linear independence of rows and of columns of R , which form the basis for the minimal representation, must be treated, then, by statistical means. Following the procedures suggested in the preceding section, the order and a maximal set of linearly independent rows and linearly independent columns will be selected simultaneously by performing sequential tests on the ranks of the matrices H_i . In the sequel we derive an approximate distribution for such tests.

Let H represent the matrix H_i which is to be checked for having a full rank in the procedure described in the previous section. Let us denote by $\lambda_i = \lambda_i(H)$, $i = 1, 2, \dots$, the eigenvalues of $H^T H$ (or the squared singular values of H) in descending order. Suppose that the smaller dimension of H is k (i.e., that H has k singular values). Clearly, H has a full rank if and only if λ_k is nonzero. A test on H having a full rank can then be formulated as a test on λ_k having a nonzero value. Since our derivation of the test distribution will require the assumption that λ_{k-1} is nonzero when testing λ_k , and since this assumption is consistent with our rank checking procedure, we shall make it implicitly when necessary.

Let r denote a vector of the distinct scalar elements of $R(P)$ arranged in some order. Let \hat{r} denote an estimate of r and denote by $\hat{\lambda}_i = \lambda_i(\hat{H})$, $i = 1, \dots, k$ the eigenvalues of $\hat{H}^T \hat{H}$, where the latter is obtained from $H^T H$ by substituting in it the values \hat{R}_j instead of R_j . The second order Taylor series expansion of $\hat{\lambda}_k$ about r is

$$(3.1) \quad \hat{\lambda}_k = \lambda_k + f^T \tilde{r} + \frac{1}{2} \tilde{r}^T F \tilde{r}$$

where

$$\tilde{r} = \hat{r} - r, \quad f = \frac{\partial \lambda_k}{\partial r}$$

and

$$F_{i,j} = \frac{\partial^2 \lambda_k}{\partial r_i \partial r_j}$$

where r_i is the i th element of r . As is well known (e.g., [13, p. 228, Thm. 14]) the sample covariance \hat{R}_i is asymptotically normally distributed with mean R_i and covariance inversely proportional to the number of measurements T . Therefore, the components of \tilde{r} are asymptotically of the order of $1/\sqrt{T}$. This justifies the second-order Taylor series approximation given above for large enough T . We note that under the tested hypothesis we have $\lambda_k = 0$. Since $H^T H \geq 0$, $\lambda_k = 0$ is a global minimum. Hence, we have $f = 0$, so (3.1) assumes the form

$$(3.2) \quad \hat{\lambda}_k = \frac{1}{2} \tilde{r}^T F \tilde{r}.$$

The vector \tilde{r} is asymptotically distributed as a normal random vector. Let us denote by Σ the covariance of \tilde{r} and by Θ the matrix of eigenvectors of $\Sigma^{1/2} F \Sigma^{1/2}$, then

$$\Theta^T \Sigma^{1/2} F \Sigma^{1/2} \Theta = L$$

where L is a diagonal matrix whose elements are denoted by α_i , $i = 1, \dots, n$, where n is the dimension of r . Defining

$$z = \Theta \Sigma^{-1/2} \tilde{r}$$

we can see that z is asymptotically distributed as a normal vector with mean zero and variance I_n .

The quadratic form (3.2) can now be written as

$$\hat{\lambda}_k = \sum_{i=1}^n \alpha_i z_i^2$$

where z_i^2 , $i = 1, \dots, n$ are independent chi-squared random variables with one degree of freedom. The characteristic function of $\hat{\lambda}_k$ is obtained as (e.g. [14])

$$\Phi(\omega) = \prod_{j=1}^n (1 - 2i\alpha_j \omega)^{-1/2}.$$

The distribution of $\hat{\lambda}_k$ can be found by integrating $\Phi(\omega)$. Several integration and approximation techniques can be found in, e.g., [14]–[16]. Given a confidence level (i.e., an acceptable probability of making the right decision) the resulting distribution defines a statistical test on the rank of the matrix H . A sequence of such tests, performed in the order of the procedure described in § 2 will provide estimates of the minimal order and maximal sets of linearly independent rows and linearly independent columns of R . The statistical properties of the rank test sequence remain a subject for further research.

It follows from the above derivation that statistical inference on the order and the linearly independent rows of R requires knowledge of F and Σ . These are derived in the following section and in the Appendix as functions of the theoretical correlations R_i . In practice, however, the values of R_i are unknown and will be replaced by estimates \hat{R}_i . Once maximal sets of linearly independent rows and linearly independent columns of R have been estimated, an estimate of the representation presented in § 2 may be obtained by substituting the estimated correlations in the corresponding matrices of (2.7).

4. Deriving the matrix F appearing in the rank test statistics. We now seek the second derivative of the eigenvalues of the matrix $H^T H$ with respect to the vector r , as defined in the previous section. Let

$$(4.1) \quad H^T H = Q \Lambda Q^T$$

be the eigenvalue/eigenvector decomposition of the symmetric matrix $H^T H$. The diagonal elements of Λ will be denoted by $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$, while the columns of Q will be denoted by $\{q_1, q_2, \dots, q_n\}$. All the eigenvalues are assumed to be distinct. Let r_i, r_j denote any two components of the vector r (which may be identical or distinct). We wish to compute

$$\left\{ \frac{\partial \lambda_l}{\partial r_i}; \frac{\partial^2 \lambda_l}{\partial r_i \partial r_j}; 1 \leq l \leq k \right\}.$$

In particular, we are interested in these derivatives for $l = k$, in the special case where $\lambda_k = 0$ and $\{\lambda_1, \lambda_2, \dots, \lambda_{k-1}\}$ are nonzero. As we will see, the computation of the second derivatives requires the computation of the derivatives

$$\left\{ \frac{\partial q_l}{\partial r_i}; 1 \leq l \leq k \right\}.$$

4.1. The first derivative of the eigenvalues of $H^T H$. Let us denote for convenience

$$M(\lambda) = \lambda I - H^T H = Q[\lambda I - \Lambda]Q^T.$$

By the definition of the eigenvalues we have

$$\det(M(\lambda))|_{\lambda=\lambda_l} = 0.$$

Differentiation yields

$$\begin{aligned} \frac{\partial}{\partial r_i} \det(M(\lambda_l)) &= \text{tr} \left\{ \text{adj}(M(\lambda_l)) \frac{\partial M(\lambda_l)}{\partial r_i} \right\} \\ &= \text{tr} \left\{ \text{adj}(M(\lambda_l)) \left(\frac{\partial \lambda_l}{\partial r_i} I - \frac{\partial}{\partial r_i} H^T H \right) \right\} = 0 \end{aligned}$$

where $\text{tr}\{\}$ denotes the trace of a matrix, and $\text{adj}(\)$ denotes the adjugate matrix. Hence we get for the first derivative,

$$(4.2) \quad \frac{\partial \lambda_l}{\partial r_i} = \frac{\text{tr} \{ \text{adj}(M(\lambda_l)) (\partial/\partial r_i) H^T H \}}{\text{tr} \{ \text{adj}(M(\lambda_l)) \}}.$$

4.2. The first derivative of the eigenvectors of H^TH . We have

$$M(\lambda_l)q_l = 0.$$

Hence,

$$\frac{\partial M(\lambda_l)}{\partial r_i} q_l + M(\lambda_l) \frac{\partial q_l}{\partial r_i} = 0,$$

or

$$(4.3) \quad M(\lambda_l) \frac{\partial q_l}{\partial r_i} = -\frac{\partial M(\lambda_l)}{\partial r_i} q_l.$$

This equation cannot be solved directly because $M(\lambda_l)$ is singular. Therefore we proceed as follows. Note that by (4.3) and by the symmetry of $M(\lambda_l)$,

$$(4.4) \quad q_l^T M(\lambda_l) \frac{\partial q_l}{\partial r_i} = -q_l^T \frac{\partial M(\lambda_l)}{\partial r_i} q_l = 0.$$

Also, since q_l is a unit vector,

$$(4.5) \quad q_l^T q_l = 1; \quad \text{hence, } q_l^T \frac{\partial q_l}{\partial r_i} = 0.$$

Consider (4.3) after premultiplying by Q^T :

$$(4.6) \quad \begin{bmatrix} \lambda_l - \lambda_1 & & & & 0 \\ & \ddots & & & \\ & & \lambda_l - \lambda_{l-1} & & \\ & & & 0 & \\ 0 & & & & \ddots \\ & & & & & \lambda_l - \lambda_k \end{bmatrix} Q^T \frac{\partial q_l}{\partial r_i} = -Q^T \frac{\partial M(\lambda_l)}{\partial r_i} q_l.$$

By (4.4), the l th equation in (4.6) is an identity $0 = 0$. On the other hand, by (4.5) we can replace the zero on the main diagonal of the matrix in the left-hand side of (4.6) by one and still get an equality. Hence, (4.6) has a unique solution

$$(4.7) \quad \frac{\partial q_l}{\partial r_i} = -Q \begin{bmatrix} \lambda_l - \lambda_1 & & & & 0 \\ & \ddots & & & \\ & & \lambda_l - \lambda_{l-1} & & \\ & & & 1 & \\ 0 & & & & \ddots \\ & & & & & \lambda_l - \lambda_k \end{bmatrix}^{-1} Q^T \left(\frac{\partial \lambda_l}{\partial r_i} I - \frac{\partial}{\partial r_i} H^T H \right) q_l.$$

4.3. An expression for $\text{adj}(M)$. In order to compute the second derivatives of the eigenvalues, we will need an explicit expression for $\text{adj}(M(\lambda_l))$. Since $M(\lambda_l)$ is singular, it is meaningless to use $\text{adj}(M(\lambda_l)) = \det(M(\lambda_l)) [M(\lambda_l)]^{-1}$, and we will have to use a limiting process. For all $\lambda \neq \lambda_1, \lambda_2, \dots, \lambda_k$ we have

$$M^{-1}(\lambda) = Q(\lambda I - \Lambda)^{-1} Q^T, \quad \det M(\lambda) = \prod_{m=1}^k (\lambda - \lambda_m).$$

Hence,

$$\text{adj}(M(\lambda)) = Q \Psi Q^T$$

where Ψ is a diagonal matrix whose l th diagonal entry is $\prod_{m \neq l} (\lambda - \lambda_m)$.

Now let λ approach λ_l , and use the fact that $\text{adj}(M(\lambda))$ is a continuous function of λ . Clearly, all the entries of Ψ go to zero, except for the l th one that goes to $\prod_{m \neq l} (\lambda_l - \lambda_m)$. Therefore we get

$$(4.8) \quad \text{adj}(M(\lambda_l)) = \left[\prod_{m \neq l} (\lambda_l - \lambda_m) \right] q_l q_l^T.$$

Since all the eigenvalues of $H^T H$ were assumed to be distinct, $\text{adj}(M(\lambda_l))$ is a matrix of rank 1. We also see from (4.8) that

$$\text{tr}[\text{adj}\{M(\lambda_l)\}] = \prod_{m \neq l} (\lambda_l - \lambda_m) \neq 0,$$

i.e., the denominator in (4.2) is nonzero.

4.4. The second derivatives of the eigenvalues. Differentiation of (4.2) with respect to r_j yields

$$(4.9) \quad \frac{\partial^2 \lambda_l}{\partial r_i \partial r_j} = \frac{\text{tr}\{(\partial \text{adj}(M(\lambda_l))/\partial r_j)(\partial/\partial r_i)H^T H + \text{adj}(M(\lambda_l))(\partial^2/\partial r_i \partial r_j)H^T H\}}{[\text{tr}\{\text{adj}(M(\lambda_l))\}]} - \frac{\text{tr}\{\partial \text{adj}(M(\lambda_l))/\partial r_j\} \cdot \text{tr}\{\text{adj}(M(\lambda_l))(\partial/\partial r_i)H^T H\}}{[\text{tr}\{\text{adj}(M(\lambda_l))\}]^2}.$$

In order to use this formula, we only need an expression for the first derivative of $\text{adj}(M(\lambda_l))$. Using (4.8) we get

$$(4.10) \quad \frac{\partial \text{adj}(M(\lambda_l))}{\partial r_j} = \left[\frac{\partial}{\partial r_j} \prod_{m \neq l} (\lambda_l - \lambda_m) \right] q_l q_l^T + \left[\prod_{m \neq l} (\lambda_l - \lambda_m) \right] \left[\frac{\partial q_l}{\partial r_j} q_l^T + q_l \frac{\partial q_l^T}{\partial r_j} \right].$$

The right-hand side of (4.10) can be computed from the previous expressions. Hence we have all the formulas needed to compute the second derivatives. We just mention that the derivatives of $H^T H$ are easy to compute, because the entries are just the various components of r . We have

$$(4.11) \quad \frac{\partial}{\partial r_i} H^T H = \frac{\partial H^T}{\partial r_i} H + H^T \frac{\partial H}{\partial r_i},$$

$$(4.12) \quad \frac{\partial^2}{\partial r_i \partial r_j} H^T H = \frac{\partial H^T}{\partial r_i} \frac{\partial H}{\partial r_j} + \frac{\partial H^T}{\partial r_j} \frac{\partial H}{\partial r_i}.$$

The matrices $\partial H/\partial r_i$ and $\partial H/\partial r_j$ are constant matrices, consisting of 1's and 0's only.

5. Conclusion. This paper has presented a method for identifying a minimal order linear system from input-output data. The crucial problem is identifying the minimal order. By identifying maximal sets of linearly independent rows and linearly independent columns of a finite dimensional Hankel matrix, both the minimal order and a basis for the representation are identified. A statistical inference technique, based on sequential tests on the ranks of submatrices of the Hankel matrix of sample correlations and on an approximate distribution for the rank test statistics has been derived. The approximate distribution converges to the true one as the number of observations becomes large. The proposed method seems to close a gap between the minimal realization problem and the problem of system identification from input-output measurement data.

Appendix. The covariance of the normal vector \tilde{r} appearing in the quadratic form. It was shown in § 3 that in order to specify the test statistics distribution, the value of $\Sigma = \text{cov}(\tilde{r})$ is needed. We have

$$E\{\hat{r}\hat{r}^T\} = E\{\hat{r}\hat{r}^T\} - r r^T.$$

Since the vector r is constructed from elements of the correlations R_i , $i = 1, 2, \dots$, the value of rr^T can be immediately obtained in terms of cross products $R_i R_j$, $i, j = 1, 2, \dots$. We next derive expressions for the second moments of the entries of \hat{R}_l and \hat{R}_k from which $E\{\hat{r}\hat{r}^T\}$ is immediately obtainable.

Let the sample covariances $\{\hat{R}_l\}$ be defined by

$$(A.1) \quad \hat{R}_l = \frac{1}{T-l} \sum_{t=l+1}^T y_t u_{t-l}^T.$$

The (i, j) th entry of \hat{R}_l will be denoted by $\hat{R}_{l,ij}$. The sample covariances thus defined are clearly unbiased, i.e.,

$$E\{\hat{R}_{l,ij}\} = \frac{1}{T-l} \sum_{t=l+1}^T E\{y_{t,i} u_{t-l,j}\} = R_{l,ij}.$$

The second moments of the sample covariances are computed as follows.

$$\begin{aligned} \hat{R}_{l,ij} \hat{R}_{k,gh} &= \left(\frac{1}{T-l} \sum_{t=l+1}^T y_{t,i} u_{t-l,j} \right) \left(\frac{1}{T-k} \sum_{s=k+1}^T y_{s,g} u_{s-k,h} \right) \\ &= \frac{1}{(T-k)(T-l)} \sum_{t=l+1}^T \sum_{s=k+1}^T y_{t,i} u_{t-l,j} y_{s,g} u_{s-k,h}. \end{aligned}$$

Since $\{u\}$ and $\{y\}$ are jointly Gaussian, we have

$$\begin{aligned} E\{\hat{R}_{l,ij} \hat{R}_{k,gh}\} &= \frac{1}{(T-k)(T-l)} \sum_{t=l+1}^T \sum_{s=k+1}^T \{E\{y_{t,i} u_{t-l,j}\} E\{y_{s,g} u_{s-k,h}\} \\ &\quad + E\{y_{t,i} y_{s,g}\} E\{u_{t-l,j} u_{s-k,h}\} + E\{y_{t,i} u_{s-k,h}\} E\{u_{t-l,j} y_{s,g}\}\}. \end{aligned}$$

Hence,

$$(A.2) \quad \begin{aligned} \text{cov}\{\hat{R}_{l,ij}, \hat{R}_{k,gh}\} &= \frac{1}{(T-l)(T-k)} \sum_{t=l+1}^T \sum_{s=k+1}^T \{E\{y_{t-s,i} y_{0,g}\} E\{u_{t-l-s+k,j} u_{0,h}\} \\ &\quad + E\{y_{t-s+k,i} u_{0,h}\} E\{u_{t-l+s,j} y_{0,g}\}\}. \end{aligned}$$

Note that, since $\{u_i\}$ is uncorrelated, we have

$$(A.3) \quad E\{u_{t-l+s-k,j} u_{0,h}\} = 0, \quad t-l+s-k \neq 0.$$

In practice, the expectations appearing in (A.2) are unknown; hence we approximate them by

$$(A.4a) \quad E\{y_{m,i} y_{0,j}\} \approx \frac{1}{T-m} \sum_{t=m+1}^T y_{t,i} y_{t-m,j},$$

$$(A.4b) \quad E\{y_{m,i} u_{0,j}\} \approx \frac{1}{T-m} \sum_{t=m+1}^T y_{t,i} u_{t-m,j},$$

$$(A.4c) \quad E\{u_{0,i} u_{0,j}\} \approx \frac{1}{T} \sum_{t=1}^T u_{t,i} u_{t,j}.$$

REFERENCES

- [1] E. G. GILBERT, *Controllability and observability in multivariable control systems*, this Journal, 1 (1963), pp. 128-151.
- [2] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152-192.

- [3] B. L. HO AND R. E. KALMAN, *Effective construction of linear state-variable models from input/output functions*, Regelungstechnik, 12 (1966), pp. 545-548.
- [4] J. RISSANEN, *Recursive identification of linear systems*, this Journal, 9 (1971), pp. 420-430.
- [5] L. M. SILVERMAN, *Realization of linear dynamical systems*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 554-567.
- [6] H. AKAIKE, *Stochastic theory of minimal realization*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 667-674.
- [7] J. C. CHOW, *On estimating the orders of an autoregressive moving-average process with uncertain observations*, IEEE Trans. Automat. Control, 17 (1972), pp. 707-709.
- [8] ———, *On the estimation of the order of a moving average process*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 386-387.
- [9] C. M. WOODSIDE, *Estimation of the order of linear systems*, presented at the IFAC Symp. Identification and System Parameter Estimation, Prague, 1970.
- [10] W. E. LARIMORE, *System identification, reduced order filtering and modelling via canonical variates analysis*, Proc. American Control Conference, H. S. Rao and T. Dorato, eds., IEEE, New York, 1983, pp. 445-451.
- [11] H. AKAIKE, *A new look at the statistical model identification*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 716-723.
- [12] T. L. FINE AND W. G. HWANG, *Consistent estimation of system order*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 387-402.
- [13] S.-Y. KUNG, *Multivariable and multidimensional systems: analysis and design*, Ph.D. dissertation, Stanford University, Stanford, CA, 1977.
- [14] E. J. HANNAN, *Multiple Time Series*, John Wiley, New York, 1970.
- [15] N. L. JOHNSON AND S. KOTZ, *Continuous Univariate Distributions—Part II*, John Wiley, New York, 1970.
- [16] J. P. IMHOFF, *Computing the distribution of quadratic forms in normal variables*, Biometrika, 48 (1961), p. 419.
- [17] D. R. JENSEN AND H. SOLOMON, *A Gaussian approximation to the distribution of a definite quadratic form*, J. Amer. Statist. Assoc., 67 (1972).

THE MINIMAL DELAY DECOUPLING PROBLEM: FEEDBACK IMPLEMENTATION WITH STABILITY*

J. M. DION† AND C. COMMAULT†

Abstract. In this paper we solve the minimal delay decoupling problem of linear multivariable systems. We look for feedback implementable solutions which moreover guarantee closed loop stability. We prove that dynamic state feedback decoupling with stability is achievable if the number of independent inputs is large enough to compensate the intrinsic “nondecouplability” of the system. We introduce new feedback invariants characterizing the minimal number of infinite and unstable zeros of the decoupled system.

Key words. linear multivariable systems, decoupling, transfer matrix approach

AMS(MOS) subject classifications. 93C05, 93C35.

1. Introduction. In this paper we solve the minimal delay decoupling problem of linear multivariable systems. When possible we look for feedback implementable solutions which moreover guarantee closed loop stability. We focus our attention on decoupling compensators achieving the minimal McMillan degree of the decoupled system. Most of our results are expressed in terms of new feedback invariants which may be useful for solving other control problems.

The decoupling problem has received a great deal of attention during the last years. For authoritative references representing important steps in the development of this theory, see [1]–[4]. Let us recall briefly the main results in this field:

—Falb and Wolovich [1] solved the decoupling problem for square systems by static state feedback, $u = Fx + Gv$, G nonsingular.

—Morse and Wonham [2], [3] solved the decoupling problem in more general cases by adding auxiliary integrators (“Extended Decoupling Problem”) in a geometric framework. The decoupling condition by a static state feedback control law on the extended system turns out to be the same as the decoupling condition by pure precompensation $u = G(s)v$. Furthermore, for square systems, a minimal stabilizing solution is provided in the sense that it requires the least possible order of dynamic compensation.

—Hautus and Heymann [4] have shown that in the square case the decoupling conditions are the same by static feedback, $u = Fx + Gv$, G nonsingular, and by dynamic state feedback $u = F(s)x + Gv$, G nonsingular. They have shown that this was not the case for nonsquare systems when G was no longer restricted to being nonsingular. An extension for systems defined over unique factorization domains is given in Datta and Hautus [5].

Discussions on the various types of decoupling compensators may be found in Hautus and Heymann [4] and Willems [6]. In our approach the exogeneous input v does not enter the system directly; v is assumed to be measurable and is used by the designer in order to steer the exogeneous output. A more general setup is considered in Willems [6]. Notice that in the particular case of row by row decoupling the admissibility conditions of output controllability preservation are identical in [3] and [4].

* Received by the editors January 20, 1986; accepted for publication (in revised form) January 20, 1987.

† Laboratoire d'Automatique de Grenoble, Ecole Nationale Supérieure d'Ingénieurs Electriciens de Grenoble, Institut Nationale Polytechnique de Grenoble, B.P. 46, 38402 Saint Martin D'Heres, France.

In the transfer matrix framework of [4] we examine in this paper the row by row decoupling problem for not necessarily injective systems. We are interested in obtaining the minimal delay and the minimal McMillan degree of the decoupled system. When possible we look for feedback implementable solutions which moreover guarantee closed loop stability. In effect, as will be shown further, the feedback solutions require that less dynamics be added than the equivalent precompensations.

Our contributions are as follows:

—We solve the problem of dynamic state feedback decoupling for noninjective systems which was left open in [4]. The decoupling control law does not contain input dynamics as in [3].

—The proposed solution achieves the minimal delay decoupling, i.e., the simplest infinite structure for the decoupled system is achieved. This corresponds to the minimum number of delays in the discrete time formulation.

—The proposed solution gives the minimal McMillan degree of the decoupled system which incidentally means that the decoupled system possesses the minimal number of unstable zeros.

—When state feedback decoupling is not possible, decoupling is achievable by dynamic precompensation with stability for full row rank transfer matrices. We also give in this case the minimum number of unstable and infinite zeros of the decoupled system.

This paper is organized as follows. In § 2 the decoupling problem is formulated and some notations and preliminaries are given. We notice that in this formulation the decoupling by precompensation is quite trivial. Some results concerning the feedback implementation of bicausal precompensators are recalled. For proper rational matrices we recall the notion of Hermite form and introduce the infinite column rank.

Section 3 is devoted to the state feedback decoupling problem. First we show that feedback solutions are advantageous because they need less auxiliary integrators than the equivalent precompensations. Then we solve the dynamic state feedback decoupling problem. This problem turns out to have a solution if the number of independent inputs is large enough to compensate the intrinsic “nondecouplability” of the system. We then introduce feedback invariants which are useful for characterizing the minimal delay decoupled system. When feedback decoupling solutions exist we show that minimal delay decoupling can be achieved by feedback.

In § 4 we study the state feedback decoupling problem with stability. In the first part of this section we briefly recall some mathematical tools concerning factorizations of transfer functions over the P.I.D. of proper rational stable functions. Then as an intermediate result we characterize all the precompensators which are feedback implementable with closed loop stability. We prove that decoupling is achievable with stability if and only if it is achievable without stability considerations. Other feedback invariants characterizing the simplest infinite and unstable structure for the decoupled system are introduced. When feedback decoupling solutions exist we exhibit such a feedback solution. Some concluding remarks end the paper.

2. Preliminaries and problem statement.

A. Problem setting. Let $R(s)$ be the field of rational functions.

$$f(s) = \frac{n(s)}{d(s)} \in R(s)$$

is said to be proper (resp. strictly proper) if $\deg(d(s)) \geq \deg(n(s))$ (resp. $\deg(d(s)) > \deg(n(s))$) where $\deg(n(s))$ denotes the polynomial degree of $n(s)$.

Denote $R_p(s)$ the ring of proper rational functions and $R_p^{p \times m}(s)$ the set of proper rational $p \times m$ transfer matrices.

The units (invertible elements) of the ring $R_p^{m \times m}(s)$ are called bicausal matrices and are characterized by the property that $B(s)$ is a bicausal matrix if and only if

$$\det \left(\lim_{s \rightarrow \infty} B(s) \right) \neq 0.$$

Throughout this paper we will consider matrices and vector spaces over the real or over the rational functions. Notions such as independence, rank, span, etc. should be understood in their usual sense over the basic field.

The system of transfer matrix $T(s) \in R_p^{p \times m}(s)$ is said to be *decoupled* if there exist nonzero positive integers $m_1 \cdots m_p$ satisfying

$$\sum_{i=1}^p m_i = m$$

such that $T(s)$ has the following diagonal form:

$$T(s) = \begin{bmatrix} T_{11}(s) & & 0 \\ & \ddots & \\ 0 & & T_{pp}(s) \end{bmatrix} = \text{diag} (T_{11}(s) \cdots T_{pp}(s))$$

with

$$T_{ii}(s) \in R_p^{l \times m_i}(s), \quad T_{ij}(s) = 0, \quad i \neq j.$$

This means that each input block defined above influences only one output. If we want this influence to be effective the $T_{ii}(s)$ must be nonnull for each i . In this case the system is called *nondegenerate*.

A system $T(s)$ is said to be *decouplable* if there exists a proper compensator $C(s)$ such that $T(s)C(s)$ is decoupled. Notice that, at this point, $C(s) = 0$ is a decoupling compensator. In order to avoid such trivialities we will require the compensated system to be as "output controllable" as the initial system. We will say that $C(s)$ is *admissible* if $\text{rank } T(s)C(s) = \text{rank } T(s)$. This admissibility condition is equivalent to the preservation of the C^∞ controlled output trajectories. This condition which is also called functional output controllability preservation was introduced in [7].

The decoupling problem above defined is always solvable for *surjective* systems ($\text{rank } T(s) = p$). A solution is given by $C(s) = (1/s^k)T^*(s)$, where $T^*(s)$ is a right inverse of $T(s)$ and k is some integer sufficiently large to ensure that $C(s)$ is proper.

In this paper we will focus our attention on decoupling compensators which are feedback implementable. This type of solution is highly desirable from a practical point of view as shown in § 3. In order to do this we need some complements about feedback compensators.

B. Precompensators and feedback. Let $T(s)$ be a $p \times m$ strictly proper rational transfer matrix and (A, B, C) be a realization of $T(s)$, i.e.,

$$\dot{x} = Ax + Bu, \quad x \in R^n, \quad u \in R^m,$$

$$y = Cx, \quad y \in R^p,$$

with $T(s) = C(sI - A)^{-1}B$.

In the following we will restrict our attention to $p \times m$ strictly proper transfer matrices $T(s)$ with null static kernel, i.e., a minimal basis of $\text{Ker}(T(s))$ contains no constant vector.

There is no serious restriction in considering only transfer matrices with zero static kernel, since this can always be achieved by eliminating the nonindependent inputs.

In the state space framework and for a strictly proper transfer matrix $T(s)$ the following can be proved: $T(s)$ possesses a null static kernel if and only if B is injective in any realization (A, B, C) of $T(s)$.

We consider the effect of a dynamic state feedback defined by $u = F(s)x + Gv$ where $F(s)$ is a $m \times n$ proper rational transfer matrix and G is a square full rank constant matrix.

The closed-loop transfer matrix is then:

$$\begin{aligned} T_{FG}(s) &= C(sI - A - BF(s))^{-1}BG \\ &= T(s)(I - F(s)(sI - A)^{-1}B)^{-1}G \\ &= T(s)B(s) \end{aligned}$$

where $B(s)$ is easily seen to be a bicausal matrix.

The converse problem was studied in [8] and [9]. More precisely we will use the following theorem from [9], which is reformulated below using a more traditional terminology.

THEOREM 1. *Let $T(s)$ be a $p \times m$ strictly proper rational matrix with null static kernel and let (A, B, C) be a realization of $T(s)$. Let $C(s)$ be an $m \times m$ proper rational compensator. There exists a $m \times n$ proper rational matrix $F(s)$ and a square full rank constant matrix G such that*

$$T(s)C(s) = C(sI - A - BF(s))^{-1}BG$$

if and only if $C(s)$ is bicausal.

The effect of a bicausal precompensator is then equivalent to the effect of a dynamic state feedback with nonsingular input transformation G . This is true in any realization and in particular in a minimal one. It is shown in [8] that the effect of a bicausal precompensator is also equivalent to the effect of a static state feedback, but when acting on a possibly nonminimal realization of $T(s)$.

A bicausal compensator is clearly admissible in the sense defined above. The decoupling problem with such compensators has been solved in [4].

As seen on the following example it is too restrictive to impose the invertibility of the G matrix; it may be possible to decouple by reducing the number of inputs.

Consider the following:

$$T(s) = \begin{bmatrix} s^{-2} & 0 & s^{-1} \\ 0 & s^{-2} & s^{-1} \end{bmatrix}.$$

It can be proven that $T(s)$ cannot be decoupled by a bicausal precompensator, while the constant admissible compensator

$$G = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

works perfectly. It appears that relaxing this invertibility condition extends significantly the class of decouplable systems.

For this reason, in the following, we will consider the effect of dynamic state feedback defined by $u = F(s)x + Gv$ where $F(s)$ is an $m \times n$ proper rational matrix and G is an $m \times l$ constant matrix. The equivalent precompensator is equal to $B(s)G$ where $B(s)$ is bicausal. The problem to be solved in this paper is the row by row decoupling

problem by this kind of compensator. For practical reasons we are looking for non-degenerate solutions; this imposes that the system under consideration be full row rank (surjective). We now state without proof a simple proposition which will allow us to restrict the set of input transformations G .

PROPOSITION 1. *With the above notations, the system of $p \times m$ surjective strictly proper transfer $T(s)$ is decouplable by an admissible dynamic state feedback if and only if the system is decouplable by an admissible dynamic state feedback where G is a $m \times p$ full rank constant matrix.*

In the following the use of Proposition 1 will allow us to restrict our attention to such compensators.

C. Hermite form and rank at infinity. In the sequel we will need Hermite forms over the ring of proper rational functions. More precisely we have the following theorem.

THEOREM 2 [10]. *Let $T(s)$ be a $p \times m$ surjective proper rational matrix. $T(s)$ can be factorized in $T(s) = H(s)B(s)$ where $B(s)$ is a bicausal matrix (unit of the ring $R_p^{m \times m}(s)$) and $H(s) = [\bar{H}(s), 0]$ where*

$$\bar{H}(s) = \begin{bmatrix} 1/\pi^{n_1} & & 0 \\ & \ddots & \\ h_{ij} & & 1/\pi^{n_p} \end{bmatrix}$$

where $\pi = s + a$ is an arbitrary polynomial of degree one and

$$h_{ij} = \gamma / \pi^{n_{ij}},$$

$n_{ij} < n_i$ with n_i, n_{ij} positive integers and γ is a polynomial.

$H(s)$ is called the π -Hermite form of $T(s)$ over the ring of proper rational functions. $H(s)$ is uniquely determined by $T(s)$ and π .

Let us call $I_\pi(s) = \bar{H}^{-1}(s)$ the π -interactor of $T(s)$. Notice that this definition coincides with the definition of the interactor $I(s)$ given in [11] where $\pi = s$; for further details see [12].

Since the bicausal matrix possesses neither poles nor zeros at infinity, $H(s)$ (or $I_\pi(s)$) contains all the information about the behaviour at infinity of $T(s)$. A related and useful notion is the notion of rank at infinity.

DEFINITION 1. Let $T(s)$ be a $p \times m$ rational matrix. Denote $T_i(s)$ as the i th column of $T(s)$. Define the integers r_i such that: $\lim_{s \rightarrow \infty} T_i(s)s^{-r_i} = t_i$, where t_i is a nonnull constant vector when $T_i(s)$ is nonnull.

The *column rank at infinity* of $T(s)$ is defined as the rank (over the real field) of the matrix $T_0 = [t_1, \dots, t_m]$.

It appears that if r_i is negative when $T_i(s)$ is nonnull, then $-r_i$ is the infinite zero order of $T_i(s)$; otherwise r_i is the infinite pole order of $T_i(s)$.

The above notion corresponds with the intuitive rank notion. We have to be careful in using this concept because, for example, the row rank at infinity may be different from the column rank at infinity. This can be seen in the following example:

$$T(s) = \begin{bmatrix} s^{-2} & 0 & 0 \\ 0 & s^{-2} & 0 \\ s^{-1} & s^{-1} & s^{-1} \end{bmatrix}$$

has infinite column rank equal to one, and infinite row rank equal to three.

3. Dynamic state feedback decoupling. This section is subdivided into three parts. In the first we stress the fact that feedback decoupling compensators are simpler, in

the sense that they need less auxiliary dynamics, than equivalent decoupling precompensators. Necessary and sufficient conditions for dynamic state feedback decoupling are given in part two. In part three we characterize the minimal McMillan degree achievable for the decoupled system. The result can also be stated in terms of the simplest infinite structure achievable or in the discrete time framework by the minimum number of delays for the decoupled system.

3.1. Why feedback implementation? Consider first the following illustrative example:

$$T(s) = \begin{bmatrix} s^{-1} & s^{-2} & s^{-3} \\ s^{-1} & 2s_2^{-x} & s^{-2} + 2s^{-3} \end{bmatrix}.$$

It can be shown that the system whose transfer matrix is $T(s)$ is decouplable neither by a constant admissible compensator nor by a dynamic state feedback $u = F(s)x + Gv$, G nonsingular.

There exists an admissible decoupling precompensator $C(s)$ of McMillan degree 2.

$$C(s) = \begin{bmatrix} 2s^{-1} & -2s^{-1} \\ -1 & 2 + s^{-1} \\ 0 & -1 \end{bmatrix}$$

then

$$T(s)C(s) = \begin{bmatrix} s^{-2} & 0 \\ 0 & s^{-2} \end{bmatrix}.$$

Consider the following minimal realization $\Sigma = (A, B, C)$ of $T(s)$:

$$\dot{x} = Ax + Bu, \quad y = Cx$$

with

$$A = \begin{bmatrix} 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}.$$

The precompensator is implementable by static state feedback $u = Fx + Gv$ with

$$F = \begin{bmatrix} 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & 0 \\ -1 & 2 \\ 0 & -1 \end{bmatrix}.$$

We can verify that

$$C(sI - A - BF)^{-1}BG = \begin{bmatrix} s^{-2} & 0 \\ 0 & s^{-2} \end{bmatrix}.$$

In this case the complexity (McMillan degree) of the feedback decoupling compensator is smaller than the complexity of any admissible decoupling precompensator. We will show that this result is general.

PROPOSITION 2. *Consider the system of $p \times m$ strictly proper transfer matrix $T(s)$ with null static kernel and (A, B, C) a minimal realization of $T(s)$. Let $C(s)$ be a feedback implementable (by static or dynamic state feedback) precompensator; then the feedback implementation of $C(s)$, $F(s)$, has McMillan degree less than or equal to that of $C(s)$.*

Proof. From Proposition 1, $C(s)$ is a piece of a bicausal matrix, that is to say that there exists a $m \times (m-p)$ constant matrix C_1 such that:

$$B(s) = [C(s), C_1] \text{ is bicausal.}$$

$B(s)$ can be written as $B(s) = (I - R(s))^{-1}G_1$ with $R(s)$ strictly proper and G_1 nonsingular.

Choose $F(s) = R(s)B'(sI - A)$ where B' is a left inverse of B and

$$G = G_1 \begin{bmatrix} I \\ 0 \end{bmatrix} \text{ where } G \text{ is an } m \times p \text{ constant matrix.}$$

Notice that $F(s)$ is proper since $R(s)$ is strictly proper. Furthermore the effect of the feedback compensator $u = F(s)x + Gv$ is equivalent to the effect of $C(s)$ because

$$(I - F(s)(sI - A)^{-1}B)^{-1}G = (I - R(s)B'(sI - A)(sI - A)^{-1}B)^{-1}G = C(s).$$

On the other hand, $B(s)$ and $C(s)$ have the same McMillan degree δ . A bicausal matrix possesses the same number of finite poles and zeros; then $B^{-1}(s)$ and $R(s) = I - G_1B^{-1}(s)$ have McMillan degree δ .

From the expression of $F(s)$ we conclude that $F(s)$ has a McMillan degree lower than or equal to δ . \square

As we have shown above, feedback implementation leads to simpler controllers, other than the well-known nice properties of feedback controllers.

3.2. State feedback decoupling. In this section we give a necessary and sufficient condition for admissible dynamic state feedback decoupling. It turns out that the problem is solvable if the number of inputs is sufficiently large in order to compensate the rank deficiency at infinity of the transfer matrix interactor. The result is expressed in the following way.

THEOREM 3. *The system whose transfer matrix $T(s)$ is a $p \times m$ strictly proper rational surjective transfer matrix with null static kernel, is decouplable by an admissible dynamic state feedback on a minimal realization of $T(s)$ if and only if*

$$m \geq 2p - k$$

where k is the column rank at infinity of $I(s)$, the interactor of $T(s)$.

Proof. Consider first the sufficiency. The proof is constructive. We will build an admissible dynamic state feedback implementable decoupling compensator. By using Proposition 1 this compensator is of the form $C(s) = B(s)G$ where $B(s)$ is a bicausal matrix and G is a $m \times p$ full rank constant matrix. As noted before $C(s)$ is a "piece" of bicausal matrix characterized by

$$\lim_{s \rightarrow \infty} C(s) = C_0,$$

C_0 being a full column rank matrix. $C(s)$ is a full column rank at infinity matrix, moreover the infinite zero orders of the columns are all equal to zero.

From Theorem 2, $T(s)$ can be factorized as follows:

$$T(s) = [\bar{H}(s), 0]B(s) = [I^{-1}(s), 0]B(s)$$

where $B(s)$ is a bicausal matrix and $I(s)$ is the interactor of $T(s)$. Consider the compensator

$$C(s) = B^{-1}(s) \begin{bmatrix} I_1(s) \\ X \end{bmatrix}$$

where $I_1(s) = I(s) \text{diag}(s^{-r_1}, \dots, s^{-r_p})$, the r_i 's being chosen as in Definition 1 such that

$$\lim_{s \rightarrow \infty} I_1(s) = I_0$$

and I_0 has no null column. By definition I_0 has a real rank k . X is an $(m-p) \times p$ constant matrix such that

$$\begin{bmatrix} I_0 \\ X \end{bmatrix}$$

has a full column rank. This is always possible from the assumption $m \geq 2p - k$. For this select k independent rows in I_0 , choose for the $p - k$ first rows of X such that with the k 's preceeding they form a $p \times p$ invertible matrix.

The above compensator $C(s)$ is such that

$$C(s) = C_0 + \bar{C}(s)$$

where

$$\begin{aligned} C_0 &= B_0^{-1} \begin{bmatrix} I_0 \\ X \end{bmatrix}, \\ B(s) &= B_0 + \bar{B}(s), \\ \lim_{s \rightarrow \infty} \bar{C}(s) &= 0, \quad \lim_{s \rightarrow \infty} \bar{B}(s) = 0. \end{aligned}$$

By construction C_0 is a full column rank constant matrix, and $C(s)$ is then a piece of a bicausal matrix. C_0 being full column rank, there exists a full column rank constant matrix Q_0 such that (C_0, Q_0) is an invertible matrix. Define $B_1(s) = (C(s), Q_0)$, $B_1(s)$ is a bicausal matrix.

Since

$$C(s) = B_1(s) \begin{bmatrix} I \\ 0 \end{bmatrix}$$

and using Theorem 1 the precompensator $C(s)$ is implementable by dynamic state feedback on a minimal realization of $T(s)$.

Furthermore,

$$T(s)C(s) = \text{diag}(s^{-r_1}, \dots, s^{-r_p}).$$

It follows that $C(s)$ is an admissible decoupling compensator.

Consider now the necessity. Suppose that the problem is solvable by Proposition 1 there exists an admissible decoupling compensator $C(s)$; dynamic state feedback implementable such that:

$$T(s)C(s) = \text{diag}(d_1(s) \cdots d_p(s)) = D(s)$$

where $d_i(s)$ are proper nonnull rational functions.

From Theorem 2 we have

$$T(s)C(s) = [I^{-1}(s), 0]B(s)C(s) = D(s).$$

It follows that

$$[I_p, 0]B(s)C(s) = I(s)D(s)$$

where I_p is the $p \times p$ identity matrix.

Denote

$$V(s) = \begin{bmatrix} V_1(s) \\ V_2(s) \end{bmatrix} = B(s)C(s),$$

where $V_1(s)$ is a $p \times p$ proper rational matrix and $V_2(s)$ is an $(m-p) \times p$ proper rational matrix. $V(s)$ is a piece of a bicausal matrix; its column rank at infinity is equal to p .

Moreover $V_1(s) = I(s)D(s)$ is not altered by a right multiplication with a nonsingular diagonal matrix.

$V(s)$ is then such that

$$\lim_{s \rightarrow \infty} V(s) = \lim_{s \rightarrow \infty} \begin{bmatrix} V_1(s) \\ V_2(s) \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = V$$

where V_1 is a constant matrix of real rank k and V is a constant matrix of real rank p .

It is then clear that V_2 must have at least $p-k$ rows, so $m-p \geq p-k$, which ends the proof. \square

Remark. When $m \geq 2p-1$ for any surjective system, decoupling is always achievable, because k is at least 1. Furthermore, as shown on the following example, in some cases we need effectively $2p-1$ independent inputs for dynamic state feedback decoupling.

$$\begin{aligned} T(s) &= \begin{bmatrix} s^{-1} & 0 & 0 & s^{-2} & 0 \\ 0 & s^{-1} & 0 & s^{-3} & 0 \\ -s^{-1} & -s^{-1} & s^{-2} & -s^{-2}-s^{-3} & s^{-3} \end{bmatrix} \\ &= \begin{bmatrix} s^{-1} & 0 & 0 & 0 & 0 \\ 0 & s^{-1} & 0 & 0 & 0 \\ -s^{-1} & -s^{-1} & s^{-2} & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & s^{-1} & 0 \\ 0 & 1 & 0 & s^{-2} & 0 \\ 0 & 0 & 1 & 0 & s^{-1} \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} = [I^{-1}(s)0]B(s) \end{aligned}$$

where

$$I(s) = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ s^2 & s^2 & s^2 \end{bmatrix}$$

and $B(s)$ is bicausal.

In this case the column rank at infinity k of $I(s)$ is one, then the five inputs are necessary for decoupling.

At this point we note that dynamic decoupling is actually related to the column rank at infinity of $I(s)$, and *not* to the row rank at infinity of the transfer matrix, which could have been reasonably conjectured from existing literature [14]. In the previous example, the row rank at infinity of $T(s)$ is two while k is one.

3.3. Minimal delay decoupling. In this section we will focus our attention on the minimal McMillan degree achievable for the decoupled system. As will be seen further what we minimize in fact is the infinite zero order of each entry of the decoupled system. In the discrete time framework this corresponds to the appealing minimal delay decoupling problem. The dual problem using a post compensator is studied in [13]. The minimal delays will be characterized in terms of new feedback invariants. Furthermore we prove that when they exist, feedback solutions lead to the same minimal delays as precompensation.

We focus our attention on the integers r_i given by Definition 1 and appearing in the proof of Theorem 3.

DEFINITION 2. Let $T(s)$ be a $p \times m$ surjective proper rational matrix and $I(s)$ its interactor. Denote $I_i(s)$ the i th column of $I(s)$. Let r_i be the integer such that

$$\lim_{s \rightarrow \infty} I_i(s)s^{-r_i} = t_i \quad \text{where } t_i \text{ is a nonnull constant vector.}$$

The integers r_1, \dots, r_p are called the *decoupling invariants* of $T(s)$. $\gamma = \sum_{i=1}^p r_i$ is called the *decoupling degree* of $T(s)$.

Notice that the decoupling invariants of $T(s)$ are invariant under the right multiplication of $T(s)$ by bicausal matrices, hence they are invariant under dynamic state feedback, with nonsingular input transformation.

Furthermore $I(s)$ (resp. the π -interactor $I_\pi(s)$) is a polynomial matrix. It follows that r_i is the maximal polynomial degree or the infinite pole order of the i th column of $I(s)$ (resp. $I_\pi(s)$).

It will be shown in the following theorem that the r_i 's are closely related with the "simplest" decoupled system achievable by admissible compensation.

As in Proposition 1 it can be proven easily that concerning McMillan degree minimality there is no loss of generality in considering $m \times p$ full rank precompensators. Then the decoupled system will be diagonal.

THEOREM 4. Consider a system whose transfer matrix $T(s)$ is a $p \times m$ strictly proper rational surjective transfer matrix with null static kernel.

(i) The system is always decouplable by an admissible proper precompensator. The minimal McMillan degree achievable for the decoupled system is γ the decoupling degree of $T(s)$. In this case the i th diagonal entry of the decoupled system has McMillan degree r_i .

(ii) If the system is decouplable by an admissible dynamic state feedback compensator on a minimal realization of $T(s)$, the minimal McMillan degree achievable for the decoupled system is γ the decoupling degree of $T(s)$. In this case the i th diagonal entry of the decoupled system has McMillan degree r_i .

Proof. Let us first prove (i). Build a precompensator satisfying the requirements of Theorem 4. Let $T(s)$ be factorized as in Theorem 2:

$$T(s) = [I^{-1}(s), 0]B(s).$$

Consider

$$C'(s) = I(s) \text{diag}(s^{-r_1}, \dots, s^{-r_p}).$$

Choose the admissible proper precompensator

$$C(s) = B^{-1}(s) \begin{bmatrix} C'(s) \\ C''(s) \end{bmatrix}$$

where $C''(s)$ is any $(m-p) \times p$ transfer matrix. We have

$$T(s)C(s) = \text{diag}(s^{-r_1}, \dots, s^{-r_p}).$$

$C(s)$ is then a decoupling compensator with the required properties. Show now that r_i is the minimal possible McMillan degree of the i th entry of the decoupled system.

Let $C(s)$ be a proper admissible decoupling compensator. We have

$$T(s)C(s) = D(s) = \text{diag}(d_1(s), \dots, d_p(s))$$

where the $d_i(s)$'s are scalar rational functions. Factorize $T(s)$ as in Theorem 2:

$$T(s) = [I^{-1}(s), 0]B(s).$$

We have

$$[I, 0]B(s)C(s) = I(s)D(s).$$

Denote

$$\begin{bmatrix} C'(s) \\ C''(s) \end{bmatrix} = B(s)C(s) \quad \text{where } C'(s) \in \mathbb{R}_p^{p \times p}(s).$$

It follows that

$$C'(s) = I(s)D(s).$$

Denote $I_i(s)$ as the i th column of $I(s)$.

$C'(s)$ is a proper transfer matrix, as is $I(s)D(s)$. Then $I_i(s)d_i(s)$ is proper. It follows that $d_i(s)$ possesses an infinite zero of order at least r_i and that $d_i(s)$ has a McMillan degree larger than or equal to r_i . γ is then the minimal possible McMillan degree.

Consider now (ii). When conditions of Theorem 3 are fulfilled, the constructive part of the proof provides us with a decoupled system of McMillan degree γ , the i th entry having McMillan degree r_i .

Since dynamic state feedback compensators are implementable by precompensation the minimality here comes from (i). \square

In order to illustrate the above theorem let us consider the following example:

$$T(s) = \begin{bmatrix} s^{-2} & 0 & s^{-5} \\ -s^{-1} & s^{-3} & 0 \end{bmatrix}.$$

The system of transfer matrix $T(s)$ is trivially decouplable by the precompensator

$$C_1(s) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix},$$

in this case

$$T(s)C_1(s) = \begin{bmatrix} s^{-5} & 0 \\ 0 & s^{-3} \end{bmatrix}.$$

The interactor of $T(s)$ is

$$I(s) = \begin{bmatrix} s^2 & 0 \\ s^4 & s^3 \end{bmatrix};$$

then $r_1 = 4$, $r_2 = 3$, $\gamma = 7$.

When we use $C_1(s)$ the McMillan degree of the decoupled system is 8. From (i) of Theorem 4 there exists an admissible proper decoupling precompensator $C_2(s)$ leading to the minimal McMillan degree $\gamma = 7$. Following the proof of (i) we get:

$$C_2(s) = \begin{bmatrix} s^{-2} & 0 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}, \quad T(s)C_2(s) = \begin{bmatrix} s^{-4} & 0 \\ 0 & s^{-3} \end{bmatrix}.$$

The system of transfer matrix $T(s)$ is decouplable by an admissible dynamic state feedback compensator on a minimal realization of $T(s)$ since

$$m = 3 \geq 2p - k = 4 - 1 \quad (\text{see Theorem 3}).$$

Then from (ii) of Theorem 4 there exists an admissible precompensator $C_3(s)$, dynamic state feedback implementable, leading to the minimal McMillan degree of the decoupled system. Following the proof of (ii) we get:

$$\begin{aligned} C_3(s) &= B^{-1}(s) \begin{bmatrix} I_1(s) \\ X \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & -s^{-3} \\ 0 & 1 & -s^{-1} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s^{-2} & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} s^{-2} & -s^{-3} \\ 1 & 1 - s^{-1} \\ 0 & 1 \end{bmatrix} \end{aligned}$$

and

$$T(s)C_3(s) = \begin{bmatrix} s^{-4} & 0 \\ 0 & s^{-3} \end{bmatrix}$$

of minimal McMillan degree.

4. Dynamic state feedback decoupling with stability. In this section we will consider the dynamic state feedback decoupling as before but with the requirement of closed loop stability. As in § 3 the problem will be solvable if the number of inputs is large enough. Surprisingly the decoupling condition is exactly the same as before (without stability considerations).

We introduce other feedback invariants called “stable decoupling invariants” which are the counterpart with stability of the r_i ’s defined in § 3.

We prove that the minimal delay decoupling problem has the same solution as without stability. But in this case the minimal McMillan degree of the decoupled system is generally larger. Minimizing the McMillan degree amounts to minimizing both the infinite and the unstable structure. In other words the proposed solution leads to a decoupled system having the least possible number of delays and unstable zeros. This is very important in practice.

We will present first some mathematical preliminaries and characterize all the precompensators that are feedback implementable with closed loop stability.

4.1. Preliminaries. As in the case of polynomials we can build on $R_{ps}(s)$, Smith forms, Hermite forms, and left and right coprime factorizations. This results from the existence on this ring of a “degree.” $R_{ps}(s)$ is a Euclidean domain, defining the degree δ of $f(s) \in R_{ps}(s)$ as follows: $\delta(f(s)) = d_1(f(s)) + d_2(f(s))$ where $d_1(f(s))$ is the infinite zero order of $f(s)$ and $d_2(f(s))$ is the number of unstable zeros of $f(s)$ counted with their multiplicity [15], [16], [17].

The stability domain under consideration is any region in the complex plane symmetrically located with respect to the real axis and including at least one point of the real axis.

Recall now the Hermite form over $R_{ps}(s)$, the ring of proper rational stable functions.

THEOREM 5 [10], [15]. Let $T(s)$ be a $p \times m$ surjective proper stable rational matrix. $T(s)$ can be factorized in $T(s) = H(s)B(s)$ where $B(s)$ is a bicausal and bistable matrix (unit of $R_{ps}(s)$) and $H(s) = [\bar{H}(s), 0]$ where

$$\bar{H}(s) = \begin{bmatrix} h_{11}(s) & & 0 \\ \vdots & \ddots & \\ h_{ij}(s) & \cdots & h_{pp}(s) \end{bmatrix}.$$

Furthermore the degree of $h_{ij}(s)$ is lower than the degree of $h_{ii}(s)$ for $j < i$. $H(s)$ is called a Hermite form of $T(s)$ over $R_{ps}(s)$.

$H(s)$ is nonunique, it is defined up to units of $R_{ps}(s)$. Uniqueness of $H(s)$ may be obtained by adding some conditions (see [10]).

Consider the following example [10]:

$$T(s) = \begin{bmatrix} \frac{1}{s+1} & \frac{1}{s+2} \\ \frac{1}{s+3} & \frac{1}{s+5} \end{bmatrix}.$$

Choose the stability domain as the open left half plane.

The Hermite form of $T(s)$ over $R_{ps}(s)$ is

$$H(s) = \begin{bmatrix} \frac{1}{\pi} & 0 \\ \frac{s+(a-1)/2}{\pi^2} & \frac{s-1}{\pi^3} \end{bmatrix}$$

where $\pi = s + a$ is any stable polynomial.

Notice that the unstable zero of $H(s)$ is the unstable zero of $T(s)$ (located at $s = 1$) and that $H(s)$ possesses two infinite zeros of order 1 and 2 as $T(s)$.

DEFINITION 3. Let $T(s)$ be a $(p \times m)$ surjective proper rational stable matrix. Consider $H(s) = [\bar{H}(s), 0]$ a Hermite form of $T(s)$ over $R_{ps}(s)$. The rational matrix $I_g(s) = \bar{H}^{-1}(s)$ is called a *generalized interactor* of $T(s)$ over $R_{ps}(s)$.

If one considers $H(s)$ the Hermite form of $T(s)$ over $R_{ps}(s)$ we will speak of the generalized interactor of $T(s)$.

We will present now a complement to Theorem 1 incorporating stability requirements.

THEOREM 6. Let $T(s)$ be a $p \times m$ strictly proper stable rational matrix with null static kernel and let (A, B, C) be a stable realization of $T(s)$. Let $C(s)$ be an $m \times m$ proper rational compensator dynamic state feedback implementable ($u = F(s)x + Gv$, G nonsingular) on the realization (A, B, C) . Then the considered feedback system is stable if and only if $C(s)$ is stable.

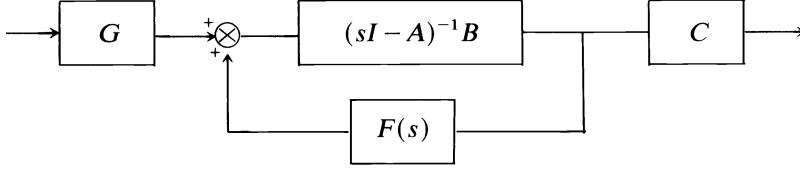
Proof. Necessity. $C(s)$ has to be a bicausal compensator from Theorem 1. Suppose $C(s)$ is unstable. Then either $T(s)C(s)$ is unstable contradicting the external stability, or unstable pole zero cancellations occur and internal stability fails.

Sufficiency. Suppose $C(s)$ bicausal and stable. By Theorem 1, $C(s)$ is implementable by dynamic state feedback. There exist $F(s)$ and G such that

$$T(s)C(s) = C(sI - A - BF(s))^{-1}BG.$$

Now prove the closed loop stability.

We have the following diagram.



We have closed loop stability if the following transfer matrix is stable (see [16]).

$$\begin{bmatrix} (I - Z(s)F(s))^{-1} & Z(s)(I - F(s)Z(s))^{-1} \\ F(s)(I - Z(s)F(s))^{-1} & (I - F(s)Z(s))^{-1} \end{bmatrix} \quad \text{where } Z(s) = (sI - A)^{-1}B.$$

By hypothesis $Z(s)$, $(I - F(s)Z(s))^{-1}$ and $Z(s)(I - F(s)Z(s))^{-1}$ are stable, where $(I - F(s)Z(s))^{-1}$ is the bicausal stable precompensator and $Z(s)(I - F(s)Z(s))^{-1}$ is the closed loop transfer matrix.

Since $\det(I - Z(s)F(s)) = \det(I - F(s)Z(s))$ and $(I - F(s)Z(s))^{-1}$ is stable, then $(I - Z(s)F(s))^{-1}$ is stable.

It remains to prove that $F(s)(I - (sI - A)^{-1}BF(s))^{-1}$ is stable. Let $N(s)D^{-1}(s)$ be a right coprime factorization of $-F(s)$.

$$\begin{aligned} (I + (sI - A)^{-1}BF(s))^{-1} &= ((sI - A) + BF(s))^{-1}(sI - A) \\ &= D(s)((sI - A)D(s) + BN(s))^{-1}(sI - A). \end{aligned}$$

We prove now that $D(s)$ and $(sI - A)D(s) + BN(s)$ are right coprime. Since $N(s)$ and $D(s)$ are right coprime there exist $U(s)$ and $V(s)$ polynomial matrices of appropriate dimensions such that $U(s)N(s) + V(s)D(s) = I$. Choose $\bar{U}(s)$ an $n \times n$ polynomial matrix such that $\bar{U}(s)B = U(s)$; this is always possible since B is full column rank.

Choose $\bar{V}(s) = V(s) - \bar{U}(s)(sI - A)$. $\bar{V}(s)$ is an $n \times n$ polynomial matrix. $\bar{U}(s)$ and $\bar{V}(s)$ satisfy

$$\bar{U}(s)((sI - A)D(s) + BN(s)) + \bar{V}(s)D(s) = U(s)N(s) + V(s)D(s) = I;$$

then $(sI - A)D(s) + BN(s)$ and $D(s)$ are right coprime. Then $D(s)((sI - A)D(s) + BN(s))^{-1}$ is stable, also. From the right coprimeness of $D(s)$ and $(sI - A)D(s) + BN(s)$ we deduce the stability of $((sI - A)D(s) + BN(s))^{-1}$. It follows that $F(s)(I - (sI - A)^{-1}BF(s))^{-1} = N(s)((sI - A)D(s) + BN(s))^{-1}(sI - A)$ is stable. \square

Consider now the stable dynamic state feedback decoupling problem.

4.2. Dynamic state feedback decoupling with stability. We will now solve the dynamic state feedback decoupling with stability. The theorem will be stated for stable systems for clarity of exposition. There is no loss of generality in making this assumption. In fact, when the system transfer matrix is not stable, we will use a static state feedback on a minimal realization of $T(s)$ which assigns all the closed loop poles to stable locations. Denote $T_1(s)$ the closed loop transfer matrix, $T(s)$ and $T_1(s)$ are feedback equivalent on a minimal realization of $T(s)$ without unstable pole zero cancellations. It follows that $T(s)$ is decouplable with stability if and only if $T_1(s)$ is. Let us state first the counterpart of Theorem 1 for stabilizing compensators.

THEOREM 7. *The system whose transfer matrix $T(s)$ is a $(p \times m)$ strictly proper stable surjective transfer matrix with null static kernel is decouplable with stability by an admissible dynamic state feedback implementable on a minimal realization of $T(s)$ if and only if*

$$m \geq 2p - k$$

where k is the infinite column rank of $I(s)$ the interactor of $T(s)$ over $R_p(s)$.

Proof. Necessity. By Theorem 3 if $m < 2p - k$ decoupling is not possible, a fortiori stable decoupling is not possible.

Sufficiency. Consider the sufficiency proof of Theorem 3. A decoupling precompensator $C(s)$ is constructed:

$$C(s) = B_1(s) \begin{bmatrix} I \\ 0 \end{bmatrix} \quad \text{where } B_1(s) \text{ is bicausal.}$$

Let $B_2(s)$ be a diagonal bicausal matrix such that $B_1(s)B_2(s)$ is stable. Choose $B_2(s) = \text{diag}(b_1(s), \dots, b_m(s))$, where the unstable zeros of $b_i(s)$ are the unstable poles of the i th column of $B_1(s)$.

Then $B_1(s)B_2(s)$ is bicausal and stable.

$$Q(s) = B_1(s)B_2(s) \begin{bmatrix} I \\ 0 \end{bmatrix}$$

is a stable admissible decoupling compensator which is implementable by dynamic state feedback with stability on a minimal realization of $T(s)$ by Theorem 6. \square

Remark that the number of necessary inputs for decoupling is the same as in Theorem 3 (without stability). As will be seen later the minimal McMillan degree of the decoupled system with stability is in general larger than in the former case (without stability consideration).

As before we will focus our interest on the minimal McMillan degree achievable by feedback. Let us define first some useful invariants.

DEFINITION 4. Let $T(s)$ be a $(p \times m)$ surjective proper stable rational matrix and $I_g(s)$ its generalized interactor. Decompose $I_g(s)$ as follows:

$$I_g(s) = \bar{I}_g(s) \text{diag}(h_1(s) \cdots h_p(s))$$

where $h_i(s)$ is a rational function such that the i th column of $I_g(s)$ possesses neither poles nor zeros in unstable and infinite locations.

Denote q_i the number of infinite and unstable zeros of $h_i^{-1}(s)$ counted with their multiplicity, i.e., the degree of $h_i^{-1}(s)$ over $R_{ps}(s)$. The integers $q_1 \cdots q_p$ are called the *stable decoupling invariants* of $T(s)$. $\mu = \sum_{i=1}^p q_i$ is called the *stable decoupling degree* of $T(s)$.

Notice that for a fixed i , $h_i(s)$ is nonunique but defines a unique q_i . Moreover the stable decoupling invariants of $T(s)$ are invariant under the right multiplication of $T(s)$ by bicausal and bistable matrices.

As shown in the following theorem the q_i 's are closely related with the "simplest" decoupled system achievable by stable admissible compensation.

THEOREM 8. Consider a system whose transfer matrix $T(s)$ is a $(p \times m)$ strictly proper stable rational surjective transfer matrix with null static kernel.

(i) The system is always decouplable by an admissible proper precompensator with stability. The minimal McMillan degree achievable for the decoupled system is μ the stable decoupling degree of $T(s)$. In this case the i th diagonal entry of the decoupled system has McMillan degree q_i .

(ii) If the system is decouplable by an admissible dynamic state feedback compensator with stability on a minimal realization of $T(s)$, the minimal McMillan degree achievable for the decoupled system is μ the stable decoupling degree of $T(s)$. In this case the i th diagonal entry of the decoupled system has McMillan degree q_i .

The proof follows the same lines as that of Theorem 4 replacing the interactor by the generalized interactor, r_i by q_i and bicausal matrices by bicausal and bistable ones.

The $h_i^{-1}(s)$ of Definition 4 now plays the same role as the s^{-r_i} of Theorem 4. $h_i^{-1}(s)$ possesses an infinite zero of order r_i and $q_i - r_i$ unstable zeros. \square

This theorem says that minimal delay decoupling is possible with the same number of input-output delays as without stability considerations. Furthermore the decoupled system possesses the minimal possible number $(\mu - \gamma)$ of unstable zeros.

5. Concluding remarks. (i) The decoupling invariants r_i and q_i can be characterized by other approaches. In particular the r_i 's are defined geometrically, via the notion of essentiality and via the infinite structure of some matrices deduced from the system transfer matrix in [18].

These decoupling invariants r_i (resp. q_i) are not related with the triangular form of the interactor (resp. the generalized interactor). Any $R(s)$ $p \times p$ full rank rational matrix such that $T(s) = [R(s), 0]B(s)$ with $B(s)$ bicausal (resp. bicausal and bistable) would do as well. For example, as proven in [19], r_i is the infinite pole order of the i th column of $R^{-1}(s)$.

(ii) When the column rank at infinity k is not equal to p this reflects the nonidentity between the infinite zero orders of the rows of $T(s)$ and the infinite zero orders of $T(s)$. In this case the control law $u = Fx + Gv$, G singular, allows us to increase the infinite structure by decreasing R^* the largest controllability subspace of $\ker C$. When R^* is "sufficiently large" static state feedback decoupling is possible [20]. In this paper we are mainly concerned with dynamic feedback decoupling; therefore only $\dim(B \cap R^*)$ has to be large enough ($\geq p - k$), since R^* can be arbitrarily extended by dynamic state feedback.

As an illustration consider the following matrix:

$$T(s) = \begin{bmatrix} s^{-2} & 0 & s^{-3} \\ -s^{-1} & s^{-3} & -s^{-2} \end{bmatrix}.$$

Since the two rows are dependent at infinity this system is not decouplable by static state feedback with regular G , but this system is decouplable by dynamic state feedback (see Theorem 3).

This system is no more decouplable by static state feedback with singular G . In fact $\dim R^* = 1$ in a minimal realization of $T(s)$ which is not sufficient to meet the condition of [20].

In [21] the dynamic state feedback decoupling problem restricting $F(s)$ to being stable is studied. This problem turns out to be solvable if $m \geq 2p - k_s$, where k_s is the infinite and unstable column rank of $I_g(s)$. k_s is generally smaller than k . A necessary condition for static state feedback with stability is then $\dim(B \cap R^*) \geq p - k_s$.

Possible extensions of the results presented here are: Block decoupling or output feedback decoupling as in [22].

Acknowledgment. The authors acknowledge a reviewer for providing them with a better formulation of Theorem 6.

REFERENCES

- [1] P. F. FALB AND W. A. WOLOVICH, *Decoupling in the design and synthesis of multivariable control systems*, IEEE Trans. Automat. Control, AC-12 (1967), pp. 651-669.
- [2] A. S. MORSE AND W. M. WONHAM, *Status of noninteracting control*, IEEE Trans. Automat. Control, AC-16 (1973), pp. 568-581.
- [3] W. M. WONHAM, *Linear Multivariable Control, a Geometric Approach*, Springer-Verlag, Berlin-New York, 1979.
- [4] M. L. J. HAUTUS AND M. HEYMANN, *Linear feedback decoupling, transfer function analysis*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 823-832.

- [5] K. B. DATTA AND M. L. J. HAUTUS, *Decoupling of multivariable control systems over unique factorization domains*, this Journal, 22 (1984), pp. 28–39.
- [6] J. C. WILLEMS, *Almost noninteracting control design using dynamic state feedback*, INRIA Conf. Proc. Springer-Verlag, Berlin–New York, 1980, pp. 555–561.
- [7] R. W. BROCKETT AND M. D. MESAROVIC, *The reproducibility of multivariable systems*, J. Math. Anal. Appl., 11 (1965), p. 548.
- [8] M. L. J. HAUTUS AND M. HEYMANN, *Linear feedback—An algebraic approach*, this Journal, 16 (1978), pp. 83–105.
- [9] J. HAMMER AND M. HEYMANN, *Causal factorization and linear feedback*, this Journal, 19 (1981), pp. 445–468.
- [10] A. S. MORSE, *Systems invariants under feedback and cascade control*, Proc. Internat. Symposium, Udine, Springer-Verlag, New York, 1975.
- [11] W. A. WOLOVICH AND P. L. FALB, *Invariants and canonical forms under dynamic compensation*, this Journal, 14 (1976), pp. 996–1008.
- [12] A. S. MORSE, *Parametrization for multivariable adaptive control*, IEEE Conf. Design and Control, pp. 970–972, 1981.
- [13] G. CONTE AND A. PERDON, *On the minimum delay problem*, Systems Control Lett., 5 (1984), pp. 213–215.
- [14] J. DESCUSSE AND J. M. DION, *On the structure at infinity of linear decouplable systems*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 971–974.
- [15] N. D. HUNG AND B. D. O. ANDERSON, *Triangularization technique for the design of multivariable control system*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 455–460.
- [16] F. M. CALLIER AND C. A. DESOER, *Multivariable feedback systems*, Springer-Verlag, New York, 1982.
- [17] M. VIDYASAGAR AND N. VISWANADHAM, *Algebraic design techniques for reliable stabilization*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 1085–1095.
- [18] C. COMMAULT, J. DESCUSSE, J. M. DION, J. F. LAFAY AND M. MALABRE, *New decoupling invariants, the essential orders*, Internat. J. Control, 44 (1986), pp. 689–700.
- [19] C. COMMAULT AND J. M. DION, *Some structural invariants within the transfer matrix approach*, IEEE Conf. Design and Control, Athens, 1986.
- [20] J. DESCUSSE, J. F. LAFAY AND M. MALABRE, *A survey on Morgan's problem*, IEEE Conf. Design and Control, Athens, 1986.
- [21] C. COMMAULT AND J. M. DION, *On dynamic state feedback decoupling*, Internal report L.A.G., 1986.
- [22] J. HAMMER AND P. KHARGONEKAR, *Decoupling of linear systems by dynamic output feedback*, Math. Systems Theory, 17 (1984), pp. 135–157.

EXACT MODELING OF A FINITE TIME SERIES*

CHRISTIAAN HEIJ†

Abstract. The problem of modeling a finite time series by means of a linear shift invariant system will be considered. To analyze this problem the concepts of complexity of a model and corroboration of a model by data are introduced. Various properties of data modeling procedures are defined. These properties are investigated for the partial realization procedure. An alternative procedure will be described which takes into account the concept of corroboration. It turns out that this procedure has many desirable properties.

Key words. data modeling, identification, modeling procedures, corroboration, finite time systems, partial realization theory

AMS(MOS) subject classifications. 93A, 93B

1. Introduction. The problem considered here is an aspect of the problem of modeling dynamical phenomena and can be described as follows.

Suppose one wants to model a dynamical phenomenon. Here the following will be assumed. First, the describing variables of the phenomenon have been identified. Second, observations of these variables over a finite time interval are available. Finally, a class of models has been specified. This class reflects a priori assumptions concerning the phenomenon, e.g., based upon theoretical considerations concerning the phenomenon or inspired by mathematical convenience.

The problem now consists in identifying (a) model(s) in the specified class to describe the phenomenon in a way which is optimal in some sense, taking into account the observations.

This problem is formalized as follows. Assume at every time instance the observation consists of an element in the set W . Let \mathbb{T} denote the time set of observation, which is assumed to be finite. So the data consist of an element of $W^{\mathbb{T}}$, i.e., a finite time series in W . Let \mathbb{M} denote the specified class of models. The problem is to construct data modeling **PROCEDURES** to assign models to data, i.e., procedures

$$P: W^{\mathbb{T}} \rightarrow 2^{\mathbb{M}}.$$

Moreover, one wants to construct procedures which are satisfactory or even optimal in some sense.

Clearly this problem is an important topic of scientific research in, e.g., statistics, econometrics and systems theory and includes problems as structure identification, estimation and data analysis.

In this paper a very special case of the data modeling problem will be investigated. This will lead to the formulation of some interesting concepts.

The central issue is the **EXACT** modeling of a **FINITE** time series of **ONE** variable by means of a (deterministic) **AUTOREGRESSIVE** model.

It is evident that the case of a multivariable time series and **APPROXIMATE** modeling are of more practical importance. Nonetheless, the case considered here is of interest in its own right and, moreover, it will lead to questions and concepts which are crucial also in the multivariable case and approximate modeling.

An outline of the paper is as follows. In § 2 we define the model class of (finite time) linear shift invariant systems and discuss some properties of this class and representation of systems by means of autoregressive equations.

* Received by the editors July 23, 1986; accepted for publication (in revised form) February 23, 1987.

† Econometrics Institute, University of Groningen, 9700 AV Groningen, The Netherlands.

In § 3 we formulate the modeling problem analyzed in this paper, i.e., modeling a finite time series by means of an autoregressive model. The concept of data modeling procedure is defined.

The concepts of simplicity and corroboration are introduced in § 4. These concepts play a crucial role and reflect the idea that on the one hand one wants the model to be as simple as possible, but that on the other hand one only wants to accept simplicity if there is some evidence for it in the data.

Some possible properties of procedures are defined in § 5.

The partial realization procedure for modeling a finite time series in one variable is described in § 6. Its properties are investigated. It turns out that this procedure, since it pays no attention to corroboration, lacks many desirable properties.

In § 7 we introduce a modification of the partial realization procedure which takes into account corroboration.

In § 8 we formulate a refinement of this procedure. The properties of this refined procedure are thoroughly investigated. It is shown that it satisfies all the properties defined in § 5.

Section 9 contains some concluding remarks. A summary of notation concludes the main text.

Proofs of the results are given mainly in the Appendix.

Some contributions closely related to the approach taken here are given in Kalman [2], Rissanen [4] and Willems [5].

2. Linear shift invariant finite time systems. The problem considered in this paper is that of modeling a finite time series by means of (a) linear shift invariant system(s). This class of models is described in this section, along with results on representation of models in this class. It turns out that the class of linear shift invariant finite time systems is a strict subclass of the models described by a finite number of autoregressive equations.

The following definition of a dynamical system is given in Willems [5].

DEFINITION. A **DYNAMICAL SYSTEM** is a triple (\mathbb{T}, W, B) with $\mathbb{T} \subset \mathbb{R}$ the time set, W the signal set, $B \subset W^{\mathbb{T}}$ the behaviour.

Remark. In the sequel we often will use the words “system” and “model” to denote the behaviour.

DEFINITION. A **FINITE (discrete) TIME system** is a dynamical system (\mathbb{T}, W, B) where \mathbb{T} is a finite subset of \mathbb{Z} .

Let $\mathbb{N} := \{1, 2, 3, \dots\}$ and for $t_1, t_2 \in \mathbb{N}$, $t_1 \leq t_2$, $[t_1, t_2] := \{t \in \mathbb{N}; t_1 \leq t \leq t_2\}$. In this paper we throughout will assume there exists $T \in \mathbb{N}$ such that $\mathbb{T} = [1, T]$ and we will take $W := \mathbb{R}^q$. So a behaviour $B \subset (\mathbb{R}^q)^{\mathbb{T}}$ is a set of sequences in \mathbb{R}^q of length T . We will identify $(\mathbb{R}^q)^{\mathbb{T}}$ and $(\mathbb{R}^q)^T$.

We restrict attention to behaviours which are linear and shift invariant. B is called linear if it is a linear subspace of $(\mathbb{R}^q)^T$.

To define shift invariance, we define the left shift operator σ as follows.

Let w be a sequence in \mathbb{R}^q of arbitrary finite length $L \geq 2$, so $w \in (\mathbb{R}^q)^L$. Then $\sigma w \in (\mathbb{R}^q)^{L-1}$ is defined by $(\sigma w)(l) := w(l+1)$, $l \in [1, L-1]$. For $B \subset (\mathbb{R}^q)^T$, $1 \leq t_1 \leq t_2 \leq T$, let $B|_{[t_1, t_2]}$ denote the restriction of B to the set $[t_1, t_2]$.

DEFINITION. Let $T \geq 2$. $B \subset (\mathbb{R}^q)^T$ is called **SHIFT INVARIANT** if $\sigma B \subset B|_{[1, T-1]}$. It is called **TRANSLATION INVARIANT** if $\sigma B = B|_{[1, T-1]}$.

The interpretation of shift invariance is as follows. Let B be shift invariant and $w \in B$. Then there exists $f \in W$ such that $\hat{w} \in B$, where $\hat{w}|_{[1, T-1]} := \sigma w$ and $\hat{w}(T) := f$. Described intuitively this means that there exists a future observation which follows

on w and which is compatible with the behaviour B . If B is shift invariant but not translation invariant this means that there exist $w \in B$ for which there is no past observation compatible with B . If B is translation invariant then for every $w \in B$ there exist a past and a future of w compatible with B .

The following proposition is immediate.

PROPOSITION 2.1. (i) *If B is shift invariant, $\Delta \in \mathbb{N}$, $1 \leq t_1 \leq t_1 + \Delta \leq t_2 + \Delta \leq T$; then $B|_{[t_1+\Delta, t_2+\Delta]} \subset B|_{[t_1, t_2]}$.*
(ii) *If B is translation invariant, $\Delta \in \mathbb{Z}$, $1 \leq t_1 \leq t_2 \leq T$, $1 \leq t_1 + \Delta \leq t_2 + \Delta \leq T$; then $B|_{[t_1+\Delta, t_2+\Delta]} = B|_{[t_1, t_2]}$.*

DEFINITION. The class $\mathbb{B}_T(\tilde{\mathbb{B}}_T)$ of linear shift (translation) invariant finite (T) time systems is given by

$$\mathbb{B}_T := \{B \subset (\mathbb{R}^q)^T; B \text{ linear, } \sigma B \subset B|_{[1, T-1]}\},$$

$$\tilde{\mathbb{B}}_T := \{B \subset (\mathbb{R}^q)^T; B \text{ linear, } \sigma B = B|_{[1, T-1]}\}.$$

It is clear that $\tilde{\mathbb{B}}_T \subset \mathbb{B}_T$ and that both classes are closed under addition, but not under union. Moreover, $\tilde{\mathbb{B}}_T$ and \mathbb{B}_T are not closed under intersection, which is seen as follows.

Example 2.1. Let $q = 1$, $T = 4$, $B_1 := \{w \in \mathbb{R}^4; w(4) = w(1)\}$, $B_2 := \{w \in \mathbb{R}^4; w(3) = -w(1) \text{ and } w(4) = -w(2)\}$. Then $B_i \in \tilde{\mathbb{B}}_T$, $i = 1, 2$, but $B := B_1 \cap B_2 = \{w \in \mathbb{R}^4; \text{there exists } \alpha \in \mathbb{R} \text{ such that } w = \alpha(1, -1, -1, 1)\}$ and $B \notin \mathbb{B}_T$.

Next we will show that systems in \mathbb{B}_T can be described by a finite number of **AUTOREGRESSIVE EQUATIONS**.

Let $g \in \mathbb{N}$ and $R \in \mathbb{R}^{g \times q}[\sigma]$ a polynomial matrix in σ with rows $r_i = (r_{i1}, \dots, r_{iq}) \in \mathbb{R}^{1 \times q}[\sigma]$, $i \in [1, g]$. Let $d(r_{ij})$ denote the degree of r_{ij} , with $d(0) := -\infty$, and $d(r_i) := \max \{d(r_{ij}), j \in [1, q]\}$, $d(R) := \max \{d(r_i), i \in [1, g]\}$. By $B_T(R)$ we will denote the behaviour where the autoregressive equations $r_i(\sigma)w = 0$ are satisfied. By definition we assume that if $d(R) \geq T$ then this equation imposes no restriction.

DEFINITION. Let $R \in \mathbb{R}^{g \times q}[\sigma]$ and let $I_T(R) := \{i \in [1, g]; d(r_i) \leq T - 1\}$. Then $B_T(R) := \{w \in (\mathbb{R}^q)^T; [r_i(\sigma)w](t) = 0, i \in I_T(R), t \in [1, T - d(r_i)]\}$.

Remark. We often will write $B(R)$ instead of $B_T(R)$ if this does not lead to confusion. Note that if $d(R) \geq T$ then $r(\sigma)w$ is undefined, as σ is only defined for sequences of length at least 2. As an example, take $q = g = 1$, $T = 3$, $R = \sigma^3$, $w \in \mathbb{R}^3$. Then $\sigma^3 w$ is undefined as $\sigma^2 w = w(3)$ has length 1.

PROPOSITION 2.2.

$$\{B \in \mathbb{B}_T\} \Rightarrow \{\exists g \exists R \in \mathbb{R}^{g \times q}[\sigma], d(R) \leq T - 1, B = B(R)\}.$$

Proof. See the Appendix.

Remark. The converse statement is not true, i.e., there exist $R \in \mathbb{R}^{g \times q}[\sigma]$ such that $B(R) \notin \mathbb{B}_T$. For example, take $q = 1$, $g = 2$, $T = 4$, $r_1(\sigma) = \sigma^3 - 1$, $r_2(\sigma) = \sigma^2 + 1$, $R = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}$. Then $B(R) = \{w \in \mathbb{R}^4; \text{there exists } \alpha \in \mathbb{R} \text{ such that } w = \alpha(1, -1, -1, 1)\} \notin \mathbb{B}_T$. So \mathbb{B}_T consists of a strict subclass of the systems described by a finite number of autoregressive equations. Proposition 2.3 will give a characterization of this subclass.

Let $0 \neq R \in \mathbb{R}^{g \times q}[\sigma]$ have rows $r_i \in \mathbb{R}^{1 \times q}[\sigma]$, $i \in [1, g]$. Let $r_i(\sigma) = \sum_{k=d_i}^{d_i} q_k^{(i)} \sigma^k$ where $d(r_i) = d_i \geq d_i'$ and $q_{d_i}^{(i)} \neq 0 \neq q_{d_i'}^{(i)}$. Let L^+ , $L^- \in \mathbb{R}^{g \times q}$ denote the matrices with rows $q_{d_i}^{(i)}$, $q_{d_i'}^{(i)}$, respectively. R is defined to be row proper if $\text{rank } L^+ = g$, bilaterally row proper if $\text{rank } L^+ = \text{rank } L^- = g$, zero order bilaterally row proper if in addition $d_i' = 0$ for all $i \in [1, g]$. We define $0 \in \mathbb{R}^{1 \times q}[\sigma]$ to be zero order bilaterally row proper.

PROPOSITION 2.3. (i) $\{B \in \mathbb{B}_T\} \Leftrightarrow \{\text{there exists } g, \text{ there exists } R \in \mathbb{R}^{g \times q}[\sigma], d(R) \leq T - 1, R \text{ row proper, } B = B(R)\}$.

(ii) $\{B \in \tilde{\mathbb{B}}_T\} \Leftrightarrow \{\text{there exists } g \text{ there exists } R \in \mathbb{R}^{g \times q}[\sigma], d(R) \leq T - 1, R \text{ zero order bilaterally row proper, } B = B(R)\}$.

Proof. See the Appendix.

Remark. Henceforth, if we write $B_T(R)$ or $B(R)$ (with T given) we will assume throughout that R is row proper (and zero order bilaterally row proper only if we explicitly assume $B(R) \in \tilde{\mathbb{B}}_T$). Note that if $B = (\mathbb{R}^q)^T$ then we can take $R = 0$.

Remark. If R is bilaterally row proper, then $B(R)$ need not belong to $\tilde{\mathbb{B}}_T$. A simple counterexample is given by taking $q = g = 2$, $T = 2$, $R = \begin{pmatrix} \sigma & 0 \\ 0 & \sigma \end{pmatrix}$.

DEFINITION. If $B = B(R) \in \mathbb{B}_T$ and R has rows r_i , $i \in [1, g]$, then $\{r_i, i \in I_T(R)\}$ are called the **LAWS** governing the behaviour B .

We finally will derive representations of the sum of two systems in \mathbb{B}_T and of the largest system in \mathbb{B}_T contained in the intersection of two systems in \mathbb{B}_T . Note that the intersection itself need not belong to \mathbb{B}_T , but for every set in $(\mathbb{R}^q)^T$ there exists a largest system in \mathbb{B}_T contained in this set, because \mathbb{B}_T is closed under addition.

Proofs of the following results are contained in the Appendix.

LEMMA 2.1. Let $R_i \in \mathbb{R}^{g_i \times q}[\sigma]$ be row proper with $d(R_i) \leq T - 1$, $i = 1, 2$. Then

$$\{B(R_1) \subset B(R_2)\} \Leftrightarrow \{\exists F \in \mathbb{R}^{g_2 \times g_1}[\sigma] R_2 = FR_1\}.$$

Remark. If we drop the condition $d(R_i) \leq T - 1$, $i = 1, 2$, then (\Rightarrow) no longer holds true. Take $q = g_1 = g_2 = 1$, $T = 3$, $R_1 = \sigma - 1$, $R_2 = \sigma^3$.

DEFINITION. Let $R_i \in \mathbb{R}^{g_i \times q}[\sigma]$, $i = 1, 2$.

(i) R is called a least common left multiple of R_1 and R_2 , notation $R \in \text{LCLM}(R_1, R_2)$ if

- (1) $\exists F_i \quad R = F_i R_i, \quad i = 1, 2;$
- (2) $\{\exists \tilde{F}_i \quad \tilde{R} = \tilde{F}_i R_i, i = 1, 2\} \Rightarrow \{\exists F \quad \tilde{R} = FR\}.$

(ii) R is called a greatest common right divisor of R_1 and R_2 , notation $R \in \text{GCRD}(R_1, R_2)$, if

- (1) $\exists F_i \quad R_i = F_i R, \quad i = 1, 2;$
- (2) $\{\exists \tilde{F}_i \quad R_i = \tilde{F}_i \tilde{R}, i = 1, 2\} \Rightarrow \{\exists F \quad R = F \tilde{R}\}.$

LEMMA 2.2. (i) For any row proper R_1, R_2 , $\text{LCLM}(R_1, R_2)$ and $\text{GCRD}(R_1, R_2)$ are nonempty and contain row proper elements.

(ii) If $R', R'' \in \text{LCLM}(R_1, R_2)$ are both row proper, then there exists a unimodular U such that $R'' = UR'$. The same holds true for GCRD .

PROPOSITION 2.4. (i) $B(R_1) + B(R_2) = B(R)$ for $R \in \text{LCLM}(R_1, R_2)$.

(ii) $B(R) = \sum \{B(\tilde{R}); B(\tilde{R}) \subset B(R_1) \cap B(R_2)\}$ for $R \in \text{GCRD}(R_1, R_2)$, $d(R_i) \leq T - 1$, $i = 1, 2$.

So a (row proper element of the) least common left multiple of R_1 and R_2 gives the laws of the sum $B(R_1) + B(R_2)$, and a (row proper element of the) greatest common right divisor of R_1 and R_2 gives the laws of the largest system in \mathbb{B}_T contained in the intersection $B(R_1) \cap B(R_2)$.

Remark. The implication $\{B(R) = B(R_1) + B(R_2)\} \Rightarrow \{R \in \text{LCLM}(R_1, R_2)\}$ does not hold true. For example, take $q = g_1 = g_2 = 1$, $T = 3$, $R_1 = \sigma^2$, $R_2 = \sigma + 1$, then $B(R_1) + B(R_2) = \mathbb{R}^3 = B(0)$ but $0 \notin \text{LCLM}(R_1, R_2)$. Moreover (ii) does not hold true generally if $d(R_i) \leq T - 1$, $i = 1, 2$ is not satisfied. For example, take $q = g_1 = g_2 = 1$, $T = 3$, $R_1 = \sigma^2(\sigma - 1)$, $R_2 = (\sigma - 1)^3$, then $\text{GCRD}(R_1, R_2) = \sigma - 1$ but $B_3(\sigma - 1) \neq \mathbb{R}^3 = B(R_1) \cap B(R_2)$.

3. Modeling of data, procedures. As stated in the Introduction, we will consider the following problem of modeling a dynamical phenomenon.

Assume we want to model a dynamical phenomenon concerning which (i) data is available in the form of a (finite) time series $w \in (\mathbb{R}^q)^T$, and (ii) an a priori specified class of models is available. The aim is to construct data modeling procedures which assign to data a model or a set of models in the prespecified class. The assigned models reflect the structure of the phenomenon which one postulates on the basis of the data.

In this paper the only a priori restrictions which we impose on the models are those of linearity and shift (or translation) invariance. So the problem becomes one of assigning linear shift (translation) invariant systems to an observed finite time series.

We now first give a definition of a data modeling procedure for deterministic modeling of a finite time series $w \in W^T$.

DEFINITION. A **DATA MODELING PROCEDURE** for t is a map $P_t: W^t \rightarrow 2^{W^t}$. A data modeling procedure on $[1, T]$ is a collection $P = \{P_t, t \in [1, T]\}$ where P_t is a data modeling procedure for t .

So a data modeling procedure P_t for every observation $w \in W^t$ assigns a collection $P_t w$ of models, i.e., a collection of subsets of W^t . The interpretation is that on observation of w , $P_t w$ consists of those models which are accepted by the procedure as a description of the phenomenon.

Here we take $W = \mathbb{R}^q$ and we impose the a priori restriction that the assigned models belong to \mathbb{B}_t . In this case a data modeling procedure for t is a map

$$P_t: (\mathbb{R}^q)^t \rightarrow 2^{\mathbb{B}_t}$$

or, if one requires translation invariance, a map $P_t: (\mathbb{R}^q)^t \rightarrow 2^{\tilde{\mathbb{B}}_t}$. As an example, we will describe the procedure P^u which assigns the set of undominated unfalsified models in \mathbb{B}_T .

DEFINITION. The model $B \in \mathbb{B}_t$ is called unfalsified by $w \in (\mathbb{R}^q)^t$ if $w \in B$. B is called undominated unfalsified if in addition $\{w \in B' \in \mathbb{B}_t, B' \subset B\} \Rightarrow \{B' = B\}$.

DEFINITION. The procedure $P^u = \{P_t^u, t \in [1, T]\}$ assigns to $w \in (\mathbb{R}^q)^t$ the set of undominated unfalsified models in \mathbb{B}_t .

Example 3.1. Let $q = 1$, $T = 4$. Using a partial Hankel matrix, described in § 6, it can be derived that $P_4^u(0, 1, 0, 1) = B(\sigma^2 - 1)$. For $w = (1, 1, 0, 1)$ the class of unfalsified models is $\{B(\sigma^2 + \sigma - 1), B(\sigma^3 + \alpha\sigma^2 + \beta\sigma - 1 - \beta), \alpha, \beta \in \mathbb{R}\}$ and $P_4^u(1, 1, 0, 1) = \{B(\sigma^2 + \sigma - 1), B(\sigma^3 + \alpha\sigma^2 + \beta\sigma - 1 - \beta), \alpha - \beta \neq 2\}$. So in this case there exists more than one undominated unfalsified model. This is essentially due to the fact that \mathbb{B}_T is not closed under intersection.

4. Simplicity and corroboration. A primary aim of modeling data is to describe the structure of the phenomenon which generated the data. For assigning structure to a phenomenon on the basis of data at least two considerations play a crucial role, which we will denote by the principles of simplicity (or parsimony) and of corroboration.

On the one hand, one wants to infer from the data as much structure for the phenomenon as possible in order to get many laws, i.e., small models. (Note that if both R and $\binom{R}{r} =: R'$ are row proper, then $B(R') \subset B(R)$.) We will call this a **FALSIFIABILITY** principle or **SIMPLICITY** principle. A model is considered to be simple if it is small in some set theoretic sense. Then a more simple model is more easily falsifiable, as it imposes more restrictions.

Note that simplicity here is connected with systems as sets of trajectories and not with the simplicity of the laws of systems. For example, $(\mathbb{R}^q)^T$ is the most complex model in \mathbb{B}_T , although from a viewpoint of laws it is very simple as there are no laws at all.

As a measure of complexity of a model in \mathbb{B}_T we here will simply take its dimension.

DEFINITION. The **COMPLEXITY** of $B \in \mathbb{B}_T$ is defined as $c(B) := \dim B$.

Although one wants to get models of low complexity, i.e., with much structure, on the other hand one wants to accept structure only if there is some evidence for it from the data. Note that data independent structure is reflected in the a priori specified model class and that the data modeling procedure reflects the additional structure which is accepted after observing the data. One has to have some reason for claiming structure on the basis of data. We call this requirement of “reason” or “evidence” a **CORROBORATION** principle.

We will now first define what we mean by corroboration and afterwards give an interpretation.

We use the following algebraic concept of genericity. $V \subset \mathbb{R}^n$ is called an algebraic variety if $V = p^{-1}(0)$ for a polynomial p in n variables. It is called a proper algebraic variety if $V \neq \mathbb{R}^n$, i.e., $p \neq 0$.

DEFINITION. $\pi \subset \mathbb{R}^n$ is called generic if there is a proper algebraic variety V such that $\pi \supset (\mathbb{R}^n \setminus V)$.

A property for points in \mathbb{R}^n is called generic if the set of points which have the property is generic.

DEFINITION. Let $B \in \mathbb{B}_T$, $P_T: (\mathbb{R}^q)^T \rightarrow 2^{\mathbb{B}_T}$.

$\{B \text{ WEAKLY CORROBORABLE by } P_T\} \Leftrightarrow \{B \notin P_T w \text{ is not generic in } w \in B\}$,

$\{B \text{ STRONGLY CORROBORABLE by } P_T\} \Leftrightarrow \{P_T w = B \text{ generically in } w \in B\}$.

If a model is not weakly corroborable by a procedure it seems natural to require that this model never will be assigned by the procedure. For suppose $B \in P_T w$ while B is not weakly corroborable by P_T . Then generically for $w \in B$, P_T assigns only wrong models. Now if $w \in B \in P_T w$, then in fact one accepts that one was very lucky in observing precisely this w from the phenomenon B , as generically one would not have “identified” B by P_T using an observation from the phenomenon.

On the other hand, if B is strongly corroborable by P_T then generically for $w \in B$ P_T assigns exactly the right model.

Remark. The relaxation that conditions are satisfied generically instead of universally is essential in the definition of corroborability as well as in many definitions which follow. This has to do with the fact that $B(R) \in \mathbb{B}_T$ contains submodels with more structure, i.e., if R' is such that there exists F such that $FR' = R$ then $B(R') \subset B(R)$ and hence exceptional observations from $B(R)$ may exhibit the stronger structure of $B(R')$.

We illustrate the concept of corroboration by means of the procedure P'' which assigns undominated unfalsified models to data.

Example 4.1. Let $q = 1$, $T = 4$ and consider P'' . Now \mathbb{R}^4 is not even weakly corroborable by P''_4 , as $\{P''_4 w = \mathbb{R}^4\} \Leftrightarrow \{\text{there exist } 0 \neq \alpha \in \mathbb{R}; w = \alpha(0, 0, 0, 1)\}$. It can be shown that every other model in \mathbb{B}_4 is weakly corroborable by P''_4 . $B(\sigma^2 - 1)$ is weakly, but not strongly corroborable by P''_4 . $B(\sigma - 1)$ is strongly corroborable by P''_4 .

In fact, with $c: \mathbb{B}_4 \rightarrow [0, 4]$, $c(B) := \dim B$ it can be shown that the class of models strongly corroborable by P''_4 is given by $c^{-1}(\{0, 1\})$ and $B(\sigma^2)$.

5. Properties of procedures. In this section we will define some possible properties of procedures and comment on their interpretation. We will investigate the properties for the procedure P'' defined in § 3. In §§ 6–8 we investigate the procedures defined there.

DEFINITION. (i) P_T is called **EXACT** if it only assigns unfalsified models, i.e., $\{B \in P_T w\} \Rightarrow \{w \in B\}$

(ii) $P = \{P_t, t \in [1, T]\}$ is called exact if P_t is exact for all $t \in [1, T]$.

Example. P'' is exact.

So a procedure is exact if it assigns only models which are not falsified by the data. This requirement is very restrictive. The case of approximate modeling in which one does not require $w \in B$ but that w is “near to” B often has more appeal in applications. Even if the data is exact, i.e., there are no errors of observation, generally one is more interested in approximate simple structure than in exact more complex structure. If the data is corrupted by noise then exact modeling is not a sound requirement. In applications generally the phenomenon under observation will not correspond to any model in the a priori class of models and also the observations will contain errors.

Nevertheless in this paper we will restrict attention to exact modeling and only will briefly comment on approximate modeling in § 9. This is a topic of current research in which the insights gained from studying the exact modeling problem play an important role.

DEFINITION. (i) P is called **MONOTONE** on $B \in \mathbb{B}_T$ if generically in $w \in B$ the following holds true for all $t \in [2, T]$:

$$\{B_{t-1} \in P_{t-1}(w|_{[1, t-1]}), B_t \in P_t(w|_{[1, t]})\} \Rightarrow \{B_t|_{[1, t-1]} \subset B_{t-1}\}.$$

(ii) P is called **monotone** if it is monotone on every $B \in \mathbb{B}_T$.

(iii) P is called **bilaterally monotone** on $B \in \tilde{\mathbb{B}}_T$ if it is monotone and if in addition generically in $w \in B$ the following holds true for all $t \in [2, T]$:

$$\{B_{T-t+1} \in P_{T-t+1}(w|_{[t, T]}), B_{T-t+2} \in P_{T-t+2}(w|_{[t-1, T]})\} \Rightarrow \{\sigma B_{T-t+2} \subset B_{T-t+1}\}.$$

So if P is monotone then, for every phenomenon $B \in \mathbb{B}_T$, on getting a new observation $w(t)$ the structure assigned on the basis of $(w(1), \dots, w(t-1))$ generically is verified or even made more tight, as $P_t(w|_{[1, t]})$ consists of models which restrictions to $[1, t-1]$ are submodels of those in $P_{t-1}(w|_{[1, t-1]})$. This is a desirable property in case the observations successively become available over time. Then the procedure assigns more structure if the observation period gets longer. This does not only hold true generically on $(\mathbb{R}^q)^T$ but even generically on every system $B \subset (\mathbb{R}^q)^T$ in \mathbb{B}^T .

Example. For $q = 1$ P^u is only monotone if $T = 2$ or on $c^{-1}(\{0, 1\})$, i.e., on $\{0\}$ and on $B(\sigma - \alpha)$, $\alpha \in \mathbb{R}$. Consider, e.g., $T = 3$, $B = B(\sigma^2) = \{(a, b, 0), a, b \in \mathbb{R}\}$ and let $w = (a, b, 0)$ with $a \neq 0 \neq b$. Then $P_2^u(w|_{[1, 2]}) = B(\sigma - (b/a))$ while $B(\sigma^2) \in P_3^u w$ and $\mathbb{R}^2 = B(\sigma^2)|_{[1, 2]} \not\subset B(\sigma - (b/a))$.

Next we will define two properties of procedures which are connected with the a priori restriction that one only assigns models in the class \mathbb{B}_T (or $\tilde{\mathbb{B}}_T$), i.e., shift (or translation) invariant and linear models. So one describes the phenomenon which generates the data as a shift (translation) invariant linear system. This leads to some desirable properties of P .

DEFINITION. (i) P is called **SHIFT INVARIANT** on $B \in \mathbb{B}_T$ if generically in $w \in B$ the following holds true for all $t \in [2, T]$:

$$\{B_{t-1}(R) \in P_{t-1}(w|_{[2, t]}), B'_t \in P_t(w|_{[1, t]})\} \Rightarrow \{B'_t \subset B_t(\sigma.R)\}.$$

(ii) P is called **shift invariant** if it is shift invariant on every $B \in \mathbb{B}_T$.

PROPOSITION 5.1. *If P is bilaterally monotone on $B \in \tilde{\mathbb{B}}_T$ then it is shift invariant on B .*

Proof. See the Appendix.

Because of this proposition we do not define a concept of translation invariance for procedures, as it would be a concept very close to bilateral monotonicity.

An interpretation of shift invariance of procedures is the following. The a priori assumption that assigned models have to be shift-invariant, i.e., $B|_{[1, t-1]} \supset B|_{[2, t]}$, does not imply any a priori restriction on $w(1)$, given $w|_{[2, t]}$. Assume one only observes

$w|_{[2,t]}$ and wants to model the phenomenon on $[1, t]$. It seems reasonable first to model $w|_{[2,t]}$ and to impose no restriction on $w(1)$, i.e., if $B_{t-1}(R) \in P_{t-1}(w|_{[2,t]})$ then take $B_t(\sigma R)$ as a model on $[1, t]$, which amounts to letting $w(1)$ be arbitrary and assigning the laws R on $[2, t]$. Now assume one also observes $w(1)$, so one has more information concerning the phenomenon. It seems reasonable to demand that one can make more accurate models and to require that $w(1)$ does not contradict any laws R which were assigned for the phenomenon on $[2, t]$ on the basis of $w|_{[2,t]}$. So if $B' \in P_t(w|_{[1,t]})$ one would like $B' \subset B_t(\sigma R)$, which implies $B'|_{[2,t]} \subset B_{t-1}(R)$.

Example. For P^u take $q=1$, $T=3$. P^u is not shift invariant on $B(\sigma^2-1)$ as for $w=(a, b, a)$, $a \neq b \neq 0$, $P_2^u(w|_{[2,3]}) = B_2(\sigma - (a/b))$ while $B_3(\sigma^2-1) \in P_3^u w$ and $B_3(\sigma^2-1) \not\subset B_3(\sigma^2 - (a/b)\sigma)$.

DEFINITION. (i) P_T is called **LINEAR** if

$$(a) \quad P_T(\alpha w) = P_T w \quad \forall w \in (\mathbb{R}^q)^T \quad \forall \alpha \neq 0.$$

$$(b) \quad \forall (B_1, B_2) \in \mathbb{B}_T^2 \text{ generically in } (w_1, w_2) \in B_1 \times B_2$$

$$\{B \in P_T(w_1 + w_2), B' \in P_T w_1, B'' \in P_T w_2\} \Rightarrow \{B' + B'' \subset B\}.$$

(ii) P is called linear if P_t is linear for all $t \in [1, T]$.

The phenomenon a priori is assumed to be linear, so (a) reflects that the observations w and αw , $\alpha \neq 0$, are in a sense equivalent. Concerning (b), consider the situation of constructing a model for the signal $w_1 + w_2$ either from observation of $w_1 + w_2$ or from observation of both w_1 and w_2 . In the latter case one has more information concerning the structure of the phenomenon which one wants to model and hence it is reasonable to demand that one can make more accurate models.

Example. P^u satisfies (a), but not (b). Take $q=1$, $T=3$, $B_i = B(\sigma - i)$, $i=1, 2$. Then generically $P_3^u w_i = B_i$, $w_i \in B_i$, $i=1, 2$. According to Proposition 2.4(i) $B_1 + B_2 = B(\sigma^2 - 3\sigma + 2)$ and if $w = (a, b, 3b - 2a) \in B_1 + B_2$, $a \neq 0$, then $B(\sigma^2 + (2a - 3b)/a) \in P_3^u w$ and $B(\sigma^2 - 3\sigma + 2) \not\subset B(\sigma^2 + (2a - 3b)/a)$, so (b) is not satisfied generically on $B_1 \times B_2$.

Finally we will define two properties which reflect the wish to accept laws only if there is some evidence for them.

We call P truthful if (generically) the laws which are accepted by P in fact are also satisfied for the phenomenon which generates the data. So the assigned models contain the phenomenon as a subset. It is not required that all laws of the phenomenon are detected.

DEFINITION. (i) P_T is called **TRUTHFUL** on $B_0 \in \mathbb{B}_T$ if generically in $w \in B_0$ $\{B \in P_T w\} \Rightarrow \{B_0 \subset B\}$.

(ii) P_T is called truthful if it is truthful on every $B_0 \in \mathbb{B}_T$.

(iii) P is called truthful if P_t is truthful for all $t \in [1, T]$.

Example. P^u is not truthful. Take $q=1$, $T=2$, $B_0 = \mathbb{R}^2$. Then generically in $w \in B_0$ $\dim(P_2^u w) = 1$.

To define the concept of prudence we will use the following notation. $\mathbb{B}_{P_T}^w := \{B \in \mathbb{B}_T; B \text{ weakly corroborable by } P_T\}$, $\mathbb{B}_{P_T}^s := \{B \in \mathbb{B}_T; B \text{ strongly corroborable by } P_T\}$ and $\text{im } P_T := \{B \in \mathbb{B}_T; \text{there exists } w \text{ such that } B \in P_T w\}$.

DEFINITION. (i) P_T is called **WEAKLY PRUDENTIAL** if $\text{im } P_T \subset \mathbb{B}_{P_T}^w$.

(ii) P_T is called **STRONGLY PRUDENTIAL** if $\text{im } P_T \subset \mathbb{B}_{P_T}^s$.

(iii) P is called weakly (strongly) prudential if P_t has this property for all $t \in [1, T]$.

The interpretation is that P_T only accepts laws which it can "verify" itself. For suppose $B \in P_T w$ while $B \notin \mathbb{B}_{P_T}^w$. Then on the basis of the data w , P_T assigns a model B of which one knows that if the phenomenon in fact coincides with B , then generically one will not assign B on the basis of an observation of this phenomenon. So if B is

assigned then it seems one was extraordinarily lucky to observe w to identify B . Assigning B is reckless. So weak prudence seems a minimum requirement. On the other hand, strongly prudential procedures are very cautious. A model B only is assigned if this model also generically will be assigned on the basis of observations of the phenomenon B . So if assigning the right model is not generic for observations from a model B , then a strongly prudential procedure does not dare to make this assignment B .

Example. Consider P'' for $q=1$, $T=4$. From Example 4.1 it follows that P''_4 is not strongly prudential and not even weakly prudential. However, every observation $w \notin \{\alpha(0, 0, 0, 1), 0 \neq \alpha \in \mathbb{R}\}$ is modeled in the class $\mathbb{B}_{P''_4}$.

Summarizing, the procedure P'' which assigns undominated unfalsified models is exact, it is not monotone, not shift invariant, not linear, not truthful and not (weakly) prudential. Moreover P'' is a complex procedure in the sense that generally it assigns many models to each observation.

In § 6 we investigate the partial realization procedure P^K for the case of a one-dimensional time series, i.e., $q=1$. This procedure is a refinement of P'' in the sense that for all $t \in [1, T]$, for all $w \in \mathbb{R}^t$, $P_t^K w \subset P_t'' w$. In fact P_t^K chooses among the undominated unfalsified models those of minimal complexity. By using a procedure P^0 which takes into account the concept of corroboration and which is described in § 7 we will define a procedure P^* in § 8 which assigns the unfalsified model of minimal complexity in the class of models for which there is some evidence from the data. This procedure is exact, monotone, shift invariant, linear, truthful and strongly prudent.

Remark. If $q=1$, $T=\infty$, then the procedures P_∞'' , P_∞^K , P_∞^0 , P_∞^* for modeling a time series $w \in (\mathbb{R})^\mathbb{N}$ can be shown to be equivalent. So P_∞'' is exact, truthful and even strongly prudent. One can show that for every system $B_\infty(R) := \{w \in (\mathbb{R})^\mathbb{N}, [R(\sigma)w](t) = 0 \text{ for all } t \in \mathbb{N}\}$ generically in $w \in B_\infty(R)$ the only undominated unfalsified model is $B_\infty(R)$ itself, so $P_\infty'' w = B_\infty(R)$ generically in $w \in B_\infty(R)$.

6. The partial realization procedure and its properties ($q=1$). From now on we throughout will restrict attention to the univariate case, i.e., $q=1$.

The “partial realization” procedure P^K is a refinement of P'' . It assigns to data the least complex model(s) unfalsified by w . This implies that all assigned models are undominated unfalsified. So in the class of exact procedures P^K is the one which maximizes the simplicity of the model, i.e., it minimizes the dimension. Stated otherwise, one determines the shortest lag AR relations exactly satisfied by the data.

DEFINITION. The partial realization procedure P^K assigns least complex unfalsified models, i.e., $P^K := \{P_t^K, t \in [1, T]\}$ where

$$P_t^K : \mathbb{R}^t \rightarrow 2^{\mathbb{B}_t} \text{ is defined by } \{B \in P_t^K w\} : \Leftrightarrow \{w \in B \in \mathbb{B}_t, \text{ and } c(B) = \min \{c(B'); w \in B' \in \mathbb{B}_t\}\}.$$

To determine $P_t^K w$ one can use an algorithm for the partial realization problem as described, e.g., in Kalman [2]. There the problem consists in realizing a partial impulse response sequence in a minimal way (i.e. with minimal dimension of the state space). Here the problem is to model an arbitrary observation over time of one variable.

The “incomplete Hankel array” for $w \in \mathbb{R}^t$ is defined by

$$H_t(w) := \begin{bmatrix} w(1) & w(2) & \cdots & w(t-1) & w(t) \\ w(2) & w(3) & & w(t) & \\ \vdots & \vdots & & & \\ w(t-1) & w(t) & & & \\ w(t) & & & & \end{bmatrix} \begin{matrix} r(1) \\ r(2) \\ \vdots \\ r(t-1) \\ r(t) \end{matrix}$$

By $r(i)$ we have denoted the i th row of $H_t(w)$.

A matrix M with elements m_{ij} , $i, j \in \mathbb{N}$ is called an extension of $H_t(w)$ if $m_{ij} = w(i+j-1)$, $i \in [1, t]$, $j \in [1, t-i+1]$. It is called a Hankel extension if M is a Hankel matrix. The rank of M is defined as the dimension of the space spanned by the columns (or rows) of M .

DEFINITION. The (Kalman) rank of $H_t(w)$, denoted by $\text{rank } H_t(w)$, is the minimal rank of extensions of $H_t(w)$.

Clearly $\text{rank } H_t(w)$ is well defined and $\text{rank } H_t(w) \leq t$ for all $w \in \mathbb{R}^t$.

We will call $r(i)$ linearly (in)dependent on $r(1), \dots, r(i-1)$ if in the $i \times (t-i+1)$ matrix consisting of the first i rows and first $t-i+1$ columns of $H_t(w)$ the last row is linearly (in)dependent on the foregoing ones.

Proofs of the following lemmas are contained in the Appendix.

LEMMA 6.1.

$$\{\text{rank } H_t(w) = n\} \Leftrightarrow \{r(n) \text{ linearly independent from } r(1), \dots, r(n-1), r(n+1) \text{ linearly dependent on } r(1), \dots, r(n)\}.$$

LEMMA 6.2. In the class of minimal rank extensions of $H_t(w)$ there is a Hankel extension.

The following proposition describes how $P_t^K w$ can be determined by using $H_t(w)$. This amounts to determining the first row in $H_t(w)$ which is linearly dependent on the foregoing ones.

PROPOSITION 6.1. (i) $\{B \in P_t^K w\} \Leftrightarrow \{w \in B \text{ and } c(B) = \text{rank } H_t(w)\}$.

(ii) If $\text{rank } H_t(w) = d$, then $P_t^K w = \{B(R)\}$; there exists a_i , $i \in [1, d]$, such that $r(d+1) = \sum_{i=1}^d a_i \cdot r(i)$ and $R = \sigma^d - \sum_{i=1}^d a_i \sigma^{i-1}$.

Proof. Let $\text{rank } H_t(w) = d$. Assume $w \in B \in \mathbb{B}_t$. Shift invariance of B implies there exists $w^e \in (\mathbb{R})^{\mathbb{N}}$, $w^e|_{[1,t]} = w$, $w^e|_{[\tau, \tau+t-1]} \in B$ for all $\tau \in \mathbb{N}$. Define an extension M of $H_t(w)$ by $m_{ij} := w^e(i+j-1)$, $i, j = 1, \dots, t$, $m_{ij} := 0$ elsewhere. Then $\text{rank } M \geq d$ and hence $w^e|_{[\tau, \tau+t-1]}$, $\tau \in [1, t]$ span a space of dimension at least d in B , hence $c(B) \geq d$. Further as $\text{rank } H_t(w) = d$ there exists $a = (a_1, \dots, a_d) \in \mathbb{R}^d$ such that $r(d+1) = \sum_{i=1}^d a_i r(i)$. Define $R_a := \sigma^d - \sum_{i=1}^d a_i \sigma^i$. Then clearly $w \in B(R_a)$ and $c(B(R_a)) = d$. Using the definition of P_t^K this proves (i) and \supset in (ii). To prove \subset in (ii), let $B(R) \in P_t^K w$, so $c(B(R)) = d$, which implies R has degree d , say $R = \sigma^d - \sum_{i=1}^d a_i \sigma^{i-1}$. Then in $H_t(w)$ $r(d+1) = \sum_{i=1}^d a_i r(i)$. \square

Remark. Note that for $q = 1$ if $B(R) \in \mathbb{B}_T$ the assumption of row properness of R implies that $R \in \mathbb{R}[\sigma]$, i.e., it consists of one row. Then for $R \neq 0$ $c(B_t(R)) := \dim B_t(R) = \min \{d, t\}$ where d is the degree of R .

The procedure P^K only takes into account simplicity, not corroboration. As a result there often is no evidence for assigned models.

Example 6.1. Consider the phenomenon $B = \mathbb{R}^T$, so there are no laws. For $x \in \mathbb{R}$ define $\text{ENT}(x) := \min \{n \in \mathbb{Z}; n \geq x\}$. Now generically in $w \in B$, $\text{rank } H_T(w) = \text{ENT}(T/2)$, hence generically in $w \in B$ $\{B \in P_T^K w\} \Rightarrow \{\dim B = \text{ENT}(T/2)\}$. So even though there are no laws at all, P^K still generically imposes them. In fact $P_T^K w = \mathbb{R}^T$ if and only if $w(t) = 0$, $t \in [1, T-1]$, and $w(T) \neq 0$.

We will illustrate P^K by giving some examples which play an important role in the sequel to construct procedures with better properties than P^K .

Examples. Take $T = 6$.

Example 6.2. $w = (1, 1, 2, 3, 5, 8)$, $\text{rank } H_6(w) = 2$, $r(3) = r(1) + r(2)$, $P_6^K w = B(\sigma^2 - \sigma - 1)$.

Example 6.3. $w = (1, 1, 1, 0, 1, 1)$, $P_6^K w = B(\sigma^3 + \sigma^2 - 2)$.

Example 6.4. $w = (1, 2, 0, 1, 1, 2)$, $P_6^K w = B(\sigma^3 - \frac{7}{5}\sigma^2 - \frac{3}{5}\sigma + \frac{1}{5})$.

Example 6.5. $w = (1, 2, 0, 4, 4)$, $P_6^K w = B(\sigma^3 - \sigma - 2)$.

Example 6.6. $w = (1, 0, 2, 0, 1, 1)$, $P_6^K w = B(\sigma^3 + \frac{1}{3}\sigma^2 - \frac{1}{2}\sigma - \frac{2}{3})$.

P^K need not assign unique models.

Example 6.7. Take $T = 3$ and the phenomenon $B = \mathbb{R}^3$. Then for generic $w = (a, b, c) \in B$ $P_3^K w = \{B(\sigma^2 - \beta\sigma - \alpha); \alpha a + \beta b = c\}$ (the condition is $ac - b^2 \neq 0$). In general, if t is odd then generically in $w \in \mathbb{R}^t$ $\text{rank } H_t(w) = (t+1)/2$ and $P_t^K w$ consists of uncountably many models, while if t is even then generically in $w \in \mathbb{R}^t$ $\text{rank } H_t(w) = t/2$ and $P_t^K w$ consists of one model.

Before we state the properties of P^K we will define the concept of a selection rule. If a procedure lacks desirable properties this could be due to the fact that assigned models are nonunique, as monotonicity, shift invariance, linearity and truthfulness are requirements which are stronger the larger the classes of assigned models for an observation. Such a procedure possibly could be improved by a selection rule. By this we mean a rule which for every observation chooses a unique model from the class of models which is assigned by the procedure.

DEFINITION. (i) A selection rule for $P_T: \mathbb{R}^T \rightarrow 2^{\mathbb{B}_T}$ is a map $S_T: \mathbb{R}^T \rightarrow \mathbb{B}_T$ such that for all $w \in \mathbb{R}^T$, $S_T w \in P_T w$.

(ii) A selection rule for $P = \{P_t, t \in [1, T]\}$ is a collection $S = \{S_t, t \in [1, T]\}$ where S_t is a selection rule for P_t .

To analyze the properties of P^K the following proposition will be helpful.

PROPOSITION 6.2. Let $0 \neq R \in \mathbb{R}[\sigma]$ have degree $d(R) = d$. Then generically in $w \in B_t(R)$ $\text{rank } H_t(w) = \min\{d, \text{ENT}(t/2)\}$.

Proof. See the Appendix.

THEOREM 1. (i) P^K is exact.

If $T \geq 3$, then (ii)–(v) hold true.

(ii) P^K is not monotone, not shift invariant, not linear.

(iii) P_T^K is not truthful.

(iv) P_T^K is not even weakly prudent, as $\text{im } P_T^K = \mathbb{B}_T$, $\mathbb{B}_{P_T^K} = \{B \in \mathbb{B}_T; c(B) \leq \text{ENT}(T/2)\}$, $\mathbb{B}_{P_T^K}^s = \mathbb{B}_{P_T^K}^w$ if T is even, $\mathbb{B}_{P_T^K}^s = \{B \in \mathbb{B}_T; c(B) \leq \text{ent}(T/2)\}$ if T is odd.

(v) There exists no selection rule for P^K which has at least one of the properties in (ii)–(iv).

Proof. See the Appendix.

Remark. It is a matter of easy verification to show that, for $T = 2$, P^K is monotone, shift invariant, not linear, not truthful and not weakly prudent.

Remark. The proof of Theorem 1 contained in the Appendix to show that P^K lacks desirable properties concentrates on the case $B = \mathbb{R}^T$. It can be shown that P^K only is monotone and shift invariant on models B for which $c(B) \in \{0, 1\}$ and that P_T^K only is truthful for models $B \in \mathbb{B}_{P_T^K}^s$. So \mathbb{R}^T is not the only model for which P^K has undesirable performance.

The analogue \tilde{P}^K of P^K for least complex modeling in $\tilde{\mathbb{B}}_t$ can be defined as follows.

DEFINITION. The procedure \tilde{P}^K assigns least complex unfalsified models which are translation invariant, i.e., $\tilde{P}_t^K: \mathbb{R}^t \rightarrow 2^{\tilde{\mathbb{B}}_t}$ is defined by

$$\{B \in \tilde{P}_t^K w\} : \Leftrightarrow \{w \in \tilde{\mathbb{B}}_t \text{ and } c(B) = \min\{c(B'); w \in B' \in \tilde{\mathbb{B}}_t\}\}.$$

Without going into details, $\tilde{P}_t^K w$ can be determined by looking in $H_t(w)$ for the first row which is linearly dependent on the foregoing ones and which explicitly involves the first row, i.e., by looking for the smallest d such that there exist $(a_1, \dots, a_d) \in \mathbb{R}^d$ with $a_1 \neq 0$ and $r(d+1) = \sum_{i=1}^d a_i r(i)$. This way of determining $\tilde{P}_t^K w$ is based upon Proposition 2.3(ii).

In the same way as P^K also \tilde{P}^K only takes into account simplicity, not corroboration. As a result \tilde{P}^K lacks desirable properties. It can be shown that \tilde{P}^K is not (bilaterally) monotone, not linear, not truthful and not even weakly prudent.

At the end of § 8 we briefly will discuss a procedure \tilde{P}^* for modeling by means of translation invariant models. Section 7 and the main part of § 8 concentrate on modeling by means of shift invariant models. In § 7 we will describe a procedure P^0 which takes corroboration into account and which has better properties than P^K . In § 8 we will refine P^0 to a procedure P^* which has all the properties defined in § 5 on every model $B \in \mathbb{B}_t$, $t \in [1, T]$, while P_t^* coincides with P_t^K on those models for which P^K is a satisfactory procedure, i.e., on $\mathbb{B}_t^K := \{B \in \mathbb{B}_t; c(B) \leq \text{ENT}(t/2) - 1\}$.

7. An alternative procedure $P^0(q=1)$. The main reason why the partial realization procedure P^K lacks desirable properties is the fact that it pays attention only to simplicity, not to corroboration. In § 8 we will describe a procedure P^* which has all the desirable properties defined in § 5 and for which the class of corroborable models satisfies

$$\mathbb{B}_{P^*}^s = \mathbb{B}_{P^*}^w \supset \{B \in \mathbb{B}_T; c(B) \leq \text{ENT}(T/2) - 1\} = \mathbb{B}_{P^K}^w \setminus \{B \in \mathbb{B}_T; c(B) = \text{ENT}(T/2)\}.$$

This procedure P^* will consist of a slight refinement of a procedure P^0 which will be described in this section. P^0 essentially consists of a refinement of P^K by taking into account corroboration.

The idea to refine P^K is to accept laws which are satisfied by the data only provided there is some evidence for them. If for an observation a nonfalsified law is of a type which also generically will be unfalsified for an observation on a phenomenon which obeys no law at all (i.e. $B = \mathbb{R}^T$), then one has reason not to accept this law. We illustrate this idea by means of an example.

Example 7.1 (cf. Example 6.4). $T=6$, $w=(1, 2, 0, 1, 1, 2)$.

$$H_T(w) = \begin{bmatrix} 1 & 2 & 0 & 1 & 1 & 2 \\ 2 & 0 & 1 & 1 & 2 & \\ 0 & 1 & 1 & 2 & & \\ 1 & 1 & 2 & & & \\ 1 & 2 & & & & \\ 2 & & & & & \end{bmatrix} \begin{matrix} r(1) \\ r(2) \\ r(3) \\ r(4) \\ r(5) \\ r(6) \end{matrix}.$$

Here $r(4) = \frac{7}{5}r(3) + \frac{3}{5}r(2) - \frac{1}{5}r(1)$, $\text{rank } H_T(w) = 3$, $P_T^K w = B(\sigma^3 - \frac{7}{5}\sigma^2 - \frac{3}{5}\sigma + \frac{1}{5})$. Note that it is not at all remarkable that $\text{rank } H_T(w) = 3$, as this generically holds true on \mathbb{R}^T for $T=6$. So one has reason not to accept this third order law, as it is of a type which generically holds true on \mathbb{R}^6 . It is not remarkable that w satisfies a law of this type.

Nonetheless, in $H_T(w)$ there is a remarkable dependence, as $r(5)$ is linearly dependent on $r(1)$ and this generically does not hold true on \mathbb{R}^T for $T=6$. One has reason to accept this remarkable law, i.e., to accept the law $w(t+4) = w(t)$, $t \in \{1, 2\}$, and to assign to w the model $B(\sigma^4 - 1)$.

In general, let $w \in \mathbb{R}^T$ and let $r(1), \dots, r(T)$ denote the rows of $H_T(w)$. Suppose $r(d+1)$ is linearly dependent on $r(i_1+1), \dots, r(i_{c-1}+1)$, where $0 \leq i_1 < i_2 < \dots < i_{c-1} < d$, say $r(d+1) = \sum_{k=1}^{c-1} a_k r(i_k+1)$, $a_k \neq 0$, $k \in [1, c-1]$. Let $R := \sigma^d - \sum_{k=1}^{c-1} a_k \sigma^{i_k}$, then $w \in B(R)$. The crucial question now is when is it remarkable to find a law such as R . We define this to be remarkable if and only if the number of elements in $r(d+1)$ is strictly larger than the number of explaining rows, i.e., if and only if $T-d > c-1$,

i.e., $c + d \leq T$. Note that this is exactly the class of laws which is not generically satisfied on \mathbb{R}^T .

We will now define a procedure P^0 which assigns to data the most simple (d minimal) unfalsified models for which it is remarkable that they are unfalsified ($c + d \leq T$).

DEFINITION. (i) For $R = \sum_{k=0}^n a_k \sigma^k \in \mathbb{R}[\sigma]$, $d(R) := \max \{k; a_k \neq 0\}$, $d(0) := -\infty$, $c(R) := \# \{k; a_k \neq 0\}$.

(ii) R is called **REMARKABLE** (for T) if $R \neq 0$ and $c(R) + d(R) \leq T$.

(iii) $\mathbb{B}_T^* := \{B_T(R); c(R) + d(R) \leq T\}$.

We will call $B_T(R)$ remarkable if R is remarkable for T . So \mathbb{B}_T^* consists of the class of remarkable models together with \mathbb{R}^T .

DEFINITION OF P^0 . (i) P_T^0 assigns least complex unfalsified remarkable models if these exist, else it assigns \mathbb{R}^T , i.e., $P_T^0: \mathbb{R}^T \rightarrow 2^{\mathbb{B}_T^*}$ is defined by $\{B \in P_T^0 w\} \Leftrightarrow \{w \in B \in \mathbb{B}_T^* \text{ and } c(B) = \min \{c(B'); w \in B' \in \mathbb{B}_T^*\}\}$.

(ii) $P^0 := \{P_t^0, t \in [1, T]\}$.

Comparing the definitions of P_T^0 and P_T^K , we see that P_T^K is refined to least complex unfalsified modeling in \mathbb{B}_T^* instead of \mathbb{B}_T , i.e., assigned laws have to be remarkable. Note that if w satisfies no remarkable law, then $P_T^0 w = \mathbb{R}^T$, so if no remarkable law is satisfied then no law is accepted. We illustrate P^0 by some examples.

Examples. Take $T = 6$.

Example 7.2 (cf. Example 6.2). $w = (1, 1, 2, 3, 5, 8)$, $P_6^0 w = B(\sigma^2 - \sigma - 1)$.

Example 7.3 (cf. Example 6.3). $w = (1, 1, 1, 0, 1, 1)$, $P_6^0 w = \{B(\sigma^4 - 1), B(\sigma^4 - \sigma)\}$.

Example 7.4 (cf. Example 6.6). $w = (1, 0, 2, 0, 1, 1)$, $P_6^0 w = \mathbb{R}^6$.

Example 7.5 (cf. Example 6.5). $w = (1, 2, 0, 4, 4, 4)$, $P_6^0 w = B(\sigma^3 - \sigma - 2)$.

Remark. Examples 7.3 and 7.5 will play a role in refining P^0 to the procedure P^* defined in § 8. This has to do with compatibility of remarkable laws, which concept is defined in § 8. For details we refer to Examples 8.1 and 8.2.

To determine $P_T^0 w$ one can investigate $H_T(w)$ and determine the first row, say $d + 1$, which is linearly dependent on $c - 1$ foregoing ones such that $c + d \leq T$. If no such row exists, then $P_T^0 w = \mathbb{R}^T$.

An important question is whether laws which are assigned by P_T^0 also (generically) are true laws, given that the data stems from a system in \mathbb{B}_T , i.e., the question of truthfulness of P^0 . An essential property of P^0 is the following.

THEOREM 2. Let $B_0 \in \mathbb{B}_T$. Then generically in $w \in B_0$, $P_T^0 w = \{B \in \mathbb{B}_T^*; B_0 \subset B, c(B) \text{ minimal}\}$.

Proof. See the Appendix.

So, generically on B_0 , P_T^0 assigns the least complex remarkable models containing B_0 and in particular generically the assigned laws also are true laws.

COROLLARY 7.1. If $B_0 \in \mathbb{B}_T^*$, then generically in $w \in B_0$ $P_T^0 w = B_0$.

Proof. See the Appendix.

Theorem 2 indicates some of the main properties of P^0 , i.e., corroboration is taken into account (modeling in \mathbb{B}_T^*), simplicity is taken into account ($c(B)$ minimal) and the procedure is truthful.

THEOREM 3. (i) P^0 is exact.

(ii) P^0 is truthful.

(iii) P^0 is strongly prudent, since $\text{im } P_t^0 = \mathbb{B}_{P_t^0}^w = \mathbb{B}_{P_t^0}^s = \mathbb{B}_t^*$, for all $t \in [1, T]$.

(iv) P^0 is not monotone and not shift invariant; the inclusion conditions for monotonicity, shift invariance and linearity are satisfied if the action of P_t^0 is restricted to models in the set \mathbb{B}_t^* .

Proof. See the Appendix.

Remark. If P^0 is linear it still is an open problem. We do not go into details of exact characterization of those models for which P^0 is monotone or shift invariant and do not answer the question of linearity. This is because (iv) implies that P^0 lacks some desirable properties, although only in quite special cases. In the following section we will slightly modify P^0 to get a procedure P^* with desirable properties everywhere.

8. The procedure P^* and its properties ($q = 1$). P^0 has some desirable properties but also lacks some of them. In this section we will refine P^0 to a procedure P^* which has all the properties which were introduced in § 5.

Before we describe P^* we will give three examples indicating in which direction P^0 could be refined. The main concepts are those of remarkability, defined in § 7, and compatibility, which we define as follows. By Λ we will denote an arbitrary index set.

DEFINITION. Let $R_\lambda \in \mathbb{R}[\sigma]$, $\lambda \in \Lambda$, and $w \in \mathbb{R}^T$. Then $\{R_\lambda, \lambda \in \Lambda\}$ and w are called **COMPATIBLE** if there exists a $B \in \mathbb{B}_T$ such that $w \in B \subset \bigcap \{B_T(R_\lambda), \lambda \in \Lambda\}$.

So a class of laws and an observation are called compatible if there exists a linear shift invariant system for which all the laws are valid and which is nonfalsified by the observation.

Example 8.1 (cf. Examples 6.3, 7.3). $T = 6$, $w = (1, 1, 1, 0, 1, 1)$, $P_6^0 w = \{B(\sigma^4 - 1), B(\sigma^4 - \sigma)\}$. Note that a priori we want to model in \mathbb{B}_6 , so the phenomenon is assumed to be linear and shift invariant. Maximizing simplicity under the restriction of corroboration leads to $P_6^0 w$ and a nonunique model being assigned to w . However, $\{\sigma^4 - 1, \sigma^4 - \sigma\}$ and w are not compatible. Note that $B(\sigma^4 - 1) \cap B(\sigma^4 - \sigma) \notin \mathbb{B}_6$ while due to Proposition 2.4(ii) the largest model in \mathbb{B}_6 contained in $B(\sigma^4 - 1) \cap B(\sigma^4 - \sigma)$ is given by $B(1) = \{0\}$, and $w \neq 0$. So given the phenomenon belongs to \mathbb{B}_6 and given the data w , the phenomenon cannot satisfy both laws $\sigma^4 - 1$ and $\sigma^4 - \sigma$, and at least one of the two assigned models has to be false. It seems reasonable to reject at least one of them, even to reject both. (In case one could wait for $w(7)$ to become available it might be sensible to store $B(\sigma^4 - 1)$ and $B(\sigma^4 - \sigma)$ and to decide on the basis of $w(7)$.)

Example 8.2 (cf. Examples 6.5, 7.5). $T = 6$, $w = (1, 2, 0, 4, 4, 4)$, $P_6^0 w = B(\sigma^3 - \sigma - 2)$. Note that in this case there is another unfalsified remarkable law, as $w \in B(\sigma^4 - \sigma^3)$, which is not accepted by P^0 . Now $\{\sigma^3 - \sigma - 2, \sigma^4 - \sigma^3\}$ and w are not compatible, as $B(\sigma^3 - \sigma - 2) \cap B(\sigma^4 - \sigma^3) = \{\alpha(1, 2, 0, 4, 4, 4), \alpha \in \mathbb{R}\} \notin \mathbb{B}_6$ while the largest model in \mathbb{B}_6 contained in this intersection is $B(1) = \{0\}$, as $\text{GCD}(\sigma^3 - \sigma - 2, \sigma^4 - \sigma^3) = 1$, and $w \neq 0$. So given the phenomenon belongs to \mathbb{B}_6 and given the data w , at least one of the remarkable laws $\sigma^3 - \sigma - 2$ and $\sigma^4 - \sigma^3$ was observed just by bad luck, because for this phenomenon not both laws can be valid. It seems reasonable to assign no law at all.

Example 8.3 (cf. the example given in the proof of Theorem 3(iv) in the Appendix). Assume the data stem from a phenomenon with law $R = \sigma^{10} + \sigma^9 + 2\sigma^8 + \sigma^7 + \sigma^6 + 2\sigma^5 + \sigma^4 + \sigma^3 + 2\sigma^2 + \sigma + 1$ and that $T = 20$, e.g., $w = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, -1, -1, 2, 0, -4, 4, 4, -12, 4, 20)$. Note $\text{rank } H_{20}(w) = 10$ and R is the only law of degree 10 which is nonfalsified by w , so $P_{20}^K w = B(R)$. However, $c(R) + d(R) = 21$ so R is not remarkable for $T = 20$. Let $R_1 := (\sigma - 1)R$, $R_2 := (\sigma - \frac{1}{2})R$; then $c(R_1) + d(R_1) = 19$, $c(R_2) + d(R_2) = 20$, $d(R_1) = d(R_2) = 11$. Here $P_{20}^0 w = \{B(R_1), B(R_2)\}$ and this also holds true generically on $B(R)$ according to Theorem 2.

Now note that $\{R_1, R_2\}$ and w are compatible, as $w \in B(R) = B(R_1) \cap B(R_2)$. This even holds true for every $w \in B(R)$. Although R itself is not remarkable, R_1 and R_2 are remarkable and moreover they are compatible with w . So it seems reasonable to assign $B(R)$ to w , not as it would be remarkable in itself, but because one has evidence for it from the remarkable and compatible laws R_1 and R_2 .

These examples indicate the ideas of how to refine P^0 .

DEFINITION. The class of remarkable laws for $w \in \mathbb{R}^T$ is defined by $L(w) := \{R \neq 0; d(R) \leq T-1 \text{ and } w \in B(R) \in \mathbb{B}_T^*\} = \{R; w \in B_T(R) \in \mathbb{B}_T^*, B_T(R) \neq \mathbb{R}^T\}$.

Remark. Because of the restriction $d(R) \leq T-1$, every law in $L(w)$ is remarkable. In the sequel we throughout will assume that in $B_T(R)$ $d(R) \leq T-1$, without always explicitly mentioning this.

If for $B \in \mathbb{B}_T$ both the laws R_1 and R_2 are valid, i.e., if $B \subset B(R_1) \cap B(R_2)$, then according to Proposition 2.4(ii) with $g = q = 1$, $B \subset B(G)$ where $G := \text{GCD}(R_1, R_2)$, the greatest common divisor of R_1 and R_2 in $\mathbb{R}[\sigma]$.

For $w \in \mathbb{R}^T$ let $R(w) := \text{GCD}\{R; R \in L(w)\}$, i.e., the least stringent law implied by all the remarkable laws for w . Define $\text{GCD}\{\emptyset\} := 0$. Now there are two possible situations.

(i) The class of all remarkable laws for w , i.e., $L(w)$, and w are compatible, i.e., there exists $B \in \mathbb{B}_T$ such that $w \in B \subset \bigcap \{B(R); R \in L(w)\}$. Note that $B(R(w))$ is the largest model in \mathbb{B}_T which satisfies this condition, i.e., $B(R(w))$ is the largest model unfalsified by w for which all remarkable laws are valid (note $B(R(w)) \subset B(R)$, for all $R \in L(w)$). Models of less complexity either are falsified by the data or are not supported by means of unfalsified remarkable laws. So one would like to accept all remarkable laws and to assign $B(R(w))$ to w .

(ii) $L(w)$ and w are not compatible, i.e., there is no $B \in \mathbb{B}_T$ which is unfalsified by w and for which all remarkable laws for w are valid. So given that the phenomenon belongs to \mathbb{B}_T , at least one of the remarkable laws has to be false. As one has no information concerning which law is false it seems reasonable to accept no law at all.

Note that (i) is equivalent to $w \in B(R(w))$ and (ii) to $w \notin B(R(w))$. We define the procedure P^* as follows.

DEFINITION OF P^* . (i) If the class of all remarkable laws is compatible with the data, P_T^* assigns the largest model for which all these laws are valid, else it assigns \mathbb{R}^T , i.e., $P_T^*: \mathbb{R}^T \rightarrow \mathbb{B}_T$ is defined by

$$P_T^* w = \begin{cases} B(R(w)) & \text{if } w \in B(R(w)), R(w) := \text{GCD}\{R; R \in L(w)\}, \\ \mathbb{R}^T & \text{if } w \notin B(R(w)). \end{cases}$$

(ii) $P^* := \{P_t^*, t \in [1, T]\}$.

Remark. In case the class of all remarkable laws is compatible with the data, P^* accepts all these laws and no other ones, as for other laws there is no evidence. Note that P^* always assigns a unique model to data.

To investigate P^* , the following proposition is crucial. It essentially says that generically all remarkable laws are true laws. That is an important motivation for our definition of remarkability.

PROPOSITION 8.1. For every $B_0 \in \mathbb{B}_T$, generically in $w \in B_0$, $B_0 \subset B(R(w))$.

Proof. See the Appendix.

Proofs of the following corollaries also are contained in the Appendix.

COROLLARY 8.1. Let $B_0 \in \mathbb{B}_T$. Then generically in $w \in B_0$ there holds $\{R \in L(w)\} \Rightarrow \{B_0 \subset B(R) \text{ and } P_T^* w \subset B(R)\}$.

COROLLARY 8.2. Let $B_0 \in \mathbb{B}_T$. Then generically in $w \in B_0$ $P_T^* w = B(R)$ where $R := \text{GCD}\{\hat{R} \neq 0; d(\hat{R}) \leq T-1 \text{ and } B_0 \subset B(\hat{R}) \in \mathbb{B}_T^*\}$.

So Corollary 8.1 states that generically all remarkable laws hold true for the phenomenon and that generally true remarkable laws also are accepted by P^* . This indicates a connection between remarkability and corroboration. Corollary 8.2 describes the generic way in which P^* models data.

The properties of P^* are stated in our main theorem.

THEOREM 4. (i) P^* is exact.

(ii) P^* is monotone, shift invariant and linear.

(iii) P^* is truthful.

(iv) P^* is strongly prudent as $\text{im } P_t^* = \mathbb{B}_{P_t^*}^w = \mathbb{B}_{P_t^*}^s = \{B(R); \exists \{R_\lambda, d(R_\lambda) \leq t-1, \lambda \in \Lambda\}, B(R_\lambda) \in \mathbb{B}_t^*, R = \text{GCD} \{R_\lambda, \lambda \in \Lambda\}\}$.

Proof. (i) The proof is obvious from the definition of P^* .

(ii) For $B \in \mathbb{B}_t$ let $R(B) := \text{GCD} \{R \neq 0; d(R) \leq t-1 \text{ and } B \subset B(R) \in \mathbb{B}_t^*\}$.

Monotonicity. Let $B_0 \in \mathbb{B}_T$, $w \in B_0$, $P_{t-1}^*(w|_{[1,t-1]}) = B_{t-1}$, $P_t^*(w|_{[1,t]}) = B_t$, then we have to prove that generically $B_t|_{[1,t-1]} \subset B_{t-1}$. If $B_{t-1} = \mathbb{R}^{t-1}$ then this is trivial, so assume $B_{t-1} \neq \mathbb{R}^{t-1}$. Then from Corollary 8.2 one can derive that generically on B_0 , $B_{t-1} = B_{t-1}(R_{t-1})$ and $B_t = B_t(R_t)$ where $R_{t-1} := R(B_0|_{[1,t-1]})$ and $R_t := R(B_0|_{[1,t]})$. Now if R is such that $B_0|_{[1,t-1]} \subset B_{t-1}(R) \in \mathbb{B}_{t-1}^*$, then $B_0|_{[2,t]} \subset B_0|_{[1,t-1]} \subset B_{t-1}(R)$ implies $B_0|_{[1,t]} \subset B_t(R) \in \mathbb{B}_t^*$ as $c(R) + d(R) \leq t-1 \leq t$; hence for these $RPB_t(R_t) \subset B_t(R)$ (use Proposition 2.4(ii)). This implies $B_t(R_t) \subset B_t(R_{t-1})$; hence also generically on B_0 , $B_t|_{[1,t-1]} = B_t(R_t)|_{[1,t-1]} \subset B_t(R_{t-1})|_{[1,t-1]} = B_{t-1}(R_{t-1}) = B_{t-1}$.

Shift invariance. Let $B_0 \in \mathbb{B}_T$, $w \in B_0$, $P_{t-1}^*(w|_{[2,t]}) = B_{t-1}(R)$, $B' = P_t^*(w|_{[1,t]})$, to prove that generically $B' \subset B_t(\sigma R)$. From Corollary 8.2 it follows that generically on B_0 , $R = R(B_0|_{[2,t]})$ and generically $B' = B_t(R')$ where $R' = R(B_0|_{[1,t]})$. Now if \tilde{R} is such that $B_0|_{[2,t]} \subset B_{t-1}(\tilde{R}) \in \mathbb{B}_{t-1}^*$, then $B_0|_{[1,t]} \subset B_t(\sigma \tilde{R}) \in \mathbb{B}_t^*$ as $c(\sigma \tilde{R}) + d(\sigma \tilde{R}) = c(\tilde{R}) + d(\tilde{R}) + 1 \leq t-1+1 = t$. So generically $B' = B_t(R') \subset B_t(\tilde{R})$ where $\tilde{R} := \text{GCD} \{\sigma \tilde{R} \neq 0; d(\tilde{R}) \leq t-1 \text{ and } B_0|_{[2,t]} \subset B_{t-1}(\tilde{R}) \in \mathbb{B}_{t-1}^*\} = \sigma R$ which proves shift invariance.

Linearity. $P_t^*(\alpha w) = P_t^*(w)$ for $\alpha \neq 0$ follows from $L(\alpha w) = L(w)$, $\alpha \neq 0$. Now let $(B_1, B_2) \in \mathbb{B}_t^2$, $B_i = B(R_i)$, $i = 1, 2$, $(w_1, w_2) \in B_1 \times B_2$, $P_t^* w_1 = B(R')$, $P_t^* w_2 = B(R'')$, $P_t^*(w_1 + w_2) = B(R)$; then we have to prove that generically $B(R') + B(R'') \subset B(R)$. According to Proposition 2.4(i) $B_1 + B_2 = B(K)$ with $K := \text{LCM}(R_1, R_2)$. According to Corollary 8.2 generically in $w_1 + w_2$, hence generically in $(w_1, w_2) \in B_1 \times B_2$, $R = R(B(K))$. Moreover, generically $R' = R(B_1)$ and $R'' = R(B_2)$. Because $B_i \subset B(K)$, $i = 1, 2$, if $B(K) \subset B(\tilde{R})$ then also $B_i \subset B(\tilde{R})$, $i = 1, 2$, so $B(R') \subset B(R)$, $B(R'') \subset B(R)$ and hence also $B(R') + B(R'') \subset B(R)$.

(iii) This is immediate from Corollary 8.2.

(iv) Let $V := \{B(R); \text{there exists } \{R_\lambda, d(R_\lambda) \leq t-1, \lambda \in \Lambda\}, B(R_\lambda) \in \mathbb{B}_t^*, R = \text{GCD} \{R_\lambda, \lambda \in \Lambda\}\}$. For every procedure $\mathbb{B}_{P_t^*}^s \subset \mathbb{B}_{P_t^*}^w \subset \text{im } P_t$ so it suffices to show $V \subset \mathbb{B}_{P_t^*}^w$ and $\text{im } P_t^* \subset V$. If $B \in V$, then Corollary 8.2 implies that generally on B , $P_t^* w = B$; hence $B \in \mathbb{B}_{P_t^*}^w$. If $B \in \text{im } P_t^*$, then either $B = \mathbb{R}^t \in V$ or there exists $w \in \mathbb{R}^t$ such that $B = B(R(w))$, $R(w) = \text{GCD} \{R; R \in L(w)\}$. Because for $R \in L(w)$, $d(R) \leq t-1$ and $B(R) \in \mathbb{B}_t^*$ it follows that $B \in V$. \square

To illustrate P^* we briefly return to some examples.

Example 8.1. $T = 6$, $w = (1, 1, 1, 0, 1, 1)$, $L(w) = \{\sigma^4 - 1, \sigma^4 - \sigma\}$, $R(w) = 1$, $w \notin B(1) = \{0\}$; hence $P_6^* w = \mathbb{R}^6$.

Example 8.2. $T = 6$, $w = (1, 2, 0, 4, 4, 4)$, $L(w) = \{\sigma^3 - \sigma - 2, \sigma^4 - \sigma^3\}$, $R(w) = 1$, $w \notin B(1) = \{0\}$; hence $P_6^* w = \mathbb{R}^6$.

Example 8.3. Take w, R, R_1, R_2 as stated before in this example. Then $L(w) = \{R_1, \sigma R_1, R_2\}$, $R(w) = R$, $w \in B(R)$; hence $P_{20}^* w = B(R)$.

Example 8.4 (cf. Examples 6.2, 7.2). $T = 6$, $w = (1, 1, 2, 3, 5, 8)$, $R := \sigma^2 - \sigma - 1$, $L(w) = \{R, \sigma R, (\sigma - 1)R, (\sigma + 1)R\}$, $R(w) = R$, $w \in B(R)$; so $P_6^* w = B(R)$.

From Theorem 2 and Corollary 8.2 we immediately get the following.

PROPOSITION 8.2. Let $B_0 \in \mathbb{B}_T$. Then generically in $w \in B_0$, $\{B \in P_T^0 w\} \Rightarrow \{P_T^* w \subset B\}$.

The following proposition indicates the relationship between P^K , P^0 , P^* and the sense in which P^* is a procedure with desirable properties everywhere and which

coincides with the partial realization procedure on those systems for which P^K is a reasonable procedure. Moreover it describes the sense in which the differences between P_T^K , P_T^0 , P_T^* disappear if $T \rightarrow \infty$. Proof of the proposition is immediate.

PROPOSITION 8.3. Let $\mathbb{B}_t^K \subset \mathbb{B}_t^*$ be defined by $\mathbb{B}_t^K := \{B \in \mathbb{B}_t; c(B) \leq \text{ENT}(t/2) - 1\}$.

(i) $\mathbb{B}_t^K = \mathbb{B}_{P_t^K}^s \subset \mathbb{B}_{P_t^0}^s \subset \mathbb{B}_{P_t^*}^s$ for t odd, $\mathbb{B}_t^K = \mathbb{B}_{P_t^K}^s \setminus \{B \in \mathbb{B}_t; c(B) = t/2\} \subset \mathbb{B}_{P_t^0}^s \subset \mathbb{B}_{P_t^*}^s$ for t even.

(ii) For $B_0 \in \mathbb{B}_t^K$, generically in $w \in B_0$, $P_t^K w = P_t^0 w = P_t^* w$; for $B_0 \in \mathbb{B}_t^*$ generically in $w \in B_0$, $P_t^0 w = P_t^* w$.

(iii) For given $R \in \mathbb{R}[\sigma]$, $P_T^K w = P_T^0 w = P_T^* w$ generically on $B_T(R)$ for T sufficiently large, i.e., $T \geq 2d(R) + 2$.

Remark. The problem of modeling an infinite time series $w \in (\mathbb{R})^\mathbb{N}$ by means of a model in the set $\mathbb{B}_\infty := \{B \subset (\mathbb{R})^\mathbb{N}; \text{there exists } R \in \mathbb{R}[\sigma] \text{ such that } w \in B \Leftrightarrow [R(\sigma)w](t) = 0, \text{ for all } t \in \mathbb{N}\}$ could be studied in an analogous way. Modeling by undominated unfalsified models, least complex unfalsified models, least complex unfalsified remarkable models or by most complex models which accept all remarkable laws if these are valid, i.e., using P_∞^u , P_∞^K , P_∞^0 or P_∞^* , turn out to be equivalent.

To conclude this section we will describe a procedure \tilde{P}^* for modeling in $\tilde{\mathbb{B}}_T$, i.e., linear translation invariant models.

Let $\tilde{\mathbb{B}}_T^* := \{B_T(R) \in \tilde{\mathbb{B}}_T; c(R) + d(R) \leq T\}$, and for $w \in \mathbb{R}^T$ let $\tilde{L}(w) := \{R \neq 0; d(R) \leq T - 1 \text{ and } w \in B(R) \in \tilde{\mathbb{B}}_T^*\}$, $\tilde{R}(w) := \text{GCD}\{R; R \in \tilde{L}(w)\}$. As $q = 1$, by using Proposition 2.3(ii) it follows that $B(R) \in \tilde{\mathbb{B}}_T$ if and only if $R = 0$, $d(R) \geq T$ or $R = \sum_{k=0}^{T-1} a_k \sigma^k$ with $a_0 \neq 0$. From this one easily gets $B(\tilde{R}(w)) \in \tilde{\mathbb{B}}_T$. Now \tilde{P}^* is defined in analogy with P^* and with the same motivation.

DEFINITION OF \tilde{P}^* .

$$\tilde{P}_T^*: \mathbb{R}^T \rightarrow \tilde{\mathbb{B}}_T \text{ is defined by } \tilde{P}_T^* w := \begin{cases} B(\tilde{R}(w)) & \text{if } w \in B(\tilde{R}(w)), \\ \mathbb{R}^T & \text{if } w \notin B(\tilde{R}(w)), \end{cases}$$

$$\tilde{P}^* := \{\tilde{P}_t^*, t \in [1, T]\}.$$

PROPOSITION 8.4. Let $B_0 \in \mathbb{B}_T$. Then generically in $w \in B_0$, $\tilde{P}_T^* w = B(R)$ where $R := \text{GCD}\{\tilde{R} \neq 0; d(\tilde{R}) \leq T - 1 \text{ and } B_0 \subset B(\tilde{R}) \in \tilde{\mathbb{B}}_T^*\}$.

Proof. See the Appendix.

The properties of \tilde{P}^* are stated in the next theorem, which can be proved by using Proposition 8.4 in a way which is completely analogous to the proof of Theorem 4 by means of Corollary 8.2.

THEOREM 5. P^* is exact, bilaterally monotone, linear, truthful and strongly prudent with $\text{im } \tilde{P}_t^* = \mathbb{B}_{\tilde{P}_t^*}^u = \mathbb{B}_{\tilde{P}_t^*}^s = \{B(R); \text{there exists } \{R_\lambda, d(R_\lambda) \leq t - 1, B(R_\lambda) \in \tilde{\mathbb{B}}_t^*, \lambda \in \Lambda\}, \text{ such that } R = \text{GCD}\{R_\lambda, \lambda \in \Lambda\}\}$.

We illustrate \tilde{P}^* by some examples.

Example 8.5 (cf. Example 8.4). $T = 6$, $w = (1, 1, 2, 3, 5, 8)$, $R := \sigma^2 - \sigma - 1$, $\tilde{L}(w) = \{R, (\sigma - 1)R, (\sigma + 1)R\}$, $\tilde{R}(w) = R$, $w \in B(R)$ so $\tilde{P}_6^* w = B(R)$.

Example 8.6. $T = 6$, $w = (1, 2, 2, 3, 2, 2)$, $P_6^K w = B(\sigma^3 - \frac{4}{3}\sigma^2 - \frac{4}{3}\sigma + \frac{7}{3})$, $P_6^0 w = P_6^* w = B(\sigma^4 - \sigma)$, $\tilde{P}_6^* w = \mathbb{R}^6$.

Example 8.7 (cf. Example 8.2). $T = 6$, $w = (1, 2, 0, 4, 4, 4)$, $\tilde{L}(w) = \{\sigma^3 - \sigma - 2\} = \tilde{R}(w)$, $w \in B(\tilde{R}(w))$ so $\tilde{P}_6^* w = B(\sigma^3 - \sigma - 2)$.

Note that in the last example the extra a priori assumption that the phenomenon is translation invariant leads to a model, $B(\sigma^3 - \sigma - 2)$, which is of less complexity than in case one only assumes the phenomenon to be shift invariant, in which case we would get $P_6^* w = \mathbb{R}^6$. This, however, is a situation which generically does not occur, as for all $w \in \mathbb{R}^T$, $\tilde{L}(w) \subset L(w)$ so $B(\tilde{R}(w)) \supset B(R(w))$. As generically $P_T^* w = B(R(w))$

and one can show that also generically $\tilde{P}_T^* w = B(\tilde{R}(w))$ this implies that generically on every $B_0 \in \mathbb{B}_T$, $\tilde{P}_T^* w \supset P_T^* w$. Even the following holds true.

PROPOSITION 8.5. (i) If $B_0 \in \mathbb{B}_T$, then generically on B_0 , $\tilde{P}_T^* w = P_T^* w$.

(ii) If $B_0 \in \mathbb{B}_T$, $B_0 \notin \mathbb{B}_T$, then generically on B_0 , $\tilde{P}_T^* w = \mathbb{R}^T$.

Proof. See the Appendix.

So if a phenomenon is not translation invariant, then generically P^* accepts no laws.

9. Concluding remarks. In the foregoing we have introduced some concepts which are relevant for the problem of modeling data. For the special case of exact modeling of a finite time series in one variable by means of an autoregressive (deterministic) model we described a procedure P^* which has many desirable properties. An interesting question is whether there is a sense in which P^* is optimal. For example, one could restrict attention to procedures which are monotone, shift invariant, linear, truthful and prudent and investigate if there exists a procedure in this class with a maximal set \mathbb{B}_P^s of strongly corroborable models. In a sense such a procedure has maximal discriminatory power on \mathbb{B}_T .

An important issue is the construction of efficient and preferably recursive numerical algorithms to compute the assigned models for a given procedure. For P^0 this amounts to checking singularity of square submatrices of the incomplete Hankel array, while for P^* also greatest common divisors have to be computed. In case the observations $w(t)$ become available over time it is desirable to have algorithms which compute $P_{t+1}(w|_{[1,t+1]})$ on the basis of $P_t(w|_{[1,t]})$, $w(t+1)$, and as small amount of additional information as possible.

Two crucial assumptions in the foregoing were the requirement of exact modeling and that $q = 1$.

If $q > 1$ then one can again define a partial realization procedure. An important question is which laws are remarkable and connected with this is the question which variables are free and which are not. For $q = 1$, the variable is declared to be free if no remarkable law holds true for it. A problem is the compatibility of laws, which essentially is due to the fact that \mathbb{B}_T is not closed under intersection. An interesting issue is the definition of appropriate measures of complexity, e.g., by introducing the concepts of input and of state; cf. Willems [5].

The case of approximate modeling, of course, is of most practical interest and it raises the question of defining appropriate measures of fit, i.e., measuring how well a model fits the data. By increasing the complexity of a model one generally will be able to increase the fit. This leads to the interesting question of which decrease in fit is low enough to make a decrease in complexity sensible. This question is connected with the concepts of corroboration and remarkability. The approximate modeling question is a topic of ongoing research.

Notation.

$\mathbb{N} := \{1, 2, 3, \dots\}$,

$\mathbb{Z} := \{\dots, -2, -1, 0, 1, 2, \dots\}$,

$T \in \mathbb{N}$: length of time interval,

$\mathbb{T} := [1, T] := \{t \in \mathbb{N}; 1 \leq t \leq T\}$; $[t_1, t_2] := \{t \in \mathbb{N}; t_1 \leq t \leq t_2\}$,

σ : left shift operator; if $w \in (\mathbb{R}^q)^L$, $L \geq 2$, then $\sigma w \in (\mathbb{R}^q)^{L-1}$ with

$(\sigma w)(l) := w(l+1)$, $l \in [1, L-1]$,

$\mathbb{B}_T := \{B \subset (\mathbb{R}^q)^T; B \text{ linear}, \sigma B \subset B|_{[1, T-1]}\}$,

$\tilde{\mathbb{B}}_T := \{B \subset (\mathbb{R}^q)^T; B \text{ linear}, \sigma B = B|_{[1, T-1]}\}$,

$R = (r_{ij}) \in \mathbb{R}^{g \times q}[\sigma]$: matrix with elements in the ring $\mathbb{R}[\sigma]$ of polynomials in σ ,

$d(r)$: degree of $r \in \mathbb{R}[\sigma]$; $d(0) := -\infty$,

$$\begin{aligned}
d(R) &:= \max \{d(r_{ij}), i \in [1, g], j \in [1, q]\}, \\
I_T(R) &:= \{i \in [1, g]; d(r_{i1}, \dots, r_{iq}) \leq T-1\}, \\
B_T(R) (= B(R)) &:= \{w \in (\mathbb{R}^q)^T; [(r_{i1}, \dots, r_{iq})(\sigma)w](t) = 0, i \in I_T(R), \\
&\quad t \in [1, T-d(r_{i1}, \dots, r_{iq})]\}, \\
c(B) &:= \dim B, B \in \mathbb{B}_T, \\
\mathbb{B}_{P_T}^w &:= \{B \in \mathbb{B}_T; B \text{ weakly corroborable by } P_T\}, \\
\mathbb{B}_{P_T}^s &:= \{B \in \mathbb{B}_T; B \text{ strongly corroborable by } P_T\}, \\
\text{im } P_T &:= \{B \in \mathbb{B}_T; \text{there exists } w \text{ such that } B \in P_T w\}, \\
\text{ENT}(x) &:= \min \{n \in \mathbb{Z}; n \geq x\}, \\
\text{ent}(x) &:= \max \{n \in \mathbb{Z}; n \leq x\}, \\
\mathbb{B}_t^K &:= \{B \in \mathbb{B}_t; c(B) \leq \text{ENT}(t/2) - 1\}, \\
d(R) &:= \max \{k; a_k \neq 0\}, c(R) := \# \{k; a_k \neq 0\} \text{ for } R = \sum_{k=0}^n a_k \sigma^k \in \mathbb{R}[\sigma], \\
\mathbb{B}_T^* &:= \{B_T(R); c(R) + d(R) \leq T\}, \\
L(w) &:= \{R \neq 0; d(R) \leq T-1 \text{ and } w \in B(R) \in \mathbb{B}_T^*\}, \\
R(w) &:= \text{GCD} \{R; R \in L(w)\} \in \mathbb{R}[\sigma]; \text{GCD} \{\emptyset\} := 0, \\
\tilde{\mathbb{B}}_T^* &:= \{B_T(R) \in \mathbb{B}_T; c(R) + d(R) \leq T\}, \\
\tilde{L}(w) &:= \{R \neq 0; d(R) \leq T-1 \text{ and } w \in B(R) \in \tilde{\mathbb{B}}_T^*\}, \\
\tilde{R}(w) &:= \text{GCD} \{R; R \in \tilde{L}(w)\} \in \mathbb{R}[\sigma], \\
I(R) &:= \{k \in [0, T-1]; a_k = 0\} \text{ for } R = \sum_{k=0}^{T-1} a_k \sigma^k \in \mathbb{R}[\sigma], \\
\mathbb{B}^*(d) &:= \{B(R) \in \mathbb{B}_T^*; d(R) = d\}, \\
\mathbb{B}^*(I) &:= \{B(R) \in \mathbb{B}_T^*; I(R) \supset I\}, \\
\mathbb{B}^*(d, I) &:= \mathbb{B}^*(d) \cap \mathbb{B}^*(I), \\
W(d) &:= \{w \in \mathbb{R}^T; \text{there exists } B(R) \in \mathbb{B}^*(d), w \in B(R)\}, \\
W(d, I) &:= \{w \in \mathbb{R}^T; \text{there exists } B(R) \in \mathbb{B}^*(d, I), w \in B(R)\}, \\
L(w; d, I) &:= \{R \in L(w); I(R) \supset I, d(R) = d\}.
\end{aligned}$$

Appendix.

Proof of Proposition 2.2. Let $B \in \mathbb{B}_T$ and define $B^e := \{w \in (\mathbb{R}^q)^{\mathbb{N}}; w|_{[t, t+T-1]} \in B, \text{ for all } t \in \mathbb{N}\}$. Shift invariance of B implies $B = B^e|_{[1, T]}$. Using the terminology and results of Willems [5], B^e is a linear, shift invariant complete system and there exist $g \in \mathbb{N}$ and row proper $R \in \mathbb{R}^{g \times q}[\sigma]$ such that $B^e = \{w \in (\mathbb{R}^q)^{\mathbb{N}}; [R(\sigma)w](t) = 0, \text{ for all } t \in \mathbb{N}\} =: B_\infty(R)$. Let R' consist of the rows of R of degree at most $T-1$. Then R' is row proper, $d(R') \leq T-1$ and $B = B^e|_{[1, T]} = B_\infty(R')|_{[1, T]}$, because laws of degree $\geq T$ do not imply any restriction on $[1, T]$.

It suffices to prove $B_\infty(R')|_{[1, T]} = B_T(R')$. Let R' have rows

$$r_i(\sigma) = \sum_{k=0}^{d_i} q_k^{(i)} \sigma^k, \quad q_{d_i}^{(i)} \neq 0, \quad i \in [1, g'],$$

and let $L^+ \in \mathbb{R}^{g' \times q}$ have rows $q_{d_i}^{(i)}$.

Let $w \in B_\infty(R')|_{[1, T]}$, so there exists a $w^e \in (\mathbb{R}^q)^{\mathbb{N}}$ such that $w^e|_{[1, T]} = w$ and $[r_i(\sigma)w^e](t) = 0, t \in \mathbb{N}$. Now $d_i \leq T-1$, so $[r_i(\sigma)w^e](t) = 0, t \in [1, T-d_i], i \in [1, g']$, which implies $w \in B_T(R')$.

Conversely, let $w \in B_T(R')$. Then define $w^e(T+1)$ as a solution of $q_{d_i}^{(i)} w^e(T+1) + q_{d_i-1}^{(i)} w(T) + \dots + q_0^{(i)} w(T-d_i+1) = 0$, for all $i \in [1, g']$. Existence of a solution is guaranteed because L^+ is surjective (R' is row proper). Next define $w^e(T+2)$ as a solution of $q_{d_i}^{(i)} w^e(T+2) + q_{d_i-1}^{(i)} w^e(T+1) + q_{d_i-2}^{(i)} w(T) + \dots + q_0^{(i)} w(T-d_i+2)$, for all $i \in [1, g']$. In this way we recursively can define a $w^e \in (\mathbb{R}^q)^{\mathbb{N}}$ with $w^e|_{[1, T]} = w$ and $[R'(\sigma)w^e](t) = 0$, for all $t \in \mathbb{N}$, so $w \in B_\infty(R')|_{[1, T]}$. \square

Proof of Proposition 2.3. First let $B \in \mathbb{B}_T$. It follows from the proof of Proposition 2.2 that there exists a row proper R with $d(R) \leq T-1$ such that $B = B_T(R)$.

Conversely, let $d(R) \leq T-1$ and R row proper. Let R have rows

$$r_i(\sigma) = \sum_{k=0}^{d_i} q_k^{(i)} \sigma^k, \quad q_{d_i}^{(i)} \neq 0, \quad i \in [1, g],$$

and let $L^+ \in \mathbb{R}^{g \times q}[\sigma]$ have rows $q_{d_i}^{(i)}$. To show that $B_T(R) \in \mathbb{B}_T$ it suffices to consider shift invariance, i.e., $\sigma B_T(R) \subset B_T(R)|_{[1, T-1]}$. Now this condition is equivalent to existence of a solution $a \in \mathbb{R}^q$ of the set of equations $q_{d_i}^{(i)} a + q_{d_{i-1}}^{(i)} w(T) + \cdots + q_0^{(i)} w(T-d_i+1) = 0$ for all $i \in [1, g]$, where $w \in B_T(R)$. Because R is row proper, L^+ is surjective and existence of a solution is guaranteed.

Next assume $d(R) \leq T-1$ with R zero order bilaterally row proper. One easily shows by a similar reasoning as before that $B_T(R) \in \hat{\mathbb{B}}_T$.

Finally let $B \in \hat{\mathbb{B}}_T$. Let $B^{ee} := \{w \in (\mathbb{R}^q)^{\mathbb{Z}}; w|_{[t, t+T-1]} \in B \text{ for all } t \in \mathbb{Z}\}$. As B is translation invariant, $B = B^{ee}|_{[1, T]}$. It follows from the results in Willems [5] that there exists a bilaterally row proper R such that $B^{ee} = B^{ee}(R) := \{w \in (\mathbb{R}^q)^{\mathbb{Z}}; [R(\sigma)w](t) = 0 \text{ for all } t \in \mathbb{Z}\}$. As $B^{ee}(R) = B^{ee}(DR)$ for any diagonal matrix $D = \text{diag}(d_1, \dots, d_g) \in \mathbb{R}^{g \times g}[\sigma]$, $d_i := \sigma^{n_i}$, $n_i \in \mathbb{Z}$, it follows that R can be chosen to be zero order bilaterally row proper.

It remains to show that $d(R) \leq T-1$ and that $B^{ee}(R)|_{[1, T]} = B_T(R)$. This follows by a reasoning completely analogous to the one given in the proof of Proposition 2.2. \square

Proof of Lemma 2.1. Let $B(R_i) \in \mathbb{B}_T$, $i = 1, 2$. Define $B_i^e := \{w \in (\mathbb{R}^q)^{\mathbb{N}}; w|_{[t, t+T-1]} \in B(R_i) \text{ for all } t \in \mathbb{N}\}$, $i = 1, 2$. Shift invariance of $B(R_i)$ implies $B_i^e|_{[1, T]} = B(R_i)$, $i = 1, 2$, and $B(R_1) \subset B(R_2)$ is equivalent to $B_1^e \subset B_2^e$.

Define $B_\infty(R_i) := \{w \in (\mathbb{R}^q)^{\mathbb{N}}; [R_i(\sigma)w](t) = 0 \text{ for all } t \in \mathbb{N}\}$, $i = 1, 2$. Because $d(R_i) \leq T-1$ there holds $B_i^e = B_\infty(R_i)$, $i = 1, 2$, which is seen as follows. Let R_i have rows $r_j^{(i)}$, $j \in [1, g_i]$. If $w \in B_\infty(R_i)$, then $[r_j^{(i)}(\sigma)w](t) = 0$ for all $j \in [1, g_i]$ for all $t \in \mathbb{N}$, especially for all $t \in [\tau, \tau + T - d(r_j^{(i)}) - 1]$, for all $\tau \in \mathbb{N}$, so $w|_{[\tau, \tau+T-1]} \in B(R_i)$ for all $\tau \in \mathbb{N}$, hence $w \in B_i^e$. Conversely, if $w \in B_i^e$ then with $w_t := w|_{[t, t+T-1]}$, $t \in \mathbb{N}$, $w_t \in B_i(R)$, so $[r_j^{(i)}(\sigma)w_t](\tau) = 0$ for all $j \in [1, g_i]$, for all $\tau \in [1, T - d(r_j^{(i)})] \neq \emptyset$, especially $[r_j^{(i)}(\sigma)w_t](1) = [r_j^{(i)}(\sigma)w](t) = 0$ for all $t \in \mathbb{N}$, hence $w \in B_\infty(R_i)$.

So to prove Lemma 2.1 it remains to prove that $\{B_\infty(R_1) \subset B_\infty(R_2)\} \Leftrightarrow \{\text{there exists an } F \text{ such that } R_2 = FR_1\}$. Now (\Leftarrow) is obvious. For (\Rightarrow) we refer to Nieuwenhuis and Willems [3]. In fact there the time axis is \mathbb{Z} while here it is \mathbb{N} , but the results easily generalize to this case. \square

Proof of Proposition 2.4 and Lemma 2.2. First we consider the results for LCLM and addition.

Let R_i be row proper, $i = 1, 2$. Define $B_\infty(R_i) := \{w \in (\mathbb{R}^q)^{\mathbb{N}}; [R_i(\sigma)w](t) = 0 \text{ for all } t \in \mathbb{N}\}$ and $B := B_\infty(R_1) + B_\infty(R_2)$. From Willems [5] it follows that there exists a row proper R_0 such that $B = B_\infty(R_0)$. We will show $R_0 \in \text{LCLM}(R_1, R_2)$.

Row properness implies that for all $T \in \mathbb{N}$ $B_T(R_i) = B_\infty(R_i)|_{[1, T]}$, $i = 1, 2$, and $B_T(R_0) = B_\infty(R_0)|_{[1, T]}$. As $[B_\infty(R_1) + B_\infty(R_2)]|_{[1, T]} = B_\infty(R_1)|_{[1, T]} + B_\infty(R_2)|_{[1, T]}$ this implies that for all $T \in \mathbb{N}$ $B_T(R_1) + B_T(R_2) = B_T(R_0)$.

Taking $T \equiv \max\{d(R_0), d(R_1), d(R_2)\} + 1$, $B_T(R_i) \subset B_T(R_0)$ by Lemma 2.1 implies there exists F_i such that $R_0 = F_i R_i$. Moreover, if for \tilde{R} there exists \tilde{F}_i such that $\tilde{R} = \tilde{F}_i R_i$, $i = 1, 2$, then let U unimodular be such that $U\tilde{R} = \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix}$ with \tilde{R} row proper. Then $B_\infty(\tilde{R}) = B_\infty(\tilde{R}) \supset B_\infty(R_1) + B_\infty(R_2) = B_\infty(R_0)$; hence for all $T \in \mathbb{N}$ $B_T(\tilde{R}) \supset B_T(R_0)$. Let $T \equiv \max\{d(R_0), d(\tilde{R})\} + 1$; then Lemma 2.1 implies there exists \tilde{F} such that $\tilde{R} = \tilde{F} R_0$, so $\tilde{R} = F R_0$ where $F := U^{-1} \begin{pmatrix} \tilde{F} \\ 0 \end{pmatrix}$. This proves $R_0 \in \text{LCLM}(R_1, R_2)$ and (i) of Lemma 2.2 for LCLM.

Next let $R \in \text{LCLM}(R_1, R_2)$ be row proper. Then there exists F_0 such that $R = F_0 R_0$ and there exists F such that $R_0 = FR$. So $B_\infty(R) = B_\infty(R_0) = B_\infty(R_1) + B_\infty(R_2)$; hence

$B_T(R) = B_T(R_1) + B_T(R_2)$ which proves (i) of Proposition 2.4. Moreover, $R = F_0FR$ and $R_0 = FF_0R_0$. Because R_0 and R have full row rank $F_0F = FF_0 = I$, which proves (ii) of Lemma 2.2 for LCLM.

Now second we consider the results for GCRD and intersection.

Let R_1, R_2 be row proper. From Willems [5] it follows that there exists a row proper R_0 such that $B_\infty(R_0) = B_\infty(R_1) \cap B_\infty(R_2)$. That $R_0 \in \text{GCRD}(R_1, R_2)$ is proved in a way analogous to the result for LCLM, and one easily gets Lemma 2.2(i) and (ii) for GCRD.

To prove Proposition 2.4(ii), let $R \in \text{GCRD}(R_1, R_2)$ row proper and $d(R_i) \leq T-1$, $i = 1, 2$. By Lemma 2.2(ii) there exists a unimodular U such that $R = UR_0$, so $B_\infty(R) = B_\infty(R_0)$ and $B_T(R) = B_T(R_0)$. So it suffices to prove that $B_T(R_0) = \sum \{B_T(\tilde{R}); B_T(\tilde{R}) \subset B_T(R_1) \cap B_T(R_2)\}$.

Now $B_\infty(R_0) \subset B_\infty(R_i)$ and row properness implies $B_T(R_0) \subset B_T(R_i)$, $i = 1, 2$; hence $B_T(R_0) \subset B_T(R_1) \cap B_T(R_2)$. So it suffices to prove that for \tilde{R} row proper $\{B_T(\tilde{R}) \subset B_T(R_1) \cap B_T(R_2)\} \Rightarrow \{B_T(\tilde{R}) \subset B_T(R_0)\}$. Let \tilde{R} consist of the rows of \tilde{R} of degree at most $T-1$, then $d(R_i) \leq T-1$, $i = 1, 2$, and Lemma 2.1 imply there exists an F_i such that $R_i = F_i\tilde{R}$; hence $B_\infty(\tilde{R}) \subset B_\infty(R_0)$ and $B_T(\tilde{R}) = B_T(\tilde{R}) \subset B_T(R_0)$. \square

Proof of Proposition 5.1. Let P be bilaterally monotone on $B \in \tilde{\mathbb{B}}_T$. By taking $t = 2$ in the definition of bilateral monotonicity we have that generically in $w \in B$, $\{B_{T-1}(R) \in P_{T-1}(w|_{[2,T]})$, $B' \in P_T w\} \Rightarrow \{\sigma B' \subset B_{T-1}(R)\} \Rightarrow \{B' \subset B_T(\sigma R)\}$, which proves the shift invariance condition for $t = T$.

To prove this condition for general $t \in [2, T]$, let $B_{t-1}(R) \in P_{t-1}(w|_{[2,t]})$ and $B' \in P_t(w|_{[1,t]})$. Now $B \in \tilde{\mathbb{B}}_T$, so by Proposition 2.1(ii) there exists $\tilde{w} \in B$, $\tilde{w}|_{[T-t+1,T]} = w|_{[1,t]}$ and $B_{t-1}(R) \in P_{t-1}(\tilde{w}|_{[T-t+2,T]})$, $B' \in P_t(\tilde{w}|_{[T-t+1,T]})$. Now bilateral monotonicity implies that generically in $\tilde{w} \in B$, $\sigma B' \subset B_{t-1}(R)$; hence $B' \subset B_t(\sigma R)$. We have to prove that this holds generically in $w \in B$. It is sufficient to construct a linear bijection $w \rightarrow \tilde{w}$. Without giving details this can be done as follows.

Let $B^{ee} := \{w \in \mathbb{R}^q\}^{\mathbb{Z}}$; $w|_{[\tau, \tau+T-1]} \in B$ for all $\tau \in \mathbb{Z}$. Because B is translation invariant $B^{ee}|_{[1,T]} = B$. It can be shown that there exists a linear injection $L: B \rightarrow B^{ee}: w \rightarrow w^{ee}$ with $w^{ee}|_{[1,T]} = w$ such that for all $\tau \in \mathbb{Z}$ $L_\tau: B \rightarrow B: w \rightarrow w^{ee}|_{[\tau+1, \tau+T]}$ is a bijection. Then for $w \in B$ take $\tilde{w} := L_{t-T} w$. \square

Proof of Lemma 6.1. (\Leftarrow) Let $r(n)$ be linearly independent from $r(1), \dots, r(n-1)$ and $r(n+1)$ linearly dependent on $r(1), \dots, r(n)$, say $r(n+1) = \sum_{i=1}^n \alpha_i r(i)$ (defined for the columns $1, \dots, t-n$ of $H_t(w)$). Define $w(\tau)$, $\tau > t$, recursively by $w(\tau) = \sum_{i=1}^n \alpha_i w(\tau - n - 1 + i)$ and define a Hankel extension M of $H_t(w)$ by $m_{ij} := w(i+j-1)$. Using the Hankel structure one gets $\text{rank } M = n$; hence $\text{rank } H_t(w) \leq n$. To prove $\text{rank } H_t(w) \geq n$, let M' be an arbitrary extension of $H_t(w)$ and let $d := \text{rank } M'$. If $d < n$ this would imply that among the rows $1, \dots, n$ of M' at least one, say row n' , is linearly dependent on the foregoing ones. This implies that $r(n')$ is linearly dependent on $r(1), \dots, r(n'-1)$, and because of the Hankel structure of $H_t(w)$ and the fact $n' \leq n$ this means that $r(n)$ would be linearly dependent on $r(1), \dots, r(n-1)$. So $\text{rank } M' \geq n$ and hence $\text{rank } H_t(w) \geq n$.

(\Rightarrow) Let $\text{rank } H_t(w) = n$. Then $r(n)$ cannot be linearly dependent on $r(1), \dots, r(n-1)$ as the construction above would give $\text{rank } H_t(w) \leq n-1$. Moreover, $r(n+1)$ cannot be linearly independent from $r(1), \dots, r(n)$ as this would imply that any extension of $H_t(w)$ would have rank at least $n+1$. \square

Proof of Lemma 6.2. A minimal rank extension which is Hankel was constructed in the proof of (\Leftarrow) of Lemma 6.1. \square

Proof of Proposition 6.2. Let $R \neq 0$ have degree d . First assume $d \leq \text{ENT}(t/2)$, so we have to show that generically in $w \in B(R)$ $\text{rank } H_t(w) = d$.

For $w \in B(R)$ row $d+1$ of $H_t(w)$ is linearly dependent on the foregoing ones, so $\text{rank } H_t(w) \leq d$. To prove that generically $\text{rank } H_t(w) = d$, it suffices to show that generically row d is linearly independent from the foregoing ones. Sufficient for this is that generically in $w \in B(R)$ $\text{rank } H_{d,d}(w) = d$, where

$$H_{d,d}(w) := \begin{bmatrix} w(1) & w(2) & \cdots & w(d) \\ w(2) & w(3) & \cdots & w(d+1) \\ \vdots & \vdots & & \vdots \\ w(d) & w(d+1) & \cdots & w(2d-1) \end{bmatrix}.$$

Note $d \leq \text{ENT}(t/2)$, so $2d-1 \leq t$ and $H_{d,d}(w)$ is well defined.

In $B(R)$, $w(\tau)$, $\tau \in [1, d]$, can be chosen arbitrarily while $w(\tau)$, $\tau \in [d+1, t]$ can be expressed as linear functions of $w(\tau)$, $\tau \in [1, d]$. So for $w \in B(R)$, $\det H_{d,d}(w)$ can be considered as a polynomial in $(w(1), \dots, w(d)) \in \mathbb{R}^d$. It suffices to show that $\det H_{d,d}(w)$ is not the zero polynomial, because then $\{w, \text{rank } H_{d,d}(w) < d\} = \{w; \det H_{d,d}(w) = 0\}$ is a proper algebraic variety and hence $\text{rank } H_{d,d}(w) = d$ generally in $w \in B(R)$.

That $\det H_{d,d}(w) \neq 0$ is seen as follows. We claim that $\det H_{d,d}(w)$ contains $\{w(d)\}^d$ as a term with coefficient ± 1 . Indeed, $\det H_{d,d}(w) = \sum_{p \in P} \text{sign}(p) \cdot \prod_{i=1}^d a_{ip(i)}$ where $H_{d,d}(w) = (a_{ij})$, P denotes the set of all permutations of $\{1, \dots, d\}$ and $\text{sign}(p) \in \{-1, +1\}$. In order to get $\{w(d)\}^d$, from every row and column in $H_{d,d}(w)$ one has to choose an element which involves $w(d)$. In the first row this is only the element $(1, d)$ so $p(1) = d$. In the second row only the elements $(2, d-1)$ and possibly $(2, d)$ contain $w(d)$, so necessarily $p(2) = d-1$. Going on in this way one gets for $\{w(d)\}^d$ the unique permutation $p = \{d, d-1, \dots, 2, 1\}$. This proves our claim and hence $\det H_{d,d}(w) \neq 0$.

Next assume $d > \text{ENT}(t/2)$. By a similar reasoning as before one can show that generically in $w \in B(R)$ $H_{\text{ENT}(t/2), \text{ENT}(t/2)}(w)$ has rank $\text{ENT}(t/2)$ and hence row $\text{ENT}(t/2)+1$ of $H_t(w)$ then is linearly dependent on the foregoing ones (its length is $\text{ENT}(t/2)$ if t is even, $\text{ENT}(t/2)-1$ if t is odd). So then $\text{rank } H_t(w) \leq \text{ENT}(t/2)$ and hence it equals $\text{ENT}(t/2)$ as row $\text{ENT}(t/2)$ is linearly independent from the foregoing ones. \square

Proof of Theorem 1. (i) The proof is obvious.

(ii) Not monotone. Let $B = \mathbb{R}^T$ and $t \in [3, T]$ odd. Then generally in $w \in B$ there holds that for $B_{t-1} \in P_{t-1}^K(w|_{[1, t-1]})$, $B_t \in P_t^K(w|_{[1, t]})$ $\dim B_{t-1} = (t-1)/2$ $\dim B_t = (t+1)/2$. For $t \geq 3$ $(t+1)/2 \leq t-1$ and $\dim B_t = (t+1)/2$ implies $\dim B_{t|_{[1, t-1]}} = (t+1)/2$ so generally $B_{t|_{[1, t-1]}} \not\subset B_{t-1}$ and P^K is not monotone. We have used the fact that $\{B \in \mathbb{B}_t, \dim B = d\} \Rightarrow \{\dim B|_{[1, \tau]} = d \text{ for all } \tau \in [d, t]\}$ which follows from $\{B \in \mathbb{B}_t, \dim B = d\} \Leftrightarrow \{\text{there exists } R \text{ of degree } d \text{ such that } B = B(R)\}$.

Not shift invariant. Let $B = \mathbb{R}^T$ and take $t = 3$, so $B|_{[2, 3]} = \mathbb{R}^2$ and $B|_{[1, 3]} = \mathbb{R}^3$. Let $w \in B$, $w|_{[1, 3]} = (a, b, c)$ with $a \neq 0$, $b \neq 0$, $ac - b^2 \neq 0$. Then $P_2^K(w|_{[2, 3]}) = B_2(\sigma - (c/b))$ and $B_3(\sigma^2 - (c/a)) \in P_3^K(w|_{[1, 3]})$. Shift invariance would require that generically $B_3(\sigma^2 - (c/a)) \subset B_3(\sigma(\sigma - (c/b))) = B_3(\sigma^2 - (c/b)\sigma)$ which clearly does not hold true.

Not linear. Take $T = 3$, $B_1 := B(\sigma^2 - 1)$, $B_2 := B(\sigma + 2)$. Then $B_1 + B_2 = \mathbb{R}^3$ so generically in $(w_1, w_2) \in B_1 \times B_2$ if $B \in P_3^K(w_1 + w_2)$ then $\dim B = 2$. Also generically $B_1 \in P_3^K w_1$ and $B_2 \in P_3^K w_2$. Linearity would require that generically $\mathbb{R}^3 = B_1 + B_2 \subset B$ which is false.

(iii) Take for example $T = 3$, $B_0 = \mathbb{R}^3$. Then generically in $w \in B_0$ if $B \in P_T^K w$ then $\dim B = 2$ so $B_0 \not\subset B$.

(iv) We will determine $\text{im } P_T^K, \mathbb{B}_{P_T^K}^w, \mathbb{B}_{P_T^K}^s$.

That $\text{im } P_T^K = \mathbb{B}_T$ is seen as follows. If $B = \mathbb{R}^T$ then take $w \in \mathbb{R}^T$ defined by $w(t) = 0$, $t \in [1, T-1]$, $w(T) = 1$, so $\text{rank } H_T(w) = T$ and by Proposition 6.1(i) $\mathbb{R}^T = P_T^K w$. If $\mathbb{R}^T \neq B \in \mathbb{B}_T$ then according to Proposition 2.3(i) there exists R with $d = d(R) \leq T-1$

such that $B = B(R)$. Choose $w \in B(R)$ by $w(\tau) = 0$, $\tau \in [1, d-1]$, $w(d) = 1$ and $w(\tau)$ for $\tau \in [d+1, t]$ computed by means of R . Then $\text{rank } H_T(w) = d$ and by Proposition 6.1 (i) $B(R) \in P_T^K w$.

Next we prove $\mathbb{B}_{P_T^K}^w = \{B \in \mathbb{B}_T; c(B) \leq \text{ENT}(T/2)\}$. From Example 6.1 we know already $\mathbb{R}^T \not\subset \mathbb{B}_{P_T^K}^w$. Now let $\mathbb{R}^T \neq B \in \mathbb{B}_T$ and $R \neq 0$ with $c(B) = d = d(R) \leq T-1$ such that $B = B(R)$. If $d > \text{ENT}(T/2)$ then from Proposition 6.2 generically in $w \in B(R)$ if $B' \in P_T w$ then $\dim B = \text{ENT}(T/2)$; hence $B(R) \notin P_T^K w$, so $B(R) \notin \mathbb{B}_{P_T^K}^w$. If $d \leq \text{ENT}(T/2)$ then generically in $w \in B(R)$ $\text{rank } H_T(w) = d$ so $B(R) \in P_T^K w$ and $B(R) \in \mathbb{B}_{P_T^K}^w$.

Finally we consider $\mathbb{B}_{P_T^K}^s$. First let T be even. Let $B \in \mathbb{B}_T$ with $d := c(B) \leq \text{ENT}(T/2)$. Then generically in $w \in B$ the first d rows of $H_T(w)$ are linearly independent and row $d+1$ has $T-d \geq T/2 \geq d$ elements, and this row is linearly dependent on the foregoing ones. Using Proposition 6.1(ii) this implies that generically in $w \in B$ $P_T^K w = B$ (i.e., assigned model is unique). So $\mathbb{B}_{P_T^K}^s \supset \mathbb{B}_{P_T^K}^w$, hence equality holds.

Next let T be odd. If $c(B) \leq \text{ENT}(T/2) - 1$ then by a reasoning as before one gets $B \in \mathbb{B}_{P_T^K}^s$. If $d := c(B) = \text{ENT}(T/2)$ then in $H_T(w)$ row $d+1$ consists of $T-d = (T-1)/2 < d = (T+1)/2$ elements and generically $P_T^K w$ is not unique; hence $B \notin \mathbb{B}_{P_T^K}^s$.

(v) As can be seen from the reasoning in (ii), (iii) and (iv), lacking the properties of monotonicity, linearity, truthfulness and prudence has not to do with possible nonunique assignment of models by P^K . We shall show that shift invariance also cannot be obtained by choice of a selection rule S .

To get shift invariance, taking the example in (ii) with $a \neq 0$, $b \neq 0$, $c \neq 0$, $ac - b^2 \neq 0$, this would require that for $B \in P_3^K(a, b, c)$, $B \subset B(\sigma^2 - (c/b)\sigma)$ while $\dim B = 2$, so this requires $S_3(a, b, c) = B(\sigma^2 - (c/b)\sigma)$. Moreover it is required that (generically) $S_4(d, a, b, c) \subset B(\sigma^3 - (c/b)\sigma^2)$. Now generically if $B \in P_4^K(d, a, b, c)$ then $\dim B = 2$. Let $B(\sigma^2 + \alpha\sigma + \beta) \in P_4^K(d, a, b, c)$. In order that $B(\sigma^2 + \alpha\sigma + \beta) \subset B(\sigma^3 - (c/b)\sigma^2)$ according to Lemma 2.1 there has to exist a γ such that $(\sigma^2 + \alpha\sigma + \beta)(\sigma + \gamma) = \sigma^3 - (c/b)\sigma^2$, which implies that $(\alpha, \beta) = (0, 0)$ or $(\alpha, \beta) = (-c/b, 0)$. But $B(\sigma^2) \notin P_4^K(d, a, b, c)$ (it requires $b = c = 0$) and $B(\sigma^2 - (c/b)\sigma) \notin P_4^K(d, a, b, c)$ (it requires $ac - b^2 = 0$). So it follows that it is impossible to construct a shift invariant selection rule for P^K . \square

Proof of Theorem 2. The proof of this theorem is quite lengthy and will be split in a number of steps. The result is proved by using a number of lemmas, some of which play a role in the proof of Proposition 8.1.

First we introduce some notation. T is assumed to be fixed throughout. For $R = \sum_{k=0}^{T-1} a_k \sigma^k \in \mathbb{R}[\sigma]$ let $I(R) := \{k \in [0, T-1]; a_k \neq 0\}$, so $\# I(R) = T - c(R)$. Let $\mathbb{B}^*(d)$, $\mathbb{B}^*(I)$ and $\mathbb{B}^*(d, I)$ as subsets of \mathbb{B}_T^* be defined as follows. $\mathbb{B}^*(d) := \{B(R) \in \mathbb{B}_T^*; d(R) = d\}$, $\mathbb{B}^*(I) := \{B(R) \in \mathbb{B}_T^*; I(R) \supset I\}$, $\mathbb{B}^*(d, I) := \mathbb{B}^*(d) \cap \mathbb{B}^*(I)$. Moreover define $W(d)$, $W(I)$ and $W(d, I)$ as subsets of \mathbb{R}^T by $W(d) := \cup \{B(R); B(R) \in \mathbb{B}^*(d)\} = \{w \in \mathbb{R}^T; \text{there exists } B(R) \in \mathbb{B}^*(d), w \in B(R)\}$, $W(I) := \cup \{B(R); B(R) \in \mathbb{B}^*(I)\}$ and $W(d, I) := \cup \{B(R); B(R) \in \mathbb{B}^*(d, I)\}$.

Let $B_0 \in \mathbb{B}_T$ be fixed, $w \in B_0$ and $H_T(w)$ its incomplete Hankel array. We now first give an outline of the proof of Theorem 2 by means of four lemmas and then will give the proof of these lemmas.

LEMMA 1. For every (d, I) either (i) $w \notin W(d, I)$ generically in $w \in B_0$, or (ii) $B_0 \subset W(d, I)$.

LEMMA 2. $\{B_0 \subset W(d, I)\} \Rightarrow \{\text{there exists } B(R(d, I)) \in \mathbb{B}^*(d, I) \text{ such that } B_0 \subset B(R(d, I))\}$.

This lemma states that if for every $w \in B_0$ there exists a model $B_w(R) \in \mathbb{B}^*(d, I)$ such that $w \in B_w(R)$, then there exists such a model independent from $w \in B_0$.

For B_0 define $d_0 \in [0, T]$ as follows. If $B_0 \not\subset W(d, I)$ for all $d \in [0, T-1]$ then $d_0 := T$, else $d_0 := \min \{d \in [0, T-1]; \text{ there exists } I, B_0 \subset W(d, I)\}$.

If $d_0 = T$, then generically in $w \in B_0$, $w \notin W(d)$ for all $d \in [0, T-1]$, which means that generically row $d+1$ of H_T is not linearly dependent on less than $T-d$ foregoing rows of H_T , so $P_T^0 w = \mathbb{R}^T$ generically on B_0 and Theorem 2 follows as obviously there is no $B(R) \in \mathbb{B}_T^*$ with $B_0 \subset B(R)$ and $c(B(R)) < T$ in this case.

For $d_0 \in [0, T-1]$ let J_0 be defined by $J_0 := \{I; B_0 \subset W(d_0, I)\}$. Because by Lemma 1 generically on B_0 , $w \notin W(d)$ for $d < d_0$ and by Lemma 2 $B_0 \subset B(R(d_0, I))$ for $I \in J_0$, it follows from the definition of P_T^0 that generally on B_0 , $\{B(R(d_0, I)); I \in J_0\} \subset P_T^0 w \subset \mathbb{B}^*(d_0)$. Indeed, on B_0 generally no remarkable laws of degree $d < d_0$ are satisfied while remarkable laws of degree d_0 always exist. Because of Lemma 1 we even have that generically on B_0 , $\{B(R(d_0, I)); I \in J_0\} \subset P_T^0 w \subset \bigcup \{\mathbb{B}^*(d_0, I); I \in J_0\}$.

LEMMA 3. For $I \in J_0$ generically on B_0 $P_T^0 w \cap \mathbb{B}^*(d_0, I)$ is a singleton, i.e., $B(R(d_0, I))$.

The generic way in which P_T^0 assigns models on the basis of data from B_0 is described in Lemma 4, which is a direct consequence of Lemma 3 and the preceding discussion.

LEMMA 4. Generically for $w \in B_0$, $P_T^0 w = \{B(R(d_0, I)); I \in J_0\}$.

Now from Lemma 2 and Lemma 4 it follows that generally on B_0 if $B \in P_T^0 w$ then $B_0 \subset B$ and $c(B) = d_0$. To conclude the proof of Theorem 2, note that by definition of d_0 , if $B \in \mathbb{B}_T^*$ with $c(B) < d_0$, then $B_0 \not\subset B$. On the other hand, if $B \in \mathbb{B}_T^*$, $c(B) = d_0$, $B_0 \subset B$, then generically $B \in P_T^0 w$. This proves Theorem 2.

Finally we will prove the foregoing lemmas.

Proof of Lemma 1. Assume that $w \notin W(d, I)$ is not generically true on B_0 ; then we have to show that $B_0 \subset W(d, I)$.

Let $[0, T-1] \setminus I = \{i_1, i_2, \dots, i_{c-1}, d\}$ with $0 \leq i_1 < i_2 < \dots < i_{c-1} < d$ and for $w \in B_0$ define $H_I(w)$ by

$$H_I(w) := \begin{bmatrix} w(i_1+1) & w(i_1+2) & \cdots & w(i_1+T-d) \\ w(i_2+1) & w(i_2+2) & \cdots & w(i_2+T-d) \\ \vdots & \vdots & & \vdots \\ w(i_{c-1}+1) & w(i_{c-1}+2) & \cdots & w(i_{c-1}+T-d) \\ w(d+1) & w(d+2) & \cdots & w(T) \end{bmatrix}.$$

Now $w \in W(d, I)$ if and only if the last row of $H_I(w)$ is linearly dependent on the foregoing ones. It is given that this is not generically false on B_0 and we have to show that it is then always true on B_0 . To do this it suffices to express the statement (S) "The last row of $M \in \mathbb{R}^{n_1 \times n_2}$ is linearly dependent on the foregoing ones" as an algebraic condition on the elements of M . If this is the case, then $w \in W(d, I)$ is an algebraic condition on $w(1), \dots, w(d_0)$ where $\bar{d}_0 := \dim B_0$, as $w(t)$, $t \in [\bar{d}_0+1, T]$, can be expressed as linear functions of $w(1), \dots, w(d_0)$. As it is given that this condition is not generically falsified it is satisfied everywhere on B_0 , which then proves Lemma 1.

An algebraic formulation of (S) can be derived as follows. Partition $M = \begin{pmatrix} \tilde{M} \\ m \end{pmatrix}$ where m denotes the last row of M . One easily verifies that $\{m \text{ linearly dependent on rows } \tilde{M}\} \Leftrightarrow \{\ker \tilde{M} = \ker M\} \Leftrightarrow \{m^T \perp \ker \tilde{M}\} \Leftrightarrow \{m^T \perp \ker \tilde{M}^T \tilde{M}\}$. Let $(\tilde{M}^T \tilde{M})^-$ denote the Moore-Penrose generalized inverse of $\tilde{M}^T \tilde{M}$. It is well known (see, e.g., Campbell and Meijer [1, Thms. 7.3.4, 7.5.1]) that $\ker \tilde{M}^T \tilde{M} = \text{im}(I - (\tilde{M}^T \tilde{M})^- \tilde{M}^T \tilde{M})$ and because $\tilde{M}^T \tilde{M}$ is symmetric there exists a polynomial p such that $(\tilde{M}^T \tilde{M})^- = p(\tilde{M}^T \tilde{M})$. Define $p_i := (I - p(\tilde{M}^T \tilde{M}) \cdot \tilde{M}^T \tilde{M}) \cdot e_i$ where e_i denotes the i th unit vector, then (S) is equivalent to $\sum_i \langle m^T, p_i \rangle^2 = 0$ which is an algebraic condition on the elements of M . \square

Proof of Lemma 2. Let $B_0 \subset W(d, I)$ and $[0, T-1] \setminus I = \{i_1, \dots, i_{c-1}, d\}$ as in the proof of Lemma 1. On B_0 the last row of $H_I(w)$ always is linearly dependent on the foregoing ones, so for all $t \in [1, T-d]$, for all $w \in B_0$, $\{w(i_k+t)=0, k \in [1, c-1]\} \Rightarrow \{w(d+t)=0\}$. Because B_0 is linear this implies that for all $t \in [1, T-d]$ there is a map $f_t: B_0|_{\{i_k+t; k \in [1, c-1]\}} \rightarrow B_0|_{\{d+t\}}: (w(i_1+t), \dots, w(i_{c-1}+t)) \rightarrow w(d+t)$ such that $(w(i_1+t), \dots, w(i_{c-1}+t), f_t(w(i_1+t), \dots, w(i_{c-1}+t))) \in B_0|_{\{i_k+t; k \in [1, c-1]\} \cup \{d+t\}}$. Linearity of B_0 implies linearity of f_t , say $f_t(w(i_1+t), \dots, w(i_{c-1}+t)) = \sum_{k=1}^{c-1} a_k(t)w(i_k+t)$. Shift invariance of B_0 implies $w(d+t) = (\sigma^{t-1}w)(d+1) = \sum_{k=1}^{c-1} a_k(1)(\sigma^{t-1}w)(i_k+1) = \sum_{k=1}^{c-1} a_k(1)w(i_k+t)$ for all $t \in [1, T-d]$. Define $R(d, I) := -\sigma^d + \sum_{k=1}^{c-1} a_k(1)\sigma^{i_k}$, then for $w \in B_0$, $w \in B(R(d, I))$. Hence $B_0 \subset B(R(d, I)) \in \mathbb{B}^*(d, I)$. \square

To prove Lemma 3 we will make use of a result stated in Lemma 5 which also plays a role in the proof of Proposition 8.1. Let $0 \leq i_1 < i_2 < \dots < i_{c-1} < d \leq T-c$ and

$$M(w) := \begin{bmatrix} w(i_1+1) & w(i_1+2) & \cdots & w(i_1+T-d) \\ w(i_2+1) & w(i_2+2) & \cdots & w(i_2+T-d) \\ \vdots & \vdots & & \vdots \\ w(i_{c-1}+1) & w(i_{c-1}+2) & \cdots & w(i_{c-1}+T-d) \end{bmatrix}.$$

LEMMA 5. If rank $M(w) \leq c-2$ everywhere on B_0 , then there exists $R \neq 0$, $d(R) \leq i_{c-1}$, $I(R) \supset [0, T-1] \setminus \{i_1, \dots, i_{c-1}\}$, such that $B_0 \subset B(R) \in \mathbb{B}_T^*$.

Proof of Lemma 5. There is at least one row of M which is not generically linearly independent from the foregoing ones, say row k_0 . Exactly analogous to the proof of Lemma 1 it follows that this row then always is linearly dependent on the foregoing ones. Exactly analogous to the proof of Lemma 2 this implies existence of a_k such that on B_0 $w(i_{k_0}+t) = \sum_{k=1}^{k_0-1} a_k w(i_k+t)$, $t \in [1, T-d]$. By shift invariance of B_0 this then also holds true on $[1, T-i_{k_0}]$. Define

$$R := \sigma^{i_{k_0}} - \sum_{k=1}^{k_0-1} a_k \sigma^{i_k},$$

then $B_0 \subset B(R) \in \mathbb{B}_T^*$ while $d(R) = i_{k_0} \leq i_{c-1}$ and $I(R) \supset [0, T-1] \setminus \{i_1, \dots, i_{c-1}\}$. \square

Proof of Lemma 3. Let $I \in J_0$, and define M as before, with $\{i_1, \dots, i_{c-1}, d_0\} := [0, T-1] \setminus I$.

We state that generically on B_0 , rank $M(w) = c-1$. For suppose this is not true; then $\det MM^T \neq 0$ is not generic, so $\det MM^T = 0$ on B_0 and rank $M(w) \leq c-2$ on B_0 . By Lemma 5 this would imply there exists $R \neq 0$ such that $B_0 \subset B(R) \in \mathbb{B}^*(d', I')$ where $d' \leq i_{c-1} < d_0$ and $I' \supset [0, T-1] \setminus \{i_1, \dots, i_{c-1}\}$. Hence $B_0 \subset B(R) \subset W(d', I')$ and $d' < d_0$, which contradicts the definition of d_0 .

Now suppose $B(R_j) \in P_T^0 w \cap \mathbb{B}^*(d_0, I)$, $j = 1, 2$. Let $R_j(\sigma) = \sigma^d + \sum_{k=1}^{c-1} a_k^j \sigma^{i_k}$, $a^j := (a_1^j, \dots, a_{c-1}^j)$, $j = 1, 2$. Using the notation of Lemma 1, this means $(a^1, 1)H_I(w) = (a^2, 1)H_I(w) = 0$, so $(a^1 - a^2)M(w) = 0$. As generically on B_0 rank $M(w) = c-1$, we get generically on B_0 $a^1 = a^2$; hence $R_1 = R_2$, i.e., generically on B_0 $P_T^0 w \cap \mathbb{B}^*(d_0, I)$ contains at most one model. From Lemma 2 and the discussion following this lemma we know that generically on B_0 $B(R(d_0, I)) \in P_T^0 w \cap \mathbb{B}^*(d_0, I)$. So generically on B_0 $P_T^0 w \cap \mathbb{B}^*(d_0, I)$ consists of a singleton, i.e., $B(R(d_0, I))$.

This concludes the proof of Theorem 2. \square

Proof of Corollary 7.1. From Theorem 2 we immediately conclude that if $B_0 \in \mathbb{B}_T^*$ then generically on B_0 , $B_0 \in P_T^0 w$. To prove the corollary, due to Theorem 2 it suffices to show that $\{B_0 \subset B \in \mathbb{B}_T^*, c(B) = c(B_0)\} \Rightarrow B = B_0$. This easily follows from Lemma 2.1 and the fact that (for $R \neq 0$) $c(B(R)) = d(R)$. \square

Proof of Theorem 3. (i) The proof is obvious from the definition of P^0 .

(ii) The proof is obvious from Theorem 2.

(iii) Obviously $\mathbb{B}_{P_t^0}^s \subset \mathbb{B}_{P_t^0}^w \subset \text{im } P_t^0 \subset \mathbb{B}_t^*$, so it suffices to show that $\mathbb{B}_t^* \subset \mathbb{B}_{P_t^0}^s$. This is immediate from Corollary 7.1.

(iv) Consider P^0 with the action of P_t^0 restricted to models in the set \mathbb{B}_t^* .

For monotonicity and shift invariance, consider the inclusion conditions on t which involve modeling $w|_{[1,t-1]}$ or $w|_{[2,t]}$, and $w|_{[1,t]}$. Assume $B = B_T(R) \in \mathbb{B}_T$ with $c(R) + d(R) \leq t - 1$.

For the monotonicity condition, observe $w|_{[1,t-1]} \in B_{t-1}(R) \in \mathbb{B}_{t-1}^*$ and $w|_{[1,t]} \in B_t(R) \in \mathbb{B}_t^*$. From Corollary 7.1 it follows that generically in $w \in B$ $P_{t-1}^0(w|_{[1,t-1]}) = B_{t-1}(R)$ and $P_t^0(w|_{[1,t]}) = B_t(R)$, and the condition $B_t(R)|_{[1,t-1]} \subset B_{t-1}(R)$ is trivial. Note that in fact Corollary 7.1 only gives that for example $P_{t-1}^0(w|_{[1,t-1]}) = B_{t-1}(R)$ generically in $w|_{[1,t-1]} \in B_{t-1}(R)$. That this also holds true generically in $w \in B$ can be derived from the fact that $d(R) \leq t - 1$ which implies that there is a linear bijection $w|_{[1,t-1]} \rightarrow w$ from $B_{t-1}(R)$ to B .

For the shift invariance condition, consider two cases for $R = \sum_{k=0}^{d(R)} a_k \sigma^k$. If $a_k \neq 0$, then $w|_{[2,t]} \in B|_{[2,t]} = B|_{[1,t-1]}$, while if $a_k = 0$, then $w|_{[2,t]} \in B|_{[2,t]} = B_{t-1}(\sigma^{-1}R)$. Generically in $w \in B$, $P_{t-1}^0(w|_{[2,t]}) = B_{t-1}(R)$ in the first case, $P_{t-1}^0(w|_{[2,t]}) = B_{t-1}(\sigma^{-1}R)$ in the second case. The shift invariance condition is trivially satisfied in both cases.

Concerning linearity, let $B_i \in \mathbb{B}_i^*$, $i = 1, 2$, then generically in $(w_1, w_2) \in B_1 \times B_2$ $P_i^0 w_i = B_i$, $i = 1, 2$, while due to Corollary 7.1 generically for $B \in P_t^0(w_1 + w_2)$ $B_1 + B_2 \subset B$, which proves linearity. Note that Corollary 7.1 in fact gives $B_1 + B_2 \subset B \in P_t^0 w$ generically in $w \in B_1 + B_2$, but one easily proves that this then also generically holds true in $(w_1, w_2) \in B_1 \times B_2$ with $w := w_1 + w_2$.

Finally we will give an example which shows that P^0 is not monotone and not shift invariant.

Example. Let $T = 20$, $R := \sigma^{10} + \sigma^9 + 2\sigma^8 + \sigma^7 + \sigma^6 + 2\sigma^5 + \sigma^4 + \sigma^3 + 2\sigma^2 + \sigma + 1$, so $c(R) + d(R) = 21$, and consider $B := B(R)$. Further define $R_1 := (\sigma - 1)R = \sigma^{11} + \sigma^9 - \sigma^8 + \sigma^6 - \sigma^5 + \sigma^3 - \sigma^2 - 1$ with $c(R_1) + d(R_1) = 19$ and $R_2 := (\sigma - \frac{1}{2})R = \sigma^{11} + \frac{1}{2}\sigma^{10} + \frac{3}{2}\sigma^9 + \frac{1}{2}\sigma^7 + \frac{3}{2}\sigma^6 + \frac{1}{2}\sigma^4 + \frac{3}{2}\sigma^3 + \frac{1}{2}\sigma - \frac{1}{2}$ with $c(R_2) + d(R_2) = 20$. Then generically in $w \in B$, $P_{19}^0(w|_{[1,19]}) = P_{19}^0(w|_{[2,20]}) = B_{19}(R_1)$ while $P_{20}^0 w = \{B_{20}(R_1), B_{20}(R_2)\}$. So P^0 is not monotone as $B_{20}(R_2)|_{[1,19]} \not\subset B_{19}(R_1)$ and P^0 is not shift invariant as $B_{20}(R_2) \subset B_{20}(\sigma R_1)$. This concludes the proof of Theorem 3. \square

Proof of Proposition 8.1. For $B_0 \in \mathbb{B}_T$ we have to prove that generically in $w \in B_0$ the following holds:

$$\{R \in L(w) := \{R \neq 0; d(R) \leq T - 1 \text{ and } w \in B(R) \in \mathbb{B}_T^*\}\} \Rightarrow \{B_0 \subset B(R)\}.$$

We will use some of the lemmas and the notation introduced in the proof of Theorem 2. Further we define $K_0 := \{(d, I); B_0 \subset W(d, I)\}$ and $L(w; d, I) := \{R \in L(w); I(R) \supset I, d(R) = d\}$.

According to Lemma 1 generically on B_0 , $L(w; d, I) = \emptyset$ if $(d, I) \notin K_0$. So generically on B_0 , $L(w) = \bigcup \{L(w; d, I); (d, I) \in K_0\}$.

The following lemma is crucial in the proof of Proposition 8.1.

LEMMA 6. *Let $(d, I) \in K_0$ be fixed. Then there exist $n \geq 0$ and $R^{(j)} \in \mathbb{R}[\sigma]$, $j \in [0, n]$, such that*

- (i) $d = d(R^{(0)}) > d(R^{(1)}) > \dots > d(R^{(n)})$, $I(R^{(j)}) \supset I$, $j \in [0, n]$;
- (ii) $B_0 \subset B(R^{(j)})$ for all $j \in [0, n]$;
- (iii) generically in $w \in B_0$ $L(w; d, I) = \text{span}_0 \{R^{(j)}, j \in [0, n]\} := \{R; \text{there exists } \alpha_j \in \mathbb{R}, j \in [0, n], \alpha_0 \neq 0, R = \sum_{j=0}^n \alpha_j R^{(j)}\}$.

We will first give an interpretation of this lemma, then give the proof of Proposition 8.1 using Lemma 6 and finally prove Lemma 6.

Lemma 6 has the following interpretation. Let $(d, I) \in K_0$; then by definition of K_0 for every $w \in B_0$ there is a remarkable law R_w with $w \in B(R_w) \in \mathbb{B}^*(d, I)$. Now the

lemma states that generically in B_0 the class of unfalsified remarkable laws in $\mathbb{B}^*(d, I)$ is independent of $w \in B_0$ and that this class is spanned by a finite number of “basic laws” $R^{(j)}$ which moreover are true laws for B_0 .

Proposition 8.1 can be proved by using Lemma 6 in the following way. Generically on B_0 , $L(w) = \bigcup \{L(w; d, I); (d, I) \in K_0\}$, and as K_0 is a finite set it suffices to prove that for $(d, I) \in K_0$ generically in $w \in B_0$ $\{R \in L(w; d, I)\} \Rightarrow \{B_0 \subset B(R)\}$.

According to Lemma 6(ii), (iii), $R \in L(w; d, I)$ generically is of the form $R = \sum_{j=0}^n \alpha_j R^{(j)}$, $\alpha_0 \neq 0$, with $B_0 \subset B(R^{(j)})$, $j \in [0, n]$. Using Lemma 6(i) this implies that for $w \in B_0$, $[R^{(j)}(\sigma)w](t) = 0$, $t \in [1, T - d(R^{(j)})] \supset [1, T - d]$. This implies $[R(\sigma)w](t) = 0$, $t \in [1, T - d]$ and hence $w \in B(R)$ as $d(R) = d$ for $\alpha_0 \neq 0$. So $B_0 \subset B(R)$, which proves Proposition 8.1.

Now finally we will prove Lemma 6. First note that if $d_0 := \min \{d; \text{there exists } I \text{ such that } (d, I) \in K_0\}$ then it follows from Lemma 3 that for $(d_0, I_0) \in K_0$, $L(w; d_0, I_0)$ generically is a singleton $B(R(d_0, I_0))$, and according to Lemma 2 $B_0 \subset B(R(d_0, I_0))$, which proves Lemma 6 for d_0 . However, in general for $(d, I) \in K_0$ $L(w; d, I)$ need not generically be a singleton. As an example, let $T = 5$, $B_0 := B(\sigma - 1)$, then $(d, I) := (2, \{3, 4\}) \in K_0$ and for all $w \in B_0$, $L(w; d, I) \supset \{R_\alpha, \alpha \in \mathbb{R}\}$ where $R_\alpha := (\sigma - \alpha)(\sigma - 1) = \sigma^2 - (\alpha + 1)\sigma + \alpha$.

Proof of Lemma 6. We give the proof by construction. First we define $R^{(j)}$ and then we show that these have the desired properties.

Part (i) and (ii). Let $(d, I) \in K_0$, $[0, T - 1] \setminus I = \{i_1, i_2, \dots, i_{c-1}, d\}$, $0 \leq i_1 < i_2 < \dots < i_{c-1} < d$ ($c + d \leq T$) and define

$$H(w) := \begin{bmatrix} w(i_1 + 1) & w(i_1 + 2) & \cdots & w(i_1 + T - d) \\ w(i_2 + 1) & w(i_2 + 2) & \cdots & w(i_2 + T - d) \\ \vdots & \vdots & & \vdots \\ w(i_{c-1} + 1) & w(i_{c-1} + 2) & \cdots & w(i_{c-1} + T - d) \\ w(d + 1) & w(d + 2) & \cdots & w(T) \end{bmatrix}$$

and let $M_k(w)$ consist of the first k rows of $H(w)$.

As $B_0 \subset W(d, I)$, Lemma 2 implies there exists $R^{(0)}$ such that $B_0 \subset B(R^{(0)}) \in \mathbb{B}^*(d, I)$.

Now note that for $R = \sum_{k=1}^{c-1} a_k \sigma^{i_k} + a_c \sigma^d$, $a_c \neq 0$, $a := (a_1, \dots, a_c)$, there holds $\{R \in L(w; d, I)\} \Leftrightarrow aH(w) = 0$. If generically on B_0 $\text{rank } M_{c-1}(w) = c - 1$, then generically a is unique up to a constant factor and hence generically $L(w; d, I) = \{\alpha R^{(0)}; 0 \neq \alpha \in \mathbb{R}\}$ and Lemma 6 is shown with $n = 0$.

So suppose not generically $\text{rank } M_{c-1}(w) = c - 1$; hence not generically $\det M_{c-1}(w) M_{c-1}(w)^T \neq 0$, so $\det M_{c-1}(w) M_{c-1}(w)^T \equiv 0$ on B_0 and on B_0 $\text{rank } M_{c-1}(w) \leq c - 2$.

Lemma 5 implies there exists R' with $B_0 \subset B(R')$ and $d(R') \leq i_{c-1} < d = d(R^{(0)})$, $I(R') \supset I \cup \{d\}$. Let $R^{(1)}$ be such a law for which $d(R^{(1)})$ is maximal. Let $d(R^{(1)}) = i_{d_1}$ and $I_1 := [0, T - 1] \setminus \{i_1, \dots, i_{d_1}\}$.

Now either generically on B_0 , $\text{rank } M_{d_1-1}(w) = d_1 - 1$, in which case generically $L(w; i_{d_1}, I_1) = \{\alpha R^{(1)}, 0 \neq \alpha \in \mathbb{R}\}$ and we stop, or $\text{rank } M_{d_1-1}(w) \leq d_1 - 2$ on B_0 . In the latter case we find, using Lemma 5, a law $R^{(2)}$ of maximal degree in the class of laws with $I(R'') \supset I_1 \cup \{i_{d_1}\}$ such that $B_0 \subset B(R'')$. So $B_0 \subset B(R^{(2)})$. Let $d(R^{(2)}) = i_{d_2} \leq i_{d_1-1} < i_{d_1} = d(R^{(1)})$ and $I_2 := [0, T - 1] \setminus \{i_1, \dots, i_{d_2}\}$.

Going on in this way we find a number $n \leq c - 1$ such that for $j \in [0, n]$ there exists $R^{(j)}$ with $B_0 \subset B(R^{(j)})$, $I(R^{(j)}) \supset I$, $d(R^{(j)}) < d(R^{(j-1)})$ while for $i_{d_n} := d(R^{(n)})$, $I_n := [0, T - 1] \setminus \{i_1, \dots, i_{d_n}\}$ generically on B_0 $\text{rank } M_{d_n-1}(w) = d_n - 1$, so generically on B_0 $L(w; i_{d_n}, I_n) = \{\alpha R^{(n)}, 0 \neq \alpha \in \mathbb{R}\}$.

In this way we have defined $n, R^{(j)}, j \in [0, n]$ with $d = d(R^{(0)}) > d(R^{(1)}) > \dots > d(R^{(n)})$, $I(R^{(j)}) \supset I$ and $B_0 \subset B(R^{(j)})$. This proves (i) and (ii).

Part (iii). If $R = \sum_{j=0}^n \alpha_j R^{(j)}$, $\alpha_0 \neq 0$, then $d(R) = d$, $I(R) \supset I$ and $B_0 \subset B(R) \in \mathbb{B}_T^*$, so $R \in L(w; d, I)$. We now have to show that generically on B_0 if $R \in L(w; d, I)$ then there exists (α_k) , $\alpha_0 \neq 0$, such that $R = \sum_{j=0}^n \alpha_j R^{(j)}$. Without loss of generality we assume R and $R^{(j)}$ to be monic, $j \in [0, n]$.

So let $R \in L(w; d, I)$ be given. If $R = R^{(0)}$ then we are done, else define $R_1 := \{R - R^{(0)}\} \cdot \beta_0$ with β_0 such that R_1 is monic. We state that generically on B_0 there exists $j \in [1, n]$ such that $d(R_1) = d(R^{(j)})$. For suppose this does not hold true, then there exists $i_k \in I \setminus \{d(R^{(j)}), j \in [0, n]\}$ such that in $H(w)$ row k is not generically linearly independent of the foregoing ones; hence by Lemma 1 it is always linearly dependent on them and Lemma 2 implies there exists R_{i_k} , $d(R_{i_k}) = i_k$, $I(R_{i_k}) \supset [0, T-1] \setminus \{i_1, \dots, i_k\}$ with $B_0 \subset B(R_{i_k})$. Now $i_k < d(R^{(n)})$ is impossible by definition of n , so $d(R^{(n)}) < i_k < d(R^{(0)})$. Let j be such that $d(R^{(j)}) < i_k < d(R^{(j-1)}) =: d_{j-1}$; then this contradicts the construction of $R^{(j)}$ as being of maximal degree in the class of laws \tilde{R} such that $I(\tilde{R}) \supset [0, T-1] \setminus \{i_1, i_2, \dots, i_{d_{j-1}-1}\}$ and $B_0 \subset B(\tilde{R})$.

So indeed generically on B_0 there exists $j \in [1, n]$, say j_1 , such that $d(R_1) = d(R^{(j_1)})$. If $R_1 = R^{(j_1)}$, then stop, else define $R_2 := \{R_1 - R^{(j_1)}\} \cdot \beta_1$ where β_1 is such that R_2 is monic. Going on in this way we generically reduce R to laws of lower degree in the set $\{d(R^{(j)}), j \in [1, n]\}$. The process generically will end either if we find a $k \in [1, n]$ such that $R_k = R^{(j_k)}$ or get R_k with $d(R_k) = d(R^{(n)})$. As $\text{rank } M_{d_n-1}(w) = d_n - 1$ generically on B_0 , also generically $R_k = R^{(n)}$ in the latter case.

In this way we have that generically on B_0 if $R \in L(w; d, I)$ then $R = R^{(0)} + \beta_0^{-1} R_1 = R^{(0)} + \beta_0^{-1} R^{(j_1)} + \beta_0^{-1} \beta_1^{-1} R_2$ and going on in this way we find α_j , $\alpha_0 = 1$, such that $R = \sum_{j=0}^n \alpha_j R^{(j)}$.

This concludes the proof of Lemma 6 and hence of Proposition 8.1. \square

Proof of Corollary 8.1. According to Proposition 8.1 generally in $w \in B_0$ $B_0 \subset B(R(w))$, so generically $w \in B(R(w))$ and hence generically $P_T^* w = B(R(w))$, $R(w) := \text{GCD}\{R; R \in L(w)\}$. So generically on B_0 if $R \in L(w)$ then $B_0 \subset B(R)$ and $P_T^* w = B(R(w)) \subset B(R)$. \square

Proof of Corollary 8.2. Let $R := \text{GCD}\{\tilde{R} \neq 0; d(\tilde{R}) \leq T-1 \text{ and } B_0 \subset B(\tilde{R}) \in \mathbb{B}_T^*\}$, $R(w) := \text{GCD}\{\tilde{R}; \tilde{R} \in L(w)\}$. From Corollary 8.1 generically on B_0 if $\tilde{R} \in L(w)$ then $B_0 \subset B(\tilde{R}) \in \mathbb{B}_T^*$; hence by using Proposition 2.4(ii) generically on B_0 , $B(R) \subset B(R(w)) = P_T^* w$. On the other hand, if $\tilde{R} \neq 0$, $d(\tilde{R}) \leq T-1$ and $B_0 \subset B(\tilde{R}) \in \mathbb{B}_T^*$, then on B_0 $\tilde{R} \in L(w)$ so $B(R(w)) \subset B(\tilde{R})$ which implies $B(R(w)) \subset B(R)$, so generically on B_0 $P_T^* w \subset B(R)$. \square

Proof of Proposition 8.4. Let $B_0 \in \mathbb{B}_T$. As $\tilde{L}(w) \subset L(w)$ we conclude from Corollary 8.1 that generically on B_0 , $\{\tilde{R} \in \tilde{L}(w)\} \Rightarrow \{B_0 \subset B(\tilde{R})\}$. So generically on B_0 , $B_0 \subset B(\tilde{R}(w))$ and hence generically $\tilde{P}_T^* w = B(\tilde{R}(w))$ where $\tilde{R}(w) := \text{GCD}\{\tilde{R}; \tilde{R} \in \tilde{L}(w)\}$.

Now define $R := \text{GCD}\{\tilde{R} \neq 0; d(\tilde{R}) \leq T-1 \text{ and } B_0 \subset B(\tilde{R}) \in \tilde{\mathbb{B}}_T^*\}$. For every $\tilde{R} \neq 0$, $d(\tilde{R}) \leq T-1$ with $B_0 \subset B(\tilde{R}) \in \tilde{\mathbb{B}}_T^*$ on B_0 $\tilde{R} \in \tilde{L}(w)$; hence $B(\tilde{R}(w)) \subset B(R)$ and generically $\tilde{P}_T^* w \subset B(R)$. On the other hand generically on B_0 for $\tilde{R} \in \tilde{L}(w)$ $B_0 \subset B(\tilde{R}) \in \tilde{\mathbb{B}}_T^*$; hence generically $B(R) \subset B(\tilde{R}(w))$ and generically $B(R) \subset \tilde{P}_T^* w$. \square

Proof of Proposition 8.5. For $R = \sum_{k=0}^{T-1} a_k \sigma^k \in \mathbb{K}[\sigma]$ let $l(R) := \min\{k; a_k \neq 0\}$. It easily follows that $\{B(R) \in \mathbb{B}_T\} \Leftrightarrow \{l(R) = 0\}$.

Let $B_0 \in \mathbb{B}_T$. From Corollary 8.2 and Proposition 8.4 it follows that generically on B_0 , $P_T^* w = B(R)$ and $\tilde{P}_T^* w = B(\tilde{R})$, where $R := \text{GCD}\{R' \neq 0; d(R') \leq T-1 \text{ and } B_0 \subset B(R') \in \mathbb{B}_T^*\}$ and $\tilde{R} := \text{GCD}\{\tilde{R}' \neq 0; d(R') \leq T-1 \text{ and } B_0 \subset B(\tilde{R}') \in \tilde{\mathbb{B}}_T^*\}$.

(i) Let $B_0 \in \mathbb{B}_T$, so $l(R_0) = 0$, and $R' \neq 0$, $d(R') \leq T-1$ such that $B_0 \subset B(R') \in \mathbb{B}_T^*$. Then according to Lemma 2.1 there exists F' such that $R' = F' R_0$. If $l := l(R') = l(F') \neq 0$

then define \tilde{F}' by $\tilde{F}' := \sigma^{-l} \cdot F'$ and $\tilde{R}' := \tilde{F}' R_0$. So $B_0 \subset B(\tilde{R}') \in \tilde{\mathbb{B}}_T^*$ as $l(\tilde{R}') = 0$ and $c(\tilde{R}') + d(\tilde{R}') = c(R') + d(R') - 1 \leq T - 1 \leq T$. Now $R' = \sigma^l \cdot \tilde{R}'$ and it follows that $R = \text{GCD} \{R' \neq 0; d(R') \leq T - 1 \text{ and } B_0 \subset B(R') \in \mathbb{B}_T^*\} = \text{GCD} \{\tilde{R}' \neq 0; d(\tilde{R}') \leq T - 1 \text{ and } B_0 \subset B(\tilde{R}') \in \tilde{\mathbb{B}}_T^*\} = \tilde{R}$ and hence generically $P_T^* w = \tilde{P}_T^* w$.

(ii) Let $B_0 \in \mathbb{B}_T$, $B_0 \in \tilde{\mathbb{B}}_T$, so $l(R_0) \geq 1$. If $d(R') \leq T - 1$ and $B_0 \subset B(R') \in \mathbb{B}_T^*$ then there exists $F' R' = F' R_0$ and hence $l(R') \geq 1$, so $B(R') \notin \tilde{\mathbb{B}}_T^*$. This implies that generically on B_0 , $\tilde{L}(w) = \emptyset$, as according to Corollary 8.1 generically on B_0 , $\{R \in \tilde{L}(w) \subset L(w)\} \Rightarrow \{B_0 \subset B(R) \in \tilde{\mathbb{B}}_T^*\}$. So generically on B_0 , $\tilde{R}(w) = 0$ and $\tilde{P}_T^* w = \mathbb{R}^T$. \square

REFERENCES

- [1] S. L. CAMPBELL AND C. D. MEIJER, *Generalized Inverses of Linear Transformations*, Pitman, London, 1979.
- [2] R. E. KALMAN, *On minimal partial realizations of a linear input/output map*, in Aspects of network and systems theory, R. E. Kalman and N. De Claris, eds., Holt, Rinehart and Winston, New York, 1971.
- [3] J. W. NIEUWENHUIS AND J. C. WILLEMS, *Deterministic ARMA models*, in Proc. of the Seventh International Conference on Analysis and Optimization of Systems, Antibes, France, 1986; Lecture Notes in Control and Information Sci., Springer, Berlin-New York, 1986.
- [4] J. RISSANEN, *Stochastic complexity and modeling*, Ann. Statist., 14 (1986), pp. 1080–1100.
- [5] J. C. WILLEMS, *From time series to linear system. Part I: Finite dimensional linear time invariant systems. Part II: Exact modeling. Part III: Approximate modeling*, Automatica, 22 (1986), pp. 561–580; 22 (1986), pp. 675–694.

ERGODIC CONTROL OF MULTIDIMENSIONAL DIFFUSIONS I: THE EXISTENCE RESULTS*

VIVEK S. BORKAR^{†‡} AND MRINAL K. GHOSH[†]

Abstract. The existence of optimal stable Markov relaxed controls for the ergodic control of multidimensional diffusions is established by direct probabilistic methods based on a characterization of a.s. limit sets of empirical measures. The optimality of the above is established in the strong (i.e., almost sure) sense among all admissible controls under very general conditions.

Key words. ergodic control, Markov controls, optimal controls, empirical measures, invariant probability measures

AMS(MOS) subject classification. 93E20

1. Introduction. The “ergodic” or “long run average cost” control problem for multidimensional diffusions is one of the few classical problems of stochastic control that still eludes a completely satisfactory treatment. The problem can be formulated as follows: Let U be a compact metric space called the control set. Let $X(\cdot)$ be an R^n -valued controlled diffusion process on some probability space satisfying the stochastic differential equation

$$(1.1) \quad dX(t) = m(X(t), u(t)) dt + \sigma(X(t)) dW(t), \quad X(0) = X_0,$$

for $t \geq 0$, where

(i) $m(\cdot, \cdot) = [m_1(\cdot, \cdot), \dots, m_n(\cdot, \cdot)]^T: R^n \times U \rightarrow R^n$ is continuous and satisfies for all $x, y \in R^n, u \in U$,

$$\begin{aligned} \|m(x, u) - m(y, u)\| &\leq K \|x - y\|, \\ \|m(x, u)\| &\leq K \end{aligned}$$

for some constant $K > 0$.

(ii) $\sigma(\cdot) = [\sigma_{ij}(\cdot)]: R^n \rightarrow R^{n \times n}$ satisfies for $x, y \in R^n$,

$$\begin{aligned} \|\sigma(x) - \sigma(y)\| &\leq K \|x - y\|, \quad \|\sigma(x)\| \leq K, \\ \|\sigma^T x\|^2 &\geq \lambda \|x\|^2 \quad (\text{uniform ellipticity}) \end{aligned}$$

for some constants $\lambda > 0, K > 0$.

(iii) X_0 is a prescribed random variable.

(iv) $W(\cdot) = [W_1(\cdot), \dots, W_n(\cdot)]^T$ is a standard n -dimensional Wiener process independent of X_0 .

(v) $u(\cdot)$ is a U -valued process with measurable sample paths satisfying the following “nonanticipativity” condition: For $t \geq s \geq y \geq 0$, $W(t) - W(s)$ is independent of $u(y)$.

A process $u(\cdot)$ as above will be called an admissible control. Of special interest is the case when $u(\cdot) = v(X(\cdot))$ for some measurable $v: R^n \rightarrow U$. In this case, (1.1) will have a strong solution [29] implying in particular that $u(\cdot)$ is admissible. $X(\cdot)$

* Received by the editors September 8, 1986; accepted for publication (in revised form) April 2, 1987.

[†] Tata Institute of Fundamental Research, Bangalore Centre, P.O. Box 1234, Bangalore, India 560012.

[‡] The research of this author was supported in part by Army Research Office contract DAAG29-84-K-0005 and Air Force Office of Scientific Research grant 85-0227.

will then be a homogeneous Markov process. Hence we call such a $u(\cdot)$ or, by abuse of terminology, the function v itself, a Markov control. A Markov control will be said to be stable if the corresponding process is positive recurrent and thus has a unique invariant measure. (The uniqueness is ensured by our uniform ellipticity condition. See, e.g., [6], [18] or [28, Chaps. 30–32].) If $u(\cdot) = v(X(\cdot), \cdot)$ for some measurable $v: R^n \times R^+ \rightarrow U$, the corresponding process will also be a Markov process, albeit not a homogeneous one. Call such a $u(\cdot)$, or again, by abuse of terminology, the map v itself, an inhomogeneous Markov control. The admissibility of these once again follows from the existence of strong solutions for the corresponding s.d.e. as in [29].

Let $c: R^n \times U \rightarrow U$ be a continuous function called the cost function. We assume that

$$(1.2) \quad c(\cdot, \cdot) \geq -K$$

for some constant K . In the ergodic control problem, one typically seeks to minimize

$$(1.3) \quad \limsup_{t \rightarrow \infty} \frac{1}{t} \int_0^t E[c(X(s), u(s))] ds$$

or a.s. minimize

$$(1.4) \quad \limsup_{t \rightarrow \infty} \frac{1}{t} \int_0^t c(X(s), u(s)) ds$$

over all admissible controls. An admissible control is said to be optimal in the mean if it minimizes (1.3) and a.s. optimal if it a.s. minimizes (1.4). The primary aims of the ergodic control problem are the following:

- (i) to show the existence of a stable Markov control which is optimal in an appropriate sense (cf. above definitions of optimality), and,
- (ii) to characterize the same via the dynamic programming equation (the “Hamilton–Jacobi–Bellman” equation).

The first attempt in this direction is perhaps [24, Chap. VI] where a one-dimensional compact state space was considered. Subsequent works considered the multidimensional case as well. An extensive survey of these appears in [25]. Here, we shall briefly recall the focus of some recent works. The traditional approach to this problem, inherited from earlier developments in discrete time and discrete state space situations, is to start with the Hamilton–Jacobi–Bellman (H.J.B.) equation and arrive at an existence result for optimal stable Markov control using this equation, the equation itself being approached by a “vanishing discount” limit argument from the corresponding H.J.B. equation for the infinite horizon discounted cost control problem. The most recent development in this direction is [27] where the H.J.B. equation is studied under a condition on the gradient of the cost. Another recent work [12] also focuses on the H.J.B. equation, but treats it as a limiting case of finite horizon problems instead of discounted cost problems on infinite time horizon. The only direct proof of existence of an optimal stable Markov control by probabilistic compactness arguments seems to be [21], which also considers the corresponding maximum principle.

These works share one or more of the following limitations:

- (a) Optimality in the mean and not a.s. optimality is considered.
- (b) Optimality is established only within the class of Markov controls and not with respect to all admissible controls.
- (c) The system model is often more restrictive than the above, e.g., it is sometimes assumed that $\sigma =$ the identity matrix and $m(x, u) = u$.
- (d) Either a blanket stability assumption is imposed or a condition on the cost function which penalizes instability is assumed.

It is clear that some condition on cost or stability must be necessary to give the desired existence of an optimal stable Markov control. For example, consider the case

$$c(x, u) = \exp(-\|x\|^2).$$

Then the cost of any stable Markov control is a.s. positive while that of an unstable Markov control is a.s. zero, making the latter optimal.

In this paper, we extend the approach of [7], [8], [11] to multidimensional diffusions. In the one-dimensional case, this was partially done in [5], [9]. These works, however, use many specificities of the one-dimensional case in a crucial manner. Here we address only the first of the two issues mentioned above, viz., the existence of stable optimal Markov controls, thus subsuming the results of [9]. The second issue, viz., the dynamic programming equations will be treated in a subsequent publication [15]. The advantages of our approach are the following:

(1) Almost sure optimality (as opposed to optimality in the mean) of a stable Markov control is established in the class of all admissible controls.

(2) The approach has a more probabilistic flavor than the previous ones and brings out certain features of the problem (e.g., asymptotics for the empirical measures) not apparent in the latter.

The main disadvantage of our approach is that we have to work with the larger class of relaxed controls. This means that we assume U to be of the form $P(V)$ = the space of probability measures on some compact metric space V with the topology of weak convergence and c, m to be of the form

$$c(x, u) = \int_V \bar{c}(x, y)u(dy), \quad m_i(x, u) = \int_V \bar{m}_i(x, y)u(dy), \quad 1 \leq i \leq n$$

for some $\bar{c}: R^n \times V \rightarrow R$ and $\bar{m}: R^n \times V \rightarrow R^n$, $\bar{m}(\cdot, \cdot) = [\bar{m}_1(\cdot, \cdot), \dots, \bar{m}_n(\cdot, \cdot)]$, which satisfy the same hypotheses as c, m , respectively, but with V replacing U . Note that any V -valued process $v(\cdot)$ can be identified with a U -valued process $u(\cdot)$ defined by $u(t)$ = the Dirac measure at $v(t)$ for $t \geq 0$. Thus relaxed controls subsume controls in the ordinary sense. In fact, if c has no explicit control dependence and $m(x, U)$ is convex for each x , each relaxed control can be identified with a control in the ordinary sense by a straightforward application of the selection theorem in Lemma 1.1 [3], as was pointed out in [9]. In [5], it was shown in the one-dimensional case that the dynamic programming equations allow us to do away with the relaxed control framework. Analogous development in the multidimensional case will be reported in [15].

The use of relaxed controls is tantamount to compactifying the space of control trajectories in a certain precise sense. A nice exposition of this can be found in [2, § 1.9, pp. 31–36]. The concept of relaxed controls was first introduced in deterministic control theory in [31]. Its use in stochastic control dates back to [14].

For a stable Markov control v , we shall denote by η_v the corresponding unique invariant probability measure for $X(\cdot)$. We assume throughout this paper that at least one stable Markov control v exists such that

$$\int c(x, v(x))\eta_v(dx) \leq \infty.$$

Thus

$$(1.5) \quad \alpha = \inf_{v \text{ stable Markov}} \int c(x, v(x))\eta_v(dx)$$

is well defined. We shall prove our existence result under two sets of assumptions. In the first one, we assume that c is near-monotone in the sense that it satisfies

$$(1.6) \quad \liminf_{\|x\| \rightarrow \infty} \inf_{u \in U} c(x, u) > \alpha.$$

The terminology is suggested by the fact that (1.5) is always satisfied when $c(x, u) = k(\|x\|)$ for a monotone increasing $k: R^+ \rightarrow R$. Such costs discourage unstable behavior for obvious reasons and arise often in practice.

The second case we shall consider is a Lyapunov-type stability condition the details of which are left to § 3. For the time being, we only mention that in particular it implies the stability of all Markov controls.

The plan of the paper is as follows: Section 2 establishes a characterization of a.s. limit sets for empirical measures of the joint state and control process along the lines of [9]. This leads to the existence result in the near-monotone case. Section 3 gives a full statement of the Lyapunov condition mentioned above and uses it to prove certain moment bounds for a class of stopping times to be defined later, which in turn implies that all Markov controls are stable and the set of their invariant probability measures is compact in $P(R^n)$. ($P(S)$ will always denote the space of probability measures on a Polish space S with the topology of weak convergence.) Section 4 proves the existence of an optimal stable Markov controls under the conditions of § 3.

2. Existence in the near-monotone case. The key result of this section is Lemma 2.2, which characterizes the a.s. limit sets of the process of empirical measures we are about to define. This immediately leads to the desired existence result for a near-monotone cost (Theorem 2.1).

Let $\bar{R}^n = R^n \cup \{\infty\}$ be the one point compactification of R^n and let $H = \{A \times B | A, B \text{ Borel subsets of } \bar{R}^n, V\}$ respectively. For $t \geq 0$, define the empirical measure $\tilde{\nu}_t$ on H by

$$\tilde{\nu}_t(A \times B) = \frac{1}{t} \int_0^t I\{X(s) \in A\} u(s, B) ds$$

for $X(\cdot)$, $u(\cdot)$ as in (1.1), with

$$u(s, B) = \int_B du(s), \quad B \subset V.$$

For each fixed sample point and fixed t , $\tilde{\nu}_t$ extends uniquely to a $\nu_t \in P(\bar{R}^n \times V)$. This defines the process of empirical measures ν_t , $t \geq 0$, taking values in $P(\bar{R}^n \times V)$. Since the latter is a compact space (because $\bar{R}^n \times V$ is compact), $\{\nu_t\}$ converges to a sample point dependent compact subset of $P(\bar{R}^n \times V)$ as $t \rightarrow \infty$.

Each $\eta \in P(\bar{R}^n \times V)$ can be decomposed as

$$(2.1) \quad \eta(A) = \delta(\eta) \eta'(A \cap (R^n \times V)) + (1 - \delta(\eta)) \eta''(A \cap (\{\infty\} \times V))$$

for A Borel in $\bar{R}^n \times V$, where $\delta(\eta) \in [0, 1]$, $\eta' \in P(R^n \times V)$ and $\eta'' \in P(\{\infty\} \times V)$. This decomposition can be rendered unique by imposing a fixed choice of $\eta' \in P(R^n \times V)$ (respectively, $\eta'' \in P(\{\infty\} \times V)$) when $\delta(\eta) = 0$ (respectively, 1). Disintegrate η' as follows:

$$(2.2) \quad \int_{R^n \times V} f(x, y) \eta'(dx, dy) = \int_{R^n} \int_V f(x, y) v_\eta(x, dy) \eta^*(dx)$$

for all bounded continuous $f: R^n \times V \rightarrow R$, where η^* is the image of η' under the projection $R^n \times V \rightarrow R^n$ and $v_\eta(x, \cdot) \in U$ for $x \in R^n$ is the regular conditional law. Then the map $x \rightarrow v_\eta(x, \cdot): R^n \rightarrow U$ can be identified with a Markov control which we also

denote by v_η (i.e., $v_\eta(x) \in U$ is defined by $v_\eta(x) = v_\eta(x, \cdot)$, the right-hand side defined as above). Note that this v_η is defined only η^* -a.s. We pick any one representative of this a.s.-equivalence class. Throughout this paper, this choice of a representative is immaterial wherever the above decomposition is used.

Thus we have associated with $\eta \in P(\bar{R}^n \times V)$ the objects $\delta(\eta) \in [0, 1]$, $\eta' \in P(R^n \times V)$, $\eta'' \in P(\{\infty\} \times V)$, $\eta^* \in P(R^n)$, $v_\eta: R^n \rightarrow U$ a Markov control. If in addition $v = v_\eta$ is stable, we also have its unique invariant probability measure η_v . This notation plays an important role in what follows.

Let C_0^2 = the Banach space of twice continuously differentiable maps $R^n \rightarrow R$ which, along with their first and second partial derivatives vanish at infinity, with the norm

$$\|f\| = \sup_x |f(x)| + \sum_{i=1}^n \sup_x \left| \frac{\partial f}{\partial x_i}(x) \right| + \sum_{i,j=1}^n \sup_x \left| \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right|.$$

For any $f \in C_0^2$, let

$$(Lf)(x, u) = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x) \bar{m}_i(x, u) + \frac{1}{2} \sum_{i,j,k=1}^n \sigma_{ik}(x) \sigma_{jk}(x) \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$$

and for any Markov control v ,

$$(L_v f)(x) = \int_V (Lf)(x, y) v(x, dy)$$

where the meaning of the right-hand side is obvious.

Let G be a countable dense subset of C_0^2 . Then G is also countable dense in $C_0 = \{f \in C(R^n) \mid \lim_{\|x\| \rightarrow \infty} f(x) = 0\}$ with supremum norm. In particular, this implies that it is a convergence determining class and hence a separating class for $P(R^n)$ (i.e., $\int f d\mu_n \rightarrow \int f d\mu_\infty$ for $f \in G$, $\{\mu_n, n = 1, 2, \dots, \infty\} \subset P(R^n)$, implies $\mu_n \rightarrow \mu_\infty$ in $P(R^n)$ and $\int f d\mu = \int f d\nu$ for $f \in G$, $\mu, \nu \in P(R^n)$ implies $\mu = \nu$).

LEMMA 2.1. *If $\nu \in P(R^n)$ satisfies*

$$(2.3) \quad \int L_v f d\nu = 0 \quad \text{for } f \in G$$

for some Markov control v , then $\nu = \eta_v$. (Recall that η_v is the unique invariant probability measure under v , whose stability is thus a part of the conclusion.)

Proof. By a standard approximation argument, (2.3) holds for all $f \in C_b^2$ = the space of twice continuously differentiable functions with bounded first and second derivatives. Let $X(\cdot)$ be a diffusion governed by the Markov control v and with initial law ν . Let ν_t = the law of $X(t)$ at time t , $t \geq 0$. Then $\{\nu_t\}$ satisfies the forward Kolmogorov equation

$$\int f d\nu_t = \int f d\nu + \int_0^t \int (L_v f) d\nu_s ds, \quad f \in C_b^2.$$

Since the solution of this p.d.e. is unique and $\nu_t \equiv \nu$ is a solution by virtue of (2.3), the claim follows. Q.E.D.

LEMMA 2.2. *Outside a set of zero probability, each limit point ν of $\{\nu_t\}$ for which $\delta(\nu) > 0$, satisfies*

$$(2.4) \quad \nu^* = \eta_{v_\eta}.$$

Remarks. Note that we do not claim pathwise tightness of $\{\nu_t\}$, which would correspond to $\delta(\nu) = 1$ a.s. This cannot be true in general, e.g., for an unstable Markov

control. Thus we must allow for the possibility $\delta(\nu) < 1$, which necessitates the compactification of the state space as done above.

Proof. For $f \in G$, Ito's formula gives

$$(2.5) \quad \begin{aligned} f(X(t)) - f(X(0)) &= \int_0^t \int_V Lf(X(s), y) u(s, dy) ds \\ &\quad + \int_0^t \langle \nabla f(X(s)), \sigma(X(s)) dW(s) \rangle. \end{aligned}$$

By standard time change arguments (see, e.g., [13, § 6.1] or [17, §§ 3.1, 4.4]), the stochastic integral term above can be shown to be of the form $B(\tau_t)$ for a standard Brownian motion $B(\cdot)$ and a process of time change τ_t satisfying

$$\limsup_{t \rightarrow \infty} \frac{\tau_t}{t} < \infty \quad \text{a.s.}$$

Since

$$\lim_{t \rightarrow \infty} \frac{B(\tau_t)}{\tau_t} = 0 \quad \text{a.s. on } \{\lim_{t \rightarrow \infty} \tau_t = \infty\} \text{ and } < \infty \text{ a.s. on } \{\lim_{t \rightarrow \infty} \tau_t < \infty\},$$

we have

$$\lim_{t \rightarrow \infty} \frac{B(\tau_t)}{t} = 0 \quad \text{a.s.}$$

Hence

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \int_V Lf(X(s), y) u(s, dy) ds = \lim_{t \rightarrow \infty} \int Lf d\nu_t = 0 \quad \text{a.s.}$$

Since G is countable, we can find a set N of zero probability outside which the above limit holds for all $f \in G$. Then outside N , each limit point ν of $\{\nu_t\}$ with $\delta(\nu) > 0$ must satisfy

$$\int Lf d\nu = 0 \quad \text{for } f \in G.$$

The claim follows from Lemma 2.1. Q.E.D.

LEMMA 2.3. *Under a stable Markov control v ,*

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t c(X(s), v(X(s))) ds = \int c(x, v(x)) \eta_v(dx).$$

See [6] for a proof using the ergodic theorem.

LEMMA 2.4. *For a near-monotone c , there exists a stable Markov control v such that*

$$\int c(x, v(x)) \eta_v(dx) = \alpha.$$

Proof. Let $\{v_n\}$ be a sequence of stable Markov controls such that

$$\int c(x, v_n(x)) \eta_{v_n}(dx) \downarrow \alpha.$$

Define $p_n \in P(\bar{R}^n \times V)$ by

$$\int_{\bar{R}^n \times V} f(x, y) p_n(dx, dy) = \int_{R^n} \int_V f(x, y) v_n(x, dy) \eta_{v_n}(dx)$$

for bounded continuous $f: \bar{R}^n \times V \rightarrow R$. Let p_∞ be a limit point of $\{p_n\}$ and let

$$v_\infty = v_{p_\infty}.$$

For $f \in G$, we have

$$\int L_{v_n} f d\eta_{v_n} = \int Lf dp_n = 0, \quad n = 1, 2, \dots$$

Letting $n \rightarrow \infty$ along an appropriate subsequence,

$$\int Lf dp_\infty = 0.$$

By Lemma 2.1 and the decomposition (2.1),

$$p_\infty^* = \eta_{v_\infty} \quad \text{if } \delta(p_\infty) > 0.$$

Now, the near-monotonicity of c implies that for some $\varepsilon > 0$,

$$\liminf_{\|x\| \rightarrow \infty} \inf_{u \in V} \bar{c}(x, u) > \alpha + \varepsilon.$$

Using this, we can construct continuous maps $c^m: \bar{R}^n \times V \rightarrow R$, $m \geq 1$, such that

$$\begin{aligned} c^m(\infty, u) &= \alpha + \varepsilon, \quad m \geq 1, \\ c^m(x, u) &\uparrow \bar{c}(x, u) \quad \text{on } R^n \times V. \end{aligned}$$

Thus

$$\int \bar{c} dp'_n \geq \int c^m dp_n.$$

Letting $n \rightarrow \infty$,

$$\lim \int \bar{c} dp'_n = \alpha \geq \left(\int c^m dp'_\infty \right) \delta(p_\infty) + (1 - \delta(p_\infty))(\alpha + \varepsilon).$$

Letting $m \rightarrow \infty$ on the right-hand side,

$$\alpha \geq \left(\int \bar{c} dp'_\infty \right) \delta(p_\infty) + (1 - \delta(p_\infty))(\alpha + \varepsilon).$$

If $\delta(p_\infty) > 0$,

$$\int \bar{c} dp'_\infty = \int c(x, v_\infty(x)) \eta_{v_\infty}(dx) \geq \alpha$$

by the definition of α . Hence we must have $\delta(p_\infty) = 1$ and

$$\int \bar{c} dp'_\infty = \int c(x, v_\infty(x)) \eta_{v_\infty}(dx) = \alpha. \quad \text{Q.E.D.}$$

As we remarked earlier, v_∞ is defined p_∞^* -a.s. and it does not matter which representative we pick.

THEOREM 2.1. *For a near-monotone c , there exists a stable a.s. optimal Markov control.*

Proof. Using Lemma 2.2 and arguments similar to those employed in the proof of the above lemma, we can show that

$$\liminf_{t \rightarrow \infty} \frac{1}{t} \int_0^t c(X(s), u(s)) ds \geq \alpha \quad \text{a.s.}$$

The claim now follows from Lemmas 2.3 and 2.4. Q.E.D.

3. Tightness of invariant probability measures. In this and the next section, we study the situation where the near-monotonicity condition on the cost is dropped, but instead we impose a Lyapunov-type stability condition which, among other things, will be shown to imply that all the Markov controls are stable and their invariant probability measures form a compact set in $P(R^n)$. This, in fact, is the principal result of this section (Theorem 3.1, Corollary 3.2), the proof of the existence of an a.s. optimal Markov control being left to § 4.

Before we give a precise statement of this condition, we mention the following technical lemma.

LEMMA 3.1. *Let $X_0 = x \in R^n$, $t > 0$, $u(\cdot)$ be an admissible control. Then the law of $X(t)$ has a density $p(t, x, \cdot)$ with respect to the Lebesgue measure on R^n , satisfying*

$$(3.1) \quad c_1 t^{-n/2} \exp(-c_2 \|x - y\|^2/t) \leq p(t, x, y) \leq c_3 t^{-n/2} \exp(-c_4 \|x - y\|^2/t)$$

for some constants $c_i > 0$, $i = 1, 2, 3, 4$, independent of $x, t, u(\cdot)$, $0 < t \leq T$.

Proof. If $u(\cdot)$ is an inhomogeneous Markov control, this is precisely the estimate of [1]. For arbitrary $u(\cdot)$, the law of $X(t)$ is the same as that under some inhomogeneous Markov control by the results of [10] and we are done. Q.E.D.

The Lyapunov-type condition we use is the following:

Assumption A. There exists a twice continuously differentiable function $w: R^n \rightarrow R$ satisfying:

$$(3.2) \quad (i) \quad \lim_{\|x\| \rightarrow \infty} w(x) = +\infty \text{ uniformly in } \|x\|,$$

$$(ii) \quad \text{there exist } a > 0, \varepsilon_0 > 0 \text{ such that whenever } \|x\| > a,$$

$$(3.3) \quad Lw(x, u) < -\varepsilon_0 \quad \text{for all } u \in U,$$

$$(3.4) \quad \|\nabla w\|^2 > \varepsilon_0,$$

$$(3.5) \quad (iii) \quad \int_0^T \int_{R^n} \|\sigma^T(x) \nabla w(x)\|^2 t^{-n/2} \exp(-c_4 \|x - y\|^2/t) dx dt < \infty \quad \forall T > 0,$$

where c_4 is as in Lemma 3.1.

Remarks. (a) Equation (3.5) is a mild technical condition that ensures (by virtue of Lemma 3.1) that the stochastic integral

$$\int_0^T \langle \nabla w(X(t)), \sigma(X(t)) dW(t) \rangle, \quad T > 0,$$

is always well defined.

(b) We have chosen the above formulation of a Lyapunov-type condition because it is easily stated and still quite general. Other variants are possible (see, e.g., [21] for one). For the general theory of stochastic Lyapunov functions, see [19]. The key consequence of the above assumption for our purposes is Lemma 3.2 below. Thus any condition that implies Lemma 3.2 will suffice. In fact, the crudeness of estimates used in proving the lemma shows that there is ample scope for improvement.

(c) As an example, consider $n = 1$, $\sigma(\cdot) \equiv 1$, $m(x, u) \leq -\varepsilon$ for x sufficiently large and $\geq \varepsilon$ for $-x$ sufficiently large for some $\varepsilon > 0$. Then $w(x) = x^2$ will do the job.

Let $B_1, B_2 \subset \mathbb{R}^n$ be concentric balls centered at zero with radii r_1, r_2 and boundaries $\delta B_1, \delta B_2$, respectively, where we choose $r_2 > r_1 > a$ such that for some $a_1 > 0$, $\{x \mid w(x) \leq a_1\}$ is nonempty and contained in B_1 . Let $a_2 = \max_{x \in \delta B_2} |w(x)|$ and $a_3 = a_1 - a_2$.

LEMMA 3.2. *Let $X_0 = x \in \delta B_2$ and $\tau = \inf \{t \geq 0 \mid X(t) \in \delta B_1\}$. Then*

$$(3.6) \quad \sup E[\tau^2] < \infty$$

where the supremum is over all $x \in \delta B_2$ and all admissible $u(\cdot)$.

Proof. For $t > 0$,

$$\begin{aligned} P(\tau \geq t) &= P\left(\min_{s \in [0, t]} w(X(s)) \geq a_1, \tau \geq t\right) \\ &\leq P\left(\min_{y \in [0, t]} \int_0^y \langle \nabla w(X(s)), \sigma(X(s)) dW(s) \rangle \geq a_3 + \varepsilon_0 t\right) \end{aligned}$$

by (3.3). Using the random time change argument we used earlier,

$$\int_0^t \langle \nabla w(X(s)), \sigma(X(s)) dW(s) \rangle = B(\xi(t))$$

for a standard Brownian motion $B(\cdot)$ with

$$\xi(t) = \int_0^t \|\sigma^T(X(s)) \nabla w(X(s))\|^2 ds \geq \lambda \varepsilon_0 t.$$

(Recall that λ is the ellipticity constant for $\sigma \sigma^T$.) Thus

$$\begin{aligned} P(\tau \geq t) &\leq 2P(B(\lambda \varepsilon_0 t) \geq \varepsilon_0 t + a_3) \\ &= 2(2\pi\lambda\varepsilon_0 t)^{-1/2} \int_{a_3 + \varepsilon_0 t}^{\infty} \exp(-y^2/2\lambda\varepsilon_0 t) dy. \end{aligned}$$

It is not hard to verify from this that

$$\int_0^{\infty} tP(\tau \geq t) dt < K < \infty$$

where the constant K is independent of the choice of x in δB_2 and of $u(\cdot)$. The claim follows. Q.E.D.

Now take $X_0 = x \in \bar{B}_2$ and define $\tau' = \inf \{t \geq 0 \mid X(t) \notin \delta B_2\}$. We have the following companion result to the above, which, however, does not need Assumption A.

LEMMA 3.3.

$$(3.7) \quad \sup E[(\tau')^2] < \infty$$

where the supremum is over $x \in \bar{B}_2$ and admissible $u(\cdot)$.

In order to prove this result, we need another technical lemma, Lemma 3.4 below, which will also be useful elsewhere in this paper. Let $\{F_t\}$ denote the natural filtration of $X(\cdot)$.

LEMMA 3.4. *For any $\{F_t\}$ -stopping time τ , the regular conditional law of $X(\tau + \cdot)$ given F_τ on $\{\tau < \infty\}$ is a.s. the law of a controlled diffusion of the same type as (1.1).*

Proof. The results of [30] (see Theorem 4.3 and the final comments on p. 632) allow us to assume without any loss of generality that $\{F_t\}$ is the canonical filtration on $C([0, \infty); R^n)$ and $u(\cdot)$ is of the form

$$u(t) = G(t, X(\cdot))$$

for some measurable $G: [0, \infty) \times C([0, \infty); R^n) \rightarrow U$ which is progressively measurable with respect to $\{F_t\}$. By Lemma 1.3.3 of [26, p. 33], a version of the regular conditional law of $X(\tau + \cdot)$ given F_τ on $\{\tau < \infty\}$ will be a.s. given by the law of a controlled diffusion $X(\cdot)$ as in (1.1), but with initial condition $X(\tau)$ and control $\tilde{u}(\cdot)$ given by $\tilde{u}(t) = G(\tau + t, X(\cdot))$ with τ and the restriction of $X(\cdot)$ to $[0, \tau]$ being held fixed as parameters. Q.E.D.

From here on, $M_i(S)$, $S \subset R^n$, $i = 1, 2$, will denote the set of $X(\cdot)$ as in (1.1) under Markov/arbitrary admissible controls, respectively, with initial law supported in S .

Proof of Lemma 3.3. By the results of [10], the law of $X(t)$ for any $t > 0$ coincides with that under some inhomogeneous Markov control and thus by the uniform ellipticity assumption on $\sigma\sigma^T$, is absolutely continuous with respect to the Lebesgue measure. (Recall (3.1).) Let $X(\cdot) \in M_2(\bar{B}_2)$, $\tau = \inf\{t \geq 0 | X(t) \in \bar{B}_2\}$. Then for $t > 0$,

$$P(\tau = t) \leq P(X(t) \in \delta B_2) = 0$$

and thus $P(\tau = t) = 0$. Fix $t > 0$. Let $\{X^n(\cdot)\}$ be a sequence in $M_2(\bar{B}_2)$ such that if $\{\tau^n\}$ denote the corresponding first exit times from \bar{B}_2 ,

$$P(\tau^n > t) \uparrow \sup_{X(\cdot) \in M_2(\bar{B}_2)} P(\tau > t).$$

As in the proof of Theorem 3.1 of [20], one can argue that $X^n(\cdot) \rightarrow X^\infty(\cdot)$ in law along a subsequence (denoted $\{n\}$ again by abuse of notation) where $X^\infty(\cdot) \in M_2(\bar{B}_2)$. (The only difference with Theorem 3.1 of [20] is the varying initial law. This can, however, be easily accommodated since the initial laws are supported on \bar{B}_2 and hence are tight.) By Skorokhod's theorem [17, p. 9], we may assume that this convergence is a.s. on a common probability space. (See [20] for an analogous argument.) Let $\tau^\infty = \inf\{t \geq 0 | X^\infty(t) \notin \bar{B}_2\}$ and $\bar{\tau} = \inf\{t \geq 0 | X^\infty(t) \in \delta B_2\}$. Path continuity of $\{X^n(\cdot)\}$ and simple geometric considerations show that for any sample point, any limit point of $\{\tau^n\}$ in $[0, \infty]$ must lie in $[\bar{\tau}, \tau^\infty]$. By our uniform ellipticity condition on $\sigma\sigma^T$, $\bar{\tau} = \tau^\infty$ a.s. Thus $\tau^n \rightarrow \tau^\infty$ a.s. Since $P(\tau^\infty = t) = 0$, $P(\tau^n > t) \rightarrow P(\tau^\infty > t)$. Since

$$P(\tau^\infty > t) \leq P(X^\infty(t) \in \bar{B}_2) < 1,$$

we have

$$\beta = \sup_{X(\cdot) \in M_2(\bar{B}_2)} P(\tau > t) < 1.$$

Hence for $X(\cdot) \in M_2(\delta B_1) \subset M_2(\bar{B}_2)$,

$$\begin{aligned} P(\tau > nt) &= E[I\{\tau > nt\}] \\ &= E[E[I\{\tau > nt\}]/F_{(n-1)t}]I\{\tau > (n-1)t\} \\ &\leq \beta E[I\{\tau > (n-1)t\}] \end{aligned}$$

by Lemma 3.4. Iterating the argument,

$$P(\tau > nt) \leq \beta^n.$$

The rest is easy. Q.E.D.

Define the extended real-valued stopping times

$$(3.8) \quad \tau_1 = \inf \{t \geq 0 \mid X(t) \in \delta B_1\},$$

$$(3.9) \quad \xi_n = \inf \{t \geq \tau_n \mid X(t) \in \delta B_2\},$$

$$(3.10) \quad \tau_{n+1} = \inf \{t \geq \xi_n \mid X(t) \in \delta B_1\}$$

for $n = 1, 2, \dots$, where as usual the quantity on the left is set equal to $+\infty$ if the set on the right is empty.

Let v be a Markov control and $X(\cdot)$ the corresponding process with initial law supported on δB_1 . By the above three lemmas, $E[\xi_i], E[\tau_i] < \infty$ for all i with $\tau_1 = 0$. Then $X(\tau_i), i = 1, 2, \dots$, is a δB_1 -valued Markov chain having a unique invariant probability measure (say, q) as argued in [18].

COROLLARY 3.1. *The measure $\eta \in P(R^n)$ defined by*

$$\int f d\eta = E \left[\int_0^{\tau_2} f(X(t)) dt \right] / E[\tau_2], \quad f \in C_b(R^n),$$

with the law of $X(0) = q$, coincides with η_v . (In particular, v is stable.)

For a proof, see [18].

Let $\{v_n\}$ be a sequence of Markov controls and $X^n(\cdot)$ the corresponding diffusions as in (1.1) for some initial laws and suppose that $X^n(\cdot) \rightarrow X^\infty(\cdot)$ in law for some process $X^\infty(\cdot)$.

LEMMA 3.5. *$X^\infty(\cdot)$ is a diffusion satisfying (1.1) for some Markov control.*

Proof. Let $T_{s,t}^n, t \geq s$, denote the transition semigroup for $X^n(\cdot)$, $n \geq 1$. Let $f \in C^2(R^n)$ with compact support and $g \in C_b(R^n \times R^n \times \dots \times R^n$ (m times)) for same $m \geq 1$. Then for $t \geq s \geq t_m \geq t_{m-1} \geq \dots \geq t_1 \geq 0$, $E[(f(X^n(t)) - T_{s,t}^n f(X^n(s)))g(X^n(t_1), \dots, X^n(t_m))] = 0$, $n = 1, 2, \dots$. For each n , $T_{s,t}^n f(\cdot)$ satisfies the appropriate backward Kolmogorov equation. From standard p.d.e. theory (see [22, Chap. III] or [32, pp. 133–134]), it follows that $T_{s,t}^n f(\cdot)$, $n = 1, 2, \dots$, are equicontinuous. Since they are clearly bounded, they form a sequentially precompact set in $C(R^n)$ with the topology of uniform convergence on compacts. Let $T_{s,t} f(\cdot)$ be a limit point of the same in $C(R^n)$. Passing to the limit in the above as $n \rightarrow \infty$, it is easily seen (e.g., using Skorokhod's theorem) that $E[(f(X^\infty(t)) - T_{s,t} f(X^\infty(s)))g(X^\infty(t_1), \dots, X^\infty(t_m))] = 0$. Since $f, g, \{t_i\}$ were arbitrary, a standard argument using the monotone class theorem establishes the Markov property of $X^\infty(\cdot)$. By Theorem 3.1 of [20], $X^\infty(\cdot)$ satisfies (1.1) for some $u(\cdot)$. Argue as in [15, pp. 184–5], to conclude that $u(\cdot)$ must be of the form $u(\cdot) = v(X^\infty(\cdot), \cdot)$ for some measurable map $v: R^n \times R^+ \rightarrow U$. Since $T_{s,t}^n f$ depends on t, s only through $t - s$ for each f and $n = 1, 2, \dots$, the same must be true for $T_{s,t} f$ in view of the above limiting argument. It follows that $X^\infty(\cdot)$ is a time-homogeneous Markov process and hence $u(\cdot)$ is in fact a Markov control. Q.E.D.

THEOREM 3.1. *The set $\{\eta_v \mid v \text{ Markov control}\}$ is compact in $P(R^n)$.*

Proof. Let $\{v_n\}$ be a sequence of Markov controls and $X^n(\cdot)$ the corresponding diffusions whose initial laws will soon be specified. Define $\{\tau_i^n\}, \{\xi_i^n\}$ as in (3.8)–(3.10) correspondingly. Let q^n be the unique invariant probability measure for the chain $\{X^n(\tau_i^n)\}$. Set the law of $X^n(0)$ equal to q^n for each $n = 1, 2, \dots$. Argue as in the proof of Theorem 3.1 of [19] to conclude that $X^n(\cdot) \rightarrow X^\infty(\cdot)$ in law along a subsequence, denoted $\{n\}$ again by abuse of notation. By Lemma 3.5, $X^\infty(\cdot)$ satisfies (1.1) for some Markov control v_∞ . Invoke Skorokhod's theorem as before to assume that the above convergence is a.s. on a common probability space. Define $\{\tau_i^\infty\}, \{\xi_i^\infty\}$ as in (3.8)–(3.10) for $X^\infty(\cdot)$. By arguments similar to those used to prove $\tau^n \rightarrow \tau^\infty$ a.s. in the

proof of Lemma 3.3, we can inductively prove that

$$(3.11) \quad \tau_i^n \rightarrow \tau_i^\infty \text{ a.s.}, \quad \xi_i^n \rightarrow \xi_i^\infty \text{ a.s.} \quad \text{for all } i.$$

Thus

$$(3.12) \quad \begin{aligned} X^n(\tau_i^n) &\rightarrow X^\infty(\tau_i^\infty) \text{ a.s.}, \\ \int_{\tau_i^n}^{\tau_{i+1}^n} f(X^n(s)) ds &\rightarrow \int_{\tau_i^\infty}^{\tau_{i+1}^\infty} f(X^\infty(s)) ds \quad \text{a.s. for all } i \end{aligned}$$

where $f \in C_b(R^n)$. By Lemmas 3.2 and 3.3,

$$(3.13) \quad \sup_n E[(\tau_2^n)^2] < \infty$$

and hence $\{\tau_2^n, n \geq 1\}$ are uniformly integrable. Thus

$$(3.14) \quad \begin{aligned} E[\tau_2^n] &\rightarrow E[\tau_2^\infty], \\ E\left[\int_0^{\tau_2^n} f(X^n(s)) ds\right] &\rightarrow E\left[\int_0^{\tau_2^\infty} f(X^\infty(s)) ds\right], \quad f \in C_b(R^n) \end{aligned}$$

by (3.11), (3.12). By Corollary 3.1,

$$(3.15) \quad \begin{aligned} \int f d\eta_{v_n} &= \frac{E[\int_0^{\tau_2^n} f(X^n(s)) ds]}{E[\tau_2^n]} \\ &\rightarrow \frac{E[\int_0^{\tau_2^\infty} f(X^\infty(s)) ds]}{E[\tau_2^\infty]}. \end{aligned}$$

Since $\tau_1^n = 0$ a.s. $n = 1, 2, \dots$, $\tau_1^\infty = 0$ a.s. Since for each $n = 1, 2, \dots$, $\{X^n(\tau_i), i = 1, 2, \dots\}$ are identically distributed, it follows that $\{X^\infty(\tau_i), i = 1, 2, \dots\}$ are identically distributed. Thus the initial law of $X^\infty(\cdot)$ equals the unique invariant probability measure for the chain $\{X^\infty(\tau_i), i = 1, 2, \dots\}$. Hence by Corollary 3.1, the right-hand side of (3.15) equals $\int f d\eta_{v_\infty}$. Thus $\eta_{v_n} \rightarrow \eta_{v_\infty}$ in $P(R^n)$. The claim follows. Q.E.D.

COROLLARY 3.2. *There exists a Markov control v such that*

$$\int c(x, v(s)) \eta_v(dx) = \alpha.$$

Proof. Pick $\{v_n\}$ above so that

$$\int c(x, v_n(x)) \eta_{v_n}(dx) \downarrow \alpha.$$

Define $p_n \in P(R^n \times V)$, $n = 1, 2, \dots$, by

$$\int f(x, y) p_n(dx, dy) = \int \int f(x, y) v_n(x, dy) \eta_{v_n}(dx), \quad f \in C_b(R^n \times V).$$

Since V is compact, the above theorem implies that $\{p_n\}$ is tight in $P(R^n \times V)$ and hence converges along a subsequence (denoted n again) to some $p_\infty \in P(R^n \times V)$. Argue as in the proof of Lemma 2.4 to conclude that p_∞ is of the form

$$p_\infty(dx, dy) = \eta_v(dx) v(x, dy)$$

for some Markov control v . Then

$$\int c(x, v(x)) \eta_v(dx) = \alpha$$

follows from Fatou's lemma and the definition of α . Q.E.D.

4. Existence of an optimal Markov control under Assumption A. In this section, we shall show that the Markov control in the statement of Corollary 3.2 is a.s. optimal. Before we get down to the main result (Theorem 4.1), we shall collect together a few minor consequences of the foregoing that will be used later.

LEMMA 4.1. $\{E[\tau_2]|X(\cdot) \in M_1(\delta B_1)\}$ is bounded from above and bounded away from zero from below.

Proof. The upper bound follows from Lemmas 3.2–3.4 in an obvious manner. An argument similar to that leading to (3.14) can be employed to show the rest. Q.E.D.

LEMMA 4.2. The set of probability measures η defined by

$$\int_{R^n} f d\eta = E \left[\int_0^{\tau_2} f(X(t)) dt \right] / E[\tau_2], \quad f \in C_b(R^n),$$

for $X(\cdot) \in M_1(\delta B_1)$ is tight in $P(R^n)$.

Proof. This can be proved the same way as Theorem 3.1 by showing that each sequence has a subsequence that converges in $P(R^n)$. Q.E.D.

Let $\{f_n\}$ be a collection of smooth maps $R^n \rightarrow [0, 1]$ such that $f_n(x) = 0$ for $\|x\| \leq n$ and $=1$ for $\|x\| \geq n+1$.

LEMMA 4.3. For any $\varepsilon \geq 0$, there exists $N_\varepsilon \geq 1$ such that for all $n \geq N_\varepsilon$ and $X(\cdot) \in M_1(\delta B_1)$,

$$E \left[\int_0^{\tau_2} f_n(X(s)) ds \right] < \varepsilon.$$

Proof. Let $Y(\cdot) = X(\xi_1 + \cdot)$. Let β_m denote the first exit time from $\{x | \|x\| \leq m\} \setminus B_1$ where m is any integer sufficiently large so that $\|x\| < m$ for $x \in B_2$. (We do not specify for which process, leaving that to depend on the context for economy of notation.) Consider the control problem for $\bar{X}(\cdot) \in M_2(\delta B_2)$ with the cost

$$E \left[\int_0^{\beta_m} f_n(\bar{X}(s)) ds \right]$$

for some n, m . By the results of [4, § IV.3, pp. 150–155], an optimal Markov control exists for this problem. Thus

$$E \left[\int_0^{\beta_m} f_n(Y(s)) ds \right] \leq \sup_{\bar{X}(\cdot) \in M_1(\delta B_2)} E \left[\int_0^{\beta_m} f_n(\bar{X}(s)) ds \right].$$

For large $n, f_n = 0$ on B_2 and hence the above is the same as

$$E \left[\int_0^{\gamma_m} f_n(X(s)) ds \right] \leq \sup_{\bar{X}(\cdot) \in M_1(\delta B_1)} E \left[\int_0^{\bar{\gamma}_m} f_n(\bar{X}(s)) ds \right]$$

where γ_m (respectively, $\bar{\gamma}_m$) = $\inf \{t \geq \xi_1 | X(t) \text{ (respectively, } \bar{X}(t) \notin \{x | \|x\| \leq m\} \setminus B_1\}$. Since $\gamma_m \uparrow \tau_2$ as $m \rightarrow \infty$, we have

$$(4.1) \quad E \left[\int_0^{\tau_2} f_n(X(s)) ds \right] \leq \sup_{\bar{X}(\cdot) \in M_1(\delta B_1)} E \left[\int_0^{\tau_2} f_n(\bar{X}(s)) ds \right] \leq \varepsilon,$$

for n sufficiently large, by virtue of Lemmas 4.1, 4.2. Q.E.D.

LEMMA 4.4. The set $\{E[\tau_2]|X(\cdot) \in M_2(\delta B_1)\}$ is bounded from above and bounded away from zero from below.

Proof. The first claim is proved by the same arguments that imply the first half of Lemma 4.1. The second claim follows by arguments similar to those used to prove a similar claim for $M_1(\delta B_1)$ in Lemma 4.1 with the following change: We consider

a sequence $\{X^n(\cdot)\}$ in $M_2(\delta B_1)$ instead of $M_1(\delta B_1)$, with initial laws arbitrary in $P(\delta B_1)$. Q.E.D.

We can now prove the main result of this section.

THEOREM 4.1. *There exists an a.s. optimal Markov control.*

Proof. Let $X(\cdot)$ be as in (1.1). By Lemmas 3.2–3.4, $\tau_i < \infty$ a.s. for all i . Thus for $\{f_n\}$ as in Lemma 4.3,

$$(4.2) \quad \limsup_{t \rightarrow \infty} \frac{1}{t} \int_0^t f_n(X(s)) ds = \limsup_{m \rightarrow \infty} \frac{\sum_{i=1}^m \int_{\tau_i}^{\tau_{i+1}} f_n(X(s)) ds}{\sum_{i=1}^m (\tau_{i+1} - \tau_i)}.$$

By Lemmas 3.4 and 4.3, for any $\varepsilon > 0$, there exists $N_\varepsilon \geq 1$ such that for all $n \geq N_\varepsilon$, $i \geq 1$,

$$(4.3) \quad E \left[\int_{\tau_i}^{\tau_{i+1}} f_n(X(s)) ds / F_{\tau_i} \right] < \varepsilon \quad \text{a.s.}$$

By Lemmas 3.2–3.4,

$$\sup_i E[(\tau_{i+1} - \tau_i)^2] < \infty.$$

Hence one can use the strong law of large numbers for square-integrable martingales [23, pp. 53] to conclude that

$$(4.4) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \left[\int_{\tau_i}^{\tau_{i+1}} f_n(X(s)) ds - E \left[\int_{\tau_i}^{\tau_{i+1}} f_n(X(s)) ds / F_{\tau_i} \right] \right] = 0 \quad \text{a.s.},$$

$$(4.5) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n [(\tau_{i+1} - \tau_i) - E[(\tau_{i+1} - \tau_i) / F_{\tau_i}]] = 0 \quad \text{a.s.}$$

From Lemmas 3.4, 4.4 and (4.2)–(4.5) above, we conclude that

$$\limsup_{t \rightarrow \infty} \frac{1}{t} \int_0^t f_n(X(s)) ds < C\varepsilon \quad \text{a.s.}$$

for n large enough, with some constant C independent of n . Recalling the definition of $\{f_n\}$, it is easily deduced from this that in the set-up of Lemma 2.2, $\delta(\nu) = 1$ for all limit points ν of $\{\nu_i\}$ outside a set of zero probability. The claim now follows as in the proof of Theorem 2.1 in view of Corollary 3.2. Q.E.D.

Remarks. Let $n = 1$. Pick Markov controls v_1, v_2 such that $m(x, v_1(x)) = \max m(x, v)$, $m(x, v_2(x)) = \min m(x, v)$. Our conditions on m and the selection theorem of Lemma 1 of [3], guarantee the existence of v_1, v_2 as above. Let $X(\cdot)$ be as in (1.1) for some admissible control $u(\cdot)$ and $X_1(\cdot), X_2(\cdot)$ be the diffusions controlled by v_1, v_2 respectively, with the same initial condition as $X(\cdot)$. (Recall that a strong solution to Markov-controlled (1.1) exists [29]. Thus we can construct $X(\cdot), X_1(\cdot), X_2(\cdot)$ on the same probability space.) By the well-known comparison theorem for one-dimensional Ito processes [17, pp. 352–355], it follows that outside a set N' of zero probability,

$$(4.6) \quad X_2(t) \leq X(t) \leq X_1(t) \quad \text{for all } t \geq 0.$$

Suppose we assume that v_1, v_2 are stable. Then (4.6) implies in a straightforward manner that

- (i) all Markov controls are stable,
- (ii) $\delta(\nu)$ in Lemma 2.2 can always be taken to be 1 outside $N \cup N'$ (N as in Lemma 2.2),
- (iii) $H = \{\eta_v | v \text{ Markov}\}$ is compact.

Thus in the one-dimensional case, we have the conclusion of Theorem 4.1 under a seemingly more general set-up than that of Assumption A.

REFERENCES

- [1] D. G. ARONSON, *Bounds for the fundamental solution of a parabolic equation*, Bull. Amer. Math. Soc., 73 (1967), pp. 890–896.
- [2] A. V. BALAKRISHNAN, *Applied Functional Analysis*, 2nd ed., Springer, New York, 1981.
- [3] V. E. BENES, *Existence of optimal strategies based on specified information, for a class of stochastic decision problems*, this Journal, 8 (1970), pp. 179–188.
- [4] A. BENSOUSSAN, *Stochastic Control by Functional Analysis Methods*, North-Holland, Amsterdam, 1982.
- [5] A. BENSOUSSAN AND V. S. BORKAR, *Ergodic control problem for one dimensional diffusions with near-monotone cost*, Systems Control Lett., 5 (1984), pp. 127–133. Errata in Control Lett., 7 (1986), pp. 233–235.
- [6] R. N. BHATTACHARYA, *Asymptotic behaviour of several dimensional diffusions*, in Stochastic Nonlinear Systems, L. Arnold and R. Lefever, eds., Springer, New York, 1981.
- [7] V. S. BORKAR, *Controlled Markov chains and stochastic networks*, this Journal, 21 (1983), pp. 652–666.
- [8] ———, *On minimum cost per unit time control of Markov chains*, this Journal, 22 (1984), pp. 965–978.
- [9] ———, *A note on controlled diffusions on line with time-averaged cost*, Systems Control Lett., 4 (1984), pp. 1–4.
- [10] ———, *A remark on the attainable distributions of controlled diffusions*, Stochastics, 18 (1986), pp. 17–23.
- [11] ———, *Control of Markov chains with long-run average cost criterion*, LIDS Report No. P-1583, Lab. for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1986.
- [12] R. M. COX AND I. KARATZAS, *Stationary control of Brownian motion in several dimensions*, Adv. in Appl. Probab., 17 (1985), pp. 53–561.
- [13] S. N. ETHIER AND T. G. KURTZ, *Markov processes: characterization and convergence*, John Wiley, New York, 1986.
- [14] W. H. FLEMING, *Generalized solutions in optimal stochastic control*, in Differential Games and Control Theory III, E. Roxin, P.-T. Liu and R. L. Sternberg, eds., Marcel Dekker, New York, 1977, pp. 147–165.
- [15] M. K. GHOSH, *Ergodic control of multidimensional diffusions II: The dynamic programming equations*, in preparation.
- [16] U. G. HAUSSMANN, *Existence of optimal Markovian controls for degenerate diffusions*, in Stochastic Differential Systems, N. Christopeit, K. Helmes and M. Kohlmann, eds., Lecture Notes in Control and Information Sciences 78, Springer, New York, 1986, pp. 171–186.
- [17] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland-Kodansha, Amsterdam-Tokyo, 1981.
- [18] R. Z. KHAS'MINSKII, *Ergodic properties of recurrent diffusion processes and stabilization of the solution to the Cauchy problem for parabolic equations*, Theory Probab. Appl., 2 (1960), pp. 179–196.
- [19] H. J. KUSHNER, *Stochastic Stability and Control*, Academic Press, New York, 1967.
- [20] ———, *Existence results for optimal stochastic control*, J. Optim. Theory Appl., 15 (1975), pp. 347–359.
- [21] ———, *Optimality conditions for the average cost per unit time problem with a diffusion model*, this Journal, 16 (1978), pp. 330–346.
- [22] O. A. LADYZENSKAJA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, Trans. Math. Monographs, 23 1968 p. 80.
- [23] M. LOEVE, *Probability Theory II*, 4th ed., Springer, New York, 1978.
- [24] P. MANDL, *Analytical Treatment of One-Dimensional Markov Processes*, Springer, New York, 1968.
- [25] M. ROBIN, *Long-term average cost control problems for continuous time Markov processes—a survey*, Acta Appl. Math., 1 (1983), pp. 281–300.
- [26] D. W. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Springer, New York, 1979.
- [27] R. TARRES, *Asymptotic evolution of a stochastic control problem*, this Journal, 23 (1985), pp. 614–631.
- [28] S. R. S. VARADHAN, *Lectures on diffusion problems and partial differential equations*, Tata Institute of Fundamental Research, Bombay, 1980.
- [29] A. JU. VERETENNIKOV, *On strong solutions and explicit formulas for solutions of stochastic integral equations*, Math. USSR-Sb., 39 (1981), pp. 387–403.
- [30] E. WONG, *Representation of martingales, quadratic variation and applications*, this Journal, 9 (1971), pp. 621–633.
- [31] L. C. YOUNG, *Calculus of Variations and Control Theory*, W. B. Saunders, Philadelphia, PA, 1969.
- [32] A. K. ZVONKIN, *A transformation of the phase space of a diffusion process that removes the drift*, Math. USSR-Sb., 22 (1974), pp. 129–149.

AN EXACT PENALTY FUNCTION AND RELAXATION APPROACH FOR SOLVING DECOMPOSABLE NONLINEAR PROGRAMS*

RAFAEL LAZIMY†

Abstract. An effective approach to solving decomposable nonlinear programs P (large-scale programs or, e.g., programs that are nonconvex in (x, y) jointly, but convex in x for each y where, e.g., y is a vector of integer variables) is based on the principles of projection and dual representation, which transform P into an equivalent master problem in y -space. The master problem may then be solved by either relaxation or implicit enumeration. However, a serious difficulty arises when solving the master problem in y -space. For a given trial solution y , the dual subproblem $D'(y)$ may be unbounded from above on its domain, in which case y is inadmissible, and a new constraint designed to eliminate y needs to be generated and added to the current relaxed master problem. The problem is that under certain conditions the constraints for eliminating inadmissible trial solutions are not producible using existing procedures. In this paper, we overcome this difficulty by employing an exact penalty function approach to derive a relaxation algorithm for solving P .

Key words. nonlinear programming, duality, relaxation, penalty functions, optimization

AMS(MOS) subject classifications. 90C30, 90C25, 90C10

1. Introduction. Consider the nonlinear program

$$P \quad \underset{(x, y)}{\text{minimize}} \quad f(x, y) \quad \text{s.t.} \quad g(x, y) \leq 0, \quad y \geq 0, \quad y \in Y,$$

where: $f: R^{n_1+n_2} \rightarrow R$, $g: R^{n_1+n_2} \rightarrow R^m$ are convex, twice continuously differentiable functions on the $(n_1 + n_2)$ real Euclidean space $R^{n_1+n_2}$. $Y \subseteq R^{n_2}$ is a compact set of arbitrarily constrained variables.

y is a vector of “complicating” variables in the sense that P is intractable and/or unmanageable in x and y jointly, but for each $y \in Y$ the subproblem $P(y) := \{\text{minimize } f(x, y), \text{ subject to } x: g(x, y) \leq 0, x \geq 0\}$ is convex and manageable. Following Geoffrion [5] we have in mind situations such as: (a) Y is a finite discrete set, so that P is not convex. For instance, y is a vector of integer-valued variables, in which case P is a mixed-integer nonlinear program. (b) P is a large-scale optimization problem, but $P(y)$ is of a manageable size. (c) $P(y)$ can be decomposed into a number of independent subprograms, each involving a different subvector of x , such that the solution of $P(y)$ can be decentralized and done in parallel for each of the smaller independent subprograms. (d) $P(y)$ has some special structure. (Obviously, y may be a vector of integer variables also in situations (b)–(d), and not only in (a).)

An effective approach to solving P is based on the principles of projection and dual representation, which transform program P into an equivalent master problem in y -space. Then, either relaxation or (if, e.g., Y is a discrete set) implicit enumeration is used in order to solve the master problem and obtain a solution for P . Benders applied these principles to solve certain programs with a structure similar to P [2]. His work, however, was limited to problems in which f and g are linear in the “complicating” variables y , and separable in x and y . Balas generalized Benders’ approach to the nonlinear case, but assumed that Y is a discrete set [1]. Also, he used implicit enumeration to solve the master problem in y -space. Geoffrion developed a somewhat different generalization of Benders’ approach [5]. He based his generalization

* Received by the editors June 30, 1986; accepted for publication (in revised form) April 8, 1987.

† Graduate School of Business, University of Wisconsin-Madison, Madison, Wisconsin 53706.

on nonlinear Lagrangian duality theory (thus, f and g are not necessarily assumed to be differentiable). Also, he used relaxation to solve the master problem.

A serious problem arises in solving the equivalent master problem in y -space (either by relaxation, or by implicit enumeration). The master problem is solved iteratively. Each iteration produces a trial solution y for the master problem; then the dual of subproblem $P(y)$ (denoted by $D'(y)$) is solved. The solution of $D'(y)$ (if it exists) is used to test the current trial solution y for optimality in the master problem, and to generate additional constraints if y is not the optimal solution. If $D'(y)$ is unbounded from above on its domain, $P(y)$ is inconsistent and, therefore, y is not admissible for P . A new constraint has to be produced, to eliminate the inadmissible point y . Balas [1, Thm. 5.3] proposed a procedure for generating constraints to eliminate inadmissible trial solutions. However, a serious flaw exists in Balas's Thm. 5.3, which makes it impossible to use it as a general procedure for producing constraints to eliminate inadmissible trial solutions y .

In this paper, we overcome the above problem by reformulating the dual pair $P(y)$ and $D'(y)$ so that: (1) The new primal subproblem $\hat{P}(y)$ is consistent for any y . (2) The new dual $\hat{D}(y)$ is bounded from above for any y . To this end, we employ an exact penalty function approach. Based on these formulations, we then present a relaxation algorithm to solve program P .

The paper is organized as follows. In § 2 we outline the relaxation approach for solving P , and identify and describe the problem that arises in the context of eliminating inadmissible points y . In § 3 we present the exact penalty function formulations, derive the dual pair $(\hat{P}(y), \hat{D}(y))$ and state the relationships between them and the original pair $(P(y), D(y))$. Finally, the modified relaxation method for solving P is presented in § 4, and some of its computational aspects are briefly discussed.

We briefly comment now on our notation. The relation $x \geq 0$ includes the case $x = 0$; however, $x \geq 0$ means that $x \neq 0$ (i.e., that $x_i > 0$ for at least one i). For $f: R^{n_1+n_2} \rightarrow R$, $\nabla_x f(x, z)$ denotes the vector of partial derivatives of f with respect to x , evaluated at (x, z) . Also, $\nabla_{x,z} f := (\nabla_x f, \nabla_z f)$. For $g: R^{n_1+n_2} \rightarrow R^m$, $\nabla_{x,z} g(x, z)$ denotes the $m \times (n_1 + n_2)$ Jacobian matrix evaluated at (x, z) . A vector of ones in any real Euclidean space is denoted by e .

2. Eliminating inadmissible solutions in the relaxation process. For some y , let $P(y)$ be the subproblem obtained by fixing y in program P :

$$P(y) \quad \pi(y) := \underset{x}{\text{minimize}} f(x, y) \quad \text{s.t. } g(x, y) \leq 0, \quad x \geq 0.$$

In order to obtain a dual of $P(y)$ whose objective function is *linear* in the “complicating” variables y and whose feasible set is *independent* of y , we employ the following subproblem $P'(y)$ which obviously is equivalent to $P(y)$ (see [1]):

$$P'(y) \quad \pi(y) := \underset{(x, z)}{\text{minimize}} f(x, z) \quad \text{s.t. } g(x, z) \leq 0, \quad z = y, \quad x \geq 0,$$

where $z \in R^{n_2}$. The dual of $P'(y)$ is (see Huard [9], Dantzig, Eisenberg and Cottle [3], Mangasarian [11], and Wolfe [16]):

$$D'(y) \quad \delta(y) := \underset{(x, z, u, v, r)}{\text{maximize}} h(x, y, z; u, v, r) \quad \text{s.t. } (x, z, u, v, r) \in \Psi,$$

where $u \in R^m$, $v \in R^{n_1}$, $r \in R^{n_1}$, and

$$(1) \quad h(x, y, z; u, v, r) := f(x, z) + ug(x, z) + v(z - y) - rx,$$

$$(2) \quad \Psi := \{(x, z, u, v, r) | \nabla_{x,z} [f(x, z) + ug(x, z) + vz - rx] = 0, (u, r) \geq 0\}.$$

Remark 2.1. If x is not restricted in sign, then the vector r does not appear in h and in Ψ .

It is assumed that g satisfies any one of the six constraint qualifications listed in [12, Thm. 7.3.7].

Define Φ to be the set of *admissible* solutions y for P :

$$(3) \quad \Phi := \{y | g(x, y) \leq 0 \text{ for some } x \geq 0\}.$$

Based on the duality relationships between $P'(y)$ and $D'(y)$ (see, e.g., [16, Thm. 2], [9], [11]), we obtain the following projection of P from (x, y) -space onto y -space (recall that Ψ is independent of y):

$$D' \quad \underset{y}{\text{minimize}} \delta(y), \quad \text{s.t. } y \in Y \cap \Phi$$

(where Y is in P), which is equivalent to

$$\begin{aligned} D' \quad & \underset{(\theta, y)}{\text{minimize}} \theta \\ & \text{s.t.} \quad (1) \quad \theta \geq f(x, z) + ug(x, z) + v(z - y) - rx \quad \text{all } (x, z, u, v, r) \in \Psi, \\ & \quad (2) \quad y \in Y \cap \Phi. \end{aligned}$$

Programs D' and P are equivalent in the sense that if (θ, y) solves D' , then (x, y) solves P , where x is the solution of $P(y)$. Thus, one can obtain the solution of P by solving D' . Program D' has an infinite number of constraints; therefore, it is natural to solve it by *relaxation*. Two programs are solved at each iteration of the relaxation process: a *relaxed* version of program D' , denoted RD, and a dual subproblem $D'(y)$. Program RD is used to generate candidate solutions (θ, y) for D' , and the dual $D'(y)$ is used to test (θ, y) for optimality in D' . If $D'(y)$ has a finite optimal solution (x, z, u, v, r) but (θ, y) is not feasible for D' , a new constraint (of type (1) of D') is added to the relaxed program RD. If $D'(y)$ is unbounded from above on Ψ (i.e., $\delta(y) \rightarrow +\infty$ along some feasible sequence $\{(x, z, u, v, r)_k\}$), then $P'(y)$ is inconsistent [16, Table on p. 242], and the trial solution y is not admissible for P . In this case, a new constraint must be added to RD in order to eliminate the inadmissible point y . The problem addressed in this paper concerns the generation of these constraints.

Balas employed the following result to generate constraints that eliminate inadmissible trial solutions y [1, Thm. 5.3]:

If $\delta(y)$ is unbounded from above on Ψ , there exist vectors (x, z, u) and $s \in R^m$ such that

$$(4a) \quad (x, z, u, v, r) \in \Psi,$$

$$(4b) \quad s \nabla_x g(x, z) \geq 0, \quad s \geq 0,$$

$$(4c) \quad sg(x, z) + t(z - y) - qx > 0,$$

where (see the definition of Ψ in (2))

$$(5) \quad v = -\nabla_z l(x, z; u), \quad r = \nabla_x l(x, z; u),$$

$$(6) \quad l(x, z; u) := f(x, z) + ug(x, z),$$

and

$$(7) \quad t := -s \nabla_z g(x, z), \quad q := s \nabla_x g(x, z).$$

Remark 2.2. If x is not restricted in sign, then the vectors r and q do not appear, and condition (4b) becomes

$$(8) \quad s \nabla_x g(x, z) = 0, \quad s \geq 0.$$

Therefore, if $D'(\bar{y})$ is unbounded from above, the above result is used to generate the following new constraint, which is added to program RD in order to eliminate the inadmissible point \bar{y} :

$$(9) \quad \bar{s}g(\bar{x}, \bar{z}) + \bar{t}(\bar{z} - y) - \bar{q}\bar{x} \leq 0,$$

where $(\bar{x}, \bar{z}, \bar{u}, \bar{v}, \bar{r})$ and $(\bar{s}, \bar{t}, \bar{q})$ are vectors that satisfy the conditions (4)–(7).

Theorem 5.3 in [1] is wrong if, for any $(x, z, u, v, r) \in \Psi$, the only solution to the system $\{s | s \nabla_x g(x, z) \geq 0, s \geq 0\}$ (if $x \geq 0$), or to $\{s | s \nabla_x g(x, z) = 0, s \geq 0\}$ (if x is not restricted in sign), is $s = 0$. Since $s = 0$ implies $t = q = 0$ (see (7)), then condition (4c) does not hold and, consequently, constraint (9) for eliminating inadmissible trial solutions y is not producible.

Using Motzkin's and Gordan's theorems of the alternatives, we obtain the following characterization of the cases when constraint (9) is not producible. Assume that $D'(y)$ is unbounded from above on Ψ . Then, for any $(x, z, u, v, r) \in \Psi$, we have¹

(a) $\{s | s \nabla_x g(x, z) \geq 0, s \geq 0\}$ has a solution s iff $\{w \in \mathbb{R}^{n_1} | \nabla_x g(x, z)w < 0, w \geq 0\}$ has no solution w (Motzkin [14]). Consequently, in the case that $x \geq 0$, if $\{w | \nabla_x g(x, z)w < 0, w \geq 0\}$ has a solution w , then constraint (9) is not producible by the procedure presented in [1].

(b) $\{s | s \nabla_x g(x, z) = 0, s \geq 0\}$ has a solution s iff $\{w \in \mathbb{R}^{n_1} | \nabla_x g(x, z)w > 0\}$ has no solution w (Gordan [6]). Consequently, in the case that x is not restricted in sign, if $\{w | \nabla_x g(x, z)w > 0\}$ has a solution w , then constraint (9) is not producible.

Remark 2.3. One particular case where constraint (9) is not producible is when x is not restricted in sign and the rows of the $m \times n_1$ matrix $\nabla_x g(x, z)$ are linearly independent for any $(x, z, u, v) \in \Psi$, since $s = 0$ is then the only solution to $s \nabla_x g(x, z) = 0$.

We next describe two examples in which $D'(y)$ is unbounded, but $s = 0$ and, thus, constraint (9) is not producible.

Example 1 (x unrestricted in sign). Consider

$$P \quad \text{minimize } -(x + y) \quad \text{s.t. } x^2 + y^2 - 1 \leq 0, \quad x \in \mathbb{R}, \quad y \in \mathbb{R}.$$

For all $|y| > 1$, the subproblem $P(y)$ is inconsistent, and the dual

$$D(y) \quad \delta(y) := \underset{(x, u)}{\text{maximize}} \quad l(x, y; u) \equiv -(x + y) + u(x^2 + y^2 - 1) \\ \text{s.t. } -1 + 2ux = 0, \quad u \geq 0,$$

is unbounded from above, since any (x, u) such that $x = 1/2u$, $u > 0$, is feasible for $D(y)$, and

$$l(x, y; u) = -y + u(y^2 - 1) - \frac{1}{4u} \rightarrow +\infty \quad \text{as } u \rightarrow +\infty$$

for all $|y| > 1$. However, $s = t = 0$ for all $(x, z, u, v) \in \Psi$, since (see (4)) the only solution s to

$$(1) \quad (x, z, u, v) \in \Psi \Leftrightarrow \nabla_x l(x, z; u) \equiv -1 + 2ux = 0, \quad u \geq 0$$

$$(\text{where } v := -\nabla_z l(x, z; u) = -1 + 2uz)$$

$$(2) \quad s \nabla_x g(x, z) \equiv 2sx = 0, \quad s \geq 0$$

¹ The difference between the notation “ \geq ” and “ \geqslant ” (as defined in the Introduction) is crucial here.

is $s = 0$, in which case $t := -s\nabla_z g(x, z) = 0$. Indeed, for any $x \in \{x | x = 1/2u, u > 0\}$, the system $\nabla_x g(x, z)w = 2xw > 0$ has a solution w , and by Gordan's theorem of the alternative it follows that, for any $x \in \{x | x = 1/2u, u > 0\}$, the system $s\nabla_x g(x, z) = 2sx = 0, s > 0$, has no solution s .

Example 2 (x restricted in sign). Consider

$$P \quad \text{minimize } x - y \quad \text{s.t. } \frac{1}{x} + y^2 - 10 \leq 0, \quad x \geq 0, \quad y \in R.$$

For all $y^2 > 10$, the subproblem $P(y)$ is infeasible and its dual $D(y)$ is unbounded from above, since

$$\begin{aligned} D(y) \quad \delta(y) &:= \max_{(x, u)} l(x, y; u) - \nabla_x l(x, y; u)x \\ &= x - y + u \left(\frac{1}{x} + y^2 - 10 \right) - x \left(1 - \frac{u}{x^2} \right) \\ \text{s.t. } 1 - \frac{u}{x^2} &\geq 0, \quad u \geq 0, \end{aligned}$$

and by taking $u = x^2$ we obtain $\delta(y) \rightarrow +\infty$ as $x \rightarrow +\infty$, provided that $y^2 > 10$. Thus, $D(y)$ is unbounded from above for any $y^2 > 10$. However, $s = t = q = 0$ for all $(x, z, u, v, r) \in \Psi$, since (see (4)) the only solution s to

$$(1) \quad (x, z, u, v, r) \in \Psi \Leftrightarrow \nabla_x l(x, z; u) \equiv 1 - \frac{u}{x^2} \geq 0, \quad u \geq 0,$$

$$\left(\text{where } (v, r) := (-\nabla_z l(x, z; u), \nabla_x l(x, z; u)) = \left(-1 + 2uz, 1 - \frac{u}{x^2} \right) \right)$$

$$(2) \quad s\nabla_x g(x, z) \equiv -\frac{s}{x^2} \geq 0, \quad s \geq 0$$

is $s = 0$, in which case $(t, q) := (-s\nabla_z g(x, z), s\nabla_x g(x, z)) = (0, 0)$. Indeed, for any $(x, u) \in \{(x, u) | 1 - u/x^2 \geq 0, u \geq 0\}$, we obtain: The system $\nabla_x g(x, z)w = -w/x^2 < 0, w \geq 0$, has a solution w , and by Motzkin's theorem of the alternative the system $s\nabla_x g(x, z) = -s/x^2 \geq 0, s > 0$, has no solution s .

Finally, it is worth noting that the problem of having $s = 0$ as the only solution to conditions (4a) and (4b) does not arise if the constraints of P are separable in x and y and linear in x , as stated in the following theorem.

THEOREM 2.1. *Suppose that $g(x, y) := Ax - k(y)$ where A is an $m \times n_1$ matrix, and $k: R^{n_2} \rightarrow R^m$ is a convex, twice continuously differentiable function on R^{n_2} . For some y , assume that $P(y)$ is inconsistent, but that its dual $D(y)$ is consistent. Then, $D(y)$ is unbounded from above on its domain, and for any dual feasible solution (x, u) the system*

$$s\nabla_x g(x, y) = 0, \quad s \geq 0 \quad (\text{if } x \in R^{n_1})$$

or

$$s\nabla_x g(x, y) \geq 0, \quad s \geq 0 \quad (\text{if } x \geq 0)$$

has a solution s .

The proof to Theorem 2.1 follows from [16, Thm. 3], the application of Gale's theorem of the alternative [4], and from the observation that, if there is a solution $s \geq 0$ to $s\nabla_x g(x, y) = 0$ (or to $s\nabla_x g(x, y) \geq 0$), then this solution exists for any dual feasible solution (x, u) , since $\nabla_x g(x, y)$ is independent of x .

In §§ 3 and 4 we present a modified projection and relaxation procedure for solving program P , under which the problem described above does not arise. Furthermore, this is done without imposing any new restrictions on the structure of P . The basic idea of the modified procedure is simple: the dual pair $(P'(y), D'(y))$ is reformulated so that, for any given y , the new primal program is always consistent, and the new dual program is always bounded from above on its domain. To this end, we employ an exact penalty function approach.

3. Exact penalty function formulation. Let $\alpha > 0$ be a constant, e , a vector of ones, $w \in R^m$, a vector of variables, and g_j be the j th constraint function. For some y , let $P(y; x, \alpha)$ be an exact penalty function of subproblem $P(y)$. (For instance, the exact l_1 penalty function may be employed.) It was noted by Mangasarian [13, Lemma 4.1] that the two programs

$$(10) \quad \underset{x \geq 0}{\text{minimize}} P(y; x, \alpha) := \underset{x \geq 0}{\text{minimize}} f(x, y) + \alpha \sum_{j=1}^m \max \{0, g_j(x, y)\},$$

$$\hat{P}(y) \quad \hat{\pi}(y) := \underset{(x, z, w)}{\text{minimize}} \hat{f}(x, z, w) := f(x, z) + \alpha ew$$

$$\text{s.t. } g(x, z) - w \leq 0, \quad z = y, \quad (x, w) \geq 0$$

are equivalent for any given \bar{y} (provided that α is larger than a threshold value), in the following sense: If \bar{x} solves (10), then $(\bar{x}, \bar{z}, \bar{w})$ solves $\hat{P}(\bar{y})$ where $\bar{z} = \bar{y}$ and $\bar{w} = g(\bar{x}, \bar{y})_+$ where $g_j(\bar{x}, \bar{y})_+ := \max \{0, g_j(\bar{x}, \bar{y})\}$; and if $(\bar{x}, \bar{z}, \bar{w})$ solves $\hat{P}(\bar{y})$, then \bar{x} solves (10).

Program $\hat{P}(y)$ is consistent for any y . Recalling that $P(y)$ is the program $\pi(y) := \underset{x \geq 0}{\text{minimize}} f(x, y)$ subject to $g(x, y) \leq 0$, and that Φ is the set of admissible solutions y of P (see (3)), we obtain the following relationships between $\hat{P}(y)$ and $P(y)$:

(a) Let $\bar{y} \notin \Phi$. Then

(1) $P(\bar{y})$ is infeasible.

(2) $\hat{P}(\bar{y})$ has an optimal solution at $(\bar{x}, \bar{z}, \bar{w})$, where $\bar{z} = \bar{y}$, \bar{x} is the global minimum of $P(\bar{y}; x, \alpha)$, $\bar{w} = g(\bar{x}, \bar{y})_+$, and $e\bar{w} := \sum_{j=1}^m g_j(\bar{x}, \bar{y})_+ > 0$.

(b) Let $\bar{y} \in \Phi$ but $P(\bar{y})$ be unbounded from below on its domain. Then $\hat{P}(\bar{y})$ is also unbounded from below. In this case the original program P is also unbounded.

(c) Let $\bar{y} \in \Phi$ and $P(\bar{y})$ have a finite optimal solution \bar{x} . Also, assume that $g(x, \bar{y})$ satisfies the Slater constraint qualification, and that f and g are convex. Then $\hat{P}(\bar{y})$ has an optimal solution at $(\bar{x}, \bar{z}, \bar{w})$ for each value α such that $\alpha > \bar{\alpha}$, where $\bar{\alpha}$ is some threshold value. (See Luenberger [10], Han and Mangasarian [8], Mangasarian [13], and Zangwill [17].) Furthermore, $\bar{z} = \bar{y}$ and $e\bar{w} := \sum_{j=1}^m g_j(\bar{x}, \bar{y})_+ = 0$.

Remark 3.1. Formulation $\hat{P}(y)$ could be considered to be an extension of the well-known big- M method of linear programming [15] to the case of convex programs. (The potential benefits of extending the big- M method to convex programming were recently investigated by Mangasarian [13]. In this study, this idea is applied to solving the decomposable nonlinear program P using a projection and relaxation approach.)

The dual of $\hat{P}(y)$ is

$$\hat{D}(y) \quad \hat{\delta}(y) := \underset{(x, z, w, u, v, r, p) \in \hat{\Psi}}{\text{maximize}} \hat{h}(x, y, z, w; u, v, r, p),$$

where $u \in R^m$, $v \in R^{n_2}$, $r \in R^{n_1}$, $p \in R^m$, and

$$(11) \quad \hat{h}(x, y, z, w; u, v, r, p) := f(x, z) + \alpha ew + u[g(x, z) - w] + v(z - y) - rx - pw,$$

$$(12) \quad \hat{\Psi} := \{(x, z, w, u, v, r, p) | \nabla_{x, z, w} \hat{h}(x, y, z, w; u, v, r, p) = 0, w \geq 0, (u, r, p) \geq 0\}.$$

Since $\nabla_w \hat{h} = \alpha e - u - p$, we can substitute $p = \alpha e - u$ in \hat{h} , to obtain the following form of $\hat{D}(y)$:

$$\hat{D}(y) \quad \hat{\delta}(y) := \underset{(x,z,u,v,r) \in \hat{\Psi}}{\text{maximize}} \quad \hat{h}(x, y, z; u, v, r),$$

$$(13) \quad \hat{h}(x, y, z; u, v, r) := f(x, z) + ug(x, z) + v(z - y) - rx,$$

$$(14) \quad \hat{\Psi} := \{(x, z, u, v, r) | \nabla_{x,z} [f(x, z) + ug(x, z) + vz - rx] = 0, u \leq \alpha e, (u, r) \geq 0\}.$$

The set $\hat{\Psi}$ is bounded and independent of y . Consequently, the dual $\hat{D}(y)$ is bounded from above (i.e., has a finite optimal solution) for *any* y , provided that $\hat{\Psi}$ is nonempty. Next, the projection and relaxation algorithm for solving program P is presented.

4. The algorithm. Using the modified dual definition $\hat{D}(y)$, the projection of program P onto y -space yields

$$\begin{aligned} \hat{D} \quad & \underset{(\theta, y)}{\text{minimize}} \quad \theta \\ \text{s.t.} \quad & (1) \quad \theta \geq f(x, z) + ug(x, z) + v(z - y) - rx \quad \text{all } (x, z, u, v, r) \in \hat{\Psi}, \\ & (2) \quad y \in Y \cap \Phi. \end{aligned}$$

Program \hat{D} is solved iteratively by relaxation. At each iteration, the current relaxed program RD generates a candidate solution (θ, y) , and the dual $\hat{D}(y)$ is solved in order to determine the status of y . First, we need the following.

THEOREM 4.1. *Assume that $\hat{\Psi}$ is nonempty. For some y , let (x, z, u, v, r) be the (finite) optimal solution of $\hat{D}(y)$, and assume that f and g satisfy either Huard's [9] or Mangasarian's [11] strict converse duality theorem conditions. Then*

$$(15) \quad (a) \quad y \in \Phi \text{ iff } ug(x, z) + v(z - y) - rx = 0,$$

$$(16) \quad (b) \quad y \notin \Phi \text{ iff } ug(x, z) + v(z - y) - rx > 0.$$

Furthermore, $u \geq 0$ in part (b).

Proof. Under the assumptions of the theorem, it follows from the strict converse duality theorem ([9], [11]), that (x, z, w) is the (finite) optimal solution of the primal program $\hat{P}(y)$, where $ew = \sum_{j=1}^m g_j(x, z)_+$. Furthermore, the optimal objective function values of $\hat{P}(y)$ and $\hat{D}(y)$ are equal: $\hat{\pi}(y) = \hat{\delta}(y)$, or, $f(x, z) + \alpha ew = f(x, z) + ug(x, z) + v(z - y) - rx$. Therefore,

$$(17) \quad ug(x, z) + v(z - y) - rx = \alpha ew$$

at the optimum.

Now, recall the relationships between programs $P(y)$ and $\hat{P}(y)$ and observe that $y \in \Phi$ iff $ew = 0$ at the optimum of $\hat{P}(y)$, and $y \notin \Phi$ iff $ew > 0$ at the optimum of $\hat{P}(y)$. This, together with (17) and the fact that $\alpha > 0$, proves parts (a) and (b) of the theorem.

It remains to show that $u \neq 0$ in part (b). Since relation (17) holds at the optimal solution (x, z, u, v, r) of $\hat{D}(y)$, and since $\alpha ew > 0$ in part (b) and $u \geq 0$, it is sufficient to show that $v(z - y) = 0$ and $rx = 0$ at the optimal solution (x, z, u, v, r) . To this end, recall that $z = y$ at the optimum of $\hat{P}(y)$. Therefore, $v(z - y) = 0$. Next, since $(x, z, u, v, r) \in \hat{\Psi}$, then (see (12)):

$$(18) \quad \begin{aligned} v &= -\nabla_z l(x, z; u), & r &= \nabla_x l(x, z; u), \\ l(x, z; u) &:= f(x, z) + ug(x, z). \end{aligned}$$

Using the complementary slackness property $\nabla_x[f(x, z) + ug(x, z) + v(z - y)]x = 0$, we obtain $\nabla_x l(x, z; u)x = 0$. By combining this result with (18) we obtain

$$(19) \quad rx = 0$$

at the optimum of $\hat{D}(y)$. This completes the proof. Q.E.D.

Letting $(\bar{\theta}, \bar{y})$ be the solution of the most recent relaxed program RD (see below), the algorithm for solving P is:

Step 1. Solve subproblem $\hat{D}(\bar{y})$.

- (a) If $\hat{D}(\bar{y})$ is inconsistent, stop: $\hat{D}(y)$ is inconsistent for any y , since its feasible set $\hat{\Psi}$ is independent of y . The original program P is either inconsistent or unbounded from below on its domain. (This situation can only be encountered during the first iteration.)
- (b) $\hat{D}(\bar{y})$ has an optimal solution $(\bar{x}, \bar{z}, \bar{u}, \bar{v}, \bar{r})$. Then
 - (i) If $\bar{u}g(\bar{x}, \bar{z}) + \bar{v}(\bar{z} - \bar{y}) - \bar{r}\bar{x} = 0$ and $\bar{\theta} \geq \hat{h}(\bar{x}, \bar{y}, \bar{z}; \bar{u}, \bar{v}, \bar{r}) =: \delta(\bar{y})$, stop: (\bar{x}, \bar{y}) is the optimal solution of P .
 - (ii) If $\bar{\theta} < \hat{h}(\bar{x}, \bar{y}, \bar{z}; \bar{u}, \bar{v}, \bar{r}) =: \delta(\bar{y})$, add the following constraint:

$$(20) \quad \theta + \bar{v}y \geq f(\bar{x}, \bar{z}) + \bar{u}g(\bar{x}, \bar{z}) + \bar{v}\bar{z}$$

to program RD.

- (iii) If $\bar{u}g(\bar{x}, \bar{z}) + \bar{v}(\bar{z} - \bar{y}) - \bar{r}\bar{x} > 0$, add the following constraint to RD:

$$(21) \quad \bar{v}y \geq \bar{u}g(\bar{x}, \bar{z}) + \bar{v}\bar{z}.$$

Go to Step 2.

Step 2. Solve the revised relaxed program RD.

RD minimize θ
(θ, y)

- s.t. (1) $\theta + v^i y \geq f(x^i, z^i) + u^i g(x^i, z^i) + v^i z^i, \quad i \in M_1,$
- (2) $v^i y \geq u^i g(x^i, z^i) + v^i z^i, \quad i \in M_2,$
- (3) $y \in Y,$

where $\{(\theta^i, y^i) | i = 1, 2, \dots\}$ and $\{(x^i, z^i, u^i, v^i, r^i) | i = 1, 2, \dots\}$, respectively, are the sequences of solutions of programs RD and $\hat{D}(y^i)$ solved so far, and

- (1) $i \in M_1$ iff $\theta^i < \hat{h}(x^i, y^i, z^i, u^i, v^i, r^i)$,
- (2) $i \in M_2$ iff $u^i g(x^i, z^i) + v^i(z^i - y^i) - r^i x^i > 0$.

Let $(\bar{\theta}, \bar{y})$ be the solution of RD, and go to Step 1. If RD has no solution, stop: program P is either inconsistent or unbounded from below.

The following comments are in order.

- (a) *Optimality conditions.* Suppose that the conditions stated in part (b)(i) of Step 1 are met. Then, the optimality of (\bar{x}, \bar{y}) in P is evident from the following observations:

- (1) Since $(\bar{x}, \bar{z}, \bar{u}, \bar{v}, \bar{r})$ is the optimal solution of $\hat{D}(\bar{y})$ and since the feasible set $\hat{\Psi}$ of $\hat{D}(y)$ is independent of y , then

$$\hat{h}(\bar{x}, \bar{y}, \bar{z}; \bar{u}, \bar{v}, \bar{r}) \geq \hat{h}(x, \bar{y}, z; u, v, r) \quad \text{for all } (x, z, u, v, r) \in \hat{\Psi}.$$

Therefore

$$\bar{\theta} \geq \hat{h}(x, \bar{y}, z; u, v, r) \quad \text{for all } (x, z, u, v, r) \in \hat{\Psi},$$

since $\bar{\theta} \geq \hat{h}(\bar{x}, \bar{y}, \bar{z}; \bar{u}, \bar{v}, \bar{r})$.

- (2) $\bar{y} \in Y$ (since $(\bar{\theta}, \bar{y})$ is a solution of RD), and $\bar{y} \in \Phi$, since $\bar{u}g(\bar{x}, \bar{z}) + \bar{v}(\bar{z} - \bar{y}) - \bar{r}\bar{x} = 0$ (Thm. 4.1(a)).

The above observations imply that $(\bar{\theta}, \bar{y})$ is a feasible solution of the *unrelaxed* program \hat{D} . However, since $(\bar{\theta}, \bar{y})$ is the optimal solution of program RD, and since RD is a *relaxed* version of \hat{D} , it follows that $(\bar{\theta}, \bar{y})$ is also the optimal solution of \hat{D} . Finally

- (3) Programs P and \hat{D} are equivalent, in the sense that if $(\bar{\theta}, \bar{y})$ is the solution of \hat{D} , then (\bar{x}, \bar{y}) is the solution of P , where $(\bar{x}, \bar{z}, \bar{u}, \bar{v}, \bar{r})$ is the solution of $\hat{D}(\bar{y})$.
- (b) *Adding new constraints to program RD.* If $\hat{\Psi}$ is nonempty and $\bar{\theta} < \hat{h}(\bar{x}, \bar{y}, \bar{z}; \bar{u}, \bar{v}, \bar{r})$ (part (b)(ii) of Step 1), then $(\bar{\theta}, \bar{y})$ does not satisfy *all* of the constraints of type (1) of program D . The new constraint (20) which is added to RD is designed to satisfy the “most violated” constraint. Next, if the conditions stated in part (b)(iii) of Step 1 are met, then $\bar{y} \notin \Phi$ (Thm. 4.1). The new constraint (21) is designed to eliminate the inadmissible point \bar{y} . (Note that the term $\bar{r}\bar{x}$ is not included in both (20) and (21), since $\bar{r}\bar{x} = 0$ at the optimum of $\hat{D}(\bar{y})$: see (19). Also, note that $u^i g(x^i, z^i) = 0$ in constraints (1) of RD if $i \in M_1 - M_2$, but that $u^i g(x^i, z^i) > 0$ if $i \in M_1 \cap M_2$.)

Finally, we briefly discuss some of the computational aspects of the algorithm. First, notice that the constraints of types (1) and (2) of program RD are all *linear* in the “complicating” variables y , although the original program P may be nonlinear in y . Consequently, if y is, e.g., a vector of integer-valued variables, then RD is a mixed-integer *linear* program; and if y is a vector of continuous variables, then RD is a standard linear program. Next, either $\hat{D}(y)$, or

$$\begin{aligned} \hat{P}_1(y) \quad \quad \quad \hat{\pi}(y) := \underset{(x, w)}{\text{minimize}} \quad \hat{f}(x, y, w) &:= f(x, y) + \alpha ew \\ \text{s.t.} \quad g(x, y) - w &\leq 0, \quad (x, w) \geq 0, \end{aligned}$$

can be employed as the subproblem in Step 1 of the algorithm. The advantages of using $\hat{D}(y)$ are: (1) Its feasible set $\hat{\Psi}$ is independent of y . (2) The objective functions of two consecutive programs $\hat{D}(y)$ differ from one another in the linear term $\bar{v}y$ only. These properties may simplify greatly the solution of consecutive programs $\hat{D}(y)$. Another obvious advantage is that the solution of $\hat{D}(y)$ provides both the primal and dual solutions. The main disadvantage of using $\hat{D}(y)$ is its size. Therefore, in some cases it may be preferable to find the solution (x, z, u, v, r) of $\hat{D}(y)$ by solving the primal subproblem $\hat{P}_1(y)$, which has less variables and constraints than $\hat{D}(y)$.

If $\hat{P}_1(y)$ is used as a subproblem in the relaxation process, then the algorithm for solving it needs to be *dual adequate*, in the sense that it provides both the primal solution (x, w) , and the optimal dual multipliers vector u . Then, the remaining components of the solution (x, z, u, v, r) of $\hat{D}(y)$ are determined by (see (18)):

$$(22) \quad z := y, \quad v := -\nabla_z l(x, z; u), \quad r := \nabla_x l(x, z; u),$$

where $l(x, z; u) := f(x, z) + ug(x, z)$. This provides us with all the information needed to execute the algorithm.

Finally, we consider the issue of determining the value α used in both $\hat{P}_1(y)$ and $\hat{D}(y)$. The parameter α has to exceed a threshold value $\bar{\alpha}$. One example of $\bar{\alpha}$ is due to Zangwill [17, p. 356]

$$(23) \quad \bar{\alpha}_1 := \frac{f(x^1, \bar{y}) - f(\bar{x}, \bar{y}) + 1}{\min_{1 \leq j \leq m} -g_j(x^1, \bar{y})},$$

and another is due to Luenberger [10] (see also Han and Mangasarian [8, Thm. 4.9])

$$(24) \quad \bar{\alpha}_2 := \|\bar{u}\|_\infty = \max_{1 \leq j \leq m} \bar{u}_j,$$

where \bar{y} is the current trial solution, \bar{x} the solution of $P(\bar{y})$, x^1 is any point satisfying the Slater constraint qualification, and \bar{u} is any optimal Lagrange multipliers vector for $P(\bar{y})$. Obviously, \bar{x} (or \bar{u}) are unknown beforehand; thus the *exact* threshold $\bar{\alpha}_1$ (or $\bar{\alpha}_2$) cannot be computed. One way to overcome this difficulty is to use an arbitrarily large penalty parameter α . In order to circumvent numerical difficulties of ill-conditioning that may result from using a large α , the numerical linear algebra has to be done carefully: see, e.g., Gould [7]. A better way (see [13]) is to compute an upper bound for the threshold value $\bar{\alpha}$ (either $\bar{\alpha}_1$, or $\bar{\alpha}_2$), and set the penalty parameter α to be equal to this upper bound. Thus, we can replace the unknown function value $f(\bar{x}, \bar{y})$ by a lower bound, and use it in (23) to compute $\alpha \geq \bar{\alpha}_1$. Similarly, the threshold $\bar{\alpha}_2$ (see (24)) can also be replaced by an upper bound. In [13, Cor. 2.2] it is shown that the following is an upper bound for $\bar{\alpha}_2$:

$$(25) \quad \|\bar{u}\|_{\infty} \leq \frac{f(x^1, \bar{y}) - f(\bar{x}, \bar{y})}{\min_{1 \leq j \leq m} -g_j(x^1, \bar{y})},$$

where \bar{x} , \bar{y} , \bar{u} , and x^1 are defined as before. By replacing the unknown value $f(\bar{x}, \bar{y})$ by a lower bound, we can obtain a penalty parameter α .

Remark 4.1. Given a trial solution y , suppose that we know beforehand whether y is admissible or inadmissible. Then

(a) If y is admissible, we employ the *original* subprogram $P(y)$ as a subproblem, rather than $\hat{P}_1(y)$:

$$P(y) \quad \underset{x}{\text{minimize}} \quad f(x, y) \quad \text{s.t.} \quad g(x, y) \leq 0, \quad x \geq 0.$$

In this case, the artificial variables w are not included. Therefore, there is no need to determine the penalty parameter α . Also, program $P(y)$ is smaller than $\hat{P}_1(y)$: it includes fewer variables.

(b) If y is inadmissible, then *any* $\alpha > 0$ can serve as a penalty parameter, and program $\hat{P}_1(y)$ will be employed as a subproblem.

The question, of course, is how to establish whether a given trial solution y is admissible or not. One approach is to use an algorithm fitted with a phase-one procedure in order to solve program $P(y)$. Thus, given \bar{y} , we first solve the phase-one program

$$\underset{(x, w)}{\text{minimize}} \quad ew \quad \text{s.t.} \quad g(x, \bar{y}) - w \leq 0, \quad (x, w) \geq 0.$$

Letting (\bar{x}, \bar{w}) be its solution, we will conclude that \bar{y} is admissible if $\bar{w} = 0$, and inadmissible if $e\bar{w} > 0$.

REFERENCES

- [1] E. BALAS, *Minimax and duality for linear and nonlinear mixed-integer programming*, in Integer and Nonlinear Programming, J. Abadie, ed., North-Holland, Amsterdam-New York, 1970, pp. 385-418.
- [2] J. F. BENDERS, *Partitioning procedures for solving mixed-variable programming problems*, Numer. Math., 21 (1963), pp. 238-252.
- [3] G. D. DANTZIG, E. EISENBERG AND R. W. COTTLE, *Symmetric dual nonlinear programs*, Pacific J. Math., 15 (1965), pp. 809-812.
- [4] D. GALE, *The Theory of Linear Economic Models*, McGraw-Hill, New York, 1960.
- [5] A. M. GEOFFRION, *Generalized Benders decomposition*, J. Optim. Theory Appl., 10 (1972), pp. 237-260.
- [6] P. GORDAN, *Über die auslösungen linearer Gleichungen mit reellen Coefficienten*, Math. Ann., 6 (1873), pp. 23-28.
- [7] N. I. M. GOULD, *On the accurate determination of search directions for simple differentiable penalty functions*, IMA J. Numer. Anal., 6 (1986), pp. 357-372.

- [8] S. P. HAN AND O. L. MANGASARIAN, *Exact penalty function in nonlinear programming*, Math. Programming, 17 (1979), pp. 251-269.
- [9] P. HUARD, *Dual Programs*, in Recent Advances in Mathematical Programming, R. Graves and P. Wolfe, eds., John Wiley, New York, 1963, pp. 55-62.
- [10] D. G. LUENBERGER, *Control problems with kinks*, IEEE Trans. Automat. Control, 15 (1970), pp. 570-575.
- [11] O. L. MANGASARIAN, *Duality in nonlinear programming*, Quart. Appl. Math., 20 (1962), pp. 300-302.
- [12] ———, *Nonlinear Programming*, McGraw-Hill, New York, 1969.
- [13] ———, *Sufficiency of exact penalty minimization*, this Journal, 23 (1985), pp. 30-37.
- [14] T. S. MOTZKIN, *Beiträge zur Theorie Der Linearen Ungleichungen*, Inaugural dissertation, Univ. of Basel, Jerusalem, 1936.
- [15] K. MURTY, *Linear and Combinatorial Programming*, John Wiley, New York, 1976.
- [16] P. WOLFE, *A duality theorem for nonlinear programming*, Quart. Appl. Math., 19 (1961), pp. 239-244.
- [17] W. ZANGWILL, *Nonlinear programming via penalty functions*, Management Sci., 13 (1967), pp. 344-358.

TRAJECTORY STABILIZING CONTROLS IN HEREDITARY LINEAR SYSTEMS*

GILEAD TADMOR†

Abstract. The following phenomenon is specific to a class of hereditary control systems: Trajectories of these systems can be stabilized (i.e., controlled to asymptotically approach the origin). However, it is impossible to obtain stabilization with controls whose rate of decay is the same as that of the trajectories. We call such systems trajectory stabilizable. We characterize the class of trajectory stabilizable systems and present both open and closed loop trajectory-stabilizing schemes. The controllability aspect of trajectory stabilizability is studied too. Examples illustrate the discussion.

Key words. stabilizability, closed loop controls, spectral controllability, control delays, reduction schemes

AMS(MOS) subject classifications. 93915, 93B05, 93C22, 49E30

1. Introduction. Consider the linear hereditary control system

$$(*) \quad \dot{x}(t) = \int_0^h d\alpha(\theta)x(t-\theta) + \int_0^h d\beta(\theta)u(t-\theta).$$

In this system $x \in R^n$, $u \in R^m$ and h is a fixed, positive number, standing for the length of the system's memory. Two matrix valued measures of bounded variations, α and β , determine the evolution's dependence on past and present values of the state and of the control. We shall be interested in stabilizing trajectories of (*).

The following phenomenon is specific to systems with control-delays: It may happen that the trajectory $x(t)$ can be stabilized, i.e., steered to the origin, at a certain asymptotic rate (say, proportional to $e^{\gamma t}$, for some negative γ), but that the decay of the corresponding control function $u(t)$, must be slower. This is contrary to the situations in systems with *no* delays in the control action, where stabilizing controls can always be chosen asymptotically proportional to $x(t)$. We call systems which might exhibit this phenomenon *trajectory stabilizable* (rigorous definitions are provided in § 3).

Stabilization of hereditary systems has been studied already by several researchers, and the articles cited below, except [2] and [6], are examples of the published results. In these papers, however, a system is termed *stabilizable*, or *regulable*, only if it is possible to steer $x(t)$ to the origin, using a control function which is asymptotically proportional to $x(t)$. We distinguish systems having this stronger property from trajectory-stabilizable systems, and term them *state stabilizable*.

The latter notion is based on the commonly accepted view that the true state of a hereditary system comprises both the history of $x(t)$ and that of the control, $u(t)$. Indeed, formally speaking, trajectory-stabilizable systems might be those in which stabilizing controls must even bear exponential growth, or ever increasing oscillations. Such systems will not be of practical use, except perhaps in rare cases of very short operation time. Excluding such anomalies, however, there remains the very rich class of systems, of real-life interest, in which it is possible (and desirable) to improve decay

* Received by the editors August 12, 1985; accepted for publication April 3, 1987. This is a revised version of a technical report, prepared at the Department of Electronic Communication, Control and Computer Systems, School of Engineering, Tel Aviv University, Tel Aviv, Israel.

† Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

rates of trajectories, but no such improvement is possible in the corresponding stabilizing controls (which remain, nonetheless, bounded, or decay at slower rates). These are the true subject of the present study. We characterize systems in this class and describe closed- and open-loop, trajectory-stabilizing controls.

The paper is organized as follows: Using a very simple motivating example, some basic observations are made in § 2. These observations will be further developed in the following sections. A characterization of trajectory-stabilizable systems, and of the relations between state- and trajectory-stabilizability, is given in Theorem A, in § 3. In the proof of Theorem A, it is shown how known open-loop state-stabilization schemes can be modified, to suit trajectory-stabilizable systems. Closed-loop schemes are developed in § 4. In § 5, a simple, and perhaps more feasible treatment is suggested for the simpler case of systems with only commensurate lags in the control term. The controllability notion which correspond to trajectory-stabilizability (versus state-stabilizability), and geometrical properties of trajectory-stabilizing controls, are studied in § 6. Some concluding remarks are brought in § 7.

Following a well-known analysis of Pandolfi [16] (see also Remark 4.4, in § 4 herein), when stabilizing a system with $x(\cdot)$ -delays, one really focuses on an ordinary system, obtained via a spectral projection of the history of the function $x(\cdot)$. All our illustrating examples will therefore have ordinary homogeneous parts, concentrating on the interesting effects of control delays.

2. A motivating simple example. Consider this next scalar system

$$(2.1) \quad \dot{x}(t) = \lambda x(t) + u(t) - e^\lambda u(t-1).$$

Assuming $u(t) = 0$ for $t \leq 0$, the Laplace transform of (2.1) becomes

$$(s - \lambda)X(s) = x(0) + (1 - e^{\lambda-s})U(s);$$

hence

$$X(s) = \frac{x(0)}{s - \lambda} + \frac{1 - e^{\lambda-s}}{s - \lambda} U(s).$$

Now, if $x(t)$ is to be stabilized (starting with $x(0) \neq 0$) at a decay rate $\gamma < \lambda$ (i.e., if the pole of $X(s)$ at $s = \lambda$ is to be removed), then $U(s)$ must have a pole at $s = \lambda$. This happens because of the zero of $B(s) \stackrel{\text{def}}{=} 1 - e^{\lambda-s}$ at $s = \lambda$. Observing this phenomenon for positive λ , Olbrot [15] suggested the concept of stabilizability which we term *state-stabilizability*, and which excludes systems such as the system (2.1), where stabilizing controls must grow exponentially.

Here is a different point of view. Suppose λ is negative. Then trajectories decay at rate λ when no control force is applied. It seems natural to ask whether this rate can be improved by use of reasonable control functions (say, square-integrable, optimal quadratic for large “ Q ” kernel, bounded etc.). In the ordinary (nondelayed) case, one obtains the answers by studying the system’s state-stabilizability properties: if $x(t)$ can be brought to the origin at rate $\gamma < \lambda$, so can a control function for that task. Here we have an example for a qualitatively different situation: although $x(t)$ can be brought to the origin at rate $\gamma < \lambda$, so can a control function for that task. Here we have an example for a qualitatively different situation: although $x(t)$ can be brought to the origin arbitrarily fast, control functions which do the job will be, at best, asymptotically proportional to $e^{\lambda t}$. Trajectory stabilizability is therefore an appropriate notion in the presence of delays.

A basic observation (Theorem A, § 3) is that (*) is γ -trajectory-stabilizable (given a decay-rate γ) if and only if it can be transformed into a γ -state-stabilizable system by factoring undesirable zeros from $B(s)$. State-stabilizing open and closed loop controls in the transformed system can be used in solving the original stabilization problem.

Here are two factorization and stabilization strategies, as displayed in the simple framework of (2.1). The first suits systems of the general form of (*). Here it goes as follows: Set $\bar{B}(s) = (1 - e^{\lambda-s})/(s - \lambda)$. Then $\bar{B}(\lambda) \neq 0$ and $\bar{B}(s)$ is the Laplace transform of a measure $d\bar{\beta}$ whose support is the interval $[0, 1]$ (as is that of $d\beta$). Precisely,

$$d\bar{\beta}(\theta) = \begin{cases} e^{\lambda\theta} d\theta & \text{for } \theta \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

Consider now a new control system

$$(2.2) \quad \dot{x}(t) = \lambda x(t) + \int_0^1 d\bar{\beta}(\theta) \bar{u}(t - \theta).$$

By Olbrot's criterion (see Theorem 3.2 in § 3), the system (2.2) is state-stabilizable. Setting $U(s) = \bar{U}(s)/(s - \lambda)$, we convert state-stabilizing controls in (2.2) into trajectory-stabilizing controls in (2.1). (Indeed, since $B(s)U(s) = \bar{B}(s)\bar{U}(s)$, the two systems share the same trajectories.) Given a state-stabilizing feedback law in (2.2), say of the form

$$\bar{u}(t) = Kx(t) + \int_0^1 k(\theta) \bar{u}(t - \theta) d\theta$$

(cf. Remark 4.4 herein), the following will be a dynamical trajectory—stabilizing feedback in (2.1)—

$$\dot{u}(t) = \lambda u(t) + \bar{u}(t), \quad \bar{u}(t) = Kx(t) + \int_0^1 k(\theta) \bar{u}(t - \theta) d\theta.$$

(The problem of defining the history of $\bar{u}(t)$ is dealt with in § 4.)

A second scheme is for systems in which only finitely many commensurate lags appear in the control element. In these systems $B(s) = P(e^{-\delta s})$, where $P(z)$ is a matrix valued polynomial, and δ is a positive constant. ($\delta \cdot \text{rank } P(z) = h$.) In (2.1) we have $P(z) = 1 - e^{\lambda z}$ and $\delta = 1$. Now one factors $B(s)$ as a product of polynomials in $e^{-\delta s}$. In (2.1) we may set $\bar{B}(s) = (1 - e^{\lambda-s})/(1 - e^{\lambda-s}) = 1$. There are two advantages to this technique: First, the measure $d\bar{\beta}$ maintains the simple structure of (fewer) commensurate lags. Second, the time domain realization of the relation between the functions $u(t)$ and $\bar{u}(t)$ is simpler. Here

$$u(t) = \bar{u}(t) + e^{\lambda} u(t - 1).$$

Both strategies are developed in detail and justified in the following sections.

3. Characterization of trajectory-stabilizable systems. We use the following terminology (compare with Olbrot [15]):

DEFINITION 3.1. Let γ be a given real number. The system (*) is γ -trajectory-stabilizable if given any initial data for (*) there exists a control $u(\cdot)$ and a real number δ such that the functions $t \rightarrow e^{-\gamma t} x(t)$ and $t \rightarrow e^{-\delta t} u(t)$ are of class $L_1(0, \infty)$. (In applications, both γ and δ will be negative.) If, in addition, one can always choose $\delta = \gamma$, then the system is γ -state-stabilizable. (It will be demonstrated hereafter that in the earlier case, an upper bound on δ depends on the structure of the system and not on the initial data.)

The following characterization is quoted from Olbrot [15, Thm. 1] and will be used later.

THEOREM 3.2. *System (*) is γ -state-stabilizable if and only if*

$$(3.1) \quad \text{rank} [\Delta(s), B(s)] = n \quad \text{for all complex } s \text{ with } \text{Re } s \geq \gamma,$$

where $\Delta(s)$ is the characteristic matrix of (*), defined as

$$\Delta(s) \stackrel{\text{def}}{=} sI - \int_0^h e^{-\theta s} d\alpha(\theta) \stackrel{\text{def}}{=} sI - A(s)$$

and

$$B(s) \stackrel{\text{def}}{=} \int_0^h e^{-\theta s} d\beta(\theta).$$

Note that $A(s)$ and $B(s)$ are the Laplace transforms of the measures α and β . Condition (3.1) comes into effect at spectral points of (*), where $\det \Delta(s) = 0$. We shall later remark on the meaning of Condition 3.1 as a spectral controllability requirement. An important fact [6, p. 181] is that the number of spectral points with $\text{Re } s \geq \gamma$ is finite.

Here is our basic observation.

THEOREM A. *The system (*) is γ -trajectory-stabilizable if and only if there exists an $m \times m$, nonsingular, rational matrix-valued function $R(s)$, with the following properties:*

(i) *$R(s)$ is of the form*

$$R(s) = \left(\frac{P_1}{s - \lambda_1} + Q_1 \right) \left(\frac{P_2}{s - \lambda_2} + Q_2 \right) \cdots \left(\frac{P_l}{s - \lambda_l} + Q_l \right),$$

where each of the matrices P_j is an orthogonal projection, $Q_j = I - P_j$, and each λ_j is an eigenvalue of the homogeneous part of (*) (i.e., $\det \Delta(\lambda_j) = 0$) with $\text{Re } \lambda_j \geq \gamma$.

(ii) *The function $\bar{B}(s) = B(s)R(s)$ is the Laplace transform of a finite measure, denoted by $\bar{\beta}$, and $\bar{\beta}$ is supported within $[0, h]$.*

(iii) *The system*

$$(3.2) \quad \dot{x}(t) = \int_0^h d\alpha(\theta)x(t-\theta) + \int_0^h d\bar{\beta}(\theta)\bar{u}(t-\theta)$$

is γ -state-stabilizable. By Olbrot's theorem this means:

$$(3.3) \quad \text{rank} [\Delta(s), \bar{B}(s)] = n \quad \text{for all complex } s \text{ with } \text{Re } s \geq \gamma,$$

Proof.

Sufficiency. Assume the existence of a matrix $R(s)$ as stated in the theorem. Seeking, for the time being, only an open loop stabilizing control, we may assume furthermore that $u(t) = 0$ for $t \leq 0$. (For otherwise, apply $u(t) = 0$ through $[0, h]$ and start with the new initial time $t = h$.)

By assumption, given an initial trajectory of $x(\cdot)$, there exists a γ -state-stabilizing control $\bar{u}(\cdot)$ in the system (3.2), with $\bar{u}(t) = 0$ for $t \leq 0$. Let $\bar{U}(s)$ be the Laplace transform of $\bar{u}(\cdot)$ and let $u(\cdot)$ be the inverse transform of $U(s) = R(s)\bar{U}(s)$. The latter is obtained via successive convolutions of $\bar{u}(\cdot)$ with exponential functions. It therefore satisfies the growth condition in Definition 3.1. The equality $\bar{B}(s)\bar{U}(s) = B(s)U(s)$ implies that the trajectory of (*) which corresponds to the control $u(\cdot)$ coincides with the trajectory of (3.2) for $\bar{u}(\cdot)$. Hence $u(\cdot)$ is a γ -trajectory-stabilizing control.

Necessity. We begin with the construction of the function $R(s)$. There are only finitely many eigenvalues of the homogeneous part of (*) within the half plane $\{s: \text{Re } s \geq \gamma\}$ at which Olbrot's condition is violated. The construction is performed successively, taking into account one such eigenvalue at a time.

Fix λ , the first spectral point of interest. The matrix $B(s)$ is entire. One can therefore find an integer $k \geq 0$ and a decomposition of R^m into three orthogonal subspaces:

$$R^m = \underline{U}_\infty \oplus \underline{U}_k \oplus \underline{Y}_k,$$

as follows: (i) For each vector $u \in \underline{U}_\infty$ the function $B(s)u$ vanishes throughout. (ii) For each $u \in \underline{U}_k$ ($u \neq 0$), the function $B(s)u$ has a zero at $s = \lambda$ of the exact order k (i.e., $(d/ds)^j B(s)u|_{s=\lambda} = 0$ for each $j = 0, 1, \dots, k-1$, but $(d/ds)^k B(s)u|_{s=\lambda} \neq 0$). (iii) If $v \in \underline{Y}_k$ ($v \neq 0$), then the function $B(s)v$ does not have a zero at $s = \lambda$ whose order exceeds $k-1$. The proof of existence of such a decomposition is straightforward and we leave it out.

Let P_k be the orthogonal projection on \underline{U}_k and let $Q_k = I - P_k$. At this first step we define first versions of $R(s)$ and $\bar{B}(s)$ (which will be revised successively at later steps) as $R(s) = P_k/(s - \lambda) + Q_k$ and $\bar{B}(s) = B(s)R(s)$. Note that with this definition, for each vector u , unless $\bar{B}(s)u \equiv 0$, the maximal possible order of zero of $\bar{B}(s)u$ at $s = \lambda$ is $k-1$.

Each of the following $k-1$ steps repeats the same idea: In the $1+k-j$ th step, R^m is decomposed into $\underline{U}_\infty \oplus \underline{U}_j \oplus \underline{Y}_j$; projections P_j and Q_j are defined and new versions of $\bar{B}(s)$ and $R(s)$ are obtained, multiplying the previous ones from the right by $P_j/(s - \lambda) + Q_j$. The new version of $\bar{B}(s)$ has the property that the largest possible finite order of zero of $\bar{B}(s)u$ at $s = \lambda$ is $j-1$. After k steps either $\bar{B}(s)u \equiv 0$ or $\bar{B}(\lambda)u \neq 0$. This completes the constructions due to the first eigenvalue, λ .

An important outcome of this process is that the k th version of $\bar{B}(s)$ maintains a constant rank in a vicinity of $s = \lambda$. In particular, it has an analytic generalized inverse, $B(s)^*$, at that vicinity. We shall need this fact immediately.

The construction continues in a completely similar manner for the rest of the eigenvalues in the half plane $\{s: \operatorname{Re} s \geq \lambda\}$. Of course, the process for each new eigenvalue starts with the latest versions of $\bar{B}(s)$ and $R(s)$, as they were obtained for the previous one.

Property (i) in the theorem's statement is satisfied by construction of $R(s)$. Properties (ii) and (iii) are left to be verified. Let us start with the latter. Suppose equation (*) is γ -trajectory-stabilizable and that condition (3.3) fails. Then there exists a vector $\eta \in \mathbb{R}^n$ and some eigenvalue λ with $\operatorname{Re} \lambda \geq \gamma$ such that $\eta' \Delta(\lambda) = 0$ and $\eta' \bar{B}(\lambda) = 0$. Let then $u(\cdot)$ be a γ -trajectory-stabilizing control in (*) for the following initial data: $x(0) = \eta$ and $x(t) \equiv 0$, $u(t) \equiv 0$ for $t < 0$. A Laplace transform of (*) yields

$$(3.4) \quad \Delta(s)X(s) - \eta = B(s)U(s).$$

Since $x(\cdot)$ is a γ -stabilized trajectory, the transform $X(s)$ is continuous on the closed half-plane $\{s: \operatorname{Re} s \geq \gamma\}$. Since near λ we have $\operatorname{Im} B(s) \subset \operatorname{Im} \bar{B}(s)$, we can substitute the right-hand side of (3.4) by $\bar{B}(s)\bar{U}(s)$, where $\bar{U}(s) = \bar{B}(s)^*(\Delta(s)X(s) - \eta)$. Thus defined, $\bar{U}(s)$ is also continuous near $s = \lambda$. Now, an inner product of (3.4) with the vector η completes the argument, for it implies that $\eta = 0$:

$$\|\eta\| = \eta' \Delta(\lambda)X(\lambda) - \eta' \bar{B}(\lambda)\bar{U}(\lambda) = 0.$$

Note. (i) The difference between our proof and its counterpart in Olbrot's paper is that here $U(s)$ is not necessarily continuous at $s = \lambda$. Thus, $\eta' B(\lambda) = 0$ does not a priori imply $\eta' B(s)U(s)|_{s=\lambda} = 0$. Substituting $B(s)U(s)$ by $\bar{B}(s)\bar{U}(s)$ with $\bar{U}(s)$ continuous is therefore essential.

(ii) If $\operatorname{Re} \lambda > \gamma$ then we can substitute "continuous" by "analytic" throughout.

It remains to establish that $\bar{B}(s)$ is the transform of a measure whose support is within the interval $[0, h]$. Since $\bar{B}(s)$ is defined via an inductive process, it suffices to

check the result of a single step construction:

$$\bar{B}(s) = B(s) \left(\frac{P}{s - \lambda} + Q \right) \quad \text{where } B(\lambda)P = 0.$$

Obviously, $\bar{B}(s)$ is the transform of this next measure:

$$d\bar{\beta}(\theta) = d\beta(\theta)Q + \int_0^\theta e^{\lambda(\theta-\tau)} d\beta(\tau)P d\theta.$$

It follows from the fact that $d\beta$ is supported within $[0, h]$ that the whole right-hand side vanishes for negative θ , and that the first term on the right is zero also for $\theta > h$. But when $\theta > h$, the second term vanishes too, since

$$\int_0^\theta e^{\lambda(\theta-\tau)} d\beta(\tau)P = e^{\lambda\theta} \int_0^h e^{-\lambda\tau} d\beta(\tau)P = e^{\lambda\theta} B(\lambda)P,$$

and by assumption, $B(\lambda)P = 0$. \square

COROLLARY 3.3. Suppose (*) is γ -trajectory-stabilizable and let $\lambda_1, \dots, \lambda_l$ be the eigenvalues appearing in the definition of the rational function $R(s)$, as done above. Set $\delta_0 = \max \{\gamma, \lambda_i, i = 1, \dots, l\}$. Then γ -trajectory-stabilizing controls can be chosen in (*) so that $e^{-\delta t}u(t)$ is an $L_1(0, \infty)$ -function for all $\delta > \delta_0$. In particular, if (*) is also η -state-stabilizable for some negative η (possibly $\eta > \gamma$), then exponentially decaying γ -trajectory-stabilizing controls do exist.

Proof. The corollary follows from the relation $U(s) = R(s)\bar{U}(s)$, and since $\bar{u}(t)$ can be chosen asymptotically proportional to $e^{\eta t}$. \square

Remark 3.4. (i) The corollary shows that, in applications, we would be interested in systems which are η -state-stabilizable for negative η , but in which γ -trajectory-stabilizability, for $\gamma < \eta$, enables improved stabilization rates.

(ii) For phrasing clarity we required in Theorem A (and in the proof) that the matrices P_j be **orthogonal** projections and $Q_j = I - P_j$. More compact presentations might be the result of **nonorthogonal** choice of P_j and Q_j . The properties which should be maintained are these: $\text{Im } P_j = U_j$, $\text{Im } Q_j = \underline{V}_j \oplus U_\infty$ and $\det(P_j + Q_j) \neq 0$. In Example 3.5, hereafter, we consider a system where nonorthogonal choice of P_j and Q_j in one of the construction steps yields a simpler " $\bar{B}(s)$ ".

Example 3.5. For simplicity and since our main interest is in the control element, we consider a system with no delays in the state

$$\begin{aligned} \dot{x}_1(t) &= -x_1(t) + \int_0^1 e^{-\theta} \cos(2\pi\theta) u_1(t-\theta) d\theta + u_2(t-1) - e^{-1} u_2(t-2), \\ \dot{x}_2(t) &= -2x_2(t) + \int_0^1 e^{-2\theta} \sin(2\pi\theta) u_2(t-\theta) d\theta. \end{aligned}$$

Here $m = n = 2$ and the matrix $B(s)$ is given by

$$B(s) = \begin{bmatrix} b_1(s) & e^{-s}(1 - e^{-1-s}) \\ 0 & b_2(s) \end{bmatrix},$$

where

$$b_1(s) = \int_0^1 e^{-\theta(1+s)} \cos(2\pi\theta) d\theta \quad \text{and} \quad b_2(s) = \int_0^1 e^{-\theta(2+s)} \sin(2\pi\theta) d\theta.$$

One easily notes that $b_1(s)$ has a zero of order 2 and the term $e^{-s}(1 - e^{-1-s})$ has a first order zero at $s = -1$, while $b_2(s)$ has a first order zero at $s = -2$. Thus, at both

spectral points of the homogeneous equation, Olbrot's rank condition ($\text{rank} [\Delta(s), B(s)] = 2$) fails. The system is therefore not γ -state-stabilizable for $\gamma < -1$.

Let us check whether the system is nonetheless γ -trajectory-stabilizable. (Obviously, the answer to this question will be one and the same for any $\gamma \leq -2$.) To that end we construct $\bar{B}(s)$, following the instructions in the proof of Theorem A and starting with the eigenvalue $\lambda = -1$: The subspace \underline{U}_∞ (on which $B(s)$ vanishes throughout) is empty. The first decomposition of R^2 is into a subspace on which $B(s)$ has a second order zero at $s = -1$, denoted \underline{U}_2 , and its orthogonal complement, \underline{Y}_2

$$\underline{U}_2 = \{(u_1, 0): u_1 \in R\} \quad \text{and} \quad \underline{Y}_2 = \{(0, u_2): u_2 \in R\}.$$

Accordingly,

$$P_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad Q_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad R(s) = \frac{P_2}{s+1} + Q_2$$

and the first version of $\bar{B}(s)$ is

$$\bar{B}(s) = \begin{bmatrix} b_1(s)/(s+1) & e^{-s}(1-e^{-1-s}) \\ 0 & b_2(s) \end{bmatrix}.$$

The second decomposition due to the eigenvalue $\lambda = -1$ is the same as the first: $\bar{B}(s)$ still has a zero (of order one) in the first column, while the second column does not vanish at $s = -1$. We thus update $R(s)$ as

$$R(s) = \left(\frac{P_2}{s+1} + Q_2 \right)^2 = \frac{P_2}{(s+1)^2} + Q_2$$

and redefine $\bar{B}(s) = B(s)R(s)$, namely

$$\bar{B}(s) = \begin{bmatrix} b_1(s)/(s+1)^2 & e^{-s}(1-e^{-1-s}) \\ 0 & b_2(s) \end{bmatrix}.$$

This completes the constructions due to $\lambda = -1$.

Next we consider $\bar{B}(s)$ at the eigenvalue $\lambda = -2$. As before, \underline{U}_∞ is the trivial subspace. $\bar{B}(s)$ has a first order zero on the subspace

$$\left\{ (u_1, u_2): u_1 = -\frac{e^2(1-e)}{b_1(-2)} u_2 \right\},$$

which is denoted \underline{U}_1 . Setting, for brevity,

$$c \stackrel{\text{def}}{=} \frac{e^2(1-e)}{b_1(-2)},$$

the orthogonal projections on \underline{U}_1 and on $\underline{Y}_1 = \underline{U}_1^\perp$ are

$$P_1 = \frac{1}{c^2+1} \begin{bmatrix} c^2 & -c \\ -c & 1 \end{bmatrix} \quad \text{and} \quad Q_1 = \frac{1}{c^2+1} \begin{bmatrix} 1 & c \\ c & c^2 \end{bmatrix}.$$

The final versions of $R(s)$ and $\bar{B}(s)$ are thereby the result of right multiplication of

the previous ones by $P_1/(s+2) + Q_1$. Explicitly, $\bar{B}(s)$ turns into

$$\bar{B}(s) = \frac{1}{c^2+1} \left[\begin{array}{c|c} \frac{1}{s+2} \left(\frac{c^2 b_1(s)}{(s+1)^2} - ce^{-s}(1-e^{-1-s}) \right) & \frac{1}{s+2} \left(\frac{-cb_1(s)}{(s+1)^2} + e^{-s}(1-e^{-1-s}) \right) \\ \hline + \frac{b_1(s)}{(s+1)^2} + ce^{-s}(1-e^{-1-s}) & + \frac{cb_1(s)}{(s+1)^2} + c^2 e^{-s}(1-e^{-1-s}) \\ \hline cb_2(s) \left(\frac{-1}{s+2} + 1 \right) & b_2(s) \left(\frac{1}{s+2} + c^2 \right) \end{array} \right].$$

Tracking back through our constructions we find out that now the rank condition (3.3) is satisfied:

$$\text{rank} [\Delta(s)\bar{B}(s)] = 2 \quad \text{for } s = -1, -2.$$

Our system is therefore γ -trajectory-stabilizable for all γ .

We shall not compute the measure $d\bar{\beta}$ which corresponds to $\bar{B}(s)$. Instead, following Remark 3.4, we shall compute a simpler form of $\bar{B}(s)$, using **nonorthogonal** projections in the last step: Let now P_1 and Q_1 be the following:

$$P_1 = \frac{1}{c} \begin{bmatrix} c & 0 \\ -1 & 0 \end{bmatrix} \quad \text{and} \quad Q_1 = \frac{1}{c} \begin{bmatrix} 0 & 1 \\ 0 & c \end{bmatrix}.$$

Then the final $\bar{B}(s)$ will be

$$\frac{1}{c} \begin{bmatrix} \frac{1}{s+2} \left(\frac{cb_1(s)}{(s+1)^2} - e^{-s}(1-e^{-1-s}) \right) & \frac{b_1(s)}{(s+1)^2} + ce^{-s}(1-e^{-1-s}) \\ -\frac{b_2(s)}{s+2} & cb_2(s) \end{bmatrix}.$$

This version too, satisfies the rank condition (3.3), and is simpler than the former. The measure $d\bar{\beta}$ in the corresponding reduced equation is

$$d\bar{\beta} = \begin{bmatrix} d\bar{\beta}_{11} & d\bar{\beta}_{12} \\ d\bar{\beta}_{21} & d\bar{\beta}_{22} \end{bmatrix},$$

where

$$d\bar{\beta}_{11}(\theta) = \frac{1}{c} \begin{cases} e^{-2\theta} \left[c \int_0^\theta e^\tau \int_0^\tau (\tau-\sigma) \cos(2\pi\sigma) d\sigma d\tau \right] d\theta, & 0 \leq \theta < 1, \\ e^{-2\theta} \left[c \int_0^1 e^\tau \int_0^\tau (\tau-\sigma) \cos(2\pi\sigma) d\sigma d\tau - e^2 \right] d\theta, & 1 \leq \theta \leq 2, \\ 0 & \text{otherwise,} \end{cases}$$

$$d\bar{\beta}_{12}(\theta) = \frac{1}{c} \begin{cases} \left[e^\theta \int_0^\theta (\theta-\tau) \cos(2\pi\tau) d\tau \right] d\theta, & 0 \leq \theta < 1, \\ c(\delta_1(\theta) - e^{-1}\delta_2(\theta)), & 1 \leq \theta \leq 2, \\ 0 & \text{otherwise} \end{cases}$$

(the symbol $\delta_\varepsilon(\theta)$ stands for Dirac's measure concentrated at $\theta = \varepsilon$),

$$d\bar{\beta}_{21}(\theta) = \frac{1}{c} \begin{cases} -e^{-2\theta} \left[\int_0^\theta \sin(2\pi\tau) d\tau \right] d\theta, & 0 \leq \theta \leq 1, \\ 0 & \text{otherwise} \end{cases}$$

and

$$d\bar{\beta}_{22}(\theta) = \begin{cases} e^{-2\theta} \sin(2\pi\theta) d\theta & 0 \leq \theta \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

4. Closed loop stabilization. The trajectory stabilization strategy suggested in the previous section (proof of sufficiency in Theorem A) is an open loop strategy. It is very much so, if we may put it this way, since it suits only the zero initial control. The purpose of this section is to demonstrate that this is not an inherent property and show that closed loop controllers can be constructed in trajectory stabilizable systems.

The choice of the zero initial control was convenient. First, it served in simplifying the proof of Theorem A: This way the control terms, both in (*) and in (3.2), admitted the forms of convolutions along $[0, \infty)$. Their transforms were $B(s)U(s)$ and $\bar{B}(s)\bar{U}(s)$ and there was no need to carry over the effects of initial data.

There was a second reason: The control $u(\cdot)$ was obtained from the auxiliary function $\bar{u}(\cdot)$ via a series of convolutions. The converse computation, i.e., that of $\bar{u}(\cdot)$, given $u(\cdot)$, requires differentiations. This fact creates a difficulty in the computation of the initial values of $\bar{u}(\cdot)$, given an initial $u(\cdot)$ which is not smooth enough. We shall develop hereby a closed-loop dynamic stabilizing controller in which there is no need to compute the initial trajectory of $\bar{u}(\cdot)$. Thus the difficulty will be overcome.

We have computed the following:

$$R(s) = \left(\frac{P_1}{s - \lambda_1} + Q_1 \right) \left(\frac{P_2}{s - \lambda_2} + Q_2 \right) \cdots \left(\frac{P_l}{s - \lambda_l} + Q_l \right).$$

(In this section we maintain the assumption that P_j are orthogonal projections and $Q_j = I - P_j$.) The equality $U(s) = R(s)\bar{U}(s)$ gives rise to the following set of differential equations and dependencies

$$(4.1) \quad u_0(t) = u(t),$$

$$(4.2) \quad \frac{d}{dt} P_{j+1} u_j(t) = \lambda_{j+1} P_{j+1} u_j(t) + P_{j+1} u_{j+1}(t), \quad j = 0, 1, \dots, l-1,$$

$$(4.3) \quad Q_{j+1} u_j(t) = Q_{j+1} u_{j+1}(t), \quad j = 0, 1, \dots, l-1,$$

$$(4.4) \quad u_l(t) = \bar{u}(t).$$

In [15], Olbrot suggested a dynamic γ -state-stabilizing controller which suits (3.2). In this scheme the evolution of the control $\bar{u}(\cdot)$ is governed by a delay equation

$$(4.5) \quad \frac{d}{dt} \bar{u}(t) = \int_0^h d\mu(\theta) x(t - \theta) + \int_0^h d\nu(\theta) \bar{u}(t - \theta).$$

Our plan is to derive a well-posed set of differential equations from (4.1)–(4.5), together with our main system (*). These differential equations will involve no delayed terms other than in $x(\cdot)$ and in $u(\cdot)$. They will form the desired dynamic compensator for (*).

As an intermediate step, we assert the following.

PROPOSITION 4.1. *If functions $u_j(\cdot)$, $0 \leq j \leq l$, satisfy (4.1)–(4.4), then we have*

$$\int_0^h d\beta(\theta) u(t - \theta) = \int_0^h d\bar{\beta}(\theta) \bar{u}(t - \theta).$$

PROPOSITION 4.2. *There exist $m \times m$ matrices K_j , $0 \leq j \leq l$, and a finite matrix valued measure $dk(\theta)$, such that if functions $u_j(\cdot)$, $0 \leq j \leq l$ satisfy (4.1)–(4.4), then the following representation is valid:*

$$(4.6) \quad \int_0^h d\nu(\theta) \bar{u}(t-\theta) = K_0 u(t) + \sum_{j=1}^l K_j P_j u_j(t) + \int_0^h dk(\theta) u(t-\theta).$$

PROPOSITION 4.3. *Let Y be the product space*

$$Y \stackrel{\text{def}}{=} P_1 R^m \times P_2 R^m \times \cdots \times P R^m.$$

Then there exist linear transformations $C: Y \rightarrow R^m$, D and $E: R^m \rightarrow R^m$, $F: Y \rightarrow Y$, and G and $H: R^m \rightarrow Y$, with the following properties:

(i) *This next set of differential equations (4.7)–(4.8) form, together with (*), a well-posed system for the initial data*

$$(x(0), u(0), y(0), x(\cdot), u(\cdot)) \in R^n \times R^m \times Y \times L_2([-h, 0], R^n) \times L_2([-h, 0], R^m),$$

$$(4.7) \quad \dot{u}(t) = Cy(t) + Du(t) + E \left(\int_0^h d\mu(\theta) x(t-\theta) + K_0 u(t) + \int_0^h dk(\theta) u(t-\theta) \right),$$

$$(4.8) \quad \dot{y}(t) = Fy(t) + Gu(t) + H \left(\int_0^h d\mu(\theta) x(t-\theta) + K_0 u(t) + \int_0^h dk(\theta) u(t-\theta) \right).$$

(ii) Given $u(t)$ and $y(t)$, let $u_j(t)$, $0 \leq j \leq l$, be as follows: $u_0 = u(t)$, $P_j u_j(t) = y(t)$ (=the component of $y(t)$ in the space $P_j R^m$) and $Q_j u_j(t)$ be defined successively by (4.3) for $j = 1, 2, \dots, l$. Suppose $x(\cdot)$, $u(\cdot)$ and $y(\cdot)$ form a solution of (*), (4.7) and (4.8). Then the function $x(\cdot)$ and $u_j(\cdot)$, $j = 0, 1, \dots, l$ satisfy equations (*) and (4.1)–(4.5) for $t \geq l \cdot h$. Conversely, assume that the latter system is solved by $x(\cdot)$ and $u_j(\cdot)$, $0 \leq j \leq l$; then the corresponding triple $x(\cdot)$, $u(\cdot)$ and $y(\cdot)$ solve (*), (4.7) and (4.8).

In the proof of Proposition 4.3, hereafter, a precise recipe for building the operators C , D , E , F , G and H is given. But first we summarize our findings.

THEOREM B. *Suppose (*) is a γ -trajectory-stabilizable system. Here is a dynamic feedback scheme which generates γ -trajectory-stabilizing controls:*

- (i) *Determine the initial values $u(0)$ and $y(0)$, arbitrarily.*
- (ii) *At each $t \geq 0$ let the dynamics of the control $u(\cdot)$ and of the auxiliary function $y(\cdot)$ be determined by (4.7) and (4.8).*
- (iii) *Substitute $u(\cdot)$ into (*).*

We give the proofs in a reverse order.

Proof of Theorem B. As is borne out by Proposition 4.3, if $x(\cdot)$, $u(\cdot)$ and $y(\cdot)$ satisfy (*), (4.7) and (4.8), then the corresponding functions $u_j(\cdot)$ satisfy (4.1)–(4.5). Furthermore, by Proposition 4.1 the pair $x(\cdot)$ and $\bar{u}(\cdot)$ then defines a solution of (3.2) and (4.5). By Olbrot's construction (that is, by the choice of the measures $d\mu(\theta)$ and $d\nu(\theta)$ in (4.5)) the latter system is γ -state-stabilized. In particular, the trajectory of $x(\cdot)$ is γ -stabilized. \square

Proof of Proposition 4.3. No matter what the operators (in fact, the matrices) C , D , E , F , G and H are, part (i) of the proposition is obvious: The well-established theory (see, e.g., [2], [6]) tells us that the product space presentation is well posed.

The following two decompositions will be handy along with the proof of part (ii): Suppose that functions $u_j(\cdot)$, $0 \leq j \leq l$, satisfy (4.1)–(4.5). Then

$$(4.9) \quad u_j(t) = P_j u_j(t) + \sum_{i=1}^{j-1} Q_j \cdots Q_{i+1} P_i u_i(t) + Q_j \cdots Q_1 u_0(t), \quad 1 \leq j \leq l$$

(the term with the \sum does not appear when $j = 1$) and

$$(4.10) \quad u_j(t) = P_{j+1}u_j(t) + \sum_{i=j+1}^{l-1} Q_{j+1} \cdots Q_i P_{i+1}u_i(t) + Q_{j+1} \cdots Q_l u_l(t), \quad 0 \leq j \leq l$$

(the \sum expression disappears now for $j = l-1$). Both (4.9) and (4.10) are obtained by successive applications of (4.3), which is the relation $Q_{j+1}u_{j+1} = Q_{j+1}u_j$.

Thereby, for each j between 0 and $l-1$, our assumptions imply

$$(4.10) \quad \Rightarrow \frac{d}{dt} P_j u_j(t) = \frac{d}{dt} P_j \left(P_{j+1} u_j(t) + \sum_{i=j+1}^{l-1} Q_{j+1} \cdots Q_i P_{i+1} u_i(t) \right) + \frac{d}{dt} P_j Q_{j+1} \cdots Q_l u_l(t)$$

$$(4.2) \quad \Rightarrow = P_j \left(\lambda_{j+1} P_{j+1} u_j(t) + P_{j+1} u_{j+1}(t) + \sum_{i=j+1}^{l-1} Q_{j+1} \cdots Q_i (\lambda_{i+1} P_{i+1} u_i(t) + P_{i+1} u_{i+1}(t)) \right)$$

$$(4.5) \quad \Rightarrow + P_j Q_{j+1} \cdots Q_l \left(\int_0^h d\mu(\theta) x(t-\theta) + \int_0^h d\nu(\theta) u(t-\theta) \right) = P_j P_{j+1} u_{j+1}(t) + \sum_{i=j+1}^{l-1} P_j Q_{j+1} \cdots Q_i P_{i+1} u_{i+1}(t),$$

$$(4.9) \quad \Rightarrow + \lambda_{j+1} P_j P_{j+1} \left(P_j u_j(t) + \sum_{k=1}^{j-1} Q_j \cdots Q_{k+1} P_k u_k(t) + Q_j \cdots Q_1 u_0(t) \right) + \sum_{i=j+1}^{l-1} \lambda_{i+1} P_j Q_{j+1} \cdots Q_i P_{i+1} \cdot \left(P_i u_i(t) + \sum_{k=1}^{i-1} Q_i \cdots Q_{k+1} P_k u_k(t) + Q_i \cdots Q_1 u_0(t) \right)$$

$$\text{Proposition 4.2} \quad \Rightarrow + (P_j Q_{j+1} \cdots Q_l) \left(\int_0^h d\mu(\theta) x(t-\theta) + K_0 u_0(t) + \sum_{i=1}^l K_i P_i u_i(t) + \int_0^h dk(\theta) u_0(t-\theta) \right).$$

The last expression involves only the functions $x(\cdot)$ and $u_0(\cdot)$, and the auxiliary vectors $y_i(t) = P_i u_i(t)$, $1 \leq i \leq l$, which is the desired formulation. Letting j run through $0, \dots, l-1$, we obtain explicitly most entries in the operators C, D, E, F, G and H . (Obviously, a more compact and orderly representation is obtained when the coefficients of each $y_i = P_i u_i$ are gathered. We shall not do it here.)

For the missing entries it remains to consider the derivatives of $Q_1 u_0(t)$ and of $P_i u_i(t)$. The computations of these terms follow exactly the same lines of those we have just displayed and we leave them to the interested reader.

Finally, one should note that the various decompositions we made above go both ways. Thus, if $x(\cdot)$, $u(\cdot)$ and $y(\cdot)$ satisfy (4.7) and (4.8), then $y(\cdot)$ can be decomposed

as $(P_1 u_1(\cdot), \dots, P_l u_l(\cdot))'$, so that with the definitions $u_0(\cdot) = u(\cdot)$ and $Q_{j+1} u_{j+1}(\cdot) = Q_{j+1} u_j(\cdot)$, a set of functions which satisfy (4.1)–(4.5) is formed. \square

Proof of Proposition 4.2. We shall not bring at this point the interesting details, the justifications, or the exact notations of Olbrot's and Pandolfi's developments. These can be found in [15] and in [16], respectively. What concerns us here is the following conclusion of their findings: The measure $d\nu(\theta)$ in the dynamic stabilizer (4.5) can be built in the form

$$(4.11) \quad d\nu(\theta) = F_0 \delta_0(\theta) + \int_0^h F(\tau - \theta) d\bar{\beta}(\tau) d\theta.$$

Here F_0 and $F(\theta)$ are $m \times m$ matrices, $\delta_0(\theta)$ is the Dirac measure which is concentrated at $\theta = 0$, and $\theta \rightarrow F(\theta)$ is an analytic function.

Here is an outline of the proof: The structure (4.11) of $d\nu(\theta)$ implies that

$$(4.12) \quad \begin{aligned} \int_0^h d\nu(\theta) \bar{u}(t - \theta) &= \int_0^h d\nu(\theta) u_l(t - \theta) \\ &= F_0 u_l(t) + \int_0^h \int_0^h F(\tau - \theta) d\bar{\beta}(\tau) u_l(t - \theta) d\theta \\ &= F_0 u_l(t) + \int_0^h F(\theta) \int_0^h d\bar{\beta}(\tau) u_l(t + \theta - \tau) d\theta. \end{aligned}$$

Using formula (4.9) we decompose the first term on the right-hand side of (4.12) into a sum which involves $u_0(t)$ and $y_j(t) = P_j u_j(t)$. With the aid of integration by parts, $d\bar{\beta}(\tau) u_l(t + \theta - \tau)$ will be substituted by $d\beta(\tau) u_0(t + \theta - \tau)$, in the second expression, while more terms, involving $u_0(t)$ and $y_j(t)$, will be created as side products (similar, yet even simpler computations will serve later in establishing Proposition 4.1).

The following notations should clarify our computations: Set first $d\beta_0(\theta) = d\beta(\theta)$. Then define recursively

$$d\beta_j(\theta) = d\beta_{j-1}(\theta) Q_j + \int_0^\theta e^{\lambda_j(\theta-\tau)} d\beta_{j-1}(\tau) P_j d\theta$$

for $j = 1, \dots, l$. Consequently, $d\bar{\beta}(\theta) = d\beta_l(\theta)$.

Now we begin the iterative process of integrations by parts. The first step goes this way:

$$(4.13) \quad \begin{aligned} &\int_0^h d\beta_l(\tau) u_l(t + \theta - \tau) \\ &= \int_0^h d\beta_{l-1}(\tau) Q_l u_l(t + \theta - \tau) + \int_0^h \int_0^\tau e^{\lambda(\tau-\sigma)} d\beta_{l-1}(\sigma) P_l u_l(t + \theta - \tau) d\tau \\ &= \int_0^h d\beta_{l-1}(\tau) Q_l u_{l-1}(t + \theta - \tau) \\ &\quad + \int_0^h \int_0^\tau e^{\lambda(\tau-\sigma)} d\beta_{l-1}(\sigma) P_l (\dot{u}_{l-1}(t + \theta - \tau) - \lambda u_{l-1}(t + \theta - \tau)) d\tau \\ &= \int_0^h d\beta_{l-1}(\tau) Q_l u_{l-1}(t + \theta - \tau) + \int_0^h d\beta_{l-1}(\tau) P_l u_{l-1}(t + \theta - \tau) \\ &\quad - \int_0^\tau e^{\lambda(\tau-\sigma)} d\beta_{l-1}(\sigma) P_l u_{l-1}(t + \theta - \tau) \Big|_{\tau=\theta}^{\tau=h} \\ &= \int_0^h d\beta_{l-1}(\tau) u_{l-1}(t + \theta - \tau) + \left(\int_0^\theta e^{\lambda(\theta-\sigma)} d\beta_{l-1}(\sigma) P_l \right) u_{l-1}(t). \end{aligned}$$

(Recall that the integral $\int_0^h e^{\lambda(\theta-\sigma)} d\beta_{l-1}(\sigma)P_l$ vanishes; this was explained in the proof of Theorem A.) The term which involves $u_{l-1}(t)$ will be decomposed, using (4.9), into a sum of expressions in $u_0(t)$ and $P_j u_j(t)$, $1 \leq j \leq l-1$.

The second iteration is applied to the integral on the right-hand side (4.13). After l steps the process terminates with the desired result. \square

Proof of Proposition 4.1. Using the notations of the previous proof, we have to verify the equality

$$\int_0^h d\beta_l(\tau)u_l(t-\tau) = \int_0^h d\beta_0(\tau)u_0(t-\tau).$$

It follows from iterative application of (4.13), choosing $\theta = 0$. \square

Remark 4.4. The basis for our constructions of a closed-loop controller in this section is Olbrot's **dynamic** controller (4.5). One may prefer to start with a "**static**" controller for system (3.2), i.e., with a closed-loop relation of the form

$$(4.14) \quad \bar{u}(t) = \int_0^h d\xi(\theta)x(t-\theta) + \int_0^h d\zeta(\theta)\bar{u}(t-\theta).$$

The extensive literature on the subject (and most of the papers cited below) offer stabilizing controllers of the type (4.14), for various subclasses of hereditary control systems. We wish to point out that also, in the general case of Olbrot's theorem, the stabilizer can be constructed that way. Instead of presenting a rigorous proof, let us mention just the main arguments.

As in Pandolfi [16], one considers the spectral projection of the γ -state-stabilizable system on the united generalized eigensubspace for the eigenvalues in the half plane $\{s: \operatorname{Re} s \geq \gamma\}$. Using Pandolfi's notation, it is this next ODE,

$$(4.15) \quad \dot{c}(t) = Hc(t) + \phi(0) \int_0^h d\bar{\beta}(\theta)\bar{u}(t-\theta).$$

Olbrot's spectral condition (3.1) implies that the system (4.15) is state-stabilizable. It thus has (see, e.g., [12], [1]) a feedback stabilizing controller of the form

$$(4.16) \quad \bar{u}(t) = F(c(t) + \int_0^h \int_0^h e^{H(\theta-\tau)} \phi(0) d\bar{\beta}(\tau)\bar{u}(t-\theta) d\theta).$$

Finally, since $c(t)$ is given by the spectral projection of the h -history of $x(\cdot)$ at t (i.e., of the function $\theta \rightarrow x(t-\theta)$, $0 \leq \theta \leq h$) on the relevant eigenmode, the desired form of (4.14) is obtained.

5. Systems with commensurate lags. A significant simplification can be obtained, both in the reduction of β to $\bar{\beta}$, and in the stabilization scheme, when the delays in the control term consist of commensurate lags. The main idea for the reduction has already been presented in the analysis of the simple example in § 2. We repeat it here in more detail, using the terminology of Theorem A and its proof.

Let δ now stand for the basic lag in the control term. Then $B(s)$ can be written in the form

$$(5.1) \quad B(s) = P(e^{-\delta s}),$$

where $P(z)$ is a matrix valued polynomial. This representation suggests the change of parameters $z = e^{-\delta s}$, which maps the half plane $\{s: \operatorname{Re} s \geq \gamma\}$ onto the punctured disc

$\{z: e^{-\delta\gamma} \geq |z| > 0\}$. Here is a variant of Theorem A for the present particular case.

THEOREM C. *Suppose $B(s)$ can be represented in the form (5.1) and set $z = e^{-\delta s}$. Then (*) is γ -state-stabilizable if and only if there exists an $m \times m$ nonsingular rational matrix-valued function $S(z)$, with the following properties:*

(i) $S(z)$ is of the form

$$S(z) = \left(\frac{P_1}{z_1 - z} + Q_1 \right) \cdots \left(\frac{P_l}{z_l - z} + Q_l \right),$$

where each of the matrices P_j is an orthogonal projection, $Q_j = I - P_j$, and each $z_j = e^{-\delta\lambda_j}$ where λ_j is a spectral point of (*) within $\{s: \operatorname{Re} s \geq \gamma\}$.

(ii) The function $\bar{P}(z) = P(z)S(z)$ is a matrix polynomial.

(iii) $\operatorname{rank} [\Delta(z), \bar{P}(z)] = n$ for all z with $e^{-\delta\gamma} \geq |z| > 0$.

The proof is by now obvious.

As pointed out in § 2, one of the advantages of the present case is the existence of simple recursive relations between the auxiliary control $\bar{u}(\cdot)$ (in the reduced system, where $\bar{P}(z)$ substitutes $P(z)$) and the true control function $u(\cdot)$. These relations, as implied by the equality $U(s) = S(e^{-\delta s})\bar{U}(s)$, can be stated in a fashion resembling (4.1)–(4.4):

$$(5.2) \quad u_0(t) = u(t),$$

$$(5.3) \quad P_{j+1}u_j(t) = e^{\delta\lambda_{j+1}}P_{j+1}(u_{j+1}(t) + u_j(t - \delta)), \quad j = 0, 1, \dots, l-1,$$

$$(5.4) \quad Q_{j+1}u_j(t) = Q_{j+1}u_{j+1}(t), \quad j = 0, 1, \dots, l-1,$$

$$(5.5) \quad \bar{u}(t) = u_l(t).$$

By repeated application of (5.4) into (5.3) we obtain a recursive definition of the form

$$u(t) = L_0\bar{u}(t) + \sum_{j=1}^l L_j u(t - j\delta).$$

Formulas (5.2)–(5.5) also enable the computation of appropriate initial values of $\bar{u}(\cdot)$ from those of $u(\cdot)$. Thus, any feedback stabilizer for the auxiliary system easily translates into a feedback stabilizer for (*).

6. Geometrical interpretations. Equivalence between controllability and stabilizability properties is a basic feature of ordinary, as well as of hereditary systems. Here are geometrical-controllability interpretations of state- and trajectory-stabilizability.

PROPOSITION 6.1. *The system (*) is γ -state-stabilizable if and only if given any positive T , γ -state-stabilizing controls can be chosen with $u(t) \equiv 0$ for $t > T$.*

Proof. The claim follows since Olbrot's condition is equivalent to spectral controllability of the united eigenspace for eigenvalues λ with $\operatorname{Re} \lambda \geq \gamma$. Let us bring some detail for completeness.

The spectral projection of (*) on the said eigenspace is the following (recall Pandolfi's construction, as described in Remark 4.4):

$$(6.1) \quad \dot{c}(t) = Hc(t) + \psi(0) \int_0^h d\beta(\theta) u(t - \theta).$$

Here the eigenvalues of H are those of (*) within the half-plane of interest, and $c(t)$ is the projection onto the corresponding eigenspace of the h -history of $x(\cdot)$ at t . Hence $x(\cdot)$ is γ -stabilized if $c(t)$ is brought to rest. It follows from [12, Thm. 2.2] that Olbrot's criterion (i.e., γ -state-stabilizability of (*)) implies controllability of this next ordinary system

$$(6.2) \quad \dot{y}(t) = Hy(t) + \tilde{B}u(t),$$

where

$$\tilde{B} \stackrel{\text{def}}{=} \int_0^h e^{-H\theta} \psi(0) d\beta(\theta).$$

From [1, Thm. 3.3] we have that if $y(0) = c(0)$ and $u(t) = 0$ for $t > T$, then $y(t)$ and $c(t)$ coincide for $t > T + h$. In particular, if $y(t) = 0$ for $t > T$, which is possible when (6.2) is controllable, $c(t)$ can be brought to rest at $t = T + h$ by a control which vanishes past $t = T$. This proves the “only if” part. The “if” claim is obvious. \square

In trajectory-stabilizable systems we made weaker assumptions on the behaviours of stabilizing controls. This fact is reflected in the corresponding controllability property.

THEOREM D. *The system (*) is γ -trajectory-stabilizable if and only if given any positive T and $\delta > \delta_0$ ($\delta_0 = \max \{ \gamma, \lambda : \lambda \text{ is an eigenvalue of } (*) \text{ with } \operatorname{Re} \lambda > \gamma \text{ and } \operatorname{rank} [\Delta(\lambda), B(\lambda)] < n \}$), as defined in Corollary 3.3), the spectral projection $c(t)$ can be brought to rest past $t = T + h$, using a control $u(t)$ with $e^{-\lambda t} u(t) \in L_1[0, \infty)$; in fact, past $t = T$ the values of $u(t)$ will be governed by an autonomous differential (or difference) equation.*

Proof. Sufficiency is, again, obvious. For the proof of necessity, derive a γ -state-stabilizable system from (*), following the instructions in the proofs of Theorems A and C. Given any initial values for $c(t)$ and $\bar{u}(t)$, it is possible to bring $c(t)$ to rest at $t = T + h$, using nonzero $\bar{u}(t)$ only through $[0, T]$. (This follows from Proposition 6.1.) The obvious idea is thereby to find the appropriate $\bar{u}(t)$ and convert it into $u(t)$, using one of the two schemes mentioned above.

The procedure is simpler when the technique of § 5 is applied, since initial values for $\bar{u}(t)$ are readily obtained from those of $u(t)$. As mentioned in the beginning of § 4, this might be harder in the more general case, when the scheme of § 3 is employed. The difficulty is circumvented as follows: an observation which is complementary to [1, Thm. 3.3] (which is quoted in the proof of Prop. 6.1) is that for $t \geq h$, the effect of the initial control can be absorbed in that of $c(0)$. Precisely, when substituting $c(0)$ by $\bar{c}(0) = c(0) + \int_0^h \int_\theta^h e^{H(\theta-\tau)} \phi(0) d\beta(\tau) u(-\theta) d\theta$, and then $u(t) = 0$ for $t \leq 0$, the new trajectory of $c(t)$ will not be different from the original one (i.e., for $c(0)$ and the true initial control) past $t = h$. So one can choose $\bar{u}(t)$ which brings the transformed system to rest at $t = T + h$, given the initial values $\bar{c}(0)$ and $\bar{u}(t) = 0$, $t \leq 0$. The inverse transform of $R(s)\bar{U}(s)$ will provide the needed values of $u(t)$ for $t \geq 0$. \square

Example 6.2. Consider again the simple example in § 2:

$$\dot{x}(t) = \lambda x(t) + u(t) - e^\lambda u(t-1).$$

This system is already in the form of a spectral projection on the eigenspace of the eigenvalue λ . Employing the technique of § 5, it converts to

$$\dot{\bar{x}}(t) = \lambda \bar{x}(t) + \bar{u}(t)$$

with

$$u(t) = \bar{u}(t) + e^\lambda u(t-1).$$

Take $\bar{u}(t)$ to be, e.g., this next step function:

$$\bar{u}(t) = \begin{cases} -\frac{\lambda x(0)}{1 - e^{-\lambda T}}, & 0 \leq t \leq T, \\ 0, & T < t. \end{cases}$$

Then for $t > T$ the control function $u(t)$ satisfies the difference equation $u(t-1) = e^\lambda u(t)$, and $x(t)$ will vanish past $t = T$. (This situation is better than that stated in the

theorem, since in the present case there is no delay in $\bar{u}(t)$.)

When using the technique of § 3 (which suits the general form of (*)), the transformed system is

$$\dot{x}(t) = \lambda x(t) + \int_0^1 e^{\lambda\theta} \bar{u}(t - \theta) d\theta.$$

If $x(0)$ and $u(t)$, $-h \leq t \leq 0$, are the prescribed initial values, we do the computations for

$$\bar{x}(0) = x(0) - \int_0^1 e^{\lambda\theta} u(-\theta) d\theta,$$

and the zero initial control. Following the recipe in the proof of Proposition 6.1, we set

$$\tilde{B} = \int_0^1 e^{-H\theta} \phi(0) d\tilde{\beta}(\theta) = \int_0^1 e^{-\lambda\theta} e^{\lambda\theta} d\theta = 1.$$

The current counterpart of (6.2) is

$$\dot{y}(t) = \lambda y(t) + \bar{u}(t),$$

with $y(0) = \bar{x}(0)$. This system is brought to zero at $t = T$, using, e.g., the control function

$$\bar{u}(t) = \begin{cases} -\frac{\lambda \bar{x}(0)}{1 - e^{-\lambda T}}, & 0 \leq t \leq T, \\ 0, & T < t. \end{cases}$$

So with this $\bar{u}(t)$, the true trajectory (in the transformed system) is brought to rest at $t = T + 1$. The appropriate control $u(t)$ is the inverse transform of $\bar{U}(s)/(s - \lambda)$, i.e., it is given by

$$u(t) = \int_0^t e^{\lambda(t-\tau)} \bar{u}(\tau) d\tau.$$

For $t \geq T$, the function $u(t)$ satisfies the homogeneous ODE $\dot{u}(t) = \lambda u(t)$.

7. Conclusions. Trajectory-stabilizability is a phenomenon which is specific to systems with delayed control action. It is of interest both from an application-oriented and from a purely mathematical point of view. We provide general characterizations of trajectory-stabilizable systems and of their trajectory-stabilizing control functions.

It is often very difficult to handle problems of "spectral-type" (e.g., to locate eigenvalues, find zeros of $B(s)$ etc.) in hereditary systems of unconstrained form, such as (*). Our analysis of the commensurate, multiple-lag case might therefore be of greater value in applications. Let us mention also that similar difficulties, which are inherent to delay systems, are treated in the literature, in different contexts (see, e.g., [12], [13]), and that a variety of available techniques could be adopted in particular classes of systems.

REFERENCES

- [1] Z. ARTSTEIN, *Linear systems with delayed controls: A reduction*, IEEE Trans. Automat. Control, 27 (1982), pp. 869-879.
- [2] JU. BORISOVIC AND A. S. TURBABIN, *On the Cauchy problem for linear nonhomogeneous differential equations with retarded arguments*, Soviet Math. Dokl., 10 (1969), pp. 401-405.
- [3] M. C. DELFOUR, C. MCCALLA AND S. K. MITTER, *Stability and the infinite time quadratic cost problem for linear hereditary differential equations*, this Journal, 13 (1975), pp. 48-88.

- [4] E. EMRE, *On necessary and sufficient conditions for regulation of linear systems over rings*, this Journal, 20 (1982), pp. 155–160.
- [5] E. EMRE AND G. J. KNOWLES, *Control of linear systems with fixed noncommensurate point delays*, IEEE Trans. Automat. Control, 29 (1984), pp. 1083–1090.
- [6] J. K. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, Berlin-New York, 1977.
- [7] E. W. KAMEN, *Linear systems with commensurate time delays: Stability and stabilization independent of delay*, IEEE Trans. Automat. Control, 27 (1982), pp. 367–375.
- [8] ———, *Correction to “Linear systems with commensurate time delays: Stability and stabilization independent of delay,”* IEEE Trans. Automat. Control, 28 (1983), pp. 248–249.
- [9] E. W. KAMEN, P. P. KHARGONEKAR AND A. TANNENBAUM, *Pointwise stability and feedback control of linear systems with noncommensurate time delays*, Acta Math. Appl., 2 (1984), pp. 159–184.
- [10] W. H. KWON AND E. PEARSON, *Feedback stabilization of linear systems with delayed control*, IEEE Trans. Automat. Control, 25 (1980), pp. 266–269.
- [11] R. H. KWONG, *Stability theory for the linear-quadratic-Gaussian problem with delays in the state, control and observation*, this Journal, 18 (1980), pp. 49–75.
- [12] A. MANITIUS AND A. W. OLBROT, *Finite spectrum assignment problem for systems with delays*, IEEE Trans. Automat. Control, 24 (1979), pp. 541–553.
- [13] A. MANITIUS AND R. TRIGGIANI, *Function space controllability and feedback stabilizability of linear retarded systems*, IEEE Trans. Automat. Control, 23 (1978), pp. 659–665.
- [14] D. A. O’CONNER AND T. L. TARN, *On stabilization by state feedback for neutral difference differential equations*, IEEE Trans. Automat. Control, 28 (1983), pp. 615–618.
- [15] A. W. OLBROT, *Stabilizability, detectability and spectrum assignment for linear autonomous systems with general time delays*, IEEE Trans. Automat. Control, 23 (1978), pp. 887–890.
- [16] I. PANDOLFI, *On feedback stabilization of functional differential equations*, Boll. Un. Mat. Ital., 4 (1975), pp. 626–635.

SUFFICIENT CONDITIONS IN FREE FINAL TIME OPTIMAL CONTROL PROBLEMS*

ATLE SEIERSTAD†

Abstract. A direct sufficient condition for the optimality of a control in standard control problems with free final time is proved. The basic tool is superderivative properties of the optimal value of the criterion as a function of the final time and the final point. Sufficient conditions arise if the superderivatives have proper signs.

Key words. sufficient conditions, optimal control

AMS(MOS) subject classification. 49B10

1. Introduction. Sufficient conditions in free final time control problems are not as easily constructed as in fixed final time problems. Some conditions have been given by Mereau and Powers [1] (see also Peterson and Zalkind [2]).

In this paper new conditions are given, which in our opinion are more useful. They are more closely related to actual solution procedures, and they seem to be able to decide optimality in broader classes of problems. In [3], results in less general circumstances have been proved in the case of no restrictions on the time path of the system.

While sufficient conditions in suitably concave, fixed time problems are “almost” necessary, in free time problems sufficient conditions cannot come that close to necessary conditions, due to an inherent lack of concavity properties.

2. First we shall consider a control problem with no time path restrictions. Let the abbreviation v.e. (virtually everywhere) mean “for all t except a finite number.” Consider the problem

$$(1) \quad \text{Maximize } \int_{t_0}^{\tau} f^0(x(t), u(t), t) dt, \quad \tau \in I = [\tau_1, \tau_2]$$

where t_0 , τ_1 and τ_2 are fixed points, $t_0 < \tau_1 < \tau_2$.

$$(2) \quad \dot{x} = f(x, u(t), t) \quad \text{v.e.}, \quad x(t_0) = x_0, \quad x_0 \text{ fixed in } R^n,$$

$$(3) \quad u(t) \in U, \quad U \text{ a fixed subset of } R^r,$$

$$x^i(\tau) = x_1^i, \quad i = 1, \dots, l,$$

$$(4) \quad x^i(\tau) \geq x_1^i, \quad i = l+1, \dots, m,$$

$$x^i(\tau) \text{ free}, \quad i = m+1, \dots, n.$$

The optimization problem consists of choosing a piecewise continuous¹ control of function and its domain of definition $[t_0, \tau]$, $\tau \in I$, in such a way that the triple $(x(\cdot), u(\cdot), \tau)$ is admissible (i.e., satisfies (2), (3) and (4) and such that it maximizes the integral in (1). The functions f^0 , f , f_x^0 and f_x are assumed to be continuous on $R^n \times R^r \times R$ (f^0 takes values in R , f in R^n , f_x^0 , f_x are partial derivatives with respect to x , assumed to exist for all (x, u, t)).

* Received by the editors February 18, 1985; accepted for publication (in revised form) June 24, 1986.

† Institute of Economics, University of Oslo, Blindern, Oslo 3, Norway.

¹ Piecewise continuous means having one-sided limits everywhere, and a finite number of discontinuity points.

If $(\bar{x}(\cdot), \bar{u}(\cdot), \tau^*)$ is an optimal triple in the above problem, it satisfies the following necessary conditions: For some number $p_0 \geq 0$, and some function $p(t)$, $(p_0, p(t)) \neq 0$ everywhere, we have

$$(5) \quad H(\bar{x}(t), \bar{u}(t), p(t), t) = \max_{u \in U} H(\bar{x}(t), u, p(t), t) \quad \text{for v.e. } t$$

where $H(x, u, p, t) = p_0 f^0(x, u, t) + p f(x, u, t)$. The function $p(\cdot)$ satisfies

$$(6) \quad \dot{p} = -H_x(\bar{x}(t), \bar{u}(t), p(t), p) \quad \text{v.e.}$$

Finally,

$$(7) \quad \begin{aligned} p^i(\tau^*) & \text{ no condition,} & i = 1, \dots, l, \\ p^i(\tau^*) & \geq 0 (=0 \text{ if } \bar{x}^i(\tau^*) > x_1^i), & i = l+1, \dots, m, \\ p^i(\tau^*) & = 0, & i = m+1, \dots, n, \end{aligned}$$

$$(8) \quad H(\bar{x}(\tau^*), \bar{u}(\tau^*), p(\tau^*), \tau^*)[\tau - \tau^*] \leq 0 \quad \text{for all } \tau \in I.$$

When two additional conditions are added to (5)–(7), we get a set of conditions that are sufficient in the problem where τ is kept fixed equal to τ^* (see [5]:²

$$(9) \quad p_0 = 1.$$

The supremum

$$(10) \quad H^*(x, p(t), t) = \sup_{u \in U} H(x, u, p(t), t)$$

is finite for all (x, t) and is a concave function of x , for each t .

3. Consider for a moment the fixed final time control problem that arises if we choose a fixed τ in I . Assume that we have found a triple $(x^\tau(\cdot), u^\tau(\cdot), p^\tau(\cdot))$ defined on $[0, \tau]$ such that (2)–(7), (9), and (10) hold for τ^* replaced by τ . Then the pair $(x^\tau(\cdot), u^\tau(\cdot))$ is optimal in this fixed final time problem.

Hence, if $(x(\cdot), u(\cdot))$ is an arbitrary admissible pair defined on $[0, \tau]$, the value of the criterion for this pair is smaller than or equal to the value $V(\tau)$ of the criterion for the pair $(x^\tau(\cdot), u^\tau(\cdot))$. Let us assume that such a pair $(x^\tau(\cdot), u^\tau(\cdot))$ exists for all $\tau \in I$. Let us furthermore assume that for some $\tau^* \in I$, $V(\tau) \leq V(\tau^*)$ for all $\tau \in I$. Then evidently the triple $(x^{\tau^*}(\cdot), u^{\tau^*}(\cdot), \tau^*)$ is optimal.

For the last inequality to hold, a natural sufficient condition to introduce would be $(d/d\tau)V(\tau) \leq 0$ if $\tau \geq \tau^*$, $(d/d\tau)V(\tau) \geq 0$ if $\tau \leq \tau^*$. Implicitly, such a condition requires differentiability of $V(\tau)$. What we actually need is that a superderivative (in some sense) satisfies such inequalities. Condition (12) below is of this type. Furthermore, it turns out that V is always “superdifferentiable”; no ad hoc assumption for this end is needed.

THEOREM 1. Assume that for each $\tau \in I$ there exists a triple $(x^\tau(\cdot), u^\tau(\cdot), p^\tau(\cdot))$ defined on $[t_0, \tau]$ satisfying (2)–(7), (9), (10) (i.e., $u^\tau(\cdot)$ is optimal for τ fixed).

Assume also that all $u^\tau(\cdot)$, $\tau \in I$ take values in a fixed bounded subset of U , that $\{p^\tau(\tau): \tau \in I\}$ is a bounded set, and that $\tau \rightarrow x^\tau(\tau)$ is continuous. Assume finally that the function

$$(11) \quad d(\tau) = H(x^\tau(\tau), u^\tau(\tau), p^\tau(\tau), \tau) \quad (\text{with } p_0 = 1)$$

has the property that there exists a $\tau^* \in I$, such that

$$(12) \quad \begin{aligned} d(\tau) & \geq 0 \quad \text{for } \tau \leq \tau^* & \text{if } \tau_1 < \tau^*, \\ d(\tau) & \leq 0 \quad \text{for } \tau \geq \tau^* & \text{if } \tau_2 > \tau^*. \end{aligned}$$

Then the triple $(x^{\tau^*}(\cdot), u^{\tau^*}(\cdot), \tau^*)$ is optimal.

² In [5] max, instead of sup, is used in (10); the proof, however, remains the same.

Remark 1. If τ is required to belong to an interval $[\tau_1, \infty)$ instead of I , and if $\tau^* \in [\tau_1, \infty)$ has the property that the conditions of Theorem 1 are satisfied for each bounded subinterval I of $[\tau_1, \infty)$ containing τ^* , then the triple $(x^{\tau^*}(\cdot), u^{\tau^*}(\cdot), \tau^*)$ is optimal. (An example showing the application of the conditions in this remark can be found at the end of this paper.)

Proof. From the remarks preceding the theorem, it suffices to show that $V(\tau) \leq V(\tau^*)$ for all $\tau \in I$. The essential part of a proof is made up of the following theorem and the subsequent Remark 2. (To obtain this remark, we have had to repeat—in a more elaborate form—arguments briefly sketched in [4, Remark].)

In the theorem below, $V(y, \tau)$ is the supremum (possibly $= \infty$) of the criterion evaluated for all pairs $x(\cdot), u(\cdot)$ defined on $[t_0, \tau]$, satisfying (2), (3) and $x(\tau) = y$. V is thus defined only at points (y, τ) for which such pairs exists. Points (y, τ) with this property are called attainable points.

THEOREM 2. *Let U be bounded. Let $(\bar{x}(\cdot), \bar{u}(\cdot), p(\cdot))$ be a triple defined on $[t_0, \tau]$ for which (2), (3), (5), (6), (9) and (10) hold and let $\bar{H}(\tau) = H(\bar{x}(\tau), \bar{u}(\tau-), p(\tau), \tau)$. Then $V(y, \tau')$ is finite at all attainable points (y, τ') in some ball $B((\bar{x}(\tau), \tau), \delta)$ around $(\bar{x}(\tau), \tau)$ and for such points (y, τ')*

$$(13) \quad \begin{aligned} V(y, \tau') - V(\bar{x}(\tau), \tau) &\leq \bar{H}(\tau)(\tau' - \tau) - p(\tau)(y - \bar{x}(\tau)) \\ &\quad + q(|\tau - \tau'|, \|y - \bar{x}(\tau)\|)|\tau - \tau'| \end{aligned}$$

where q converges to zero when both arguments of q converge to zero.

Proof of Theorem 2. Let $(x(\cdot), u(\cdot))$ be an arbitrary admissible pair defined on $[t_0, \tau]$. If $\Delta(\tau)$ is the difference between the values of the criterion in (1) for $(\bar{x}(\cdot), \bar{u}(\cdot))$ and for $(x(\cdot), u(\cdot))$, we have

$$(14) \quad \Delta(\tau) \geq \varphi(\tau) - \varphi(t_0)$$

where $\varphi(t) = p(t)(x(t) - \bar{x}(t))$. See [5, p. 372]. Actually, when this property was derived in [5], the only properties used were the facts that $(x(\cdot), u(\cdot))$ and $(\bar{x}(\cdot), \bar{u}(\cdot))$ satisfy $\dot{x} = f$ and (3), that $(\bar{x}(\cdot), \bar{u}(\cdot))$ satisfies (5) and (6) and that (9) and (10) hold, for $t \in [t_0, \tau]$.

Next, we need an auxiliary system arising out of a modification of f^0 and f . Define \check{f}^0 and \check{f} to be equal to $f^0(\bar{x}_1, u, \tau)$ and $f(\bar{x}_1, u, \tau)$, respectively, in $(\tau, \tau_2]$, where $\bar{x}_1 = \bar{x}(\tau)$. For the optimal value function corresponding to \check{f}^0 and \check{f} , we write \check{V} and for the Hamiltonian, we write \check{H} .

Extend $\bar{u}(\cdot)$ to $(\tau, \tau_2]$ by letting $\bar{u}(t) = \bar{u}(\tau-)$ here. Extend $\bar{x}(\cdot)$ to $(\tau, \tau_2]$ as a solution in the present auxiliary system. The extension of $p(\cdot)$ simply becomes $p(t) = p(\tau)$, for $t > \tau$. Let τ' be any point in I .

Note that in the present auxiliary system, the supremum of the Hamiltonian is concave in x , for $t \in [t_0, \tau_2]$. When $(\bar{x}(t), \bar{u}(t), p(t))$ is inserted in \check{H} , denote it $\bar{H}(t)$. $\bar{H}(t)$ becomes a continuous function of t and equals $\check{H}^*(\bar{x}(t), p(t), t)$. If $(\check{x}(\cdot), u(\cdot))$ defined on $[0, \tau']$ is any admissible pair in the auxiliary system, then $\Delta(\tau') \geq \varphi(\tau') - \varphi(t_0)$, if $\Delta(\cdot)$ and φ now refer to entities of the auxiliary system. (See the sufficient properties mentioned in connection with (14)). In addition $\varphi(t_0) = 0$. Thus if x is attainable at $t = \tau'$ in this system (i.e., equals $\check{x}(\tau')$ for some $\check{x}(\cdot)$), we have that³

$$(15) \quad A = \check{V}(x, \tau') - \check{V}(\bar{x}(\tau'), \tau') \leq -p(\tau')(x - \bar{x}(\tau')).$$

³ Use the inequality for $\Delta(\tau')$.

Next, note that $B = \check{V}(\bar{x}(\tau'), \tau') - \check{V}(\bar{x}_1, \tau) = \int_{\tau}^{\tau'} f^0(\bar{x}(t), \bar{u}(t), t) dt = \int_{\tau}^{\tau'} \bar{H}(t) dt + C$, where $C = -\int_{\tau}^{\tau'} p(t) \check{f}(\bar{x}(t), \bar{u}(t), t) dt$.⁴

Let us now write $s = \tau' - \tau$. We shall, in turn, consider two cases: (i) $s < 0$ and (ii) $s \geq 0$.

Case (i). Here we get $C \leq -p(\tau)(\bar{x}(\tau') - \bar{x}_1) + \sigma_1(|s|)$, for $|s| < \check{\delta}$, where generally, here and below $\sigma_i(\cdot)$, $i = 1, 2, \dots$ means a second-order term in the variable(s) it contains. In fact, $\sigma_1(|s|)$ can be chosen equal to $K(\sup_{t \in J} \|p(t) - p(\tau)\|) \cdot |s|$, where $J = [\tau', \tau]$ and K is defined (also for later needs) by $K = \sup \|f(x, u, t)\|$ for $(x, t) \in G$, $u \in U$, G a bounded set in R^{n+1} containing $B(\bar{x}_1, 2) \times [t_0, \tau_2]$. We can choose $\check{\delta} = 2/K$ (then $\bar{x}(t)$ belongs to $B(\bar{x}_1, 2)$ for all $t \in (\tau - \check{\delta}, \tau) \cap [t_0, \tau]$).

Now, $A \leq -p(\tau)(x - \bar{x}(\tau')) + \sigma_2(|s|, \|x - \bar{x}(\tau')\|)$, where $\sigma_2(|s|, \|x - \bar{x}(\tau')\|) = \|p(\tau) - p(\tau')\| \cdot \|x - \bar{x}(\tau')\|$. Then we get $A + B = V(x, \tau') - V(\bar{x}_1, \tau) \leq \int_{\tau}^{\tau'} \bar{H}(t) dt - p(\tau)(x - \bar{x}_1) + \sigma_1(|s|) + \sigma_2(|s|, \|x - \bar{x}(\tau')\|)$ for $|s| < \check{\delta}$. Note that $\|p(\tau) - p(\tau')\| \leq L|s|$, where $L = \sup_{t \in J} \|\dot{p}(t)\|$. If we define $\sigma_3(|s|) = (\sup_{t \in J} |\bar{H}(t) - \bar{H}(\tau)|) \cdot |s|$, then $A + B \leq \bar{H}(\tau)s - p(\tau)(x - \bar{x}_1) + \sigma_1 + \sigma_2 + \sigma_3$, for $|s| < \check{\delta}$. Hence, for $y = x$, (13) is proved in this case.

Case (ii). In this case, $B = \bar{H}(\tau)s - p(\tau) \cdot (\bar{x}(\tau') - \bar{x}(\tau))$; hence $A + B \leq \bar{H}(\tau)s - p(\tau)(x - \bar{x}_1) = D$, by (15) and the fact that $p(\tau') = p(\tau)$.

Define $\Gamma = B(\bar{x}_1, 1) \times [\tau, \tau + \delta']$, where δ' is chosen so small that if $(x(\cdot), u(\cdot), \tau')$ is any triple in the original system such that $(x(\tau'), \tau') \in \Gamma$, then $x(t) \in B(\bar{x}_1, 2)$ for all $t \in [\tau, \tau']$. (In fact, δ' can be chosen as $\delta' = 1/K$.) In what follows, let $(x(\cdot), u(\cdot), \tau')$ be a triple for which $(x(\tau'), \tau') \in \Gamma$. Let $\check{x}(\cdot)$ correspond to this triple. (Then $\check{x}(\cdot) = x(\cdot)$ on $[\tau_0, \tau]$ while $\check{x}(\cdot) = f(\bar{x}_1, u(t), \tau) = \check{f}(\check{x}(t), u(t), t)$ on $(\tau, \tau']$). As before, $x = \check{x}(\tau')$.

Define $x_1 = x(\tau')$. When x_1 is near \bar{x}_1 , also $x(t)$ is near \bar{x}_1 for $t \in [\tau, \tau']$, more precisely, $\|\bar{x}_1 - x(t)\| \leq \|\bar{x}_1 - x_1\| + \|x_1 - x(t)\| \leq \|\bar{x}_1 - x_1\| + K \cdot s$. Choose a term $\alpha(\|z - y\|, |t' - t|)$ such that $\alpha(\|z - y\|, |t' - t|) \geq |f^0(z, u, t') - f^0(y, u, t)|$ and $\alpha(\|z - y\|, |t' - t|) \geq \|f(z, u, t') - f(y, u, t)\|$ for all $u \in U$, $(y, t), (z, t') \in G$, $\alpha(\cdot, \cdot)$ nondecreasing in both arguments and converging to zero if the arguments converge to zero. Since

$$\begin{aligned} x - x_1 &= \int_{\tau}^{\tau'} (\check{f}(\check{x}(t), u(t), t) - f(x(t), u(t), t)) dt \\ &= \int_{\tau}^{\tau'} (f(\bar{x}_1, u(t), \tau) - f(x(t), u(t), t)) dt, \end{aligned}$$

we get

$$\|x - x_1\| \leq \int_{\tau}^{\tau'} \alpha(\|\bar{x}_1 - x(t)\|, |\tau - t|) dt \leq \alpha(\|\bar{x}_1 - x_1\| + K \cdot s, s) \cdot s = \sigma_4(s, \|\bar{x}_1 - x_1\|)$$

(a definition of σ_4), when $x_1 \in B(\bar{x}_1, 1)$, $\tau' \in [\tau, \tau + \delta']$.

Similarly,

$$\begin{aligned} E &= \int_{t_0}^{\tau'} f^0(x(t), u(t), t) dt \\ &\leq \int_{t_0}^{\tau'} \check{f}^0(\check{x}(t), u(t), t) dt + \sigma_4(s, \|\bar{x}_1 - x_1\|) \\ &\leq \check{V}(x, \tau') + \sigma_4(s, \|\bar{x}_1 - x_1\|) \quad \text{when } x_1 \in B(\bar{x}_1, 1), \quad \tau' \in [\tau, \tau + \delta']. \end{aligned}$$

⁴ The second equality is valid, since $\bar{u}(\cdot)$, restricted to $[0, \tau']$ is optimal in the fixed end problem $\check{x}(\tau') = \bar{x}(\tau')$. (See the inequality for $\Delta(\tau')$ above.)

Note that $V(\bar{x}_1, \tau) = \check{V}(\bar{x}_1, \tau)$. Hence, $E - V(\bar{x}_1, \tau) \leq \check{V}(x, \tau') - \check{V}(\bar{x}_1, \tau) + \sigma_4 = A + B + \sigma_4 \leq D + \sigma_4$. Furthermore, if the term $-p(\tau)(x - \bar{x}_1)$ in the expression D is replaced by $-p(\tau)(x_1 - \bar{x}_1)$, the error is less than $\|p(\tau)\|\sigma_4$. Thus,

$$(16) \quad E - V(\bar{x}_1, \tau) \leq \bar{H}(\tau) \cdot s - p(\tau)(x_1 - \bar{x}_1) + (1 + \|p(\tau)\|)\sigma_4.$$

Given any pair (y, τ') in $B((\bar{x}_1, \tau), \delta)$, $\delta = \min(1, \delta')$. If (y, τ') is attainable, then (16) holds for all triples $(x(\cdot), u(\cdot), \tau')$ in the original system for which $x(\tau') = x_1 = y$. Since the right-hand side of (16) does not depend on the pair $(x(\cdot), u(\cdot))$, the inequality (13) is obtained also in this case.

Remark 2. Let G' be a bounded set in $R^n \times [t_0, \tau_2]$, and let $K'' > 0$. Assume that for each $(x, \tau) \in G'$, there exists a triple $(\bar{x}(\cdot), \bar{u}(\cdot), p(\cdot))$ defined on $[t_0, \tau]$ such that $x = \bar{x}(\tau)$, $\bar{x}(\cdot)$, $\bar{u}(\cdot)$, $p(\cdot)$ satisfying (2), (3), (5), (6), (9), (10), with $\|p(\tau)\| < K''$. Then $q(\cdot, \cdot)$ and δ in (13) can be taken to be one and the same for all such triples.

To prove this assertion, let $B(0, b) \times [t_0, \tau_2]$ be the set G in the proof, the number b so chosen that $B(0, b) \supset B(x, 2)$ for all $(x, \tau') \in G'$, and let $\delta = \min(1, 1/K, 1/K')$, where $K' = \sup \| (f_x^0(x, u, t) + pf_x(x, u, t)) \|$ for $(x, t) \in G, u \in U, p \in B(0, K'' + 1)$. Note that $p(t) \in B(0, K'' + 1)$ if $t \in [\tau - \delta, \tau] = I_\delta$. Thus $\|p(t) - p(\tau)\| \leq L|t - \tau| \leq K' \cdot |\tau' - \tau|$ for $t, \tau' \in I_\delta, t > \tau'$. Hence σ_1 and σ_2 can be taken to be independent of the triple $(\bar{x}(\cdot), \bar{u}(\cdot), p(\cdot))$.

Both $\alpha(\cdot, \cdot)$ and σ_4 are independent of $(\bar{x}(\cdot), \bar{u}(\cdot), p(\cdot))$. Finally, the function $\hat{H}(x, p, t) = \max_{u \in \bar{U}} H(x, u, p, t)$ is continuous in (x, p, t) , hence uniformly continuous for $(x, t) \in G, p \in B(0, K'' + 1)$: Since $\|p(t) - p(\tau)\| \leq K'|t - \tau|$ and $\|\bar{x}(t) - \bar{x}(\tau)\| \leq K \cdot |t - \tau|$ for $t \in I_\delta$, the limit $\lim_{t \rightarrow \tau^-} \bar{H}(t) = \bar{H}(\tau)$ is uniform in the triples $(\bar{x}(\cdot), \bar{u}(\cdot), p(\cdot))$, which shows that even σ_3 can be chosen independent of these triples.

Remark 3. Assume that U is unbounded, and let U' be a given bounded subset of U . Assume that the restriction $u(t) \in U$ (see (3)) is replaced by $u(t) \in U'$, and that $V(\cdot, \cdot)$ and the word "attainable" refer to this new system. Assume furthermore that $\bar{u}(t)$, while taking values in U' , nevertheless maximizes H even in U , and that the supremum of H , still taken for $u \in U$, is concave. Then Theorem 2 and Remark 2 are still valid.

The proofs can be kept unchanged, provided K and K' and $\alpha(\cdot, \cdot)$ are defined by restricting u to U' and defining $\hat{H}(x, p, t) = \max_{u \in \bar{U}'} H(x, u, p, t)$.

In the proof of Theorem 1, a set U' is chosen such that $u^\tau(t) \in U'$ for all t and τ . Since $u^\tau(\cdot)$ is optimal in the original problem, $V(\cdot, \cdot)$ has the same value at all points $(x^\tau(\tau), \tau)$ in the present redefined system as in the original one.

Proof of Theorem 1. For $y = x^{\tau'}(\tau')$, $x^\tau(\cdot) = \bar{x}(\cdot)$, $u^\tau(\cdot) = \bar{u}(\cdot)$, from (13) and Remark 2, we get that for $|\tau' - \tau| \leq \delta$

$$(16') \quad \begin{aligned} V(\tau') - V(\tau) &\leq d(\tau)(\tau' - \tau) - p^\tau(\tau)[x^{\tau'}(\tau') - x^\tau(\tau)] + q|\tau' - \tau| \\ &\leq d(\tau)(\tau' - \tau) + q(|\tau' - \tau|, \|x^{\tau'}(\tau') - x^\tau(\tau)\|)|\tau' - \tau| \end{aligned}$$

where we have written out the arguments of q only once, and where we have used (see (4) and (7)), that

$$\begin{aligned} p^\tau(\tau)x_{\tau'}(\tau') &= \sum_{i=1}^m p^{\pi_i}(\tau)x^{\tau'i}(\tau') \geq \sum_{i=1}^m p^{\pi_i}(\tau)x_i^i \\ &= \sum_{i=1}^m p^{\pi_i}(\tau)x^{\pi_i}(\tau) = p^\tau(\tau) \cdot x^\tau(\tau). \end{aligned}$$

Note that $\tau \rightarrow x^\tau(\tau)$ is uniformly continuous. Given an arbitrary $\varepsilon > 0$, let δ' and δ'' , $0 < \delta' < \delta''$ be chosen such that

if $|\tau' - \tau| < \delta''$ and $\|x^{\tau'}(\tau') - x^\tau(\tau)\| < \delta''$, then $q < \varepsilon$ and if $|\tau' - \tau| < \delta'$ then $\|x^{\tau'}(\tau') - x^\tau(\tau)\| < \delta''$.

Observe that δ' is independent of $\tau \in [\tau_1, \tau_2]$. Let $\check{\tau} \in [\tau_1, \tau^*)$. There exists a finite partition of $[\check{\tau}, \tau^*]$ into intervals $[\tau^i, \tau^{i+1}]$, $i = 1, \dots, i^*$ of lengths $< \delta'$. In $[\check{\tau}, \tau^*]$, $d(\tau) \geq 0$ (see (12)); thus, by (16')

$$V(\check{\tau}) - V(\tau^*) = \sum_{i=1}^{i^*} V(\tau^i) - V(\tau^{i+1}) \leq \sum_i \varepsilon(\tau^{i+1} - \tau^i) \leq \varepsilon(\tau^* - \check{\tau}).$$

Since ε is arbitrary, $V(\check{\tau}) \leq V(\tau^*)$. A symmetric argument also gives that if $\check{\tau} \in (\tau^*, \tau_2]$, $V(\check{\tau}) \geq V(\tau^*)$, and the proof is finished.

(By a slight modification of the last arguments, it may be seen that it suffices to assume that (12) holds for v.e. τ .)

4. In this section, we generalize the results to problems with state and mixed state/control restrictions and more general initial and terminal conditions.

$$(17) \quad \max \left\{ \int_{t_0}^{\tau} f^0(x(t), u(t), t) dt + S^0(x(t_0)) + S^1(x(\tau), \tau) \right\}.$$

$$(18) \quad \dot{x}(t) = f(x(t), u(t), t) \quad \text{v.e.,}$$

$$(19) \quad R_k^0(x(t_0)) = 0, \quad k = 1, \dots, r'_0, \quad t_0 \text{ fixed,}$$

$$R_k^0(x(t_0)) \geq 0, \quad k = r'_0 + 1, \dots, r_0,$$

$$(20) \quad R_k^1(x(\tau), \tau) = 0, \quad k = 1, \dots, r'_1, \quad \tau \in [\tau_1, \tau_2],$$

$$R_k^1(x(\tau), \tau) \geq 0, \quad k = r'_1 + 1, \dots, r_1$$

where $t_0 < \tau_1 < \tau_2$, τ_1, τ_2 fixed.

$$(21) \quad u(t) \in U, \quad U \text{ fixed, convex set in } R'.$$

For all t ,

$$(22) \quad \begin{aligned} g^j(x(t), u(t), t) &\geq 0, & j = 1, \dots, s', \\ g^j(x(t), u(t), t) &\equiv \bar{g}^j(x(t), t) \geq 0, & j = s' + 1, \dots, s. \end{aligned}$$

(In (22) the assumption is that g^j does not contain u for $j > s'$.)

$$(23) \quad f^0, f \text{ and their partial derivatives with respect to } x \text{ and } u \text{ exist and are continuous; } g, S^0, S^1, R^0 = (R_1^0, \dots, R_{r_0}^0), R^1 = (R_1^1, \dots, R_{r_1}^1) \text{ and their derivatives exist and are continuous.}$$

We first formulate as a theorem sufficient conditions for this problem for the case where τ is kept fixed. In the theorem, we assume that $p(\cdot)$ is continuous in (t_0, τ) . When the mixed constraints $g^j, j = 1, \dots, s'$ satisfy certain constraints qualifications, these constraints do not cause jumps in $p(\cdot)$. Neither do the pure state constraints $g^j, j > s'$ in the case where $\bar{g}_x^j(\bar{x}(t), t)f(\bar{x}(t), \bar{u}(t), t)$ is discontinuous at points $t \in (t_0, \tau)$ at which $\bar{g}^j(\bar{x}(t), t) = 0$, or when $u \rightarrow H(\bar{x}(t), u, p(t), t)$ has a unique maximum and a standard constraints qualification is satisfied.

Hence, the following sufficient conditions, which assume no jumps in $p(\cdot)$ in (t_0, τ) , have fairly wide applicability.

As before, we define the Hamiltonian as

$$(24) \quad H(x, u, p, t) = p_0 f^0(x, u, t) + pf(x, u, t)$$

and we define the Lagrangian as

$$(25) \quad L(x, u, p, q, t) = H(x, u, p, t) + qg(x, u, t).$$

The following theorem will serve as a reference for our next free final time theorem. (See [5, Thm. 9], [6, Thm. 6.7.1].)

THEOREM 3. (Final time τ fixed.) *Let $(\bar{x}(t), \bar{u}(t))$ be an admissible⁵ pair in problem (17)–(23). Assume that there exist a continuous and piecewise continuously differentiable function $p(t) = (p_1(t), \dots, p_n(t))$, a piecewise continuous function $q(t) = (q_1(t), \dots, q_s(t))$ and vectors $\beta = (\beta_1, \dots, \beta_s)$, $\beta^0 = (\beta_1^0, \dots, \beta_s^0)$, $\gamma = (\gamma_1, \dots, \gamma_r)$, $\delta = (\delta_1, \dots, \delta_{r_0})$, such that the following properties hold for $p_0 = 1$:*

$$(26) \quad L_u^*(u - \bar{u}(t)) \leq 0 \quad \text{for all } u \in U \quad \text{for v.e.t.}$$

(The $*$, here and below, indicates that the derivative is evaluated along $(\bar{x}(t), \bar{u}(t), p(t), q(t))$).

$$(27) \quad \dot{p}(t) = -L_x^* \quad \text{v.e.,}$$

$$(28) \quad p(\tau) = \beta g_x(\bar{x}(\tau), \bar{u}(\tau), \tau) + S_x^1(\bar{x}(\tau), \tau) + \gamma R_x^1(\bar{x}(\tau), \tau)$$

where for $k > r'_1$, $\gamma_k \geq 0$ ($= 0$ if $R_k^1(\bar{x}(\tau), \tau) > 0$) and where the vector $\beta = (\beta_1, \dots, \beta_s)$ satisfies

$$(29) \quad \begin{aligned} \beta_j &= 0, & j &= 1, \dots, s', \\ \beta_j &\geq 0 \quad (= 0 \text{ if } g^j(\bar{x}(\tau), \bar{u}(\tau), \tau) > 0), & j &= s' + 1, \dots, s, \end{aligned}$$

$$(30) \quad p(t_0) = -\beta^0 g_x(\bar{x}(t_0), \bar{u}(t_0), t_0) - S_x^0(\bar{x}(t_0)) - \delta R_x^0(\bar{x}(t_0))$$

where for $k > r'_0$, $\delta_k \geq 0$ ($= 0$ if $R_k^0(\bar{x}(t_0)) > 0$), and where the vector $\beta = (\beta_1^0, \dots, \beta_s^0)$ satisfies

$$(31) \quad \begin{aligned} \text{(i)} \quad \beta_j^0 &= 0, & j &\leq s', \\ \text{(ii)} \quad \beta_j^0 &\geq 0 \quad (= 0 \text{ if } g^j(\bar{x}(t_0), \bar{u}(t_0), t_0) > 0), & j &> s', \end{aligned}$$

$$(32) \quad q_j(t) \geq 0 \quad (= 0 \text{ if } g^j(\bar{x}(t), \bar{u}(t), t) > 0), \quad j = 1, \dots, s,$$

$$(33) \quad H(x, u, p(t), t) \text{ is concave in } (x, u) \in R^n \times R^r \text{ for each } t;$$

$$(34) \quad g^j(x, u, t) \text{ is quasiconcave in } (x, u) \in R^n \times R^r \text{ for each } t;$$

$$(35) \quad \text{For each } t, S^0 \text{ and } S^1 \text{ are concave in } x \text{ and } \delta_k R_k^0 \text{ and } \gamma_k R_k^1 \text{ are quasiconcave in } x, \text{ for each } k.$$

Then $(\bar{x}(\cdot), \bar{u}(\cdot))$ is an optimal pair.

In the next theorem, τ is again subject to choice in $[\tau_1, \tau_2]$.

THEOREM 4. (Final time τ free.) *Consider problem (17)–(23). Suppose that τ is free to vary in $[\tau_1, \tau_2]$, $t_0 < \tau_1 < \tau_2$.*

⁵ An admissible pair $x(\cdot), u(\cdot)$ is a pair satisfying (18)–(22), $u(\cdot)$ piecewise continuous.

Assume that for each $\tau \in [\tau_1, \tau_2]$ there exists an admissible pair $(x^\tau(t), u^\tau(t))$, defined on $[t_0, \tau]$, with associated multipliers $p^\tau(t), q^\tau(t), \delta^\tau = (\delta_1^\tau, \dots, \delta_{r_0}^\tau), \gamma^\tau = (\gamma_1^\tau, \dots, \gamma_{r_1}^\tau), \beta^{0\tau} = (\beta_1^{0\tau}, \dots, \beta_s^{0\tau}), \beta^\tau = (\beta_1^\tau, \dots, \beta_s^\tau)$ that satisfy the conditions (26)–(35) for $t \in [t_0, \tau]$ with $p_0 = 1$. Assume, moreover, that $u^\tau(\cdot)$ and $q^\tau(\cdot)$ take values in fixed, bounded subsets being independent of $\tau \in [\tau_1, \tau_2]$, and that $\tau \rightarrow x^\tau(\tau)$ is Lipschitz continuous. Assume also that the functions $\tau \rightarrow \beta^\tau$ and $\tau \rightarrow \gamma^\tau$ are piecewise continuous. Assume next that $u^\tau(\tau)$ belongs to the closure of the set $\{u \in U, g^j(x^\tau(t), u, \tau) > 0 \text{ for all } j \leq s'\} \text{ for all } \tau$. Finally, assume that the function

$$(36) \quad F(\tau) = H(x^\tau(\tau), u^\tau(\tau), p^\tau(\tau), \tau) + \beta^\tau \cdot g_t(x^\tau(\tau), u^\tau(\tau), \tau) \\ + S_t^1(x^\tau(\tau), \tau) + \gamma^\tau \cdot R_t^1(x^\tau(\tau), \tau)$$

has the property that there exists a $\tau^* \in [\tau_1, \tau_2]$ such that

$$(37) \quad \begin{aligned} F(\tau) &\geq 0 \quad \text{for } \tau < \tau^* \quad \text{if } \tau_1 < \tau^*, \\ F(\tau) &\leq 0 \quad \text{for } \tau > \tau^* \quad \text{if } \tau_2 > \tau^*. \end{aligned}$$

Then the pair $(x^{\tau^*}(t), u^{\tau^*}(t))$ defined on $[t_0, \tau^*]$ is optimal.

Proof. Again $V(\tau)$ is the value of the criterion, see (17), for $(x^\tau(\cdot), u^\tau(\cdot), \tau)$. And as before, $(x^\tau(\cdot), u^\tau(\cdot))$ is optimal in the problem where τ is the fixed terminal time. Hence, to show that $V(\tau^*)$ is no less than the value of the criterion for an arbitrary admissible pair $x(\cdot), u(\cdot)$ defined on $[0, \tau]$, $\tau \in [\tau_1, \tau_2]$, it suffices to show that $V(\tau) \leq V(\tau^*)$.

To give a proof of this property, we shall refer to the proofs of Theorems 1 and 2. Hence, let $(\bar{x}(\cdot), \bar{u}(\cdot)) = (x^\tau(\cdot), u^\tau(\cdot))$ for some given τ . Choose an open ball B' such that $u^\tau(t) \in B'$ for all τ and t , and define $U' = U \cap B'$. Note that for some K'' , we have $K'' \geq \sup \|p^\tau(\tau)\|$, $(\tau \rightarrow x^\tau(\tau), \tau \rightarrow \gamma^\tau, \tau \rightarrow \beta^\tau)$ are piecewise continuous (see (28), (29)). Define $K' = \sup \|f_x^0(x, u, t) + pf_x(x, u, t) + qg_x(x, u, t)\|$, the supremum taken for $(x, t) \in G = B(0, b) \times [t_0, \tau_2]$, $u \in U'$, $p \in B(0, K'' + 1)$, $q \in Q$, where Q is a set that has the property that $q^\tau(t) \in Q$ for all t and τ and $B(0, b) \supset B(x^\tau(\tau), 2)$ for all τ . Finally, let $K = \sup \|f(x, u, t)\|$ for $u \in U'$, $(x, t) \in G$ and let $\delta = \min(1, 1/K, 1/K')$. For these definitions, K, K', K'' and δ can play the same role as in the proof of Theorem 1.

$$(37') \quad \text{Define } H^*(x, p, t) = \sup \{H(x, u, p, t) : u \in U, g(x, u, t) \geq 0\}.$$

In [5], property (14) was obtained by establishing the inequality

$$(38) \quad H(x(t), u(t), p(t), t) - H^*(\bar{x}(t), p(t), t) \leq -\dot{p}(t)(x(t) - \bar{x}(t)).$$

For the arguments needed to obtain (38), see the proof of Theorem 6 and Note 3 in [5]. (The properties used in the proof in [5] are that $(x(\cdot), u(\cdot))$ and $(\bar{x}(\cdot), \bar{u}(\cdot))$ satisfy $\dot{x} = f$, $u(t) \in U$, $\dot{p} = -L_x^*, L_u^*(u - \bar{u}(t)) \leq 0$ for $u \in U$, $qg^* = 0$, $q \geq 0$, H concave in (x, u) , g quasiconcave in (x, u)).

Let us first consider the case $\tau' < \tau$. Evidently,

$$(38') \quad \int_{t_0}^{\tau'} f^0(\bar{x}(t), \bar{u}(t), t) dt - \int_{t_0}^{\tau'} f^0(x^{\tau'}(t), u^{\tau'}(t), t) dt \geq \phi^{\tau'}(\tau') - \phi^{\tau'}(t_0)$$

where $\phi^{\tau'}(t) = p^{\tau'}(t)(x^{\tau'}(t) - x^\tau(t))$ (see property (14)).

With the same arguments that lead from (15) to (13) above, we obtain from (38') that if $(x^{\tau'}(\tau'), \tau') \in B((x^\tau(\tau), \tau), \delta)$, then

$$(38'') \quad -\phi^{\tau'}(t_0) + \Delta''(\tau', \tau) \leq d(\tau)(\tau' - \tau) - p^\tau(\tau)(x^{\tau'}(\tau') - x^\tau(\tau)) + \sigma$$

where $\Delta''(\tau', \tau) = \int_{t_0}^{\tau'} f^0(x^{\tau'}(t), u^{\tau'}(t), t) dt - \int_{t_0}^{\tau} f^0(x^{\tau}(t), u^{\tau}(t), t) dt$ and σ is the second-order term of the proof of Theorem 2 (i.e., $\sigma = q|\tau' - \tau|$).

Even the independence of σ and δ on $\bar{x}(\cdot)$ (i.e., on τ) is obtained in the same way.

For $\tau' > \tau$, the argument runs as follows: Since $\tau \rightarrow \beta^{\tau}$ and $\tau \rightarrow \gamma^{\tau}$ are piecewise continuous, by (28) and (29), $\tau \rightarrow p^{\tau}(\tau)$ is piecewise continuous. Next, note that for $x(\cdot) = x^{\tau'}(\cdot)$, (38) can be obtained for all t in (t_0, τ) . This means that we can obtain an inequality similar to (38') also in this case, namely,

$$\int_{t_0}^{\tau} f^0(\bar{x}(t), \bar{u}(t), t) dt - \int_{t_0}^{\tau} f^0(x^{\tau'}(t), u^{\tau'}(t), t) dt \geq \phi^{\tau}(\tau) - \phi^{\tau}(t_0).$$

Defining Δ'' as above, we get

$$-\Delta''(\tau', \tau) + \int_{\tau}^{\tau'} f^0(x^{\tau'}(t), u^{\tau'}(t), t) dt \geq \phi^{\tau}(\tau) - \phi^{\tau}(t_0).$$

The integrand can be rewritten as $H(x^{\tau'}(t), u^{\tau'}(t), \Phi(\tau), t) - \Phi(\tau)f(x^{\tau'}(t), u^{\tau'}(t), t)$, where $\Phi(\tau) = p^{\tau}(\tau)$. Thus,

$$-\Delta''(\tau', \tau) + \int_{\tau}^{\tau'} H(x^{\tau'}(t), u^{\tau'}(t), \Phi(\tau), t) dt - \Phi(\tau)(x^{\tau'}(\tau') - x^{\tau'}(\tau)) \geq \phi^{\tau}(\tau) - \phi^{\tau}(t_0).$$

Using the definition of $\phi^{\tau}(\tau)$ and approximating the last integrand by $d(\tau)$, we get

$$(38''') \quad -\phi^{\tau}(t_0) + \Delta''(\tau', \tau) \leq d(\tau)(\tau' - \tau) - p^{\tau}(\tau)(x^{\tau'}(\tau') - x^{\tau'}(\tau)) + \sigma$$

where $\sigma = \int_{\tau}^{\tau'} \lambda(t, \tau', \tau) dt$, $\lambda(t, \tau', \tau) = |d(\tau) - H(x^{\tau'}(t), u^{\tau'}(t), \Phi(\tau), t)|$. We are going to show that $\lambda(t, \tau', \tau)$ is uniformly small within each interval of continuity of $\tau \rightarrow p^{\tau}(\tau)$, in a sense to be made precise.

Define $A = \sup \{|H(x, u, p, t)|: \text{for } (x, t) \in G, u \in U', p \in B(0, K''+1)\}$ and let (a_i, a_{i+1}) , $i = 1, \dots, i'$ be the intervals of continuity of $\tau \rightarrow p^{\tau}(\tau)$.

Let $\varepsilon > 0$. Define $a = \varepsilon/8Ai'$, and let $I_i = [a_i + a, a_{i+1} - a]$. Define $h(\tau', t) = H^*(x^{\tau'}(t), p^{\tau'}(t), t)$. Note that $\|p^{\tau'}(\tau') - p^{\tau'}(t)\| \leq K'(\tau' - t)$ and that $\|x^{\tau'}(\tau') - x^{\tau'}(t)\| \leq K(\tau' - t)$ when $\tau', t \in I_i$, $\tau' - \delta \leq t \leq \tau'$. Hence for any $\tau \in I_i$, $\lim x^{\tau'}(t) = x^{\tau'}(\tau)$ and $\lim p^{\tau'}(t) = p^{\tau'}(\tau)$ when $\tau' \rightarrow \tau$, $t \rightarrow \tau$, $t \leq \tau'$, $\tau', t \in I_i$. We now assert that *there exists a $\delta \in (0, \delta)$ such that $\lambda(t, \tau', \tau) = \|h(\tau', t) - h(\tau, \tau)\| < \varepsilon$, when $\tau \leq t \leq \tau' \leq \tau + \delta$; $t, \tau', \tau \in I_i$* . By contradiction, assume that, for all $n = 1, 2, \dots$, there exist points $t_n, \tau'_n, \tau_n \in I_i$, $\tau_n \leq t_n \leq \tau'_n \leq \tau_n + 1/n$ such that $\lambda(t_n, \tau'_n, \tau_n) \geq \varepsilon$. By compactness of I_i , we may assume that $\tau_n \rightarrow \tau$, for some $\tau \in I_i$. Then also $t_n \rightarrow \tau$, and $\tau'_n \rightarrow \tau$. Next, observe that the set \bar{U}' is compact. Note furthermore that $h(\tau', t) = \max \{H(x^{\tau'}(t), u, p^{\tau'}(t), t): u \in \bar{U}', g(x^{\tau'}(t), u, t) \geq 0\}$, $h(\tau, \tau) = \sup \{H(x^{\tau}(\tau), u, p^{\tau}(\tau), \tau): u \in \bar{U}', g^j(x^{\tau}(\tau), u, \tau) > 0 \text{ for all } j \leq s'\}$. The first equality gives that $\limsup h(\tau'_n, t_n) \leq h(\tau, \tau)$, the second one that $\liminf h(\tau'_n, t_n) \geq h(\tau, \tau)$. Thus, $\lim h(\tau'_n, t_n) = h(\tau, \tau)$. But this contradicts the fact that $\lambda(t_n, \tau'_n, \tau_n) \geq \varepsilon$. Hence, the assertion made about $\lambda(t, \tau', \tau)$ above is valid. Exactly the same assertion can also be made about $\lambda(t, \tau', \tau)$ for the following reason: Observe first that $d(\tau) = h(\tau, \tau)$, while $h(\tau', t) - H(x^{\tau'}(t), u^{\tau'}(t), \Phi(t), t) = (p^{\tau'}(t) - \Phi(\tau))f(x^{\tau'}(t), u^{\tau'}(t), t)$. In the last expression, f is bounded by K if $\tau \leq t \leq \tau' \leq \tau + 2/K$, while $p^{\tau'}(t) \rightarrow p^{\tau}(\tau) = \Phi(\tau)$ when τ' (and t) $\rightarrow \tau$, uniformly in $\tau \in I_i$, by a simple compactness argument again.

Since $\lambda(t, \tau', \tau)$ has the asserted property, $\sigma \leq \varepsilon(\tau' - \tau)$ when $\tau' - \tau \leq \tilde{\delta}$, $\tau' \geq \tau$, $\tau', \tau \in I_i$ for some i .

Let $\tilde{\tau}$ be some point in $[\tau_1, \tau^*)$ (assuming $\tau_1 < \tau^*$), and let $\varepsilon > 0$. Choose a $\delta' \in (0, \delta)$ such that the term σ of (38'') has the property that $\sigma \leq \varepsilon|\tau' - \tau|$ when $|\tau' - \tau| < \delta'$ (cf. the proof of Theorem 1). Choose a partition $\{t^j\}$, $j = 1, \dots, j'$ of $[\tilde{\tau}, \tau^*]$, $t^1 - \tilde{\tau}, t^{j'} = \tau^*$,

such that $0 < t^{j+1} - t^j < \delta'$. Let $\tau' = t^j$, $\tau = t^{j+1}$, $j = 1, 2, \dots$, and sum the left, respectively right, side of (38''). For $y(t) = x'(t)$, $\Phi(t) = p'(t)$, we then get

$$(39) \quad - \sum_{j=1}^{j'-1} \phi^{t^{j+1}}(t_0) + \Delta''(\check{\tau}, \tau^*) \leq \sum_j d(t^{j+1})(t^j - t^{j+1}) - \sum_j \Phi(t^{j+1})(y(t^j) - y(t^{j+1})) \\ + \sum_j \varepsilon(t^{j+1} - t^j).$$

For the t^j 's spaced closely enough (all $t^{j+1} - t^j$ small enough), the two former sums above approximate the integrals $\int_{\tau^*}^{\check{\tau}} d(t) dt$ and $\int_{\tau^*}^{\check{\tau}} \Phi(t) dy(t)$, respectively. Evidently, it is possible to obtain

$$(39') \quad - \sum_j \phi^{t^{j+1}}(t_0) + \Delta''(\check{\tau}, \tau^*) \leq - \int_{\check{\tau}}^{\tau^*} d(t) dt + \int_{\check{\tau}}^{\tau^*} \Phi(t) dy(t) + \varepsilon'$$

where $\varepsilon' = \varepsilon(\tau_2 - \tau_1) + \varepsilon$.

A similar argument also works when $\check{\tau} \in (\tau^*, \tau_2]$. In this case, the term σ is the one defined subsequently to (38'''). Since $\lambda(t, \tau', \tau) \leq 2A$ when $\tau \leq t \leq \tau' \leq \tau + \tilde{\delta}$, then $\sigma \leq 2A(\tau' - \tau)$, ($\sigma \leq \varepsilon(\tau' - \tau)$ if $\tau', \tau \in I_i$). Choose a partition $\{t^j\}$, $j = 1, \dots, j'$ of $[\tau^*, \check{\tau}]$ such that $0 < t^{j+1} - t^j < \tilde{\delta}$, and such that any of the intervals $[t^j, t^{j+1}]$ either belongs to some I_i or equals one of the intervals $I_i \cap [\tau^*, \check{\tau}]$, with $t^1 = \tau^*$, $t^{j'} = \check{\tau}$.

Otherwise, by letting $\tau' = t^{j+1}$, $\tau = t^j$, and using the facts that $\sigma \leq \varepsilon(t^{j+1} - t^j)$ when $[t^j, t^{j+1}] \in I_i$, for some i , and that $\sigma \leq 2A(t^{j+1} - t^j)$, we get

$$(40) \quad - \sum_{j=1}^{j'-1} \phi^{t^j}(t_0) + \Delta''(\check{\tau}, \tau^*) \leq \sum_j d(t^j)(t^{j+1} - t^j) - \sum_j \Phi(t^j)(y(t^{j+1}) - y(t^j)) \\ + \sum_j \varepsilon(t^{j+1} - t^j) + i'2A2a$$

by summing the left and right sides of (38'''). Finally, when the t^j 's are densely distributed, the sums can be replaced by integrals, with the effect of introducing at most an additional error $\varepsilon/2$. Hence

$$(41) \quad - \sum_j \phi^{t^j}(t_0) + \Delta''(\check{\tau}, \tau^*) \leq \int_{\tau^*}^{\check{\tau}} d(t) dt - \int_{\tau^*}^{\check{\tau}} \Phi(t) dy(t) + \varepsilon''$$

where $\varepsilon'' = \varepsilon(\tau_2 - \tau_1) + \varepsilon$.

Next, let τ' be either smaller or greater than τ . Let $\bar{x}(\cdot) = x^\tau(\cdot)$. Then, for $\delta = \delta^\tau$, $\delta_k R_k^0(x^\tau(t_0)) \geq \delta_k R_k^0(\bar{x}(t_0))$ when $\bar{x}(t_0)$ satisfies (19); hence by quasiconcavity, $\delta_k R_{kx}^0(\bar{x}(t_0))(x^\tau(t_0) - \bar{x}(t_0)) \geq 0$. Similarly, for $\beta^0 = \beta^{0\tau}$, $\beta_j^0 g_x^j(\bar{x}(t_0), \bar{u}(t_0), t_0)(x^\tau(t_0) - \bar{x}(t_0)) \geq 0$.

Also, $S^0(x^\tau(t_0)) - S^0(\bar{x}(t_0)) \leq S_x^0(\bar{x}(t_0))(x^\tau(t_0) - \bar{x}(t_0))$. Thus, $\phi^\tau(t_0) + S^0(x^\tau(t_0)) - S^0(\bar{x}(t_0)) \leq \phi^\tau(t_0) + S_x^0(\bar{x}(t_0))(x^\tau(t_0) - \bar{x}(t_0)) + \delta R_x^0(\bar{x}(t_0))(x^\tau(t_0) - \bar{x}(t_0)) + \beta^0 g_x(\bar{x}(t_0), \bar{u}(t_0), t_0)(x^\tau(t_0) - \bar{x}(t_0)) = 0$, using condition (30). Define $\Delta'(\tau', \tau) = S^0(x^\tau(t_0)) - S^0(\bar{x}(t_0))$. Then, by the last inequality, $\Delta' \leq -\phi^\tau(t_0)$. Hence, $\Delta'(\check{\tau}, \tau^*) = \sum_j \Delta'(t^j, t^{j+1}) \leq -\sum \phi^{t^{j+1}}(t_0)$ (respectively, $\leq -\sum \phi^{t^j}(t_0)$), when $\check{\tau} < \tau^*$ (respectively, $\check{\tau} > \tau^*$). Consider now the case $\check{\tau} > \tau^*$. Using (39'), we get

$$(42) \quad \tilde{\Delta} \leq - \int_{\check{\tau}}^{\tau^*} d(t) dt + \int_{\check{\tau}}^{\tau^*} \Phi(t) dy(t) + \varepsilon'$$

where $\varepsilon' = \varepsilon(\tau_2 - \tau_1) + \varepsilon$, and $\tilde{\Delta} = \Delta'(\check{\tau}, \tau^*) + \Delta''(\check{\tau}, \tau^*)$.

Note that $y(t)$ is absolutely continuous. Hence, $V(\tilde{\tau}) - V(\tau^*) = \tilde{\Delta} + S^1(y(\tilde{\tau}), \tilde{\tau}) - S^1(y(\tau^*), \tau^*) = \tilde{\Delta} + \int_{\tilde{\tau}}^{\tau^*} (d/dt) S^1(y(t), t) dt$. Thus, from (42) we get

$$(43) \quad V(\tilde{\tau}) - V(\tau^*) \leq - \int_{\tilde{\tau}}^{\tau^*} d(t) dt + \int_{\tilde{\tau}}^{\tau^*} \Phi(t) \dot{y}(t) dt - \int_{\tilde{\tau}}^{\tau^*} (S_x^1(y(t), t) \dot{y}(t) + S_t^1(y(t), t)) dt + \varepsilon'.$$

Next, note that at each point $\bar{t} \in (\tau_1, \tau_2)$, the function $t \rightarrow \gamma^{\bar{t}} R^1(y(t), t)$ has a minimum at $t = \bar{t}$, i.e., $\gamma^{\bar{t}} (R_x^1(y(\bar{t}), \bar{t}) \cdot \dot{y}(\bar{t}) + R_t^1(y(\bar{t}), \bar{t})) = 0$ a.e. Similarly, $\beta^{\bar{t}} g_x(y(\bar{t}), u^{\bar{t}}(\bar{t}), t) \cdot \dot{y}(\bar{t}) + \beta^{\bar{t}} g_t(y(\bar{t}), u^{\bar{t}}(\bar{t}), \bar{t}) = 0$ a.e. Then, in a shorthand notation,

$$(44) \quad V(\tilde{\tau}) - V(\tau^*) \leq - \int_{\tilde{\tau}}^{\tau^*} (d(t) + \beta^{\bar{t}} g_t + S_t^1 + \gamma^{\bar{t}} R_t^1) dt - \int_{\tilde{\tau}}^{\tau^*} (-\Phi \dot{y} + \beta^{\bar{t}} g_x \dot{y} + S_x^1 \dot{y} + \gamma^{\bar{t}} R_x^1 \dot{y}) dt + \varepsilon'.$$

Using (28) and (37), we get $V(\tilde{\tau}) - V(\tau^*) \leq \varepsilon(\tau_2 - \tau_1) + \varepsilon$, i.e., $V(\tilde{\tau}) \leq V(\tau^*)$, since ε was arbitrary. The case $\tilde{\tau} \in (\tau^*, \tau_2]$, in the case where $\tau^* < \tau_2$, is treated completely symmetrically. Hence the proof of Theorem 4 is finished. Note that the Lipschitz continuity of $\tau \rightarrow x^\tau(\tau)$ is used only to establish (43) and (44). Mere continuity is needed when $S^1 \equiv 0$, and (7) is the terminal condition and $s = s'$.

Remark 4. Theorem 4 can be generalized to the case where the functions $p^\tau(\cdot)$ are allowed to have a finite number of discontinuities caused by the constraints $g^j, j > s'$, of the type (77), (78) of [5], in either of two cases: (i) there exists a $\delta > 0$ such that for all τ , $p^\tau(\cdot)$ has no discontinuity point in $(\tau - \delta, \tau)$; (ii) $V(\tau)$ is a continuous function.

The conditions of Theorems 1 and 4 imply that $V(\tau)$ is continuous: In the situation of Theorem 1, this follows from the boundedness of $d(\cdot)$ and (16'), which also holds for τ' and τ interchanged. In the situation of Theorem 4, (43) holds for arbitrary $\tilde{\tau}$ and $\tau^*(\varepsilon')$ independent of $\tilde{\tau}$ and τ^* , which again implies continuity of $V(\tau)$.

Remark 5. Assume that $U = R'$. Then, in Theorem 3, the concavity condition on H and $g^j, j \leq s'$ can be replaced by the following conditions: $H(\bar{x}(t), \bar{u}(t), p(t), t) = H^*(\bar{x}(t), p(t), t)$, where H^* is defined in (37'). Furthermore, $H^*(x, p(t), t)$ has, for all t , a concave extension to the set $\text{co}\{x: \text{for some } u \in R', g^j(x, u, t) \geq 0 \text{ for all } j \leq s'\}$ as a function of x . Finally, if we define $I(t) = \{j: g^j(\bar{x}(t), \bar{u}(t), t) = 0, j \leq s'\}$, the matrix with elements $g_{u^i}^{j,i}(\bar{x}(t), \bar{u}(t), t), i = 1, \dots, r, j \in I(t)$ has a rank equal to the number of elements in $I(t)$, for v.e. t .

(The proof of Theorem 4 remains the same, since (38) holds even for the present conditions; see the proof of Theorem 9 in [5].)

Remark 6. Theorems 1 and 4 also hold if $\tau_1 = t_0$, and it is then sufficient that the conditions of the theorems are satisfied for $\tau \in (\tau_1, \tau_2]$. This assertion follows at once from the fact that $V(\tau)$ is continuous at $\tau = \tau_1 = t_0$. This continuity is a consequence of the existence of bounded sets U' and D such that $u^\tau(t) \in U'$ for all τ and t , and $x^\tau(t) \in D$ for $t_0 \leq t \leq \tau < t', t'$ near enough t_0 . (In Theorem 4, the Lipschitz continuity is needed for the last assertion).

Remark 7. A remark exactly analogous to Remark 1 pertains also to Theorem 4.

Remark 8. Assume for Theorems 1 and 4 that instead of (12), respectively (37), holding for some τ^* for all $\tau \in [\tau_1, \tau_2]$, there exist subintervals $[a_{i-1}, a_i]$ partitioning $[\tau_1, \tau_2]$ and points $\tau_i^* \in [a_{i-1}, a_i]$ such that (12), respectively (37), holds for $\tau^* = \tau_i^*$ in the subinterval $[a_{i-1}, a_i]$, for all i . Then one of the points τ_i^* is optimal.

Remark 9. Consider the case where $\tau_2 = \infty$ and let us enlarge the set of admissible pairs in problems (1)–(4) by adding pairs $(x(t), u(t))$ defined on $[t_0, \infty)$ satisfying (1)–(4) with $x^i(\tau)$ in (4) replaced by $\lim_{\tau \rightarrow \infty} x^i(\tau)$ (assumed to exist), $i = 1, \dots, l$ and by $\liminf_{\tau \rightarrow \infty} x^i(\tau)$, $i = l+1, \dots, m$. For simplicity, assume that the integral in (1) is convergent for $\tau = \infty$, for any admissible pair defined on an infinite interval. (If the so-called catching-up criterion is used (see [5]), then this last assumption is not needed.)

We then have the following two results: (I) If $\liminf_{\tau \rightarrow \infty} p^\tau(\tau)(x(\tau) - x^\tau(\tau)) \geq 0$ for each admissible solution defined on $[t_0, \infty)$ and the conditions of Remark 1 are satisfied, then τ^* is again optimal. (II) If $(\bar{x}(\cdot), \bar{u}(\cdot))$ defined on $[t_0, \infty)$ together with an adjoint function $p(\cdot)$ satisfy the conditions (5), (6), (9), (10) and the conditions $\liminf_{\tau \rightarrow \infty} p(\tau)(x^\tau(\tau) - \bar{x}(\tau)) \geq 0$ and $\liminf_{\tau \rightarrow \infty} p(\tau)(x(\tau) - \bar{x}(\tau)) \geq 0$ for all admissible solutions $x(\cdot)$ defined on $[t_0, \infty)$, and finally, the condition $d(\tau) \geq 0$ for all τ , then $(\bar{x}(\cdot), \bar{u}(\cdot))$ is optimal.

The two assertions above easily follow from an application of (14).

Example. Consider the problem

$$(i) \quad \max_{u(\cdot), \tau} \int_0^\tau (qu - c(u)) e^{-rt} dt,$$

$$(ii) \quad \dot{x} = -u, \quad x(0) = \bar{x} > 0, \quad x(\tau) \geq 0, \quad u \geq 0$$

where c is a C^2 function, $c' > 0$, $c'' > 0$, $q > 0$, $r > 0$, $c'(0) < q < c'(\infty)$, $c(0) > 0$, $\max_{u \geq 0} qu - c(u) = k > 0$. The problem can be given the following interpretations: x is a stock of a natural resource, u the extraction rate (the control variable), q the price of the resource, $c(u)$ the extraction cost. With $H = (qu - c(u)) e^{-rt} - p(t)u$, we get that $\dot{p} = 0$, i.e., $p(t) \equiv \bar{p} \geq 0$. By concavity, the Hamiltonian is maximized by $u \geq 0$ if and only if u satisfies

$$(iii) \quad (q - c'(u)) e^{-rt} - \bar{p} \leq 0 \quad (\text{if } u > 0).$$

Note that condition (iii) defines u as a continuous function $u(t; \bar{p})$ being nonincreasing in \bar{p} . For $\bar{p} = 0$, $u(t; 0) \equiv u_0 > 0$.

Define τ' by $\int_0^{\tau'} u_0 dt = \bar{x}$. A candidate satisfying the fixed final time necessary conditions (for $p_0 = 1$), with $\tau < \tau'$, must satisfy $x(\tau) \geq 0$ with strict inequality ($u_0 \geq u(t; \bar{p})$). On the other hand, for $\tau > \tau'$, \bar{p} has to be > 0 in order for $\int_0^\tau u(t; \bar{p}) dt$ not to exceed \bar{x} , i.e., $x(\tau) \geq 0$ is satisfied with equality in this case.

Thus for $\tau \geq \tau'$, \bar{p} is determined as a continuous nondecreasing function $\bar{p}(\tau)$ of τ by

$$(iv) \quad \int_0^\tau u(t; \bar{p}) = \bar{x}.$$

(Note that for $\bar{p} = 0$, the left side is ≥ 0 , while for $\bar{p} = q - c'(0)$, $u(t; \bar{p}) \equiv 0$ by (iii), and the left side of (4) is $< \bar{x}$).

For $\tau < \tau'$, let $\bar{p}(\tau) = 0$, and write $u_\tau(t) = u(t; \bar{p}(\tau))$, $d(\tau) = (qu_\tau(\tau) - c(u_\tau(\tau))) e^{-r\tau} - \bar{p}u_\tau(\tau)$.

Now, $u(t; \bar{p})$ is nonincreasing in t . Hence, by (4), $u_\tau(\tau) \rightarrow 0$ when $\tau \rightarrow \infty$. (Actually, for some $\tilde{\tau}$, $u_\tau(t) = 0$ for $\tau \geq t \geq \tilde{\tau}$). Thus $d(\tau) < 0$ for all $\tau \geq \tau''$ if τ'' is large enough, ($c(0) > 0$). For $\tau = 0$, $\bar{p} = 0$ and $d(0) = k > 0$, by one of our assumptions. Since $d(\tau)$ is continuous, for some $\tau^* > 0$, $d(\tau^*) = 0$.

Let τ^* be largest possible. Then $0 = d(\tau^*)$ implies $(qu_{\tau^*}(\tau^*) - c(u_{\tau^*}(\tau^*))) e^{-r\tau^*} = \bar{p}u_{\tau^*}(\tau^*) \geq 0$. That is, by the maximum condition, for $\tilde{\tau} < \tau^*$, $d(\tilde{\tau}) \geq (qu_{\tau^*}(\tau^*) - c(u_{\tau^*}(\tau^*))) e^{-r\tilde{\tau}} - \bar{p}u_{\tau^*}(\tau^*) \geq d(\tau^*)$. By definition of τ^* , $0 = d(\tau^*) \geq d(\tilde{\tau})$ if $\tilde{\tau} > \tau^*$. (We can even show that the two last, weak inequalities are in fact strict ones.)

It is now easily seen that the conditions in Remark 1 to Theorem 1 are satisfied. That is, $u_{\tau^*}(\cdot) = u(\cdot; \bar{p}(\tau^*))$ is optimal.

Here, what might be called qualitative arguments were used in order to show the existence of a candidate satisfying all the sufficient conditions. Such arguments are needed in the numerous cases where no explicit formulas can be obtained. Although exceedingly simple, this example, we hope, serves as an illustration of the fact that the tools presented may also be useful in cases of this type.

REFERENCES

- [1] P. M. MEREAU AND W. F. POWERS, *A direct sufficient condition for free final time optimal control problems*, this Journal, 14 (1976), pp. 613-622.
- [2] D. W. PETERSON AND J. H. ZALKIND, *A review of direct sufficient conditions in optimal control theory*, Internat. J. Control, 28 (1978), pp. 589-610.
- [3] A. SEIERSTAD, *Sufficient conditions in free final time optimal control problems. A comment*, J. Econom. Theory, 32 (1984), pp. 367-370.
- [4] ———, *Differentiability properties of the optimal value function in control theory*, J. Econom. Dynamics Control, 4 (1982), pp. 303-310.
- [5] A. SEIERSTAD AND K. SYDSAETER, *Sufficient conditions in optimal control theory*, Internat. Econom. Rev., 18 (1977), pp. 367-391.
- [6] ———, *Optimal Control Theory With Economic Applications*, North-Holland, Amsterdam-New York, 1987.

DISTURBANCE DECOUPLING PROBLEMS BY MEASUREMENT FEEDBACK: A CHARACTERIZATION OF ALL SOLUTIONS AND FIXED MODES*

VASFİ ELDEM† AND A. BÜLENT ÖZGÜLER†

Abstract. This paper presents a new approach to disturbance decoupling problems by measurement feedback (DDPM). The novelty in this approach is that it yields a characterization of the sets of all solutions to DDPM and DDPM with internal stability and that the set of fixed modes of DDPM with respect to dynamic output feedback is easily identified. The central object used in deriving these results is a minimal polynomial basis of a rational vector space which conveniently represents the problem data.

Key words. disturbance decoupling, dynamic measurement feedback, internal stability, pole-placement, minimal polynomial basis of rational vector spaces

AMS(MOS) subject classifications. 93B25, 93B50

1. Introduction. Disturbance decoupling problems via dynamic measured output (measurement) feedback (DDPM) have been studied extensively in the last decade. Solutions to the problem without internal stability constraint have been obtained by Akashi and Imai [1979] and Schumacher [1980]. The problems with internal stability and pole placement have been solved by Willems and Commault [1981] and Imai and Akashi [1981]. In these papers, the concepts from the geometric approach to linear system theory was primarily used. Through a transfer matrix approach Ohm, Bhattacharyya and Howze [1984] have obtained an alternative solution to DDPM and using a stable rational fractional approach Pernebo [1981] has obtained solutions to the problems with or without internal stability. In Özgüler and Eldem [1985] a polynomial fractional approach has been taken to reduce the problems to the solvability of matrix equations of the type $AXB = C$ where a solution X is being sought on the various subrings of the field of rational functions. This approach has also yielded alternative solutions to DDPM with internal stability or pole-placement and the results have been used in Özgüler [1986] to make the relation between the geometric and polynomial fractional solutions to DDPM explicit.

Despite the existence of a rich variety of solutions (solvability conditions and synthesis procedures) and a thorough understanding of the structural aspects of disturbance decoupling problems, there are at least a few more problems that are worth studying in the same context. Among these are (i) a characterization of the set of all solutions to disturbance decoupling problems, (ii) finding an explicit expression for the set of fixed modes of DDPM in terms of the problem data, (iii) determination of a minimal McMillan degree (order) solution to DDPM, (iv) obtaining “approximate” solutions to DDPM which achieve maximal disturbance attenuation when the “exact” problem is not solvable and (v) examining the robustness issues involved in DDPM. Recall that the “state-feedback” version of the characterization problem has been solved by Forney [1975]. However, no such result exists in DDPM; in fact “a computationally attractive parametrization of the set of all solutions to DDPM” has been stated as an open problem in Ohm, Howze and Bhattacharyya [1984]. For the fixed mode characterization of the state-feedback version of DDPM, the work of Wolovich,

* Received by the editors January 6, 1986; accepted for publication (in revised form) August 11, 1986.

† Division of Applied Mathematics, Research Institute for Basic Sciences, P.O. Box 74 Gebze, Kocaeli, Turkey.

Antsaklis and Elliot [1977] can be mentioned. As to the characterization of fixed modes in DDPM, it can be easily seen that the earlier studies concerned with DDPM with internal stability all come very close to the resolution of this problem. However, this point is not directly addressed in any of these works. The minimal McMillan degree solution of the state-feedback version of disturbance decoupling problems has been obtained by Wang and Davison [1973] and later in a more systematic setting by Forney [1975]. Concerning the same point in DDPM, one finds nothing more than a brief discussion in for example Schumacher [1980] and Willems and Commault [1981]. In the present paper, we do not directly address this problem. We should add, however, that the approach taken here has already led to some preliminary results towards the determination of minimal order solutions for DDPM (see Eldem, Özgüler and Başer [1986]). Concerning issues (iv) and (v), we mention the works by Willems [1982] on “almost disturbance decoupling” and by Bhattacharyya, del Nero Gromes and Howze [1983] on “robustness” respectively.

This paper is primarily concerned with points (i) and (ii). In characterizing the set of all solutions to DDPM with or without internal stability and in identifying the fixed modes of DDPM (with respect to dynamic measurement feedback), we carry out the following conceptual program: (1) Identify a rational vector space which conveniently represents the problem data. (2) State the solvability condition of various disturbance decoupling problems on a minimal polynomial basis of this vector space. (3) Determine the sources of nonuniqueness that enrich the class of solutions and describe the set of all solutions accordingly. (4) Identify the fixed modes again on a minimal polynomial basis of the same vector space. Hence it is clear that this work rests heavily on the work of Forney [1975] on minimal polynomial bases of rational vector spaces. In identifying the relevant rational vector space the main source of insight has been the work of Kučera [1983] that considers the same problem for the scalar case. (The reader may note the similarity between the crucial matrix $[A: -B]$ in § 3 and matrix (8) in Kučera [1983].)

It might be argued that some recent concepts developed through a series of papers by Hammer and Heymann [1981], Vardoulakis and Karcianas [1984a], [1984b] of proper and proper stable minimal bases for rational vector spaces might provide a more compact setup compared to the minimal polynomial bases concept of Forney [1975]. However since the “system zeros” play a primary role in disturbance decoupling (see Özgüler and Eldem [1985]) we believe that using stable proper fractions and/or stable proper minimal basis will bring little improvement, if any, to the approach taken in this paper.

We would also like to mention the fact that the concept of internal stability employed here is slightly different, and in fact more general, than the corresponding concept in the papers by Imai and Akashi [1981], Willems and Commault [1981], Özgüler and Eldem [1985] and Ohm, Howze and Bhattacharyya [1984]. This point is discussed in more detail at the end of § 2.

The paper is structured as follows: In § 2 some results in Forney [1975] concerning minimal basis and solutions of the one-sided matrix equations $A = XB$ are stated in a suitable form for our purposes. The precise definitions of various disturbance decoupling problems also appear in § 2. Section 3 contains the solvability conditions, stated on the properties of a minimal basis of the relevant vector space, for the problem without internal stability constraint. A characterization for the set of all solutions of this problem is also included in § 3. In § 4 we identify “the fixed modes of disturbance decoupling under dynamic measurement feedback” through the employment of minimal basis and obtain solutions for the problem with internal stability and pole

placement. We then characterize the set of all solutions to the problem with internal stability in a similar manner to that of § 3. Section 4 is closed with the specialization of the results of the previous sections to the two-sided matrix equations $\Pi_4 = \Pi_3 X \Pi_2$, where Π_2 , Π_3 and Π_4 are arbitrary rational matrices and a solution X is sought over various subrings of $\underline{\mathbb{R}}(z)$. The last section is on conclusions.

2. Preliminaries and problem formulation. This section contains a summary of certain results concerning rational vector spaces, with special emphasis on the concept of minimal basis and proper solutions X to the rational matrix equation $A = XB$. Our exposition of the minimal (polynomial) basis and its relevance to model matching problems closely follows Forney [1975]. The disturbance decoupling problems that we consider are defined towards the end of this section, where we also emphasize the difference between the stable disturbance decoupling formulated here and in some earlier works. For various undefined concepts related to polynomial and rational matrices such as “right unimodular, proper, greatest left factor, etc.” the reader is referred to Özgüler and Eldem [1985].

As usual $\underline{\mathbb{R}}$, $\underline{\mathbb{R}}[z]$ and $\underline{\mathbb{R}}(z)$ denote the field of real numbers, the ring of polynomials in the indeterminate z with coefficients from $\underline{\mathbb{R}}$, and the field of real rational functions of z , respectively. A rational vector space \mathcal{V} of k -tuples of rational functions is an $\underline{\mathbb{R}}(z)$ -linear set which admits a rational basis consisting of v ($=\dim \mathcal{V}$) $\underline{\mathbb{R}}(z)$ -linearly independent elements. A $k \times v$ full column rank rational matrix, the columns of which span \mathcal{V} , will be called a rational basis of \mathcal{V} . Among the rational basis of \mathcal{V} there are polynomial ones which can, for instance, be obtained by multiplying a given rational basis by the least common denominator of all its entries. Such a basis is called a polynomial basis of \mathcal{V} . A *minimal basis* of \mathcal{V} is defined as a polynomial basis which is (i) *right unimodular* and (ii) *column proper*. Recall that a $k \times v$ polynomial matrix V is column proper iff its high column coefficient matrix V_h has full column rank, i.e., $\rho[V_h] = v$. The order $\mu(V)$ of V is defined as $\mu = \sum_{i=1}^v \mu_i$, where μ_i 's are the column degrees of V . A minimal base V of \mathcal{V} has the least order among all polynomial bases of \mathcal{V} (Forney [1975]). Furthermore, given two minimal bases V and \hat{V} of \mathcal{V} , they satisfy (i) $V = \hat{V}U$ for some unimodular matrix U , (ii) $\mu(V) = \mu(\hat{V})$ and (iii) they have the same column degrees up to a permutation of their columns. In other words, μ and μ_i 's are invariant under different representations of the vector space \mathcal{V} ; thus they are usually attributed to \mathcal{V} , i.e., μ is called the *order of \mathcal{V}* and μ_i 's are called the *dynamical indices of \mathcal{V}* (Forney [1975]).

The notion of minimal basis proved most useful in solving the following problem: Let Z_g and Z_d be $p \times l$ and $m \times l$ rational matrices respectively and determine a proper rational matrix Z_c such that $Z_d = Z_c Z_g$, i.e.,

$$(2.1) \quad [Z_c : -I] \begin{bmatrix} Z_g \\ Z_d \end{bmatrix} = 0.$$

Further, if a solution exists, then find one with the least McMillan degree. This problem, known as the *exact model matching problem*, was first solved by Wolovich [1974] and minimal McMillan degree solution was first obtained by Wang and Davison [1973]; later, the minimal basis approach of Forney provided a convenient framework for such problems. We therefore briefly review this approach below.

Let \mathcal{V} denote the rational vector space spanned by the columns of $[Z_g' : Z_d']'$ and V be a minimal basis for \mathcal{V} . The orthogonal complement in $\underline{\mathbb{R}}^{p+m}(z)$ of \mathcal{V} is given by

$$\mathcal{V}^\perp := \{x \text{ in } \underline{\mathbb{R}}^{p+m}(z) : x'y = 0 \text{ for all } y \text{ in } \mathcal{V}\}.$$

Let W be a minimal basis for \mathcal{V}^\perp , where $w := \dim \mathcal{V}^\perp = p + m - \dim \mathcal{V}$.

LEMMA 2.1 (Forney [1975]). (i) *There exists a proper rational Z_c satisfying (2.1) (equivalently $[Z_c: -I]V = 0$), iff the last m rows of W_h are linearly independent.*

(ii) *If a solution exists, then the set of all solutions of (2.1) is given by $\{Z_c := Q_c^{-1}R_c: [R_c: -Q_c]' = WT\}$, where T is a $w \times m$ right unimodular polynomial matrix such that WT is column proper and the last m rows of $(WT)_h$ are linearly independent.*

It is possible to state the solvability conditions for (2.1) more directly on the problem data. For this aim, we first establish the following simple results concerning column proper polynomial matrices.

LEMMA 2.2. *Let W and V be column proper polynomial matrices of sizes $k \times w$ and $k \times v$, respectively. Then,*

(i) *$V = WT$ for some polynomial matrix T implies that $V_h = W_h T_0$ for some constant matrix T_0 ,*

(ii) *$W'V = 0$ implies $W'_h V_h = 0$.*

Proof. Both statements follow by employing the representation $V = V_h \Lambda_v + L_v$ of a column proper polynomial matrix (see Kailath [1980, p. 384]), where Λ_v is a diagonal polynomial matrix with the same column degrees as V and L_v is a polynomial matrix such that $L_v \Lambda_v^{-1}$ is strictly proper. We omit the details of this easy proof. \square

By using Lemma 2.2 and the equation $[Z_c: -I]V = 0$, where V is a minimal basis for \mathcal{V} , the column span of $[Z'_g: Z'_d]'$, the solvability conditions for (2.1) can be alternatively expressed as follows.

LEMMA 2.3. *There exists a proper solution Z_c of $[Z_c: -I]V = 0$ iff the submatrix consisting of the first p rows of V_h has full column rank $v := \dim \mathcal{V}$.*

Proof. If $[Z_c: -I]V = 0$ admits a proper solution Z_c , then by Lemma 2.1 the last m rows of W_h are linearly independent. Furthermore, Lemma 2.2(ii) implies $W'_h V_h = 0$. Partitioning W_h and V_h compatibly, we have

$$W'_{h1} V'_{h1} = -W'_{h2} V_{h2},$$

where W'_{h2} has full column rank, i.e., there exist a constant matrix L such that $LW'_{h2} = I$. Thus $V_{h2} = -LW'_{h1} V_{h1}$ which implies that $V_h = [I: -W_{h1}L']V_{h1}$. Now since $\rho[V_h] = v$, it follows that $\rho[V_{h1}] = v$.

Conversely, let \bar{V}_{h1} be a $v \times (p - v)$ constant matrix such that $[V_{h1}: \bar{V}_{h1}]$ is nonsingular and consider $Z_c := [V_2: 0][V_1: \bar{V}_{h1}]^{-1}$, where $V := [V'_1: V'_2]'$. Clearly, $Z_c V_1 = V_2$, i.e., $[Z_c: -I]V = 0$. Properness of Z_c follows by the fact that $\partial_i[V_1: \bar{V}_{h1}] \geq \partial_i[V_2: 0]$, where $\partial_i[\cdot]$ denotes the i th column degree. \square

In this paper, we consider *the disturbance decoupling problem with measurement feedback* (DDPM), *the disturbance decoupling problem with measurement feedback and internal stability* (DDPMS) and *the disturbance decoupling problem with measurement feedback and pole placement* (DDPMP), all of which will be defined in a transfer matrix (input-output) setting.

Let Z_1 , Z_2 , Z_3 , and Z_4 be $p \times m$, $p \times s$, $q \times m$, and $q \times s$ rational matrices, respectively, with Z_1 strictly proper, and consider the following composite system:

$$(2.2) \quad \begin{bmatrix} y_m \\ y \end{bmatrix} = \begin{bmatrix} Z_1 & Z_2 \\ Z_3 & Z_4 \end{bmatrix} \begin{bmatrix} u \\ d \end{bmatrix},$$

where y_m denotes *the measured outputs*, y *the controlled outputs*, u *the control inputs*, and d *the disturbance inputs*.

DEFINITION 2.1. For the composite system given by (2.2) determine an $m \times p$ proper rational matrix Z_c such that under the control law

$$(2.3) \quad y = -Z_c y_m + u_e,$$

where u_e is a possible external input to the closed-loop system,

(DDPM) The closed-loop transfer matrix $Z_{dy} = Z_4 - Z_3 Z_c (I + Z_1 Z_c)^{-1} Z_2$ from d to y is identically zero.

(DDPMS) Given an admissible stability region Ω , which is a conjugate symmetric region of the complex plane with at least one point on the real axis, Z_{dy} is identically zero and the poles of all for rational matrices $\Delta := Z_c(I + Z_1 Z_c)^{-1}$, $Z_1 \Delta$, ΔZ_1 , and $Z_1 \Delta Z_1 - Z_1$ are in Ω . (Note that as Z_1 is strictly proper and Z_c is proper, $I + Z_1 Z_c$ is bicausal and hence $(I + Z_1 Z_c)^{-1}$ exists.)

(DDPMP) Find the solvability conditions for DDPMS for an arbitrarily given admissible stability region Ω .

Our definition of DDPMS differs from the earlier definitions in the requirement of internal stability. Note that in DDPMS all four transfer matrices Δ , $Z_1 \Delta$, ΔZ_1 and $Z_1 \Delta Z_1 - Z_1$ should have all their poles in Ω . This is well known to be equivalent in state space terminology to the condition that a *canonical realization of Z_c internally stabilizes a canonical realization of Z_1 under the control law (2.3)* (see Desoer and Chan [1975]). Let

$$\hat{Z} := \begin{bmatrix} Z_1 & Z_2 \\ Z_3 & Z_4 \end{bmatrix}$$

and consider a slightly different version DDPMS' of DDPMS as determining a proper Z_c such that, under (2.3), (i) DDPM is solved and (ii) a canonical realization of Z_c internally stabilizes a canonical realization of \hat{Z} . This is, in essence, the problem considered in Ohm, Howze and Bhattacharyya [1984] in a transfer matrix setting, in Imai and Akashi [1981] and Willems and Commault [1981] in a state space setting and in Özgüler and Eldem [1985] in a polynomial fractional setting. The problems DDPMS and DDPMS' are simply related as expressed by the following result: Let $Z = Z_s + Z_u$ be the partial fraction expansion of Z into rational matrices Z_s and Z_u , where Z_s has all its poles in Ω and Z_u has all its poles outside Ω . Also let $\nu^+[Z]$ denote the McMillan degree of Z_u .

PROPOSITION 2.1. *DDPMS' for (2.2) is solvable iff (i) $\nu^+[Z_1] = \nu^+[\hat{Z}]$ and (ii) DDPMS is solvable for (2.2). Further, if (i) holds, then any solution to DDPMS is also a solution to DDPMS', and conversely.*

Proof. The result follows easily by the definition of DDPMS and by Ohm, Howze and Bhattacharyya [1984, Thm. 3.1] where one lets $Z_1 = M$, $Z_2 = N$, $Z_3 = G$ and $Z_4 = H_d - H$. \square

Note that in cases where $\nu^+[Z_1] \neq \nu^+[\hat{Z}]$, DDPMS' has no solution whereas it may still be possible to solve DDPMS. In this sense, DDPMS is a more general problem definition and in cases where "the plant" is represented by the transfer matrix Z_1 between the control inputs and the measured outputs, a more realistic approach would be to solve DDPMS rather than DDPMS'.

We close this section by giving a first set of solvability conditions for DDPMS. The result follows directly by the characterization of internal stability of the pair (Z_1, Z_c) given by Desoer and Chan [1975].

PROPOSITION 2.2. *Let $Z_1 = Q^{-1}R$ be a left coprime polynomial fractional representation of Z_c . Then, DDPMS is solvable iff there exists a proper Z_c such that for some right coprime polynomial fractional representation $Z_c = P_c Q_c^{-1}$, both of the following conditions hold: (i) $\det(QQ_c + RP_c)$ has all its zeros in Ω and (ii) $Z_{dy} = Z_4 - Z_3 Z_c (I + Z_1 Z_c)^{-1} Z_2$ is identically zero.*

3. A characterization of solutions to DDPM. In this section, we present a new set of solvability conditions for DDPM, through which it is possible to characterize the set of all solutions to DDPM.

Let us define

$$A := \begin{bmatrix} I & Z_1 \\ 0 & Z_3 \end{bmatrix}, \quad B := \begin{bmatrix} Z_2 \\ Z_4 \end{bmatrix},$$

where Z_1, Z_2, Z_3, Z_4 are as given in (2.2). Also, for convenience let

$$\bar{m} := \dim(\text{Ker } A), \quad \bar{s} := \dim(\text{Im } B),$$

where $\text{Ker } A$ denotes the kernel of A considered as a map from $\underline{R}(z)^{p+m}$ to $\underline{R}(z)^{p+q}$ and $\text{Im } B$ denotes the image of B considered as a map from $\underline{R}(z)^s$ into $\underline{R}(z)^{p+q}$. Throughout this section, we will be mainly concerned with the rational vector space $[\underline{R}(z)\text{-linear set}] A^{-1}(\text{Im } B)$. Since $\text{Ker } A \subset A^{-1}(\text{Im } B)$, $A^{-1}(\text{Im } B)$ can be decomposed, in general nonuniquely, as

$$A^{-1}(\text{Im } B) = \mathcal{V} \oplus \text{Ker } A.$$

This decomposition implies the existence of a polynomial basis $W := [V : N]$ of $A^{-1}(\text{Im } B)$, where V is a polynomial basis of \mathcal{V} and N is a minimal basis for $\text{Ker } A$. Let $Z_1 := Q^{-1}R$ be a left coprime polynomial fractional representation of Z_1 with Q row proper. It follows by strict properness of Z_1 that $[Q : R]_h' = [Q_h : 0]'$ with Q_h nonsingular. The equality $[Q : R]N = 0$ now implies by Lemma 2.2(ii) that

$$N_h = [0 : N'_{h2}]',$$

for some $m \times \bar{m}$ full column rank constant matrix N_{h2} .

In order to determine a polynomial basis of the form $W = [V : N]$ for $A^{-1}(\text{Im } B)$, it is enough to consider a polynomial basis for $\text{Ker } S$, where

$$S := [A : -B] = \begin{bmatrix} I & Z_1 & -Z_2 \\ 0 & Z_3 & -Z_4 \end{bmatrix}.$$

In terms of the $\underline{R}(z)$ -linear sets \mathcal{A}_0 and \mathcal{B}_0 given by

$$\mathcal{A}_0 := \left\{ \begin{bmatrix} x \\ 0 \end{bmatrix} \text{ in } \underline{R}(z)^{p+m+s} : x \text{ is in } \text{Ker } A \right\},$$

$$\mathcal{B}_0 := \left\{ \begin{bmatrix} 0 \\ x \end{bmatrix} \text{ in } \underline{R}(z)^{p+m+s} : x \text{ is in } \text{Ker } B \right\},$$

$\text{Ker } S$ can be written (in general nonuniquely) as

$$\text{Ker } S = \mathcal{V}_0 \oplus \mathcal{A}_0 \oplus \mathcal{B}_0$$

for some $\underline{R}(z)$ -linear set \mathcal{V}_0 . Clearly, $\dim \mathcal{V}_0 = \dim(\text{Im } A \cap \text{Im } B)$. Consequently, there exists a polynomial basis W_0 of $\text{Ker } S$ in the form,

$$(3.1) \quad W_0 = \begin{bmatrix} V & N & 0 \\ M_0 & 0 & M \end{bmatrix},$$

where N is a minimal basis for $\text{Ker } A$, M is a minimal basis for $\text{Ker } B$ and $[V' : M'_0]'$ is a minimal basis for \mathcal{V}_0 . We will now show that $W := [V : N]$ is a polynomial basis of $A^{-1}(\text{Im } B)$. First note that W has full column rank; otherwise, there would exist a nonzero rational vector $\alpha = [\alpha'_1 : \alpha'_2]'$ satisfying $V\alpha_1 = -N\alpha_2$. Since N has full column rank, α_1 is nonzero. Now note that $AV\alpha_1 = -AN\alpha_2 = BM_0\alpha_1 = 0$, which implies that

$M_0\alpha_1$ is in $\text{Ker } B$ while $V\alpha_1$ is in $\text{Ker } A$. Consequently, $[V':M'_0]'\alpha_1$ is in $\mathcal{A}_0 \oplus \mathcal{B}_0$, which is a contradiction. On the other hand since $AV = BM_0$ and $AN = 0$, the column span of W is in $A^{-1}(\text{Im } B)$. By the fact that $\dim \mathcal{V}_0 = \dim (\text{Im } A \cap \text{Im } B)$, it now follows that the column span of W is precisely $A^{-1}(\text{Im } B)$. This proves the underlined statement above.

Remark 3.1. The argument above is based on the assumption that there exists a polynomial basis W_0 of $\text{Ker } S$ of the form given by (3.1). In order to calculate such a basis for $\text{Ker } S$, one might start with a minimal basis $\tilde{W}_0 := [\tilde{W}':\tilde{M}']'$ of $\text{Ker } S$, where \tilde{W} and \tilde{M} have $p+m$ and s rows, respectively; and follow the steps given below:

- (1) Find a unimodular matrix U_1 such that

$$\tilde{W}U_1 = [W:0], \quad \tilde{M}U_1 = [\tilde{M}_0:M]$$

where W and M are of full column rank and M is column proper.

- (2) Find a unimodular matrix U_2 such that

$$\tilde{M}_0U_2 = [M_0:0], \quad WU_2 = [V:N],$$

where M_0 and N are of full column rank and N is column proper. Clearly, M and N are minimal bases for $\text{Ker } B$ and $\text{Ker } A$, respectively. Thus,

$$\begin{bmatrix} V & N & 0 \\ M_0 & 0 & M \end{bmatrix}$$

is the desired basis for $\text{Ker } S$.

Given a polynomial basis $W = [V:N]$ of $A^{-1}(\text{Im } B)$, a minimal basis \bar{W} of $A^{-1}(\text{Im } B)$ can be obtained by extracting a greatest right factor T of W so that

$$\bar{W} := WT^{-1}$$

is column proper (and right unimodular).

Our first main result below yields a solvability condition for DDPM in terms of a minimal basis for the rational vector space $A^{-1}(\text{Im } B)$.

THEOREM 3.1. *DDPM for the composite system model (2.2) is solvable if and only if (i) $\rho[Z_2] = \bar{s}$ and (ii) the submatrix consisting of the first p rows of \bar{W}_h has rank \bar{s} , where \bar{W} is a minimal basis for $A^{-1}(\text{Im } B)$.*

Proof. Necessity: Let Z_c be a solution to DDPM so that by Definition 2.1, $Z_4 = Z_3\Delta Z_2$ with $\Delta := Z_c(I + Z_1Z_c)^{-1}$. It follows at once that

$$(3.2) \quad \text{Im } Z_4 \subset \text{Im } Z_3, \quad \text{Ker } Z_2 \subset \text{Ker } Z_4,$$

and also that

$$\begin{bmatrix} I & Z_1 \\ 0 & Z_3 \end{bmatrix} \begin{bmatrix} I \\ Z_c \end{bmatrix} (I + Z_1Z_c)^{-1} Z_2 = \begin{bmatrix} Z_2 \\ Z_4 \end{bmatrix}.$$

The first inclusion in (3.2) is equivalent to $\text{Im } B \subset \text{Im } A$ and the second, to $\rho[Z_2] = \rho[B] = \bar{s}$. Now let U be an $s \times s$ unimodular polynomial matrix such that

$$Z_2U = [\tilde{Z}_2:0],$$

where \tilde{Z}_2 has full column rank \bar{s} . In view of $\text{Ker } Z_2 \subset \text{Ker } Z_4$, we have $Z_4U = [\tilde{Z}_4:0]$ for some $q \times \bar{s}$ rational matrix \tilde{Z}_4 . Let $V = [V'_1:V'_2]'$ be a minimal basis for the column span over $\mathbb{R}(z)$ of $[I:Z'_c]'(I + Z_1Z_c)^{-1}\tilde{Z}_2$; note as the latter matrix is of full column rank that \tilde{V} is $(p+m) \times \bar{s}$ and it satisfies

$$(3.3) \quad \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} Y = \begin{bmatrix} I \\ Z_c \end{bmatrix} (I + Z_1Z_c)^{-1} \tilde{Z}_2.$$

For a nonsingular $\bar{s} \times \bar{s}$ rational matrix Y ; equation (3.3) now yields

$$(3.4) \quad \begin{bmatrix} I & Z_1 \\ 0 & Z_3 \end{bmatrix} \begin{bmatrix} V_1 & N_1 \\ V_2 & N_2 \end{bmatrix} = \begin{bmatrix} Z_2 \\ Z_4 \end{bmatrix} U \begin{bmatrix} Y^{-1} & 0 \\ 0 & I \end{bmatrix},$$

where $N := [N'_1 : N'_2]'$ is a minimal basis for $\text{Ker } A$. This equality, in turn, implies that the column span of $W := [V : N]$ is in $A^{-1}(\text{Im } B)$. Also note that, by (3.3), $[Z_c : -I]V = 0$, where Z_c is proper. By Lemma 2.3, it follows that the submatrix consisting of the first p rows of V_h has full column rank \bar{s} . By the property of a minimal basis N of $\text{Ker } A$, we now have

$$W_h = \begin{bmatrix} V_{1h} & 0 \\ V_{2h} & N_{2h} \end{bmatrix},$$

where V_{1h} and N_{2h} are of full column rank. This implies that W_h is of full column rank and, in particular, $\rho[W] = \bar{s} + \bar{m}$, i.e., W is also of full column rank. Since $\dim[A^{-1}(\text{Im } B)] = \dim(\text{Im } B \cap \text{Im } A) + \dim(\text{Ker } A) = \bar{s} + \bar{m}$, it now follows that W is a polynomial basis of $A^{-1}(\text{Im } B)$. If \bar{W} is any minimal basis for $A^{-1}(\text{Im } B)$, it is related to W by $W = \bar{W}T$, where T is a nonsingular polynomial matrix. By Lemma 2.2(i), and by column properness of both W and \bar{W} , it now follows that $W_h = \bar{W}_h T_0$ for some constant nonsingular matrix T_0 . Consequently, the submatrix consisting of the first p rows of \bar{W}_h is of rank equal to $\rho[V_{h1} : 0] = \bar{s}$.

Sufficiency: Let \bar{W} be a minimal basis for $A^{-1}(\text{Im } B)$ so that, by hypothesis, the submatrix consisting of the first p rows of \bar{W}_h is of rank \bar{s} . Let T_0 be a $(\bar{s} + \bar{m}) \times \bar{s}$ constant matrix which picks up those \bar{s} linearly independent columns of \bar{W}_h , i.e., T_0 is such that $V := \bar{W}T_0$ has the property that the submatrix consisting of the first p rows of V_h has rank \bar{s} . Let $W := [V : N] = [\bar{W}T_0 : N]$, where N is a minimal basis for $\text{Ker } A$. Clearly, W is a polynomial basis of $A^{-1}(\text{Im } B)$ and

$$W_h = \begin{bmatrix} V_{h1} & 0 \\ V_{h2} & N_{h2} \end{bmatrix},$$

where V_{h1} is the submatrix consisting of the first p rows of V_h , and V_{h2}, N_{h2} are the submatrices consisting of the last m rows of V_h, N_h , respectively. It follows that W is column proper, is of full column rank, and that $\text{Im } V \cap \text{Ker } A = \{0\}$. By Lemma 2.3, there exists a proper $m \times p$ Z_c such that $[Z_c : -I]V = 0$ or equivalently, $Z_c V_1 = V_2$, where $V = [V'_1 : V'_2]'$. By the fact that W is a polynomial basis for $A^{-1}(\text{Im } B)$, we have $\text{Im } V \subset A^{-1}(\text{Im } B)$ and hence there exists a $s \times \bar{s}$ rational matrix Y satisfying

$$\begin{bmatrix} I & Z_1 \\ 0 & Z_3 \end{bmatrix} \begin{bmatrix} I \\ Z_c \end{bmatrix} V_1 = \begin{bmatrix} Z_2 \\ Z_4 \end{bmatrix} Y.$$

The matrix BY is of full column rank; since, if $BY\alpha = 0$ for some nonzero rational vector α , then $AV\alpha = 0$ which contradicts, as V is of full column rank, the fact that $\text{Im } V \cap \text{Ker } A = \{0\}$. Therefore, $\text{Im } Y \cap \text{Ker } B = \{0\}$ and further if \tilde{Y} is a basis for $\text{Ker } B$, the rational matrix $[Y : \tilde{Y}]$ is nonsingular and satisfies

$$\begin{bmatrix} I & Z_1 \\ 0 & Z_3 \end{bmatrix} \begin{bmatrix} I \\ Z_c \end{bmatrix} [V_1 : 0] = \begin{bmatrix} Z_2 \\ Z_4 \end{bmatrix} [Y : \tilde{Y}].$$

This equality implies, as $I + Z_1 Z_c$ is bicausal, that

$$Z_3 Z_c (I + Z_1 Z_c)^{-1} Z_2 [Y : \tilde{Y}] = Z_4 [Y : \tilde{Y}].$$

The nonsingularity of $[Y : \tilde{Y}]$ further implies that Z_c is actually a solution to DDPM. \square

The solvability condition of Theorem 3.1 is in terms of an arbitrary minimal basis for $A^{-1}(\text{Im } B)$. To obtain a useful characterization of the set of all solutions to DDPM, we fix a particular minimal basis W^* of $A^{-1}(\text{Im } B)$, the high coefficient matrix of which is in a special form:

$$W_h^* = \begin{bmatrix} V_{h1}^* & 0 \\ V_{h2}^* & N_{h2}^* \end{bmatrix},$$

where V_{h1}^* , V_{h2}^* , and N_{h2}^* are $p \times \bar{s}$, $m \times \bar{s}$, and $m \times \bar{m}$ constant matrices with V_{h1}^* and N_{h2}^* of full column rank. Note that an arbitrary minimal basis \bar{W} can be brought to the desired form W^* by multiplication on the right with a unimodular matrix. In subsequent discussions on the characterization of all solutions to DDPM, we assume that DDPM is solvable and hence the solvability conditions Theorem 3.1(i) and 3.1(ii) hold and W^* as above exists.

Let $W^* = [V^*; N^*]$, where V^* is the submatrix consisting of the first \bar{s} columns of W^* and N^* is the submatrix representing the last \bar{m} columns of W^* . (Caution: N^* may no longer be a minimal basis for $\text{Ker } A$!). Let us further consider the representations

$$(3.5) \quad V^* = V_h^* \Lambda_1 + Y_1, \quad N^* = N_h^* \Lambda_2 + Y_2$$

of the column proper polynomial matrices V^* , N^* , respectively, where Λ_1, Λ_2 are diagonal polynomial matrices with entries z^{μ_i}, z^{ν_i} , where μ_i and ν_i denote the column degrees of V^* and N^* , respectively. The polynomial matrices Y_1 and Y_2 have the property that $Y_1 \Lambda_1^{-1}$ and $Y_2 \Lambda_2^{-1}$ are strictly proper rational matrixes, respectively. (For the existence and uniqueness of the representations (3.5), see Kailath [1980]). Consider the set

$$\mathcal{L}(W^*) := \{L \text{ in } \underline{R}(z)^{\bar{m} \times \bar{s}} : \Lambda_2 L \Lambda_1^{-1} \text{ is proper}\},$$

which can equivalently be described as the set of all $\bar{m} \times \bar{s}$ rational matrices whose ij th entry l_{ij} satisfies the degree constraint

$$\partial(l_{ij}) \leq \partial_j(V^*) - \partial_i(N^*); \quad i = 1, \dots, \bar{m}, \quad j = 1, \dots, \bar{s}.$$

COROLLARY 3.1. *A proper rational $m \times p$ matrix Z_c is a solution to DDPM if and only if there exists L in $\mathcal{L}(W^*)$ such that*

$$(3.6) \quad [Z_c : -I](V^* + N^* L) = 0.$$

Proof. Necessity: It is clear by the necessity part of the proof of Theorem 3.1 that any solution Z_c to DDPM satisfies (3.3) and (3.4) for some nonsingular $\bar{s} \times \bar{s}$ rational matrix Y , where $V = [V_1' : V_2']'$ is a minimal basis for the column span of $[I : Z_c'](I + Z_1 Z_c)^{-1} \tilde{Z}_2$. Since $\text{Im } V \subset A^{-1}(\text{Im } B)$ by (3.4), it follows that there exist polynomial matrices L_1 and L_2 satisfying

$$(3.7) \quad V = V^* L_1 + N^* L_2 = W^* T,$$

where $T := [L_1' : L_2']'$. Let V and W^* be written similar to the representations (3.5) as

$$V = V_h \Lambda + Y, \quad W^* = W_h^* \Lambda^* + Y^*,$$

where Λ is a diagonal matrix whose nonzero entries are z^{λ_i} with λ_i denoting the degree of the i th column of V , and Y is such that $Y \Lambda^{-1}$ is strictly proper. Note that Λ^* , Y^* , Λ_i , Y_i are related by

$$\Lambda^* = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}, \quad Y^* = [Y_1 : Y_2].$$

By (3.7), we now have

$$V_h + Y\Lambda^{-1} = (W_h^* + Y^*\Lambda^{*-1})\Lambda^*T\Lambda^{-1}$$

where W_h^* is of full column rank. By strict properness of $Y\Lambda^{-1}$ and $Y^*\Lambda^{*-1}$, it easily follows from the above equality that $\Lambda^*T\Lambda^{-1}$ is proper and that

$$V_h = W_h^*(\Lambda^*T\Lambda^{-1})_0$$

where $(\cdot)_0$ denotes the constant term of the argument in its formal Laurent series expansion in z^{-1} . Letting $[V'_{h1} : V'_{h2}]' = V_h$ be a compatible partitioning of V_h , we have

$$\begin{bmatrix} V_{h1} \\ V_{h2} \end{bmatrix} = \begin{bmatrix} V_{h1}^* & 0 \\ V_{h2}^* & N_{h2}^* \end{bmatrix} \begin{bmatrix} (\Lambda_1 L_1 \Lambda^{-1})_0 \\ (\Lambda_2 L_2 \Lambda^{-1})_0 \end{bmatrix},$$

where, by Lemma 2.3, V_{h1} has full column rank. We now have $V_{h1} = V_{h1}^*(\Lambda_1 L_1 \Lambda^{-1})_0$, where V_{h1} , V_{h1}^* are both of full column rank. It follows that $(\Lambda_1 L_1 \Lambda^{-1})_0$ is nonsingular and hence $\Lambda_1 L_1 \Lambda^{-1}$ is bicausal. Therefore, if we let $L := L_2 L_1^{-1}$, the rational matrix

$$\Lambda_2 L \Lambda_1^{-1} = \Lambda_2 L_2 \Lambda^{-1} (\Lambda_1 L_1 \Lambda^{-1})^{-1}$$

is proper, proving the fact that L is in $\mathcal{L}(W^*)$.

Sufficiency: Let Z_c be a proper $m \times p$ matrix such that (3.6) holds. Let $L = L_2 L_1^{-1}$ be a right coprime polynomial fractional representation of L such that $V := V^* L_1 + N^* L_2$ is column proper. Let $V_h = [V'_{h1} : V'_{h2}]'$ where V_{h1} , V_{h2} are of sizes $p \times \bar{s}$, $m \times \bar{s}$, respectively. Note, by $[Z_c : -I]V = 0$ and properness of Z_c , that V_{h1} is of full column rank and $Z_c V_1 = V_2$, where $V = [V'_1 : V'_2]'$. Since $V = W^* T$, where $T := [L'_1 : L'_2]'$, $\text{Im } V \subset A^{-1}(\text{Im } B)$. The fact that $\rho[V_{h1}] = \bar{s}$ implies by Lemma 2.2(ii) that $\text{Im } V \cap \text{Ker } A = \{0\}$. Thus, there exists a rational $s \times \bar{s}$ matrix Y such that $A[I : Z'_c]' V_1 = BY$, where $\rho[BY] = \rho[B] = \bar{s}$. The rest of the proof now follows the steps of the sufficiency proof of Theorem 3.1. \square

The above result shows in effect that to obtain the set of all solutions to DDPM, it is enough to consider the set of all proper solutions Z_c of (3.6) as L runs over the elements of $\mathcal{L}(W^*)$. For a fixed L in $\mathcal{L}(W^*)$, it is clear by the definition of the set $\mathcal{L}(W^*)$ that (3.6) has at least one proper solution. (To see this, let $V := V^* L_1 + N^* L_2$, where $L_2 L_1^{-1}$ is a right coprime polynomial fractional representation of L , such that V_h has full column rank \bar{s} . In view of the fact that $\Lambda_2 L \Lambda_1^{-1}$ is proper, it follows that the submatrix consisting of the first p rows of V_h also has rank \bar{s} . Now, by Lemma 2.3, it holds that (3.6) has at least one proper solution Z_c .) In order to generate all proper solutions Z_c to (3.6) for a fixed L in $\mathcal{L}(W^*)$, characterization of Lemma 2.1(ii) can now be used. We remark here that the characterization Lemma 2.1(ii) can actually be made more explicit by fixing a special minimal basis for $(\text{Im } V)^\perp$, obtaining a similar characterization to that provided by $\mathcal{L}(W^*)$. We omit the details of this complementary part of the characterization. *The important point here is that Corollary 3.1 reduces the characterization of all solutions to DDPM to the characterization of all proper solutions of a one-sided equation (3.6).*

Remark 3.2. It is interesting to note that two distinct elements L and \tilde{L} ($L \neq \tilde{L}$) of $\mathcal{L}(W^*)$ generate two disjoint classes of solutions to DDPM. To see this suppose that Z_c is a proper solution to (3.6) for both L and \tilde{L} . Then $[Z_c : -I](V^* + N^* L) = 0$ and $[Z_c : -I](V^* + N^* \tilde{L}) = 0$, which yield $[Z_c : -I]N^*(L - \tilde{L}) = 0$. By properness of Z_c and by the fact that $N_h^* = [0 : N_{h2}^{*'}]'$, it is easy to see that $[Z_c : -I]N^*$ has full column rank \bar{m} . This implies that $L = \tilde{L}$, proving our claim. Therefore, *our characterization for $\mathcal{L}(W^*)$ has no redundancy.*

Remark 3.3. The characterization provided by Corollary 3.1 not only yields the set of all solutions to DDPM; (i) when we set $Z_1 = 0$, it yields the set of all proper solutions to the two-sided matrix equation $Z_4 = Z_3 X Z_2$; (ii) when we set $Z_1 = 0$, $Z_4 = 0$, it yields the set of all solutions to $Z_3 X Z_2 = 0$; (iii) when we set $Z_1 = 0$, $Z_2 = I$ ($Z_3 = I$), it yields the set of all solutions to the one-sided matrix equations $Z_4 = Z_3 X$ ($Z_4 = X Z_2$).

Remark 3.4. In consistency with the two-sided nature of DDPM, the nonuniqueness of solution stems from two sources; first, from nontriviality of $\text{Ker } A$, or equivalently, of $\text{Ker } Z_3$ and second from nontriviality of $(\text{Im } B)^\perp$, or equivalently, of $(\text{Im } Z_2)^\perp$. Our approach to the characterization of all solutions to DDPM handles these two sources of nonuniqueness in two consecutive stages. Thus, $\mathcal{L}(W^*)$ is nonempty iff $\text{Ker } A \neq \{0\}$ and each fixed L in $\mathcal{L}(W^*)$ yields a unique corresponding solution Z_c iff $(\text{Im } Z_2)^\perp = \{0\}$. Since there is a bijective correspondence between the set of all proper solutions X to

$$(3.8) \quad Z_4 = Z_3 X Z_2,$$

and the set of all solutions $Z_c = Z_1(I - XZ_1)^{-1}$ to DDPM, the above observations are also apparent from (3.8). This suggests a slightly different alternative to our method of characterization of all solutions to DDPM:

- (i) Determine the set of all proper solutions X to (3.8);
- (ii) Compute for each X , $Z_c = Z_1(I + XZ_1)^{-1}$. Step (i) of this alternative method can be performed by setting $Z_1 = 0$ in our procedure above (see Remark 3.3). Although with this alternative method some amount of simplification may result in performing step (i), altogether, both methods seem to require the same amount of computation.

4. Fixed modes of DDPM. The subject of this section is the determination of the “fixed modes” of disturbance decoupling problem under dynamic measurement feedback, from which the solutions of both DDPMs and DDMP immediately follow. Throughout this section we assume that DDPM is solvable and $Q^{-1}R = Z_1$ is a fixed left coprime polynomial fractional representation of Z_1 with Q row proper. Also recall from § 2 that $\Phi := QQ_c + RP_c$ represents the internal modes of the closed loop system, where $P_c Q_c^{-1}$ is any right coprime fractional representation of a solution Z_c to DDPM. The *fixed modes* of DDPM can now be defined as

$$(4.1) \quad \sigma := \bigcap_{Z_c} \text{zeros of } \det(QQ_c + RP_c),$$

where the intersection ranges over all solutions $Z_c = P_c Q_c^{-1}$ of DDPM in right coprime representation. Note that σ is independent of the particular representation chosen for Z_c . Let us also define

$$(4.2) \quad \sigma_0 := \text{invariant zeros of } [Q : R] \bar{W},$$

where \bar{W} is a minimal basis for $A^{-1}(\text{Im } B)$. Also note that σ_0 is independent of the particular minimal basis chosen for $A^{-1}(\text{Im } B)$, since minimal bases are related via post multiplication by unimodular matrices. Thus, σ_0 is also equal to the set of invariant zeros of $[Q : R] W^*$, where $W^* = [V^* : N^*]$ is a special minimal basis for $A^{-1}(\text{Im } B)$ of the type given in § 3, i.e., one such that

$$(4.3) \quad W_h^* = \begin{bmatrix} V_{h1}^* & 0 \\ V_{h2}^* & N_{h2}^* \end{bmatrix}.$$

The following technical lemmas will be used in establishing the main results of this section.

LEMMA 4.1. (i) *There exists a unimodular matrix U such that*

$$(4.4) \quad U[Q:R]W^* = \begin{bmatrix} C_1 & C_2 \\ 0 & 0 \end{bmatrix},$$

where C_1 and C_2 are $\bar{s} \times \bar{s}$ and $\bar{s} \times \bar{m}$ polynomial matrices with C_1 nonsingular.

(ii) *Given L in $\mathcal{L}(W^*)$, let $L = L_2 L_1^{-1}$ be its some right coprime fractional representation. For every solution $Z_c = P_c Q_c^{-1}$ of (3.6), there exist a unimodular matrix U and a right unimodular matrix \tilde{U} such that*

$$(4.5) \quad U[Q:R](V^* L_1 + N^* L_2) = U(QQ_c + RP_c)\tilde{U}.$$

Proof. (i) There clearly exists a U such that (4.4) holds and $[C_1: C_2]$ is of full row rank. Since $[Q:R]_h' = [Q_h: 0]'$ for some nonsingular constant Q_h and $\rho[V_{h1}^*] = \bar{s}$ in expression (4.3), it easily follows that

$$(4.6) \quad \rho[[Q:R]V^*] = \rho[V^*] = \bar{s}.$$

Using the fact that $\rho[W^*] = \dim[A^{-1}(\text{Im } B)]$ we can also show by simple linear algebraic arguments that

$$\rho[[Q:R]W^*] = \dim[A^{-1}(\text{Im } B)] - \dim\{A^{-1}(\text{Im } B) \cap \text{Ker}[Q:R]\}.$$

Note that $\text{Ker}[Q:R] = \text{Ker}[I:Z_1]$ contains $\text{Ker } A$. By (4.6) and by the modular distributive rule, it follows that $A^{-1}(\text{Im } B) \cap \text{Ker}[Q:R] = \text{Ker } A$. Hence,

$$\rho[[Q:R]W^*] = \dim[A^{-1}(\text{Im } B)] - \dim(\text{Ker } A) = \bar{s}.$$

This implies that $\rho[C_1: C_2] = \bar{s}$ and by (4.6) nonsingularity of C_1 follows.

(ii) Let $V := V^* L_1 + N^* L_2$. As $\text{Ker}[Z_c: -I]$ is spanned by $[Q'_c: P'_c]'$, by (3.6) we have $V = [Q'_c: P'_c]'\tilde{U}$, for some right unimodular \tilde{U} . Substituting for V in the left-hand side of (4.5) we obtain the right-hand side. \square

LEMMA 4.2. *Given an admissible stability region Ω , there exists an L in $\mathcal{L}(W^*)$ such that the set of invariant zeros of $[Q:R](V^* L_1 + N^* L_2)$ is the union of σ_0 and a subset of Ω , where $L_2 L_1^{-1}$ is a right coprime polynomial fractional representation of L .*

Proof. Let C_0 be a greatest common left factor of C_1 and C_2 so that $C_1 = C_0 G_1$, $C_2 = C_0 G_2$ for some left coprime polynomial matrices G_1 and G_2 , where C_1, C_2 are obtained through Lemma 4.1(i). Clearly, the set of zeros of $\det C_0$ is precisely σ_0 . By Khargonekar and Özgüler [1984, Lemma 2.9], there exist polynomial matrices L_1, L_2 , and G with L_1 and G nonsingular such that the set of zeros of $\det G$ are in Ω , the causality degree of $L_2 L_1^{-1}$, $\partial(L_2 L_1^{-1})$, satisfies $\partial(L_2 L_1^{-1}) < -k$ for a given nonnegative integer k , and

$$(4.7) \quad G_1 L_1 + G_2 L_2 = G.$$

If k is chosen to satisfy $k \geq \max\{\partial_i(N^*)\}$, then it is easy to see that $L := L_2 L_1^{-1}$ is in $\mathcal{L}(W^*)$. By (4.7) and by Lemma 4.1(i), it now follows that

$$U[Q:R](V^* L_1 + N^* L_2) = \begin{bmatrix} C_0 G \\ 0 \end{bmatrix}.$$

Since the set of zeros of $\det G$ are in Ω , our claim follows. \square

THEOREM 4.1. *The set of fixed modes of DDPM is equal to the set of invariant zeros of $[Q:R]\tilde{W}$, i.e., $\sigma = \sigma_0$.*

Proof. Let $Z_c = P_c Q_c^{-1}$ be a solution to DDPM. By Corollary 3.1, there exists an $L = L_2 L_1^{-1}$ in $\mathcal{L}(W^*)$ satisfying (3.6). By Lemma 4.1(ii), there exists a unimodular U and a right unimodular \tilde{U} such that (4.5) holds. Let $\tilde{G} := C_1 L_1 + C_2 L_2$, where C_1 and

C_2 are defined via Lemma 4.1(i). Note that $\tilde{G} = C_0 G$ with C_0 denoting a greatest common left factor of C_1 and C_2 and with $G := C_0^{-1} \tilde{G}$ a polynomial matrix. We now have $U\Phi\tilde{U} = [\tilde{G}': 0]'$ which implies by right unimodularity of \tilde{U} that $\det \tilde{G}$, and hence, that $\det C_0$ divides $\det \Phi$. This establishes the inclusion $\sigma_0 \subset \sigma$. We will establish the reverse inclusion by showing that the zeros of $\det \Phi$ can be assigned arbitrarily except those which coincide with the zeros of $\det C_0$ by an appropriate choice of Z_c .

Given stability region Ω , by Lemmas 4.1(i) and 4.2, there exist $L = L_2 L_1^{-1}$ in $\mathcal{L}(W^*)$ and a unimodular U such that

$$(4.8) \quad U[Q:R](V^*L_1 + N^*L_2) = \begin{bmatrix} C_0 & G \\ 0 & \end{bmatrix},$$

where the set of zeros of $\det C_0$ is σ_0 and the set of zeros of $\det G$ is in Ω . Let $V := V^*L_1 + N^*L_2$ and partition V and V_h compatibly with $[Q:R]$ as $V := [V_1': V_2']'$ and $V_h := [V_{h1}': V_{h2}']'$. Since $\rho[V_{h1}] = \bar{s}$, there exists a $p \times p$ constant nonsingular J such that

$$J^{-1}V_{h1} = [I:0]'$$

Also let

$$(4.9) \quad UQJ := \begin{bmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{bmatrix}, \quad UR := \begin{bmatrix} R_1 \\ R_2 \end{bmatrix},$$

where Q_1 is $\bar{s} \times \bar{s}$ and R_1 is $\bar{s} \times m$. We assume here without loss of generality that $[Q_3:Q_4:R_2]$ is row proper. (This can always be achieved by further premultiplying UQJ and UR by a unimodular matrix without altering Q_1 , Q_2 and R_1 .) As Z_1 is strictly proper and $[Q_3:Q_4:R_2][(J^{-1}V_1)': V_2']' = 0$, Lemma 2.2(ii) implies that $[Q_3:Q_4:R_2]'_h = [0:\tilde{T}':0]'$ for some nonsingular \tilde{T} . Consequently Q_4 is nonsingular and hence by Khargonekar and Özgüler [1984, Lemma 2.9], there exist polynomial matrices P_3 , P_4 , P_2 , and P_0 with P_4 and P_0 nonsingular, the zeros of $\det P_0$ in Ω , $[P_2':P_3']'P_4^{-1}$ strictly proper, and $[P_2':P_3':P_4']'$ column proper such that

$$(4.10) \quad Q_3P_3 + Q_4P_4 + R_2P_2 = P_0.$$

Let $P := [P_3:P_4]$ and note that $[P_2':P_3':P_4']'_h = [0:0:T']'$ for some nonsingular T . Now consider

$$Q_c := [V_1:JP], \quad P_c := [V_2:P_2].$$

By writing the highest column coefficient matrix of $[(J^{-1}Q_c)':P_c']'$ explicitly in terms of V_h and T , we can easily show that Q_c is nonsingular and $P_cQ_c^{-1}J$, and hence $Z_c := P_cQ_c^{-1}$ are proper. Further, as $Z_cV_1 = V_2$, it follows by Corollary 3.1 that Z_c is a solution of DDPM. Finally, by (4.8)–(4.10) we have

$$U(QQ_c + RP_c) = \begin{bmatrix} C_0G & P_1 \\ 0 & P_0 \end{bmatrix},$$

where $P_1 := [Q_1:Q_3]JP + R_1P_2$. It follows that

$$\det \Phi = \det C_0 \det G \det P_0.$$

As the zeros of both $\det G$ and $\det P_0$ are in Ω , Z_c places all the zeros of $\det \Phi$ into Ω , except those which coincide with the zeros of $\det C_0$, i.e., σ_0 . Note that this can also be achieved for two given nonintersecting stability regions; consequently $\sigma \subset \sigma_0$ and the result follows. \square

COROLLARY 4.1. (i) DDPMS with respect to a given stability region Ω is solvable iff $\sigma_0 \subset \Omega$; (ii) DDPMP is solvable iff σ_0 is empty.

Proof. The proof immediately follows from Theorem 4.1 and the definition of σ_0 . \square

A characterization of all solutions of DDPMS is now in order. A solution Z_c of DDPMS can be constructed in two stages. First, one determines an L in $\mathcal{L}(W^*)$ such that the invariant zeros of $[Q:R]V$ are stable, where $V := V^*L_1 + N^*L_2$ and $L_2L_1^{-1}$ is a right coprime polynomial fractional representation of L . In the sequel, the class of such L will be denoted as $\mathcal{L}_\Omega(W^*)$. More specifically,

$$(4.11) \quad \mathcal{L}_\Omega(W^*) := \{L := L_2L_1^{-1} \text{ in } \mathcal{L}(W^*): \text{the invariant zeros of } [Q:R] \\ (V^*L_1 + N^*L_2) \text{ are in } \Omega.\}$$

Second, one constructs all possible solutions Z_c of DDPMS, which satisfy $[Z_c: -I] \cdot (V^* + N^*L) = 0$, by the synthesis procedure given in the sufficiency proof of Theorem 4.1. Now let \bar{L} be a particular element of $\mathcal{L}_\Omega(W^*)$ and $\bar{L} := S_2S_1^{-1}$ be a Ω -stable rational fractional representation of \bar{L} such that

$$(4.12) \quad G_1S_1 + G_2S_2 = I$$

Here, $C_0G_1 = C_1$ and $C_0G_2 = C_2$, where C_1 and C_2 are polynomial matrices as defined in Lemma 4.1(i). In view of Lemma 4.2 and (4.7) such S_1 and S_2 clearly exist. Also let K_1 and K_2 be $\bar{s} \times \bar{m}$ and $\bar{m} \times \bar{m}$ polynomial matrices such that

$$(4.13) \quad V^*K_1 + N^*K_2 = N,$$

where N is a minimal basis for $\text{Ker } A$. As $\text{Ker } A \subset A^{-1}(\text{Im } B)$ such K_1 and K_2 obviously exist. Furthermore, since $\text{Ker } [Q:R] \cap A^{-1}(\text{Im } B) = \text{Ker } A$ (immediate consequence of Lemma 4.1(i)), it follows that $[K'_1: K'_2]'$ also span $\text{Ker } [G_1: G_2]$. Substituting the representations of V^* and N^* as in (3.5) and a similar representation of N into the equation above, we obtain

$$(V_h^* + Y_1\Lambda_1^{-1})\Lambda_1K_1 + (N_h^* + Y_2\Lambda_2^{-1})\Lambda_2K_2 = (N_h + Y_3\Lambda_3^{-1})\Lambda_3,$$

where $N = N_h\Lambda_3 + Y_3$ with Λ_3 diagonal and $Y_3\Lambda_3^{-1}$ strictly proper. Since $\text{Im } N_h^* = \text{Im } N_h$, it is straightforward to show that K_2 is nonsingular, $\Lambda_1K_1K_2^{-1}\Lambda_2^{-1}$ is strictly proper and $\Lambda_2K_2\Lambda_3^{-1}$ is bicausal. (See the necessity proof for Corollary 3.1. We can now give an explicit characterization of $\mathcal{L}_\Omega(W^*)$ in terms of the particular element \bar{L} of $\mathcal{L}_\Omega(W^*)$ satisfying (4.12).

LEMMA 4.3. An $\bar{m} \times \bar{s}$ rational matrix L is an element of $\mathcal{L}_\Omega(W^*)$ iff there exists a Ω -stable rational matrix M such that $\Lambda_3MS_1^{-1}\Lambda_1^{-1}$ is proper and

$$(4.14) \quad L = (S_2 + K_2M)(S_1 + K_1M)^{-1}.$$

Proof. Let L and M be rational matrices satisfying (4.14) and with $\Lambda_3MS_1^{-1}\Lambda_1^{-1}$ proper. Consider

$$\Lambda_2L\Lambda_1^{-1} = (\Lambda_2S_2S_1^{-1}\Lambda_1^{-1} + \Lambda_2K_2MS_1^{-1}\Lambda_1^{-1})(I + \Lambda_1K_1MS_1^{-1}\Lambda_1^{-1})^{-1}.$$

By properness of $\Lambda_3MS_1^{-1}\Lambda_1^{-1}$ and bicausality of $\Lambda_2K_2\Lambda_3^{-1}$ it follows that $\Lambda_2K_2MS_1^{-1}\Lambda_1^{-1}$ is proper. Hence, the first term on the right-hand side of the above equation is proper. For the second term note that

$$\Lambda_1K_1MS_1^{-1}\Lambda_1^{-1} = \Lambda_1K_1K_2^{-1}\Lambda_2^{-1}(\Lambda_2K_2MS_1^{-1}\Lambda_1^{-1}).$$

As $\Lambda_1K_1K_2^{-1}\Lambda_2^{-1}$ is strictly proper, it follows that $I + \Lambda_1K_1MS_1^{-1}\Lambda_1^{-1}$ is bicausal and that $\Lambda_2L\Lambda_1^{-1}$ is proper, i.e., L is in $\mathcal{L}(W^*)$. Also note that

$$G_1(S_1 + K_1M) + G_2(S_2 + K_2M) = I,$$

where both $S_1 + K_1 M$ and $S_2 + K_2 M$ are Ω -stable rational matrices as M is Ω -stable rational. Therefore, if $L_2 L_1^{-1}$ is a right coprime polynomial fractional representation of L and $G_1 L_1 + G_2 L_2 =: G$, then the zeros of $\det G$ are in Ω , which implies that L is in $\mathcal{L}_\Omega(W^*)$.

For the converse, let L be an element of $\mathcal{L}_\Omega(W^*)$ and $L_2 L_1^{-1}$ be its right coprime polynomial fractional representation. Define $G := G_1 L_1 + G_2 L_2$. Clearly, $L_1 G^{-1}$ and $L_2 G^{-1}$ are Ω -stable rational matrices. Furthermore, as $G_1 L_1 G^{-1} + G_2 L_2 G^{-1} = I$, it holds that $L_1 G^{-1} = S_1 + K_1 M$ and $L_2 G^{-1} = S_2 + K_2 M$ for some Ω -stable rational M (Recall that the right unimodular polynomial matrix $[K_1' : K_2']$ spans $\text{Ker } [G_1 : G_2]$ and S_1 and S_2 are Ω -stable rational matrices.) Since $\Lambda_2 L \Lambda_1^{-1}$ is proper, it follows that

$$\Lambda_2 L \Lambda_1^{-1} = (\Lambda_2 S_2 S_1^{-1} \Lambda_1^{-1} + \Lambda_2 K_2 M S_1^{-1} \Lambda_1^{-1})(I + \Lambda_1 K_1 M S_1^{-1} \Lambda_1^{-1})^{-1}$$

is proper. After carrying out appropriate manipulations and rearranging the terms of the equation above, we obtain

$$\Lambda_2(L - \bar{L})\Lambda_1^{-1} = [I - (\Lambda_2 L \Lambda_1^{-1})\Lambda_1 K_1 K_2^{-1} \Lambda_2^{-1}](\Lambda_2 K_2 \Lambda_3^{-1})\Lambda_3 M S_1^{-1} \Lambda_1^{-1}.$$

Since both L and \bar{L} are in $\mathcal{L}(W^*)$, the left-hand side of the equality above is proper. By strict properness of $\Lambda_1 K_1 K_2^{-1} \Lambda_2^{-1}$ the term in the parentheses is bicausal. Now, bicausality of $\Lambda_2 K_2 \Lambda_3^{-1}$ implies that $\Lambda_3 M S_1^{-1} \Lambda_1^{-1}$ is proper. \square

THEOREM 4.2. *A proper Z_c is a solution of DDPMS only if there exists an L in $\mathcal{L}_\Omega(W^*)$ such that*

$$(4.15) \quad [Z_c : -I](V^* + N^* L) = 0.$$

Conversely, for any L in $\mathcal{L}_\Omega(W^)$ there exists a proper solution Z_c of (4.15) which is also a solution to DDPMS.*

Proof. If Z_c is a solution, then by Corollary 3.1, there exists an L in $\mathcal{L}(W^*)$ satisfying (4.15). We just need to show that L is also in $\mathcal{L}_\Omega(W^*)$. In order to do this, note that by Lemma 4.1(i)–(ii) there exist a unimodular matrix U and a right unimodular polynomial matrix \tilde{U} such that

$$U[Q : R](V^* L_1 + N^* L_2) = U \Phi \tilde{U},$$

where $L_2 L_1^{-1}$ is a right coprime polynomial fractional representation of L and $\Phi := Q Q_c + R P_c$ with P_c, Q_c right coprime and $Z_c := P_c Q_c^{-1}$. Clearly, as U is unimodular and \tilde{U} is right unimodular, the set of invariant zeros of $[Q : R](V^* L_1 + N^* L_2)$ are included in the set of zeros of $\det \Phi$ which are, in turn, all in Ω as Z_c is a solution of DDPMS. Thus, by the definition of $\mathcal{L}_\Omega(W^*)$, it follows that L is in $\mathcal{L}_\Omega(W^*)$.

For the converse, let L be an element of $\mathcal{L}_\Omega(W^*)$. Then, defining $V := V^* L_1 + N^* L_2$ and following the synthesis procedure given in the proof of Theorem 4.1, one can easily construct a proper Z_c , satisfying (4.15), which is also a solution of DDPMS. \square

Remark 4.1 (Dual development). Up to now, the central object used in solving the disturbance decoupling problems being considered has been a minimal basis for $A^{-1}(\text{Im } B)$. A dual development is also possible where now the central object is a minimal basis for $(C \text{ Ker } D)^\perp$ with C and D defined as

$$C := \begin{bmatrix} Z_1 & Z_2 \\ -I & 0 \end{bmatrix}, \quad D := [Z_3 : Z_4].$$

Duality in these two approaches can easily be seen by noticing that $A^{-1}(\text{Im } B) = C \text{ Ker } D$.

Before concluding this section, we extend our Remark 3.2 to consider the solutions of the two-sided matrix equation

$$(4.16) \quad \Pi_4 = \Pi_3 X \Pi_2,$$

where Π_4 , Π_3 and Π_2 are arbitrary rational matrices and a solution X is sought over various subrings of $\underline{R}(z)$. Note that (4.16) is actually the starting point for obtaining solutions to various disturbance decoupling problems in Özgüler and Eldem [1985]. Our minimal basis approach yields alternative solvability conditions to the one given in Özgüler and Eldem [1985] and in (4.16), and has the further advantage of yielding the set of all possible solutions. The latter has been stated as an open problem in Ohm, Howze and Bhattacharyya [1984]. The main point is that a solution to (4.16) over various subrings of $\underline{R}(z)$ can easily be obtained by specialization of our procedure in obtaining a solution and the set of all solutions to DDPM, DDPMS and DDPMP to the case $Z_1 = 0$. (Note that both in Theorem 3.1 and Corollary 4.2 we imposed no restriction of properness on Z_2 , Z_3 , and Z_4 .) Let

$$A := \begin{bmatrix} I & 0 \\ 0 & \Pi_3 \end{bmatrix}, \quad B := \begin{bmatrix} \Pi_2 \\ \Pi_4 \end{bmatrix}$$

and $W := [W'_1 : W'_2]'$ be a minimal basis for $A^{-1}(\text{Im } B)$. Also partition W_h similarly so $W_h = [W'_{h1} : W'_{h2}]'$. Consider the following condition A , B and W :

- (C1) $\text{Im } B \subset \text{Im } A$ and $\rho[B] = \rho[\Pi_2]$;
- (C2) $\rho[W^*_{h1}] = \rho[B]$;
- (C3) The invariant zeros of W_1 are stable;
- (C4) W_1 has no nontrivial invariant zeros, i.e., W_1 is right unimodular.

Let $\underline{R}_p(z)$ and $\underline{R}_{ps}(z)$ denote the rings of proper and proper stable rational functions, respectively.

THEOREM 4.3. *The two-sided matching equation $\Pi_4 = \Pi_3 X \Pi_2$ has a solution over*

- (i) $\underline{R}(z)$ *iff* (C1) *holds*;
- (ii) $\underline{R}_p(z)$ *iff* (C1) *and* (C2) *hold*;
- (iii) $\underline{R}_{ps}(z)$ *iff* (C1), (C2) *and* (C3) *hold*;
- (iv) $\underline{R}(z)$ *and* $\underline{R}[z]$ *iff* (C1), (C2) *and* (C4) *hold*.

Proof. If $\Pi_4 = \Pi_3 X \Pi_2$ has a solution, then clearly $\text{Im } B \subset \text{Im } A$ and $\text{Ker } \Pi_2 \subset \text{Ker } \Pi_4$. Consequently, $\rho[B] = \rho[\Pi_2]$. Conversely, note that $\text{Im } B \subset \text{Im } A$ implies the existence of Y_1 and Y_2 such that

$$\begin{bmatrix} I & 0 \\ 0 & \Pi_3 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \Pi_2 \\ \Pi_4 \end{bmatrix}.$$

Let U be a unimodular matrix such that $\Pi_2 U = [\tilde{\Pi}_2 : 0]$ and $\Pi_4 U = [\tilde{\Pi}_4 : 0]$. Define $[\tilde{Y}_1 : 0] := Y_1 U$, $[\tilde{Y}_{21} : \tilde{Y}_{22}] := Y_2 U$. Clearly $\text{Im } \tilde{Y}_{22} \subset \text{Ker } Z_3$. As \tilde{Y}_1 has full column rank, there exists a rational X such that $X \tilde{Y}_1 = \tilde{Y}_{21}$. Then, $\Pi_3 X \tilde{Y}_1 = \tilde{\Pi}_4$, which implies that $\Pi_3 X \tilde{\Pi}_2 = \tilde{\Pi}_4$ and $\Pi_3 X \Pi_2 = \Pi_4$.

For the proofs of (ii)–(iv), set $Z_1 = 0$ in the proofs of the solvability conditions of DDPM, DDPMS and DDPMP, respectively, and the results follow from the equivalence between these two sets of problems.

We conclude this section by noting that the characterization of the solutions of $\Pi_4 = \Pi_3 X \Pi_2$ over $\underline{R}_p(z)$ and $\underline{R}_{ps}(z)$ can be established along the same lines as the characterizations of the solutions of DDPM and DDPMS.

5. Conclusions. In this paper we have employed a “minimal polynomial basis approach” to obtain solvability conditions and a characterization of all solutions of

DDPM and DDPMS. By this approach DDPM and DDPMS, which are inherently two-sided problems, are treated by consecutively solving two one-sided problems, thus availing us of the results by Forney [1975]. In fact, this has been the main convenience in obtaining the characterizations given in Corollary 3.1 and Theorem 4.2. With a slight alteration in the problem data (i.e. setting $Z_1 = 0$) the same approach applies to the two-sided matrix equation $A = BXC$, where X is to be determined over the rings of proper and stable proper rational functions. The principal contribution of this paper is the characterization in Corollary 3.1 and Theorem 4.2, and identification of the fixed modes. Since the solvability conditions for DDPM and DDPMS and the characterization of "the fixed modes with respect to dynamic output feedback in DDPM" can be easily attained via a minimal polynomial basis for the rational vector space $A^{-1}(\text{Im } B)$, this paper presents a compact analysis and synthesis method for disturbance decoupling problems.

REFERENCES

- H. AKASHI AND H. IMAI, *Disturbance localization and output dead-beat control through an observer in discrete time, linear multivariable systems*, IEEE Trans. Automat. Control., AC-24 (1979), pp. 621-627.
- S. P. BHATTACHARYYA, A. C. DEL NERO GROMES AND J. W. HOWZE, *The structure of robust disturbance rejection control*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 874-881.
- C. A. DESOER AND W. S. CHAN, *The feedback interconnection of lumped linear time-invariant systems*, J. Franklin Inst., 300 (1975), pp. 335-351.
- V. ELDEM, A. B. ÖZGÜLER AND U. BASER, *On the minimal McMillan degree solutions of disturbance decoupling problems via measurement feedback*, IFAC/IMACS International Symposium on Simulation of Control Systems, September 22-26, 1986, Vienna, Austria.
- G. D. FORNEY, *Minimal bases of rational vector spaces with applications to multivariable linear systems*, this Journal, 13 (1975), pp. 493-520.
- J. HAMMER AND M. HEYMANN, *Causal factorization and linear feedback*, this Journal, 19 (1981), pp. 465-468.
- H. IMAI AND H. AKASHI, *Disturbance localization and pole shifting by dynamic compensation*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 226-235.
- T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
- P. P. KHARGONEKAR AND A. B. ÖZGÜLER, *Regulator problem with internal stability: A frequency domain solution*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 331-343.
- V. KUCERA, *Disturbance resection: A polynomial approach*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 508-511.
- D. OHM, S. P. BHATTACHARYYA AND J. W. HOWZE, *Transfer matrix conditions for $(C'A, B)$ -pairs*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 172-174.
- D. OHM, J. W. HOWZE AND S. P. BHATTACHARYYA, *Structural synthesis of multivariable controllers*, Proc. 9th World Congress of IFAC, Budapest, Hungary, 2-6 July, 1984.
- A. B. ÖZGÜLER AND V. ELDEM, *Disturbance decoupling problems via dynamic output feedback*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 756-764.
- A. B. ÖZGÜLER, *Polynomial characterizations of (H, F) -invariant subspaces with applications*, Linear Algebra Appl., 73 (1986), pp. 1-31.
- L. PERNEBO, *An algebraic theory for design of controllers for linear multivariable systems—Part I: Feedback realizations and feedback design*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 183-193.
- H. SCHUMACHER, *Compensator synthesis using (C, A, B) -pairs*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 1133-1138.
- A. I. G. VARDULAKIS AND N. KARCANIAS (1984a), *On the stable exact model matching and stable minimal design problems in Multivariable Control*, D. Reidel, Boston, MA, 1984, Chap. 13, pp. 223-263.
- (1984b), *Proper and stable, minimal McMillan degree bases of rational vector spaces*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 1118-1120.
- S. H. WANG AND E. J. DAVISON, *A minimization algorithm for the design of linear multivariable systems*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 220-225.
- J. C. WILLEMS AND C. COMMAULT, *Disturbance decoupling by measurement feedback with stability or pole-placement*, this Journal, 19 (1981), pp. 409-504.

- J. C. WILLEMS, *Almost invariant subspaces: An approach to high gain feedback design—Part II: Almost conditionally invariant subspaces*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 1071–1085.
- W. A. WOLOVICH, *Linear multivariable systems*, Springer-Verlag, Berlin, New York, 1974.
- W. A. WOLOVICH, P. J. ANTSAKLIS AND H. ELLIOT, *On the stability of solutions to minimal and nonminimal design problems*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 88–94.

ON THE LOCAL STRUCTURE OF TIME-OPTIMAL BANG-BANG TRAJECTORIES IN \mathbb{R}^3 *

HEINZ SCHÄTTLER†

Abstract. We consider the problem of time-optimal control for systems of the form $\dot{x} = f(x) + ug(x)$ where f and g are smooth vector fields and admissible controls are measurable functions u with values in $-1 \leq u \leq 1$. Under the assumption that f , g and $[f, g]$ are independent, we prove that generically every point has a neighborhood \mathcal{U} such that bang-bang trajectories that lie in \mathcal{U} and have more than 7 switchings are not time-optimal.

Key words. nonlinear systems, time-optimal control, bang-bang trajectories, Baker–Campbell–Hausdorff formula

AMS(MOS) subject classifications. 49B10, 93B10

1. Introduction. We study the problem of time-optimal control for a system

$$\Sigma: \dot{x} = f(x) + ug(x), \quad |u| \leq 1, \quad x \in \mathbb{R}^3, \quad u \in \mathbb{R}$$

where f and g are smooth vector fields. Admissible controls are measurable functions with values in $[-1, 1]$ and a trajectory of the system corresponding to a control $u(\cdot)$ is an absolutely continuous curve such that $\dot{x}(t) = f(x(t)) + u(t)g(x(t))$ holds almost everywhere.

As in every optimal control problem, the question of regularity properties of an optimal solution comes up very naturally. The standard existence results only prove existence of an optimal solution within the class of measurable functions. The necessary conditions for optimality given in the Pontryagin Maximum Principle [9] and other high order conditions (e.g. [7]) imply certain restrictions on the possible structures of optimal controls, in particular on their regularity properties. But, in general, these conditions are not strong enough to prove that optimal controls are nice in the sense that they are for instance piecewise continuous. And, in fact, they do not have to be. Fuller's example [6] shows that time-optimal controls with countably many discontinuities (switchings) can occur even for analytic systems in \mathbb{R}^3 . A much more irregular structure seems possible. In principle, the set of discontinuities of an optimal control could even be a Cantor-like set with positive measure. The currently known necessary conditions for optimality fail to exclude such a pathological behaviour of optimal controls.

On the other hand, certain regularity properties of optimal controls are needed if one wants to obtain sufficient conditions in the form of a "regular synthesis" [2]. Even though the strict conditions given originally by Boltyansky have subsequently been substantially relaxed (Brunovsky [5], Sussmann [13]), mainly due to the introduction of the powerful theory of subanalytic sets to control theory by Brunovsky, the basic concept of synthesizing an optimal control from local knowledge does require a certain degree of regularity. One of the key hypotheses for any type of regular synthesis using the concepts that are currently discussed in the literature is still the following regularity property of optimal trajectories: for every compact set K there exists an integer $N = N(K)$ such that any optimal trajectory that lies in K is a concatenation of at

* Received by the editors June 16, 1986; accepted for publication (in revised form) March 3, 1987.

† Department of Mathematics, University of California at Davis, Davis, California 95616. Present address: Department of Systems Science and Mathematics, School of Engineering and Applied Science, Washington University, St. Louis, Missouri 63130.

most N “nice” pieces [5]. For our problem “nice” simply means trajectories corresponding to the constant controls $u \equiv +1$ or $u \equiv -1$ (bang arcs) or to a so-called singular control (singular arc), which is a control usually with values in the interior of the control set and which satisfies certain compatibility conditions. (These are the possible candidates for time-optimal controls to which the necessary conditions of the Pontryagin Maximum Principle lead.)

Fuller’s example shows that such a regularity result about optimal trajectories need not hold. It has been an open problem in control theory for years to understand the phenomenon of optimal trajectories with infinitely many switchings in arbitrarily small times. Recently, important contributions to this problem have been made by Kupka [8], who showed that Fuller-like extremals necessarily exist if certain conditions are satisfied, and that these conditions are even generic in spaces of sufficiently high dimension. For small dimensions, however, it is still not even known whether the Fuller-phenomenon is generic or not.

In case of an analytic system, i.e., when f and g are analytic vector fields, all the local properties of Σ , in particular the possible structures of time-optimal trajectories, are determined by the Lie algebra generated by f and g and, at least in principle, can be characterized by the Lie bracket configuration. Roughly, the “Lie bracket configuration” consists of all the “Lie relations” that hold between f , g and brackets of f and g . A precise definition is technical and we skip it since it is not relevant for our purposes. The interested reader may consult [14, § 4]. Therefore it makes sense to pose the following *Question*:

Given the Lie bracket configuration at a reference point p , what can be said about the structure of time-optimal trajectories that lie in a sufficiently small neighborhood of p ?

In view of our remarks above, it is clear that one is interested in conditions under which time-optimal trajectories are finite concatenations of bang and singular arcs with a bound on the number of switchings. As far as Fuller’s problem is concerned, this includes conditions under which bang-bang trajectories with too many switchings are not time-optimal near p , i.e., conditions when the Fuller phenomenon cannot occur.

For a generic system in the plane a classification of the local structure of time-optimal trajectories was given by Sussmann [12], [17]. From his results he proved the existence of a regular synthesis for basically arbitrary analytic systems in the plane, only subject to a mild “nonexplosion” condition [18], [19]. However, the reasoning used to exclude optimality of bang-bang trajectories with a large number of switchings in small time depended heavily on their being in \mathbb{R}^2 ; only recently was a general geometric version of this argument given [15]. Therefore, so far even for the three-dimensional case, only partial results are known, due to Bressan [4] and Bonnard [3].

Here we examine what in a certain sense are the least degenerate general cases possible in \mathbb{R}^3 . All our considerations are local near a reference point p where we assume that $f(p)$, $g(p)$ and $[f, g](p)$, the Lie bracket of f and g at p , are independent. In this paper we will exclusively consider bang-bang trajectories. Under additional assumptions on low order brackets of f and g at p we will show that bang-bang trajectories with too many switchings are not optimal near p , i.e., if they lie in a sufficiently small neighborhood of p . So the Fuller phenomenon does not occur. In a forthcoming paper we will then continue the local analysis of optimal trajectories by incorporating singular arcs. In particular, we will show there that in the cases under

consideration here optimal trajectories are indeed finite concatenations of bang and singular arcs with a bound on the number of switchings.

In § 2 we will give a summarizing version of our results in terms of standard genericity notions. There we will also describe the notation and terminology we use. In §§ 3 and 4 we develop new necessary conditions for time-optimality of bang-bang trajectories which form the core of our results. Whereas the conditions of § 4 are related to concepts used to prove the Pontryagin Maximum Principle—they will follow from point variations—we present in § 3 a new technique to compute necessary conditions for optimality by making variations along a reference trajectory. This technique is not limited to dimension 3, but in higher dimensions the computational complexity increases rapidly. We then use these new necessary conditions to analyze bang-bang trajectories near a reference point in § 5. Due to the length and technicality of some of the arguments we cannot give a full proof for all generic cases in this paper, but have to restrict ourselves to a few selected examples.

2. Statement of the result. We identify Σ with the C^∞ map $\Sigma = (f, g): \mathbb{R}^3 \rightarrow \mathbb{R}^6$ and equip $C^\infty(\mathbb{R}^3, \mathbb{R}^6)$ with the Whitney C^∞ topology. Let \mathcal{A} be the open subset of all systems Σ for which f, g and $[f, g]$ are independent everywhere. Then we have the following.

THEOREM. *For a generic system $\Sigma \in \mathcal{A}$ every point has a neighborhood \mathcal{U} such that bang-bang trajectories that lie in \mathcal{U} and have more than seven switchings are not time-optimal.*

The genericity concepts are only used to have an easy way to summarize our results. By Thom's Transversality Theorem it is necessary to analyze bang-bang trajectories near a reference point p in all cases where the Lie bracket configuration at p is of codimension ≤ 3 . As mentioned earlier, the "Lie bracket configuration" consists of all the Lie relations at p . The "codimension" is the number of independent equality type constraints. An elaborate account of these ideas is given in [12]; precise definitions are long and technical and so we omit them (cf. for instance [14]). Our theorem combines all generic cases in which the Lie bracket configuration is characterized by at most three equality constraints. Some more detailed statements of our results are given in the propositions of § 5.

In this paper we will not prove the theorem completely. We will establish all new necessary conditions for time-optimality of trajectories that will be needed, but we will then restrict ourselves to analyzing some selected generic cases. In order to keep the paper within reasonable length we have to omit all the cases which have long and technical proofs. Also, we will not prove the genericity of our conditions. Aside from being almost obvious, these are straightforward computations. Complete and detailed proofs of all the results are contained in the author's doctoral dissertation [11].

2.1. Notation and terminology. Throughout this paper all our considerations are local, i.e., hold only on a sufficiently small neighborhood of a reference point p , where we assume that f, g and $[f, g]$ are independent. We always tacitly assume this. Therefore we can express all higher-order brackets of f and g as linear combinations of f, g and $[f, g]$ with smooth coefficients. Equivalently, if we let $X = f - g$ and $Y = f + g$, we can write them in terms of X, Y and $[X, Y]$. We also use $\text{ad } X(Y) = [X, Y]$.

$$[X, [X, Y]] = a_1 X + a_2 Y + a_3 [X, Y] = \alpha f + \cdots,$$

$$[Y, [X, Y]] = b_1 X + b_2 Y + b_3 [X, Y] = \beta f + \cdots,$$

$$[X, [X, [X, Y]]] = c_1 X + c_2 Y + c_3 [X, Y] = \gamma f + \cdots,$$

$$[X, [Y, [X, Y]]] = d_1 X + d_2 Y + d_3 [X, Y] = \delta f + \dots,$$

$$[Y, [Y, [X, Y]]] = e_1 X + e_2 Y + e_3 [X, Y] = \eta f + \dots,$$

$$\text{ad}^4 X(Y) = m_1 X + m_2 Y + m_3 [X, Y] = \mu f + \dots,$$

$$\text{ad}^5 X(Y) = k_1 X + k_2 Y + k_3 [X, Y] = \kappa f + \dots,$$

$$-\text{ad}^4 Y(X) = n_1 X + n_2 Y + n_3 [X, Y] = \nu f + \dots,$$

$$-\text{ad}^5 Y(X) = r_1 X + r_2 Y + r_3 [X, Y] = \rho f + \dots.$$

We denote the Lie derivative of a function Φ in direction of X by $L_X \Phi$. Observe that

$$L_X \alpha = \gamma - a_3 \alpha, \quad L_Y \alpha = \delta - a_3 \beta,$$

$$L_X \beta = \delta - b_3 \alpha, \quad L_Y \beta = \eta - b_3 \beta,$$

$$L_X \gamma = \mu - c_3 \alpha,$$

$$L_X \mu = \kappa - m_3 \alpha.$$

We use exponential notation for the flows of vector fields, i.e., we write $t \mapsto p e^{tX}$ for the integral curve of the vector field X starting at p at time 0. We let the diffeomorphism act on the right to make certain algebraic manipulations involving the Baker–Campbell–Hausdorff formula come out right later.

Finally, for a trajectory which is a concatenation of X and Y arcs, we use the corresponding letter combination to denote the trajectory, i.e., an XYX -trajectory is a concatenation of an X -arc, followed by a Y -arc and another X -arc. Sometimes, in particular when we talk of subtrajectories of a long concatenation, we indicate the first switching point in parentheses after the structure, i.e., a $YXY(p_0)$ -trajectory is a YXY -trajectory with first switching point p_0 .

3. Necessary conditions for time-optimality of bang-bang trajectories: A variation along the trajectory. In this section we will develop new necessary conditions for optimality of bang-bang trajectories.

3.1. Construction of a variation along a bang-bang trajectory. The idea is to take a bang-bang trajectory Γ with three switching points and compare it to a suitably constructed bang-bang trajectory with only two switching points that has the same initial and terminal points as Γ . We will then compute an asymptotic expansion for the difference in time between these two trajectories in terms of the data along Γ . This will give us new necessary conditions for optimality of Γ .

Let Γ be a $YXYX$ -trajectory with initial point \tilde{p} and switching points p_0, p_1 and p_2 . Let τ_1 and τ_2 be the times along the intermediate X and Y arcs respectively (we assume they are positive) and let s be a small positive number.

LEMMA 1. *The equation*

$$(3.1) \quad p_0 e^{\tau_1 X} e^{\tau_2 Y} e^{\tau_1 s X} = p_0 e^{t_1 Y} e^{t_2 X} e^{t_3 Y}$$

is solvable by smooth functions t_1, t_2, t_3 of τ_1, τ_2 and s near the origin.

Proof. The equation

$$p_0 e^{t_1 Y} e^{t_2 X} e^{t_3 Y} = p_0 e^{aY} e^{bX} e^{c[X,Y]}$$

is solvable by smooth functions a, b, c of t_1, t_2, t_3 near the origin by the Implicit Function Theorem. Notice that $a = t_1 + t_3 + t_2 t_3 \tilde{a}$, $b = t_2 + t_2 t_3 \tilde{b}$ and $c = t_2 t_3 \tilde{c}$ where \tilde{a}, \tilde{b} and \tilde{c} are smooth functions. Therefore also

$$d := \frac{c}{b} = \frac{t_3 \tilde{c}}{1 + t_3 \tilde{b}}$$

is smooth. An elementary computation shows that the Jacobian of the map $(t_1, t_2, t_3) \xrightarrow{\Phi} (a, b, d)$ is nonsingular at the origin and so we can invert locally.

Also, a, b, c are smooth functions of τ_1, τ_2 and s via

$$p_0 e^{\tau_1 X} e^{\tau_2 Y} e^{\tau_1 s X} = p_0 e^{aY} e^{bX} e^{c[X, Y]}$$

and $b = \tau_1 + s\tau_1 + \tau_1 \tau_2 \tilde{b}$, $c = \tau_1 \tau_2 \tilde{c}$, $d = c/b = \tau_2 \tilde{c}/(1 + s + \tau_2 \tilde{b})$. Hence $(\tau_1, \tau_2, s) \xrightarrow{\Psi} (a, b, d)$ is smooth and so then is $\Phi^{-1} \circ \Psi$. \square

For $s = 0$ the solutions are $t_1 = 0$, $t_2 = \tau_1$ and $t_3 = \tau_2$ and so t_2 and t_3 are positive for small s whereas t_1 is small. Therefore these times define a YXY -trajectory of Σ which starts at \tilde{p} and steers the system to some point $q = p_2 e^{\tau_1 s X}$ on Γ just after the third switching point. We are interested in the difference in time. Let

$$(3.2) \quad \Delta(s; \tau_1, \tau_2) := t_1 + t_2 + t_3 - \tau_1 - \tau_2 - \tau_1 s.$$

By Lemma 1 this function is well defined for small τ_1 and τ_2 . We can choose a neighborhood of $(\tau_1, \tau_2) = (0, 0)$ such that this is true uniformly for points p_0 in a small neighborhood V of p . Then Δ is well defined for any $YXYX$ -trajectory lying in V with small total time T . But since $f(p)$ and $g(p)$ are independent, we can choose a neighborhood \mathcal{U} of p so small that any trajectory of Σ has to leave \mathcal{U} within time T . Then Δ is well defined for any $YXYX$ -trajectory which lies in \mathcal{U} . From now on we will always assume that we work on such a neighborhood \mathcal{U} .

LEMMA 2. *If Γ is time-optimal, we have $\dot{\Delta}(0; \tau_1, \tau_2) = 0$ and $\ddot{\Delta}(0; \tau_1, \tau_2) \geq 0$ where the dot denotes differentiation with respect to s .*

Proof. Since Γ is time-optimal, by the Pontryagin Maximum Principle, there exists a nontrivial adjoint multiplier $\lambda : [0, T] \rightarrow \mathbb{R}^3$, $\dot{\lambda} = -\lambda(Df + uDg)$, such that $\langle \lambda(t), g(x(t)) \rangle u(t) = \text{Min}_{|v| \leq 1} \langle \lambda(t), g(x(t)) \rangle v$. In particular $\langle \lambda, g \rangle$ vanishes at the switching points. The first relation, $\dot{\Delta} = 0$, is a consequence of this. We simply have to compute the derivatives of the functions t_i at $s = 0$. If we differentiate (3.1) we get

$$(*) \quad \tau_1(p_2 X) = \dot{t}_1 v_1 + \dot{t}_2 v_2 + \dot{t}_3 v_3,$$

where $\dot{t}_i = (d/ds)|_{s=0} t_i(s; \tau_1, \tau_2)$ and $v_1 = p_0 Y e^{\tau_1 X} e^{\tau_2 Y}$, $v_2 = p_1 X e^{\tau_2 Y}$, $v_3 = p_2 X$. Now $\langle \lambda, g \rangle$ vanishes at p_0, p_1 and p_2 and if we transport these equations to p_2 , we have with $\tilde{\lambda} = \lambda(p_2)$:

$$\langle \tilde{\lambda}, g(p_2) \rangle = 0, \quad \langle \tilde{\lambda}, e^{-\tau_2 \text{ad } Y} g(p_1) \rangle = 0$$

and

$$\langle \tilde{\lambda}, e^{-\tau_2 \text{ad } Y} e^{-\tau_1 \text{ad } X} g(p_0) \rangle = 0.$$

Since λ is nontrivial, these three vectors are dependent. (In the terminology of [14] the points p_0, p_1 and p_2 form a conjugate triple.) Let S be the two-dimensional subspace spanned by these vectors and observe that $X(p_2)$ and $Y(p_2)$ do not lie in S due to our independence assumption on f, g and $[f, g]$. Also $X(p_2) \equiv Y(p_2) \text{ mod } S$ since $g(p_2) \in S$. An easy computation shows that in fact $v_1 \equiv v_2 \equiv v_3 \equiv X(p_2) \text{ mod } S$ and since $X(p_2) \notin S$, we obtain from (*) that $\tau_1 = \dot{t}_1 + \dot{t}_2 + \dot{t}_3$. Therefore $\dot{\Delta}(0; \tau_1, \tau_2) = 0$.

We call this a *conjugate point relation*. Since Γ is time-optimal we have $\Delta(s; \tau_1, \tau_2) \geq 0$ for small $s \geq 0$ and hence $\ddot{\Delta}(0; \tau_1, \tau_2) \geq 0$. \square

To get an applicable criterion from this we compute an asymptotic expansion for Δ . To state it more easily we introduce the following notation: for a smooth function ϕ we write $\tilde{\phi}$ for terms which are of the form $\phi(1 + O(1))$ where $O(k)$ denotes terms in the times τ_1 and τ_2 which are of order $\leq k$. Also we write $0 \cdot \tilde{\phi}$ for terms of order $\phi \cdot O(1)$.

LEMMA 3. Δ has the following asymptotic expansion

$$(3.3) \quad \begin{aligned} \Delta = \tau_1 \tau_2 \{ & -\frac{1}{2}s(\tilde{\alpha}\tau_1 + \tilde{\beta}\tau_2 + \frac{2}{3}\tilde{\gamma}\tau_1^2 + \tilde{\delta}\tau_1\tau_2 + \frac{1}{3}\tilde{\eta}\tau_2^2 + \frac{1}{4}\tilde{\mu}\tau_1^3 + \frac{1}{12}\tilde{\nu}\tau_2^3 \\ & + \frac{1}{15}\tilde{\kappa}\tau_1^4 + \tau_1\tau_2 \cdot O(1) + O(\tau_2^4) + O(5)) \\ & + \frac{1}{2}s^2 \cdot (0 \cdot \tilde{\alpha}\tau_1 + \tilde{\beta}\tau_2 - \frac{1}{3}\tilde{\gamma}\tau_1^2 + 0 \cdot \tilde{\delta}\tau_1\tau_2 + 0 \cdot \tilde{\eta}\tau_2^2 - \frac{1}{4}\tilde{\mu}\tau_1^3 \\ & + 0 \cdot \tilde{\nu}\tau_2^3 - \frac{1}{10}\tilde{\kappa}\tau_1^4 + \tau_1\tau_2 \cdot O(1) + O(\tau_2^4) + O(5)) \} \end{aligned}$$

where all the functions are evaluated at the first switching point p_0 .

In the expansion we listed only those terms which we actually need later on. A proof of Lemma 3 will be given in § 3.3. First we draw the obvious *corollaries* to get as necessary conditions for optimality of Γ :

- the “conjugate point relation”

$$(3.4) \quad 0 = \tilde{\alpha}\tau_1 + \tilde{\beta}\tau_2 + \frac{2}{3}\tilde{\gamma}\tau_1^2 + \tilde{\delta}\tau_1\tau_2 + \frac{1}{3}\tilde{\eta}\tau_2^2 + \frac{1}{4}\tilde{\mu}\tau_1^3 + \frac{1}{12}\tilde{\nu}\tau_2^3 + \frac{1}{15}\tilde{\kappa}\tau_1^4 + \dots$$

- and the “optimality condition”

$$(3.5) \quad 0 \leq 0 \cdot \tilde{\alpha}\tau_1 + \tilde{\beta}\tau_2 - \frac{1}{3}\tilde{\gamma}\tau_1^2 + 0 \cdot \tilde{\delta}\tau_1\tau_2 + 0 \cdot \tilde{\eta}\tau_2^2 - \frac{1}{4}\tilde{\mu}\tau_1^3 + 0 \cdot \tilde{\nu}\tau_2^3 - \frac{1}{10}\tilde{\kappa}\tau_1^4 + \dots$$

3.2. Input symmetries. For later use we need the equivalent versions of (3.4) and (3.5) when we evaluate all the functions at the last switching point p_2 and we also need the corresponding optimality conditions for a $YXYX$ -trajectory. All that can be obtained easily by exploiting input symmetries of the system Σ .

Let $\text{Sym } 1$ denote the input symmetry which formally interchanges X and Y , i.e., $\text{Sym } 1(X) = Y$ and $\text{Sym } 1(Y) = X$. If we apply $\text{Sym } 1$ to the equation

$$[X, [X, Y]] = \alpha f + (a_2 - a_1)g + a_3[X, Y],$$

we obtain

$$-[Y, [X, Y]] = \text{Sym } 1(\alpha) \cdot f - \text{Sym } 1(a_2 - a_1) \cdot g - \text{Sym } 1(a_3)[X, Y]$$

and so $\text{Sym } 1(\alpha) = -\beta$. Analogously it follows that

$$(3.6) \quad \begin{aligned} \text{Sym } 1(\beta) &= -\alpha, & \text{Sym } 1(\gamma) &= -\eta, & \text{Sym } 1(\delta) &= -\delta, \\ \text{Sym } 1(\eta) &= -\gamma, & \text{Sym } 1(\mu) &= -\nu, & \text{Sym } 1(\kappa) &= -\rho. \end{aligned}$$

Let $\text{Sym } 2$ be the symmetry which formally interchanges X and $-Y$. Then

$$(3.7) \quad \begin{aligned} \text{Sym } 2(\alpha) &= -\beta, & \text{Sym } 2(\beta) &= -\alpha, \\ \text{Sym } 2(\gamma) &= \eta, & \text{Sym } 2(\delta) &= \delta, & \text{Sym } 2(\eta) &= \gamma, \\ \text{Sym } 2(\mu) &= -\nu, & \text{Sym } 2(\kappa) &= \rho. \end{aligned}$$

Finally, let $\text{Sym } 3$ be time reversal, i.e., $\text{Sym } 3(X) = -X$, $\text{Sym } 3(Y) = -Y$. Then $\text{Sym } 3$ is just $\text{Sym } 1$ composed with $\text{Sym } 2$, and thus α, β, μ and ν are fixed, whereas γ, δ, η and ρ are transformed into $-\gamma, -\delta, -\eta$ and $-\rho$.

We now apply these input symmetries along a $YXYX$ -trajectory. Note that instead of (3.1) we might as well have considered

$$p_2 e^{-\tau_2 Y} e^{-\tau_1 X} e^{-\tau_2 Y} = p_2 e^{-t_1 X} e^{-t_2 Y} e^{-t_3 X}.$$

This equation is identical with (3.1) if we identify p_2 with p_0 , interchange X with $-Y$ and also interchange τ_1 and τ_2 . Therefore we simply get the corresponding optimality condition when we apply Sym 2, which interchanges X and Y and reverses time, to (3.4) and (3.5). This gives as conjugate point relation

$$(3.8) \quad \begin{aligned} 0 = & \tilde{\alpha}(p_2)\tau_1 + \tilde{\beta}(p_2)\tau_2 - \frac{1}{3}\tilde{\gamma}(p_2)\tau_1^2 - \tilde{\delta}(p_2)\tau_1\tau_2 - \frac{2}{3}\tilde{\eta}(p_2)\tau_2^2 \\ & + \frac{1}{12}\tilde{\mu}(p_2)\tau_1^3 + \frac{1}{4}\tilde{\nu}(p_2)\tau_2^3 + \tau_1\tau_2 \cdot O(1) + O(4) \end{aligned}$$

and as optimality condition

$$(3.9) \quad \begin{aligned} 0 \leq & -\tilde{\alpha}(p_2)\tau_1 + 0 \cdot \tilde{\beta}(p_2)\tau_2 + 0 \cdot \tilde{\gamma}(p_2)\tau_1^2 + 0 \cdot \tilde{\delta}(p_2)\tau_1\tau_2 \\ & - \frac{1}{3}\tilde{\eta}(p_2)\tau_2^2 + 0 \cdot \tilde{\mu}(p_2)\tau_1^3 + \frac{1}{4}\tilde{\nu}(p_2)\tau_2^3 + \tau_1\tau_2 \cdot O(1) + O(4). \end{aligned}$$

Since here all the functions are evaluated at the last switching point, we refer to these relations as “backward conditions,” whereas we call (3.4) and (3.5) “forward conditions.” It is clear that (3.8) and (3.9) are equivalent to (3.4) and (3.5). This can also easily be seen if one transports all the functions from p_2 to p_0 via Taylor’s theorem. Both versions will be needed later on.

Finally, we also compute necessary conditions for optimality of an $XYXY$ -trajectory. We call the time along the intermediate X arc τ_3 and consider now

$$p_1 e^{\tau_2 Y} e^{\tau_3 X} e^{\tau_2 Y} = p_1 e^{t_1 X} e^{t_2 Y} e^{t_3 X}.$$

Again this is of the form (3.1) with the obvious identifications and we get the corresponding optimality conditions by applying Sym 1, which interchanges X and Y , to (3.4) and (3.5). This yields as forward conditions

$$(3.10) \quad \begin{aligned} 0 = & \tilde{\alpha}(p_1)\tau_3 + \tilde{\beta}(p_1)\tau_2 + \frac{1}{3}\tilde{\gamma}(p_1)\tau_3^2 + \tilde{\delta}(p_1)\tau_2\tau_3 + \frac{2}{3}\tilde{\eta}(p_1)\tau_1^2 \\ & + \frac{1}{12}\tilde{\mu}(p_1)\tau_3^3 + \frac{1}{4}\tilde{\nu}(p_1)\tau_2^3 + \tau_2\tau_3 \cdot O(1) + O(4), \end{aligned}$$

$$(3.11) \quad \begin{aligned} 0 \leq & -\tilde{\alpha}(p_1)\tau_3 + 0 \cdot \tilde{\beta}(p_1)\tau_2 + 0 \cdot \tilde{\gamma}(p_1)\tau_3^2 + 0 \cdot \tilde{\delta}(p_1)\tau_2\tau_3 \\ & + \frac{1}{3}\tilde{\eta}(p_1)\tau_2^2 + 0 \cdot \tilde{\mu}(p_1)\tau_3^3 + \frac{1}{4}\tilde{\nu}(p_1)\tau_2^3 + \tau_2\tau_3 \cdot O(1) + O(4). \end{aligned}$$

The backward conditions follow by time reversal from (3.4) and (3.5):

$$(3.12) \quad \begin{aligned} 0 = & \tilde{\alpha}(p_3)\tau_3 + \tilde{\beta}(p_3)\tau_2 - \frac{2}{3}\tilde{\gamma}(p_3)\tau_3^2 - \tilde{\delta}(p_3)\tau_2\tau_3 - \frac{1}{3}\tilde{\eta}(p_3)\tau_2^2 + \frac{1}{4}\tilde{\mu}(p_3)\tau_3^3 \\ & + \frac{1}{12}\tilde{\nu}(p_3)\tau_2^3 + \tau_2\tau_3 \cdot O(1) - \frac{1}{15}\tilde{\kappa}(p_3)\tau_3^4 + O(\tau_2^4) + O(5), \end{aligned}$$

$$(3.13) \quad \begin{aligned} 0 \leq & 0 \cdot \tilde{\alpha}(p_3)\tau_3 + \tilde{\beta}(p_3)\tau_2 + \frac{1}{3}\tilde{\gamma}(p_3)\tau_3^2 + 0 \cdot \tilde{\delta}(p_3)\tau_2\tau_3 + 0 \cdot \tilde{\eta}(p_3)\tau_2^2 - \frac{1}{4}\tilde{\mu}(p_3)\tau_3^3 \\ & - 0 \cdot \tilde{\nu}(p_3)\tau_2^3 + \tau_2\tau_3 \cdot O(1) + \frac{1}{10}\tilde{\kappa}(p_3)\tau_3^4 + O(\tau_2^4) + O(5). \end{aligned}$$

3.3. Proof of the asymptotic expansion for Δ . Here we give the proof of Lemma 3. This section is independent of the remaining parts of the paper and may be skipped at first reading. It contains the proof of the core of the argument. We rewrite both sides of (3.1) in terms of canonical coordinates of the second kind as $p_0 e^{z_1[X,Y]} e^{z_2 Y} e^{z_3 X}$ and then we compare the coefficients to obtain the asymptotic expansion (3.3).

Our computations are based on the following commutator formula [1]:

$$(3.14) \quad \begin{aligned} e^A \cdot e^B = & e^R \cdot e^{(1/5!) \operatorname{ad}^5 A(B)} \cdot e^{(1/4!) \operatorname{ad}^4 A(B)} \cdot e^{-(1/4!) \operatorname{ad}^4 B(A)} \\ & \cdot e^{(1/6)[A,[A,[A,B]]]} \cdot e^{(1/4)[A,[B,[A,B]]]} \cdot e^{(1/6)[B,[B,[A,B]]]} \\ & \cdot e^{(1/2)[A,[A,B]]} \cdot e^{(1/2)[B,[A,B]]} \cdot e^{[A,B]} \cdot e^B \cdot e^A \end{aligned}$$

where e^R is the remainder term and R contains only brackets of A and B of length ≥ 5 , but none of $\operatorname{ad}^4 A(B)$, $\operatorname{ad}^5 A(B)$ or $\operatorname{ad}^4 B(A)$. This formula can be obtained by

iterating the Campbell–Hausdorff formula or by a technique known as variation of constants for formal power series [14]. It is no problem to write down the expansion which contains all brackets of length 6, but most of these brackets are irrelevant for our purpose, and we listed only the ones which we do need later on.

If we apply (3.14) to $p_0 e^{\tau_1 X} e^{\tau_2 Y} e^{\tau_1 s X}$ we obtain

$$(3.15) \quad \begin{aligned} e^{\tau_1 X} e^{\tau_2 Y} e^{\tau_1 s X} = & \dots e^{(1/5!) \tau_1^5 \tau_2 \text{ad}^5 X(Y)} \cdot e^{(1/4!) \tau_1^4 \tau_2 \text{ad}^4 X(Y)} \cdot e^{-(1/4!) \tau_1 \tau_2^4 \text{ad}^4 Y(X)} \\ & \cdot e^{(1/6) \tau_1^3 \tau_2 [X, [X, [X, Y]]]} \cdot e^{(1/4) \tau_1^2 \tau_2^2 [X, [Y, [X, Y]]]} \cdot e^{(1/6) \tau_1 \tau_2^3 [Y, [Y, [X, Y]]]} \\ & \cdot e^{(1/2) \tau_1^2 \tau_2 [X, [X, Y]]} \cdot e^{(1/2) \tau_1 \tau_2^2 [Y, [X, Y]]} \cdot e^{\tau_1 \tau_2 [X, Y]} e^{\tau_2 Y} e^{\tau_1 (1+s) X}. \end{aligned}$$

We express the higher-order brackets as linear combinations of X , Y and $[X, Y]$. Let us explain how to do this in a specific example. We write

$$(3.16) \quad [X, [X, Y]] = a_1(p_0)X + a_2(p_0)Y + a_3(p_0)[X, Y] + W_a,$$

where W_a is defined by this expression. So $W_a(p_0) = 0$ and hence $p_0 e^{W_a} = p_0$. Now substitute (3.16) into (3.15) and change from the exponential of the sum to a product of exponentials. All commutator terms which come up in this process are negligible within our desired accuracy and so we get terms

$$p_0 \dots e^{(1/2) \tau_1^2 \tau_2 W_a} \cdot e^{(1/2) \tau_1^2 \tau_2 a_3(p_0)[X, Y]} e^{(1/2) \tau_1^2 \tau_2 a_2(p_0)Y} e^{(1/2) \tau_1^2 \tau_2 a_1(p_0)X} \dots$$

Next we commute the error term all the way up to p_0 where it drops out of the calculation. Again, the commutators are of higher order and so we can simply replace

$$e^{(1/2) \tau_1^2 \tau_2 [X, [X, Y]]} \quad \text{by} \quad e^{(1/2) \tau_1^2 \tau_2 a_3(p_0)[X, Y]} e^{(1/2) \tau_1^2 \tau_2 a_2(p_0)Y} e^{(1/2) \tau_1^2 \tau_2 a_1(p_0)X}.$$

The same is true for all brackets of order ≥ 3 and we get as intermediate step

$$\begin{aligned} p_0 e^{\tau_1 X} e^{\tau_2 Y} e^{\tau_1 s X} = & p_0 \dots e^{(1/5!) \tau_1^5 \tau_2 k_3(p_0)[X, Y]} e^{(1/5!) \tau_1^5 \tau_2 k_2(p_0)Y} e^{(1/5!) \tau_1^5 \tau_2 k_1(p_0)X} \\ & \cdot e^{(1/24) \tau_1^4 \tau_2 m_3(p_0)[X, Y]} e^{(1/24) \tau_1^4 \tau_2 m_2(p_0)Y} e^{(1/24) \tau_1^4 \tau_2 m_1(p_0)X} \\ & \cdot e^{(1/24) \tau_1 \tau_2^4 n_3(p_0)[X, Y]} \cdot e^{(1/24) \tau_1 \tau_2^4 n_2(p_0)Y} e^{(1/24) \tau_1 \tau_2^4 n_1(p_0)X} \\ & \cdot e^{(1/6) \tau_1^3 \tau_2 c_3(p_0)[X, Y]} e^{(1/6) \tau_1^3 \tau_2 c_2(p_0)Y} e^{(1/6) \tau_1^3 \tau_2 c_1(p_0)X} \\ & \cdot e^{(1/4) \tau_1^2 \tau_2^2 d_3(p_0)[X, Y]} \cdot e^{(1/4) \tau_1^2 \tau_2^2 d_2(p_0)Y} e^{(1/4) \tau_1^2 \tau_2^2 d_1(p_0)X} \\ & \cdot e^{(1/6) \tau_1 \tau_2^3 e_3(p_0)[X, Y]} \cdot e^{(1/6) \tau_1 \tau_2^3 e_2(p_0)Y} e^{(1/6) \tau_1 \tau_2^3 e_1(p_0)X} \\ & \cdot e^{(1/2) \tau_1^2 \tau_2 a_3(p_0)[X, Y]} \cdot e^{(1/2) \tau_1^2 \tau_2 a_2(p_0)Y} \cdot e^{(1/2) \tau_1^2 \tau_2 a_1(p_0)X} \\ & \cdot e^{(1/2) \tau_1 \tau_2^2 b_3(p_0)[X, Y]} \cdot e^{(1/2) \tau_1 \tau_2^2 b_2(p_0)Y} \cdot e^{(1/2) \tau_1 \tau_2^2 b_1(p_0)X} \\ & \cdot e^{\tau_1 \tau_2 [X, Y]} e^{\tau_2 Y} e^{\tau_1 (1+s) X}. \end{aligned}$$

Now we have to rearrange terms. Here we find commutators which lie in the range of our required accuracy. For instance, if we commute $\exp(\frac{1}{2} \tau_1^2 \tau_2 a_1(p_0)X)$ and $\exp(\tau_2 Y)$, we get a term $\exp(\frac{1}{2} \tau_1^2 \tau_2^2 a_1(p_0)[X, Y])$. But also none of these terms matters. The reason is that all such terms are small relative to terms which appear already in the expansion such as $\frac{1}{2} \tau_1^2 \tau_2^2 a_1(p_0) \ll \tau_1 \tau_2$ at $[X, Y]$. These terms are only higher-order perturbations and their exact structure is irrelevant. We indicate these perturbations

by a tilde on the functions. So we get

$$\begin{aligned}
 p_0 e^{\tau_1 X} e^{\tau_2 Y} e^{\tau_1 s X} &= p_0 \exp ([X, Y] \{ \tau_1 \tau_2 (1 + O(2)) + \frac{1}{2} \tau_1^2 \tau_2 \tilde{a}_3(p_0) + \frac{1}{2} \tau_1 \tau_2^2 \tilde{b}_3(p_0) + O(4) \}) \\
 &\cdot \exp (Y \{ \tau_2 + \frac{1}{2} \tau_1^2 \tau_2 \tilde{a}_2(p_0) + \frac{1}{2} \tau_1 \tau_2^2 \tilde{b}_2(p_0) + \frac{1}{6} \tau_1^3 \tau_2 \tilde{c}_2(p_0) + \frac{1}{4} \tau_1^2 \tau_2^2 \tilde{d}_2(p_0) \\
 (3.17) \quad &+ \frac{1}{6} \tau_1 \tau_2^3 \tilde{e}_3(p_0) + \frac{1}{24} \tau_1^4 \tau_2 m_2(p_0) + \frac{1}{24} \tau_1 \tau_2^4 n_2(p_0) \\
 &+ \frac{1}{5!} \tau_1^5 \tau_2 k_2(p_0) + \dots \}) \\
 &\cdot \exp (X \{ \tau_1 (1 + s) + \frac{1}{2} \tau_1^2 \tau_2 \tilde{a}_1(p_0) + \frac{1}{2} \tau_1 \tau_2^2 \tilde{b}_1(p_0) + \frac{1}{6} \tau_1^3 \tau_2 \tilde{c}_1(p_0) \\
 &+ \frac{1}{4} \tau_1^2 \tau_2^2 \tilde{d}_1(p_0) + \frac{1}{6} \tau_1 \tau_2^3 \tilde{e}_1(p_0) \\
 &+ \frac{1}{24} \tau_1^4 \tau_2 m_1(p_0) + \frac{1}{24} \tau_1 \tau_2^4 n_1(p_0) + \frac{1}{5!} \tau_1^5 \tau_2 k_1(p_0) + \dots \}).
 \end{aligned}$$

The same calculation has to be done for $p_0 e^{t_1 Y} e^{t_2 X} e^{t_3 Y}$. It follows from (3.1) that t_2 is divisible by τ_1 and that t_1 and t_3 are divisible by τ_2 . Also t_1 is divisible by s . Therefore any term of order $\geq k$ in the t -variables is also of order $\geq k$ in the τ -variables. To stay within the desired accuracy we have to pick up the fourth-order terms plus terms of orders $t_1 t_2^4$, $t_3 t_2^4$, $t_1 t_2^5$, $t_3 t_2^5$ and $t_2 \cdot O(4, \text{ in } t_1, t_3)$. The formal procedures are exactly as above and the result is

$$\begin{aligned}
 p_0 e^{t_1 Y} e^{t_2 X} e^{t_3 Y} &= p_0 \exp ([X, Y] \{ t_2 t_3 (1 + O(2)) + \frac{1}{2} t_2^2 t_3 \tilde{a}_3(p_0) + t_2 t_3 (t_1 + \frac{1}{2} t_3) \tilde{b}_3(p_0) + O(4) \}) \\
 &\cdot \exp (Y \{ t_1 + t_3 + \frac{1}{2} t_2^2 t_3 \tilde{a}_2(p_0) + t_2 t_3 (t_1 + \frac{1}{2} t_3) \tilde{b}_2(p_0) + \frac{1}{6} t_2^3 t_3 \tilde{c}_2(p_0) \\
 &+ \frac{1}{2} t_2^2 t_3 (t_1 + \frac{1}{2} t_3) \tilde{d}_2(p_0) + \frac{1}{2} t_2 t_3 (t_1^2 + t_1 t_3 + \frac{1}{3} t_3^2) \tilde{e}_2(p_0) \\
 (3.18) \quad &+ \frac{1}{24} t_2^4 t_3 m_2(p_0) + \frac{1}{6} t_2 t_3 (t_1^3 + \frac{3}{2} t_1^2 t_3 + t_1 t_3^2 + \frac{1}{4} t_3^3) n_2(p_0) \\
 &+ \frac{1}{5!} t_2^5 t_3 k_2(p_0) + \dots \}) \\
 &\cdot \exp (X \{ t_2 + \frac{1}{2} t_2^2 t_3 \tilde{a}_1(p_0) + t_2 t_3 (t_1 + \frac{1}{2} t_3) \tilde{b}_1(p_0) + \frac{1}{6} t_2^3 t_3 \tilde{c}_1(p_0) \\
 &+ \frac{1}{2} t_2^2 t_3 (t_1 + \frac{1}{2} t_3) \tilde{d}_1(p_0) + \frac{1}{2} t_2 t_3 (t_1^2 + t_1 t_3 + \frac{1}{3} t_3^2) \tilde{e}_1(p_0) \\
 &+ \frac{1}{24} t_2^4 t_3 m_1(p_0) + \frac{1}{6} t_2 t_3 (t_1^3 + \frac{3}{2} t_1^2 t_3 + t_1 t_3^2 + \frac{1}{4} t_3^3) n_1(p_0) \\
 &+ \frac{1}{5!} t_2^5 t_3 k_1(p_0) + \dots \}).
 \end{aligned}$$

Now we compare the coefficients in (3.17) and (3.18). Adding the X and Y equations we get

$$\begin{aligned}
 \Delta &= t_1 + t_2 + t_3 - \tau_1 - \tau_2 - \tau_1 s \\
 &= \tilde{\alpha}(p_0) (\frac{1}{2} \tau_1^2 \tau_2 - \frac{1}{2} t_2^2 t_3) + \tilde{\beta}(p_0) (\frac{1}{2} \tau_1 \tau_2^2 - t_2 t_3 (t_1 + \frac{1}{2} t_3)) \\
 (3.19) \quad &+ \tilde{\gamma}(p_0) (\frac{1}{6} \tau_1^3 \tau_2 - \frac{1}{6} t_2^3 t_3) + \tilde{\delta}(p_0) (\frac{1}{4} \tau_1^2 \tau_2^2 - \frac{1}{2} t_2^2 t_3 (t_2 + \frac{1}{2} t_3)) \\
 &+ \tilde{\eta}(p_0) (\frac{1}{6} \tau_1 \tau_2^3 - \frac{1}{2} t_2 t_3 (t_1^2 + t_1 t_3 + \frac{1}{3} t_3^2)) \\
 &+ \tilde{\mu}(p_0) (\frac{1}{24} \tau_1^4 \tau_2 - \frac{1}{24} t_2^4 t_3) + \tilde{\kappa}(p_0) (\frac{1}{5!} \tau_1^5 \tau_2 - \frac{1}{5!} t_2^5 t_3) \\
 &+ \tilde{\nu}(p_0) (\frac{1}{24} \tau_1 \tau_2^4 - \frac{1}{6} t_2 t_3 (t_1^3 + \frac{3}{2} t_1^2 t_3 + t_1 t_3^2 + \frac{1}{4} t_3^3)) + \dots
 \end{aligned}$$

To transform this into an equation in τ_1 , τ_2 and s we solve the equations which we get when we equate coefficients in (3.17) and (3.18) for t_1 , t_2 and t_3 . Notice that all we really need are the leading terms in these approximations since each of the summands in Δ has a certain polynomial in τ as a factor; for instance $\tau_1^3 \tau_2$ factors the

term at $\gamma(p_0)$, and we already allow for perturbations at $\gamma(p_0)$. Therefore the following three equations suffice:

- (i) $\tau_1(1+s) + O(3) = t_2 + O(3)$,
- (ii) $\tau_2 + O(3) = t_1 + t_3 + O(3)$,
- (iii) $\tau_1\tau_2(1+O(1)) = t_2t_3(1+O(1))$.

The approximate solutions are

$$t_1 = \tau_2 s(1 + O(1)), \quad t_2 = \tau_1(1 + s + O(2)), \quad t_3 = \tau_2 \left(\frac{1}{1+s} + O(1) \right).$$

We now substitute these formulas into (3.19) to compute the leading terms. For instance, let us compute the coefficient at $\tilde{\eta}(p_0)$:

$$\begin{aligned} \frac{1}{6} \tau_1 \tau_2^3 - \frac{1}{2} t_2 t_3 \left(t_1^2 + t_1 t_3 + \frac{1}{3} t_3^2 \right) &= \tau_1 \tau_2^3 \left(\frac{1}{6} - \frac{1}{2} (1 + O(1)) \right) \left(\frac{s^2 + s + 1/3}{(1+s)^2} + O(1) \right) \\ &= \tau_1 \tau_2^3 \left(-\frac{1}{6} s + 0 \cdot s^2 + O(1) \right). \end{aligned}$$

We do this for each term and this yields the asymptotic expansion (3.3). \square

Remark. Observe that the function t_1 is positive. A consequence of this is that the structure of the first junction is irrelevant and so (3.4) and (3.5) are in fact necessary conditions for time-optimality of an $\cdot XYX$ concatenation, where the dot stands for any switching.

4. Necessary conditions for time optimality of bang-bang trajectories: point variations. Let $\text{Reach}_{\Sigma, \leq T}(\tilde{p})$ be the set of all points reachable from \tilde{p} by a trajectory of Σ within time T . In this section we will compute an approximating cone for $\text{Reach}_{\Sigma, \leq T}(\tilde{p})$ by making point variations along bang-bang trajectories. This leads to additional necessary conditions for time optimality.

Let $\tilde{q} \in \text{Reach}_{\Sigma, \leq T}(\tilde{p})$. We call a cone $\tilde{K} \subseteq \mathbb{R}^3$ an almost approximating cone for $\text{Reach}_{\Sigma, \leq T}(\tilde{p})$ at \tilde{q} if for every finite set of vectors $v_1, \dots, v_n \in \tilde{K}$ and any $\delta > 0$ there exist vectors $v'_1, \dots, v'_n \in \mathbb{R}^3$ with $\|v_i - v'_i\| < \delta$ for $i = 1, \dots, n$ and a continuous map $\Psi: \{x \in \mathbb{R}^3: \|x\| < \varepsilon, x_i \geq 0\} \rightarrow \text{Reach}_{\Sigma, \leq T}(\tilde{p})$ such that $\Psi(x_1, \dots, x_n) = \tilde{q} + x_1 v_1 + \dots + x_n v_n + O(\|x\|)$ as $\|x\| \rightarrow 0$. We note that, if \tilde{K} is an almost approximating cone, so is its convex hull. The following topological fact is well known if $v'_i = v_i$ [13] and can easily be generalized to this slightly more general situation.

PROPOSITION 1. *Let Γ be a trajectory which steers \tilde{p} to \tilde{q} in time T . If \mathbb{R}^3 is an almost approximating cone for $\text{Reach}_{\Sigma, \leq T}(\tilde{p})$ at \tilde{q} , then \tilde{q} is an interior point of $\text{Reach}_{\Sigma, \leq T}(\tilde{p})$ and therefore Γ is not time-optimal beyond \tilde{q} .*

We now use this result to prove the following.

PROPOSITION 2. *A necessary condition for time optimality of a YXY -trajectory beyond its second switching point p_1 is that*

$$(4.1) \quad -\alpha(p_1)\tau_1 + \frac{1}{2}\gamma(p_1)\tau_1^2 - \frac{1}{6}\mu(p_1)\tau_1^3 + O(\tau_1^4) \leq 0$$

where τ_1 is the time along the X -trajectory.

COROLLARY. *A necessary condition for time optimality of an XYX -trajectory beyond its second switching point p_2 is that*

$$(4.2) \quad \beta(p_2)\tau_2 - \frac{1}{2}\eta(p_2)\tau_2^2 + \frac{1}{6}\nu(p_2)\tau_2^3 + O(\tau_2^4) \leq 0$$

where τ_2 is the time along the Y -arc.

The corollary follows immediately from Proposition 2 if we interchange X and Y , i.e., apply Sym 1 to (4.2). We now prove Proposition 2. First we compute some

variational vectors. Let $\Gamma = (x(\cdot), u(\cdot))$ be a YXY -trajectory with switching points p_0 and p_1 and let t_0 and t_1 be the corresponding switching times. Let $\tilde{q} = p_1 e^{rY}$ be a point on Γ just after the second switching point, i.e., for r small. Everywhere we can shorten the trajectory by deleting time. The variation is given by

$$z_1(s) = x(\bar{t}) e^{-sY} \quad \text{or} \quad z_2(s) = x(\bar{t}) e^{-sX}$$

depending on whether $\bar{t} \in (0, t_0] \cup (t_1, t_1 + r]$ or $\bar{t} \in (t_0, t_1]$. This gives the variational vectors $\dot{z}_1(0) = -Y(x(\bar{t}))$, respectively, $\dot{z}_2(0) = -X(x(\bar{t}))$. Another standard variation is to interchange X with Y . For $\bar{t} \in (0, t_0] \cup (t_1, t_1 + r]$ we have

$$z_3(s) = x(\bar{t}) e^{-sY} e^{sX}, \quad \dot{z}_3(0) = (X - Y)(x(\bar{t})) = -2g(x(\bar{t}))$$

and for $\bar{t} \in (t_0, t_1]$

$$z_4(s) = x(\bar{t}) e^{-sX} e^{sY}, \quad \dot{z}_4(0) = (Y - X)(x(\bar{t})) = 2g(x(\bar{t})).$$

Finally, we also use two variations which can only be done at the switching points, namely

$$z_5(s) = p_0 e^{-\sqrt{s}Y} e^{\sqrt{s}X} e^{\sqrt{s}Y} e^{-\sqrt{s}X}, \quad \dot{z}_5(0) = [X, Y](p_0)$$

and

$$z_6(s) = p_1 e^{-\sqrt{s}X} e^{\sqrt{s}Y} e^{\sqrt{s}X} e^{-\sqrt{s}Y}, \quad \dot{z}_6(0) = -[X, Y](p_1).$$

Now transport all these variational vectors $\dot{z}_i(0)$, $i = 1, \dots, 6$ along Γ to \tilde{q} and call the resulting vectors $v_i(r)$. We *claim* that the vectors $v_i(r)$, $i = 1, \dots, 6$, where we choose \bar{t} to be one of the switching times t_0 or t_1 , generate an almost approximating cone for $\text{Reach}_{\Sigma, \leq t_1+r}(\tilde{p})$ at \tilde{q} . To see this, observe first that variational vectors which are obtained at different times along Γ can be combined to generate an approximating cone (this is an almost approximating cone where we require in addition that $v'_i = v_i$ for all i). For instance, it is easily checked that

$$\Psi(x_1, x_2) := \tilde{p} e^{(t_0 - \sqrt{x_1})Y} e^{\sqrt{x_1}X} e^{\sqrt{x_1}Y} e^{(t_1 - \sqrt{x_1})X} e^{(r - x_2)Y}$$

is an approximating map for the vectors $v_1(r)$ (at $\bar{t} = t_1 + r$) and $v_5(r)$ (at $\bar{t} = t_0$). The general case is as simple as this one; only the notation really becomes cumbersome. Since these arguments are fairly standard to everyone in control theory by now, we omit a detailed proof. (The interested reader may consult, for instance, [7]). In our case we also use variational vectors at the same times $\bar{t} = t_0$ or $\bar{t} = t_1$. Since we can get $\pm g(x(\bar{t}))$ for times \bar{t} arbitrarily close to the switching times, it is still possible to combine the vectors $v_i(r)$ to obtain an almost approximating map. The definition is tailored such that exactly this can be done. This proves the claim.

Let $\tilde{K}(r)$ be the convex cone generated by $v_i(r)$, $i = 1, \dots, 6$. Γ is not time-optimal beyond p_1 if $\tilde{K}(r) = \mathbb{R}^3$ for sufficiently small r . Let K be the cone generated by the vectors $v_i(0)$, i.e., when we set $r = 0$. Equivalently K is generated by the vectors

$$-f(p_1), \quad \pm g(p_1), \quad e^{-\tau_1 \text{ad} X}([X, Y](p_0)) \quad \text{and} \quad -[X, Y](p_1).$$

Observe that $e^{-\tau_1 \text{ad} X}([X, Y](p_0)) = [X, Y](p_1) + O(\tau_1)$ and that

$$\begin{aligned} & e^{-\tau_1 \text{ad} X}([X, Y](p_0)) - [X, Y](p_1) \\ &= -\tau_1[X, [X, Y]](p_1) + \frac{1}{2}\tau_1^2 \text{ad}^3 X(Y)(p_1) - \frac{1}{6}\tau_1^3 \text{ad}^4 X(Y)(p_1) + O(\tau_1^4) \\ &= (-\alpha(p_1)\tau_1 + \frac{1}{2}\gamma(p_1)\tau_1^2 - \frac{1}{6}\mu(p_1)\tau_1^3 + O(\tau_1^4)) \cdot f(p_1) \\ & \quad + \text{a linear combination of } g(p_1) \text{ and } [X, Y](p_1). \end{aligned} \tag{4.3}$$

If (4.1) is violated, the f -coefficient in (4.3) is positive. Also, $e^{-\tau_1 \text{ad} X}([X, Y](p_0))$ is only a small perturbation of $[X, Y](p_1)$. Since f , g and $[X, Y]$ are independent, we

can therefore generate all of \mathbb{R}^3 in the positive span of these vectors. This does not change if we perturb the vectors of order r and so $\tilde{K}(r) = \mathbb{R}^3$ for sufficiently small r . Thus (4.1) is a necessary condition for time optimality of Γ beyond p_1 . \square

In Proposition 2 we used only the most obvious variational vectors to generate an almost approximating cone. Doing a little bit more sophisticated combinations, we get a different result which will be needed in § 5.

PROPOSITION 3. *Suppose $\gamma(p) > 0$ and let $\varepsilon > 0$ be small. There exists a neighborhood $\mathcal{U} = \mathcal{U}(\varepsilon)$ of p such that a necessary condition for time optimality of a $YXY(p_0)$ -trajectory which lies in \mathcal{U} is that*

$$(4.4) \quad -\alpha(p_1) + \left(\frac{2}{3} - \varepsilon\right) \tilde{\gamma}(p_1) \tau_1 \leq 0.$$

Proof. We start with a small neighborhood V of p such that $\gamma(q) > \frac{1}{2}\gamma(p)$ on V . Let M be an upper bound for $|a_3|$ and $|c_3|$ on V . Choose $N \in \mathbb{N}$ such that $1 \leq 3\varepsilon(N+1)$ and choose T so small that $TNM < \frac{1}{2}$, $TN \ll \frac{1}{2}$, $T \ll \gamma(p)$. Let \mathcal{U} be a neighborhood of p such that any trajectory of Σ leaves \mathcal{U} within time T . We claim that this choice of \mathcal{U} does what we want.

As in the proof of Proposition 2 we will show that $K = \mathbb{R}^3$ if (4.4) is violated. Recall that $-f(p_1)$, $\pm g(p_1)$ and $-[X, Y](p_1)$ lie in K . Therefore it suffices to exhibit a vector in K which has positive coefficients at both $f(p_1)$ and $[X, Y](p_1)$. Since $(Y - X)(p_0)$, $[X, Y](p_0)$ and $(X - Y)(p_1)$ are variational vectors, the following vector lies in K :

$$\begin{aligned} W &= N \cdot e^{-\tau_1 \text{ad}^X}([X, Y](p_0)) + (N-1) e^{-\tau_1 \text{ad}^X}(Y - X)(p_0) + (X - Y)(p_1) \\ &= [X, Y](p_1) + N \cdot \left(-\frac{1}{2} \tau_1 \left(1 + \frac{1}{N} \right) [X, [X, Y]](p_1) + \frac{1}{3} \tau_1^2 \left(1 + \frac{1}{2N} \right) [X, [X, [X, Y]] \right. \\ &\quad \left. (p_1) + O(\tau_1^3) \right). \end{aligned}$$

Now we express the higher order brackets as linear combinations of f , g and $[X, Y]$. The coefficient at $[X, Y](p_1)$ is of the form $1 + N\tau_1 h$, where $|h| \leq M$. By construction this term is positive. If (4.4) now is violated then also

$$-\frac{1}{2} \alpha(p_1) + \frac{1}{3} \frac{2N+1}{2N+2} \gamma(p_1) \tau_1 + O(\tau_1^2) > 0$$

and so W also has a positive f -coefficient. Hence $K = \mathbb{R}^3$ and Γ is not time optimal beyond p_1 . \square

5. The analysis of time-optimal bang-bang trajectories in generic cases. We now use the necessary conditions of §§ 3 and 4 to analyze the time optimality of bang-bang trajectories near a reference point p in some generic cases. All our considerations apply only on a sufficiently small neighborhood \mathcal{U} of p which we adjust whenever necessary for the argument without changing the name. For instance, if $\alpha(p) \neq 0$, then we can choose \mathcal{U} so small that the time T along any trajectory that lies in \mathcal{U} is dominated by $\alpha(q)$ for any $q \in \mathcal{U}$. This is possible since $f(p)$ and $g(p)$ are independent and therefore we can make T arbitrarily small by choosing \mathcal{U} small. Henceforth we will tacitly assume that any function which does not vanish at p dominates on \mathcal{U} all terms of order $O(1)$ that come up. Also we will not refer explicitly to \mathcal{U} from now on. It is understood that all statements are valid only for trajectories that lie in a sufficiently small neighborhood of p and we use the phrase “near p ” to indicate this.

5.1. The generic cases where at most one of $\alpha(p)$ and $\beta(p)$ vanishes. Several of the cases can be analyzed directly with the necessary conditions of §§ 3 and 4. If $\alpha(p) < 0$, then YXY -trajectories are not time-optimal by (4.1) and if $\beta(p) > 0$, then (4.2) excludes XYX -trajectories. The only nontrivial cases occur therefore when $\alpha(p) \geq 0$ and $\beta(p) \leq 0$. In the codimension 0 case, $\alpha(p) > 0$ and $\beta(p) < 0$, the conjugate point relation (3.4) implies that τ_1 and τ_2 are comparable in the sense that there exists a constant C , only depending on the size of a small neighborhood of p , but not on the specific trajectory under consideration, such that $\tau_1 \leq C\tau_2$ and $\tau_2 \leq C\tau_1$. Therefore the optimality condition (3.5) excludes the time optimality of $YXYX$ -concatenations. Since the assumptions $\alpha(p) > 0$ and $\beta(p) < 0$ are invariant when we apply Sym 1, the input symmetry which interchanges X and Y , this excludes also $XYXY$ -concatenations. Equivalently, this can be seen from (3.12) and (3.13) as well. So we have the following.

PROPOSITION 1. *Near p time-optimal bang-bang trajectories have at most the following form:*

- (a) XYX if $\alpha(p) < 0$; YXY if $\beta(p) > 0$.
- (b) XY or YX if $\alpha(p) < 0$ and $\beta(p) > 0$; XYX or YXY if $\alpha(p) > 0$ and $\beta(p) < 0$.

For the more degenerate generic cases we can assume without loss of generality that $\alpha(p) = 0$ and $\beta(p) < 0$. The case $\alpha(p) > 0$ and $\beta(p) = 0$ is then a direct consequence of our results when we apply Sym 1.

PROPOSITION 2. *Let $\alpha(p) = 0$ and $\beta(p) < 0$. Then near p time-optimal bang-bang trajectories are at most concatenations of the form:*

- (i) $XYXY$ if $\gamma(p) > 0$; $YXYX$ if $\gamma(p) < 0$.
- (ii) $XYXYX$ if $\gamma(p) = 0$, $\mu(p) < 0$; $YXYXY$ if $\gamma(p) = 0$, $\mu(p) > 0$.
- (iii) $XYXYXY$ if $\gamma(p) = 0$, $\mu(p) = 0$, $\kappa(p) > 0$; $YXYXYX$ if $\gamma(p) = 0$, $\mu(p) = 0$, $\kappa(p) < 0$.

COROLLARY. *Let $\alpha(p) > 0$ and $\beta(p) = 0$. Then near p time-optimal bang-bang trajectories are at most concatenations of the form:*

- (i) $XYXY$ if $\eta(p) > 0$; $YXYX$ if $\eta(p) < 0$.
- (ii) $XYXYX$ if $\eta(p) = 0$, $\nu(p) < 0$; $YXYXY$ if $\eta(p) = 0$, $\nu(p) > 0$.
- (iii) $XYXYXY$ if $\eta(p) = 0$, $\nu(p) = 0$, $\rho(p) > 0$; $YXYXYX$ if $\eta(p) = 0$, $\nu(p) = 0$, $\rho(p) < 0$.

Since this is the first time we use input symmetries to exclude the optimality of trajectories we give a proof of the corollary. Later on we will omit these arguments. Let Σ^* be the system where we replace g with $-g$, i.e., $\dot{x} = f(x) - ug(x)$. Then $\Sigma^* = \text{Sym } 1(\Sigma)$ and $\alpha^* = -\beta$, $\beta^* = -\alpha$, $\gamma^* = -\eta$, etc. Therefore, by Proposition 1, $X^*Y^*X^*Y^*$ concatenations are not time-optimal if $\alpha^*(p) = 0$, $\beta^*(p) < 0$ and $\gamma^*(p) < 0$. For Σ this excludes $YXYX$ concatenations if $\beta(p) = 0$, $\alpha(p) > 0$ and $\eta(p) > 0$. All the other results follow exactly like this.

Remark. In a forthcoming paper we will prove that, in all the cases considered above, time-optimal trajectories are in fact bang-bang and so this gives then a precise classification of the structure of time-optimal trajectories near p .

We single out the computational part of the proof in the following lemma.

LEMMA 1. *Let Γ be a time-optimal $YXYXYX$ trajectory with switching points p_0, \dots, p_4 and let τ_i , $i = 1, \dots, 4$, be the time along the trajectory between p_{i-1} and p_i . Then the following conditions are necessary for time optimality of the sub-arcs specified:*

$$(5.1) \quad 0 \leq -\gamma(p_2) + \frac{1}{4}\tilde{\mu}(p_2)\tau_1 - \frac{1}{20}\kappa(p_2)\tau_1^2 + O(\tau_1^3) \quad \text{for } YXYX(p_0),$$

$$(5.2) \quad 0 \leq \gamma(p_2) + \frac{1}{4}\tilde{\mu}(p_2)\tau_3 + \frac{1}{20}\kappa(p_2)\tau_3^2 + O(\tau_3^3) \quad \text{for } XYXY(p_1),$$

$$(5.3) \quad 0 \leq -\gamma(p_2) - \frac{3}{4}\tilde{\mu}(p_2)\tau_3 - \frac{3}{10}\kappa(p_2)\tau_3^2 + O(\tau_3^3) \quad \text{for } YXYX(p_2).$$

Proof of the lemma. We first prove (5.1) and (5.3). We use the conjugate point relation (3.4) to eliminate the term $0 \cdot \tilde{\alpha}(p_0)\tau_1$ in (3.5) and then we bound all terms which have the factor τ_2 by $\frac{1}{3}\beta(p)\tau_2$. This gives

$$(5.4) \quad 0 \leq \frac{1}{3}\beta(p)\tau_2 - \frac{1}{3}\tilde{\gamma}(p_0)\tau_1^2 - \frac{1}{4}\tilde{\mu}(p_0)\tau_1^3 - \frac{1}{10}\tilde{\kappa}(p_0)\tau_1^4 + O(\tau_1^5).$$

Exactly the same condition holds at p_2 with τ_3 and τ_4 instead of τ_1 and τ_2 . This implies already (5.3). (We bound $\beta(p) < 0$, divide out the perturbation at $\gamma(p_2)$ and absorb the perturbation at $\kappa(p_2)$ in $O(\tau_3^5)$.) To prove (5.1), we now transport the functions γ , μ and κ from p_0 to p_2 . By Taylor's theorem

$$\begin{aligned} \gamma(p_0) &= \gamma(p_2) + O(\tau_2) - L_X\gamma(p_2)\tau_1 + \frac{1}{2}L_X^2\gamma(p_2)\tau_1^2 + O(\tau_1^3) \\ &= \tilde{\gamma}(p_2) - \mu(p_2)\tau_1 + \frac{1}{2}\kappa(p_2)\tau_1^2 + \alpha(p_2) \cdot O(1) + O(\tau_2) + O(\tau_1^3), \\ \mu(p_0) &= \mu(p_2) - L_X\mu(p_2)\tau_1 + O(\tau_2) + O(\tau_1^2) \\ &= \mu(p_2) - \kappa(p_2)\tau_1 + \alpha(p_2) \cdot O(1) + O(\tau_2) + O(\tau_1^2) \end{aligned}$$

and $\kappa(p_0) = \kappa(p_2) + O(1)$. Therefore we get from (5.4)

$$0 \leq \frac{1}{3}\beta(p)\tau_2 + \alpha(p_2)\tau_1^2 \cdot O(1) - \frac{1}{3}\tau_1^2(\tilde{\gamma}(p_2) - \frac{1}{4}\tilde{\mu}(p_2)\tau_1 + \frac{1}{20}\tilde{\kappa}(p_2)\tau_1^2 + O(\tau_1^3) + O(\tau_2)).$$

The α -term and the terms with a factor τ_2 that came up in this procedure are dominated by $\beta(p)\tau_2 < 0$ and so this implies (5.1).

We get (5.2) in the same way from (3.13). Using (3.12) to eliminate the α -terms and again bounding the τ_2 terms by $\frac{1}{3}\beta(p)\tau_2$, we obtain

$$0 \leq \frac{1}{3}\beta(p_2)\tau_2 + \frac{1}{3}\tau_3^2(\tilde{\gamma}(p_3) - \frac{3}{4}\tilde{\mu}(p_3)\tau_3 + \frac{3}{10}\tilde{\kappa}(p_3)\tau_3^2 + O(\tau_3^3)).$$

Now we transport to p_2 to get

$$0 \leq \frac{1}{3}\beta(p)\tau_2 + \alpha(p_2)\tau_3^2 \cdot O(1) + \frac{1}{3}\tau_3^2(\tilde{\gamma}(p_2) + \frac{1}{4}\tilde{\mu}(p_2)\tau_3 + \frac{1}{20}\tilde{\kappa}(p_2)\tau_3^2 + O(\tau_2) + O(\tau_3^3))$$

and this implies (5.2). \square

We now give the *proof* of the proposition. Case (i) follows immediately from (5.1) and (5.2) excluding the time optimality of $XYXY$ -trajectories if $\gamma(p) < 0$ and of $YXYX$ -trajectories if $\gamma(p) > 0$.

Now suppose $\gamma(p) = 0$. We add (5.1) and (5.2) to cancel out $\gamma(p_2)$. This leads to

$$0 \leq \mu(p_2)(\tau_1 + \tau_3) - \frac{1}{5}\tilde{\kappa}(p_2)(\tau_1^2 - \tau_3^2) + O(\tau_1^3, \tau_3^3).$$

Since $\tau_i/(\tau_1 + \tau_3) \leq 1$ for $i = 1, 3$, this implies that

$$(5.5) \quad 0 \leq \mu(p_2) + \frac{1}{5}\tilde{\kappa}(p_2)(\tau_3 - \tau_1) + O(\tau_1^2, \tau_3^2).$$

Therefore $YXYXY$ trajectories are not time-optimal if $\mu(p) < 0$. In the case $\mu(p) > 0$ we combine (5.2) and (5.3) to obtain

$$(5.6) \quad 0 \leq -\mu(p_2) - \frac{1}{2}\tilde{\kappa}(p_2)\tau_3 + O(\tau_3^2)$$

which excludes $XYXYX$ -trajectories. This proves (ii). If also $\mu(p) = 0$, we combine (5.6) and (5.5) to get

$$0 \leq \kappa(p_2)(-2\tau_1 - 3\tau_3) + O(\tau_1^2, \tau_3^2).$$

This excludes the optimality of $YXYXYX$ -trajectories if $\kappa(p) > 0$. Then we apply this to the time reversed system (i.e., use “time reversal” = Sym 3) to obtain that $XYXYXY$ -trajectories are not time-optimal in case $\kappa(p) < 0$. This proves (iii) and also the proposition. \square

5.2. $\alpha(p)=0$ and $\beta(p)=0$: The codimension 2 cases. The cases when $\alpha(p)$ and $\beta(p)$ vanish are technically more difficult due to the fact that now both linear terms in the necessary conditions of §§ 3 and 4 are small, i.e., not dominant, and we have no a priori knowledge about their signs. Our strategy is to combine several necessary conditions to cancel out the linear terms and to get applicable conditions with dominant higher-order terms.

PROPOSITION 3. *Let $\alpha(p)=0$, $\beta(p)=0$ and assume $\gamma(p)>0$, $\eta(p)\neq 0$. Then near p bang-bang trajectories of the following structures are not time-optimal:*

- (i) $XYXYX$ or $YXYXY$ if $\eta(p)<0$;
- (ii) $YXYX$ if $\eta(p)>0$ and $\delta(p)\geq 0$; $YXYXYX$ if $\eta(p)>0$ and $\delta(p)<0$.

The generic cases when $\gamma(p)<0$ follow from this by a time reversal argument (recall that $\text{Sym } 3(\gamma)=-\gamma$). We get the following corollary.

COROLLARY. *Let $\alpha(p)=0$, $\beta(p)=0$ and assume $\gamma(p)<0$, $\eta(p)\neq 0$. Then near p bang-bang trajectories of the following structures are not time-optimal:*

- (i) $XYXYX$ or $YXYXY$ if $\eta(p)>0$;
- (ii) $YXYX$ if $\eta(p)<0$ and $\delta(p)\leq 0$; $YXYXYX$ if $\eta(p)<0$ and $\delta(p)>0$.

Proof of Proposition 3. We consider a $YXYXYX$ -concatenation with switching points p_0, \dots, p_4 and we let τ_i be the time along the arc joining p_{i-1} and p_i , $i=1, \dots, 4$. An important step in the proof will be to get information on how the times τ_i relate to each other. We will write $\tau_i \leq C\tau_{i+1}$ if there exists a constant, only dependent on some neighborhood \mathcal{U} of p , but not on a specific trajectory under consideration, such that $\tau_i \leq \text{Const} \cdot \tau_{i+1}$ is a necessary condition for time optimality of any trajectory that lies in \mathcal{U} . We also have to deal with constants which can be made arbitrarily small by choosing the neighborhood \mathcal{U} of p small enough. For instance, a necessary condition like

$$\frac{1}{3}\tilde{\gamma}(p_2)\tau_3^2 \leq -\beta(p_2)\tau_2 + \frac{1}{3}\tilde{\eta}(p_2)\tau_2^2, \quad \gamma(p)>0,$$

implies that $\tau_3^2 \leq \text{Const} \cdot \tau_2$, where the constant can be made as small as we please near p since $\beta(p)=0$ and since η occurs with a factor τ_2^2 . We denote constants of this type by ε . We will not indicate changes in these constants in the notation, that is we freely use $C+C=C$, $\varepsilon \cdot C=\varepsilon$, etc. Another technical problem is to keep track of higher-order terms. Whenever possible we will absorb cubic remainders at γ or η terms. This can be done since $\gamma(p)\neq 0$ and $\eta(p)\neq 0$. Therefore these terms are large relative to the time along trajectories near p . We also use the notation $\tilde{\tau}_i$ for terms of order $\tau_i(1+O(1))$.

(i) Since the conditions of this case are invariant under $\text{Sym } 1$, it suffices to exclude $YXYXY$ -concatenations. If the first switching point is p_0 , we get from (4.1) as necessary condition $\alpha(p_1) \geq \frac{1}{2}\tilde{\gamma}(p_1)\tau_1$. We use this in (3.11) to say

$$(5.7) \quad -\frac{1}{2}\tilde{\gamma}(p_1)\tau_1\tau_3 + \frac{1}{3}\tilde{\eta}(p_1)\tau_2^2 + O(\tau_3^3, \tau_3^2\tau_2) \geq 0.$$

Notice that this condition is violated if τ_2 and τ_3 are comparable since then $\eta\tau_2^2$ dominates the remainders. We distinguish the cases

- (a) $\alpha(p_2) \leq -\frac{1}{4}\gamma(p_2)\tau_3$
- and
- (b) $\alpha(p_2) > -\frac{1}{4}\gamma(p_2)\tau_3$.

In case (a) we have

$$\begin{aligned} -\frac{1}{4}\tilde{\gamma}(p_1)\tau_3 &= -\frac{1}{4}\tilde{\gamma}(p_2)\tau_3 \geq \tilde{\alpha}(p_2) = \tilde{\alpha}(p_1) + \delta(p_1)\tau_2 + O(\tau_2^2) \\ &\geq \frac{1}{2}\tilde{\gamma}(p_1)\tau_1 + \tilde{\delta}(p_1)\tau_2 + O(\tau_2^2). \end{aligned}$$

This can hold only if $\delta(p) \leq 0$ and it then implies $\tau_3 \leq C\tau_2$. We now combine (3.11) and $-\alpha(p_3)\tau_3 + \frac{1}{2}\tilde{\gamma}(p_3)\tau_3^2 \leq 0$, which follows from (4.1). If we transport the functions back to p_1 , this gives as necessary condition for a $XYXY(p_1)$ trajectory that

$$(5.8) \quad -\frac{1}{2}\gamma(p_1)\tau_3^2 - \tilde{\delta}(p_1)\tau_2\tau_3 - \frac{1}{3}\tilde{\eta}(p_1)\tau_2^2 \leq 0.$$

This implies $\tau_2 \leq C\tau_3$ and so τ_2 and τ_3 are comparable, i.e., (5.7) is violated. We now treat case (b). Observe that the conjugate point relations (3.4) and (3.10) are equivalent to

$$(5.9) \quad 0 = \tilde{\alpha}(p_1)\tau_1 + \tilde{\beta}(p_1)\tau_2 - \frac{1}{3}\tilde{\gamma}(p_1)\tau_1^2 + \frac{1}{3}\tilde{\eta}(p_1)\tau_2^2,$$

$$(5.10) \quad 0 = \tilde{\alpha}(p_2)\tau_3 + \tilde{\beta}(p_2)\tau_2 + \frac{1}{3}\tilde{\gamma}(p_2)\tau_3^2 - \frac{1}{3}\tilde{\eta}(p_2)\tau_2^2.$$

If we let $r_1 := p_1 e^{-(1/3)\tau_1 X}$, $r_2 := p_1 e^{(1/3)\tau_2 Y}$, $s_1 := p_2 e^{-(1/3)\tau_2 Y}$ and $s_2 := p_2 e^{(1/3)\tau_3 X}$, then we can also write

$$0 = \tilde{\alpha}(r_1)\tau_1 + \tilde{\beta}(r_2)\tau_2 + O(3), \quad 0 = \tilde{\alpha}(s_2)\tau_3 + \tilde{\beta}(s_1)\tau_2 + O(3).$$

Therefore

$$(\tilde{\beta}(r_2) - \tilde{\beta}(s_1))\tau_2 = \tilde{\alpha}(s_2)\tau_3 - \tilde{\alpha}(r_1)\tau_1 + O(3) = (\tilde{\alpha}(s_2) - \tilde{\alpha}(r_1))\tau_3 + \tilde{\alpha}(r_1)(\tau_3 - \tau_1) + O(3)$$

and thus

$$(5.11) \quad -\frac{1}{3}\tilde{\eta}(p_1)\tau_2^2 = (\frac{1}{3}\tilde{\gamma}(p_1)\tau_1 + \tilde{\delta}(p_1)\tau_2 + \frac{1}{3}\tilde{\gamma}(p_1)\tau_3)\tau_3 + \tilde{\alpha}(r_1)(\tau_3 - \tau_1) + O(3).$$

In case (b) (5.10) implies $-\tilde{\beta}(p_2)\tau_2 + \frac{1}{3}\tilde{\eta}(p_2)\tau_2^2 \geq \frac{1}{12}\tilde{\gamma}(p_2)\tau_3^2$ and so $\tau_3^2 \leq \varepsilon\tau_2$. Therefore (5.7) yields both $\tau_1 \leq \varepsilon\tau_2$ and $\tau_2 \leq \varepsilon\tau_3$, in particular we can assume that $\tau_1 < \tau_3$. Since also $\tilde{\alpha}(r_1) = \tilde{\alpha}(p_1) - \frac{1}{3}\tilde{\gamma}(p_1)\tau_1 \geq \frac{1}{6}\tilde{\gamma}(p_1)\tau_1 > 0$ by (4.1), this implies

$$-\frac{1}{3}\tilde{\eta}(p_1)\tau_2^2 \geq (\tilde{\delta}(p_1)\tau_2 + \frac{1}{3}\tilde{\gamma}(p_1)\tau_3)\tau_3 \geq (\frac{1}{3}\tilde{\gamma}(p_1)\tau_3 - \varepsilon\tau_3)\tau_3 \geq \frac{1}{6}\tilde{\gamma}(p_1)\tau_3^2$$

and so $\tau_3 \leq C\tau_2$. But this contradicts $\tau_2 \leq \varepsilon\tau_3$ and thus $YXYXY$ -concatenations are not time-optimal near p .

(ii) The subcase when $\delta(p) \geq 0$ is easy. If we combine (3.5) with (4.2), where we transport the functions back to p_0 , we get as necessary condition for time optimality of $YXYX(p_0)$:

$$(5.12) \quad \frac{1}{3}\tilde{\gamma}(p_0)\tau_1^2 + \tilde{\delta}(p_0)\tau_1\tau_2 + \frac{1}{2}\tilde{\eta}(p_0)\tau_2^2 \leq 0.$$

This cannot hold if $\gamma(p) > 0$, $\eta(p) > 0$ and $\delta(p) \geq 0$. So let $\delta(p) < 0$. Then (5.12) implies that τ_1 and τ_2 are comparable. Analogously it follows that τ_3 and τ_4 are comparable. Near p also $\tau_3 \leq C\tau_1$ has to hold. This follows from (5.11) since it suffices to consider trajectories which satisfy $\tau_3 \geq \frac{4}{3}\tau_1$. Then $\tilde{\alpha}(r_1)(\tau_1 - \tau_3) > 0$ and we obtain $0 \geq \frac{1}{3}\tilde{\gamma}(p_1)\tau_1 + \tilde{\delta}(p_1)\tau_2 + \frac{1}{3}\tilde{\gamma}(p_1)\tau_3$ which implies even $\tau_1 + \tau_3 \leq C\tau_2$. We now fix some constant \hat{C} such that $\tau_3 \leq \hat{C}\tau_1$ is a necessary condition for time optimality near p . Let $\hat{\varepsilon} := \min(1/24, 1/24\hat{C})$ and choose a neighborhood of p such that (4.4) is valid for $\hat{\varepsilon}$.

This case becomes technical. Therefore we give the argument in a series of lemmas. The idea is to show that $YXYXY(p_0)$ is not time-optimal if $\tau_1 \leq \frac{3}{4}\tau_2$ and that $YXYXY(p_1)$ is not time-optimal if $\tau_4 \leq \frac{3}{4}\tau_2$. To combine the cases we will show that, if $\tau_1 > \frac{3}{4}\tau_2$, then necessarily $\tau_4 \leq \frac{3}{4}\tau_2$.

LEMMA 2. *A necessary condition for time optimality of $YXYXY(p_0)$ is*

$$(5.13) \quad \frac{1}{3}\tilde{\gamma}(p_1)\tau_1 + \tilde{\delta}(p_2)\tau_2 + \frac{1}{2}\tilde{\gamma}(p_1)\tau_3 \leq 0.$$

Proof. We subtract (5.9) from (5.10), cancelling out $\beta(p_1)\tau_2$ exactly by adjusting the error terms, and rearrange to get

$$0 = \tilde{\alpha}(p_1)(\tau_3 - \tilde{\tau}_1) + \frac{1}{3}\tilde{\gamma}(p_1)(\tau_1^2 + \tilde{\tau}_3^2) + \tilde{\delta}(p_1)\tau_2\tau_3 + \frac{1}{3}\tilde{\eta}(p_1)\tau_2^2.$$

Using $\tilde{\alpha}(p_1) \geq \frac{7}{12}\tilde{\gamma}(p_1)\tau_1 > 0$, which follows from (4.4), and (3.9), transported back to p_1 , we obtain

$$(5.14) \quad \begin{aligned} 0 &\geq \frac{7}{12}\tilde{\gamma}(p_1)\tau_1\tau_3 + \frac{1}{3}\tilde{\gamma}(p_1)(\tau_1^2 + \tilde{\tau}_3^2) + \tilde{\delta}(p_1)\tau_2(\tau_1 + \tilde{\tau}_3) + \frac{2}{3}\tilde{\eta}(p_1)\tau_2^2 \\ &= (\frac{1}{3}\tilde{\gamma}(p_1)\tau_1 + \tilde{\delta}(p_1)\tau_2 + \frac{1}{2}\tilde{\gamma}(p_1)\tau_3)\tau_1 + \frac{1}{3}\tilde{\eta}(p_1)\tau_2^2 \\ &\quad + ((\frac{1}{3} + \hat{\varepsilon})\tilde{\gamma}(p_1)\tau_3^2 + \tilde{\delta}(p_1)\tau_2\tau_3 + \frac{1}{3}\tilde{\eta}(p_1)\tau_2^2) + \tilde{\gamma}(p_1)\tau_3(\frac{1}{12}\tau_1 - \hat{\varepsilon}\tilde{\tau}_3). \end{aligned}$$

By our choice of $\hat{\varepsilon}$ the last term is positive. From (4.4) we have $\alpha(p_1) \geq (\frac{2}{3} - \hat{\varepsilon})\tilde{\gamma}(p_1)\tau_1$ and if we combine this with (3.11) we obtain

$$(5.15) \quad (\frac{1}{3} + \hat{\varepsilon})\tilde{\gamma}(p_1)\tau_3^2 + \tilde{\delta}(p_1)\tau_2\tau_3 + \frac{1}{3}\tilde{\eta}(p_1)\tau_2^2 \geq 0.$$

Clearly the perturbations at these terms need not be the same as at the corresponding terms in (5.14). But we can assume to have exactly the same expression $(\frac{1}{3} + \hat{\varepsilon})\tilde{\gamma}(p_1)\tau_3^2$ and so any higher-order perturbations which come up when we use (5.15) in (5.14) have the factor τ_2 . Since τ_2 is comparable to τ_1 we even conclude

$$0 \geq (\frac{1}{3}\tilde{\gamma}(p_1)\tau_1 + \tilde{\delta}(p_1)\tau_2 + \frac{1}{2}\tilde{\gamma}(p_1)\tau_3)\tau_1 + \frac{1}{3}\tilde{\eta}(p_1)\tau_2^2. \quad \square$$

LEMMA 3. If $\tau_1 \leq \frac{3}{4}\tau_3$, then a necessary condition for time optimality of $YXYXY(p_0)$ is

$$(5.16) \quad \frac{5}{12}\tilde{\gamma}(p_2)\tau_1 + \tilde{\delta}(p_2)\tau_2 + \frac{5}{12}\tilde{\gamma}(p_2)\tau_3 \geq 0.$$

Proof. Combining the strengthened approximating cone condition $\alpha(p_1) \geq \frac{7}{12}\tilde{\gamma}(p_1)\tau_1$ with (3.9) and with (3.11) we get

$$(5.17) \quad \frac{7}{12}\tilde{\gamma}(p_2)\tau_1^2 + \tilde{\delta}(p_2)\tau_1\tau_2 + \frac{1}{3}\tilde{\eta}(p_2)\tau_2^2 \leq 0,$$

$$(5.18) \quad -\frac{5}{12}\tilde{\gamma}(p_2)\tau_3^2 - \tilde{\delta}(p_2)\tau_3\tau_3 - \frac{1}{3}\tilde{\eta}(p_2)\tau_2^2 \leq 0.$$

We add (5.17) and (5.18), cancel out $\eta(p_2)\tau_2^2$ exactly and rearrange terms. This yields

$$(\frac{1}{12}\tilde{\gamma}(p_2)(7\tau_1 + 5\tilde{\tau}_3) + \tilde{\delta}(p_2)\tau_2)(\tau_1 - \tilde{\tau}_3) + \frac{1}{6}\tilde{\gamma}(p_2)\tau_1\tau_3 \leq 0.$$

Since $\tau_1 - \tau_3 \leq -\frac{1}{4}\tau_3 < 0$, the term at $(\tau_1 - \tilde{\tau}_3)$ is nonnegative and so (5.16) follows from $(\frac{1}{12}\tilde{\gamma}(p_2)(7\tau_1 + 5\tilde{\tau}_3) + \tilde{\delta}(p_2)\tau_2)\tau_1 \geq 0$. \square

But relations (5.13) and (5.16) are contradictory near p . Therefore $YXYXY(p_0)$ trajectories are not time-optimal if $\tau_1 \leq \frac{3}{4}\tau_3$.

Next we will show that, if $\tau_1 > \frac{3}{4}\tau_3$, then necessarily $\tau_4 \leq \frac{3}{4}\tau_2$. We first remark that (3.11) is equivalent to

$$\tilde{\alpha}(p_1)\tau_3 - \frac{1}{3}\tilde{\gamma}(p_1)\tau_3^2 - \tilde{\beta}(p_3)\tau_2 - \frac{1}{3}\tilde{\eta}(p_3)\tau_2^2 \leq 0.$$

This follows if we use the conjugate point relation (3.10) and $\beta(p_3) = \beta(p_1) + \tilde{\delta}(p_1)\tau_3 + \tilde{\eta}(p_1)\tau_2$. The approximating cone condition (4.4) implies

$$\tilde{\alpha}(p_1)\tau_3 - \frac{1}{3}\tilde{\gamma}(p_1)\tau_3^2 \geq \tilde{\gamma}(p_1)\tau_3(\frac{7}{12}\tau_1 - \frac{1}{3}\tilde{\tau}_3) > 0$$

and therefore $\beta(p_3) + \frac{1}{3}\tilde{\eta}(p_3)\tau_2 \geq 0$. Let $q := p_4 e^{-(1/3)\tau_4 Y}$. Then we have

$$\beta(q) = \beta(p_3) + \frac{2}{3}\tilde{\eta}(p_3)\tau_4 \geq \frac{1}{3}\tilde{\eta}(p_4)(2\tau_4 - \tilde{\tau}_2)$$

and from (4.4) (along Y , τ_4) we get also

$$\beta(q) = \beta(p_4) - \frac{1}{3}\tilde{\eta}(p_4)\tau_4 \leq \frac{1}{2}\tilde{\eta}(p_4)\tau_4 - \frac{1}{3}\tilde{\eta}(p_4)\tau_4 = \frac{1}{6}\tilde{\eta}(p_4)\tau_4.$$

This implies $\tau_4 \leq \frac{2}{3}\tilde{\tau}_2$, i.e., $\tau_4 \leq \frac{3}{4}\tau_2$ for small times.

LEMMA 4. If $\tau_4 \leq \frac{3}{4}\tau_2$, then a necessary condition for time optimality of $YXYXY(p_1)$ is that

$$(5.19) \quad \frac{1}{3}\tilde{\eta}(p_2)\tau_2 + \tilde{\delta}(p_2)\tau_3 + \frac{1}{4}\tilde{\eta}(p_2)\tau_4 \geq 0.$$

Proof. We now combine (5.18) with the analogous version of (5.17) for τ_3 and τ_4 in the form

$$\frac{7}{12}\tilde{\gamma}(p_2)\tau_3^2 + \tilde{\delta}(p_2)\tau_3\tau_4 + \frac{1}{3}\tilde{\eta}(p_2)\tau_4^2 \leq 0.$$

We can evaluate at p_2 since the change from p_4 to p_2 produces only higher-order perturbations. We cancel out the term $\gamma(p_2)\tau_3^2$ exactly to get

$$0 \leq (\frac{1}{3}\tilde{\eta}(p_2)\tau_2 + \tilde{\delta}(p_2)\tau_3 + \frac{1}{3}\tilde{\eta}(p_2)\tau_4)(7\tau_2 - 5\tilde{\tau}_4) - \frac{2}{3}\tilde{\eta}(p_2)\tau_2\tau_4.$$

Since $7\tau_2 - 5\tilde{\tau}_4 > 0$ we have $(\frac{1}{3}\tilde{\eta}(p_2)\tau_2 + \tilde{\delta}(p_2)\tau_3 + \frac{1}{3}\tilde{\eta}(p_2)\tau_4)(-5\tilde{\tau}_4) < 0$ and the result follows if we simply omit this term. \square

LEMMA 5. *If $\tau_4 \leq \frac{3}{4}\tau_2$, then a necessary condition for time optimality of $XYXYX(p_1)$ is that*

$$(5.20) \quad \frac{1}{3}\tilde{\eta}(p_2)\tau_2 + \tilde{\delta}(p_2)\tau_3 + \frac{1}{3}\tilde{\eta}(p_2)\tau_4 \leq 0.$$

Proof. Applying Sym 1 to (5.11), we get the corresponding relation for a $XYXYX(p_2)$ concatenation, namely

$$(5.21) \quad \frac{1}{3}\tilde{\gamma}(p_2)\tau_3^2 = -(\frac{1}{3}\tilde{\eta}(p_2)\tau_2 + \tilde{\delta}(p_2)\tau_3 + \frac{1}{3}\tilde{\eta}(p_2)\tau_4)\tau_4 - \beta(\bar{r}_1)(\tau_4 - \tau_2) + O(3)$$

where $\bar{r}_1 = p_2 e^{-(1/3)\tau_2 Y}$ and we choose to make this term exact. Here all cubic remainders can be absorbed at the quadratic terms except for those which have a factor τ_2^2 . This causes a little inconvenience and we split into the subcases (a) $-\beta(\bar{r}_1) \leq \frac{1}{6}\eta(p_2)\tau_2$ and (b) $-\beta(\bar{r}_1) > \frac{1}{6}\eta(p_2)\tau_1$. In case (b) we have in particular $\beta(\bar{r}_1) < 0$ and thus

$$-\beta(\bar{r}_1)(\tau_4 - \tau_2) + \tau_2^2 \cdot O(1) \leq \frac{1}{4}\beta(\bar{r}_1)\tau_2 + \varepsilon\tau_2^2 \leq (-\frac{1}{24}\eta(p_2) - \varepsilon)\tau_2^2 < 0.$$

Therefore (5.20) follows from (5.21). In case (a) we have

$$\frac{1}{6}\eta(p_2)\tau_2 \geq -\beta(\bar{r}_1) \geq \tilde{\delta}(p_2)\tau_3 + \frac{1}{3}\tilde{\eta}(p_2)\tau_2$$

since $\beta(p_3) \leq 0$. Thus $\tau_2 \leq C\tau_3$. Therefore the $\tau_2^2 \cdot O(1)$ terms can be compensated through $\gamma\tau_3^2$ on the left-hand side. As above, the lemma therefore follows from (5.21) if $\beta(\bar{r}_1) \leq 0$. So we assume $\beta(\bar{r}_1) > 0$. Now we need to bound $\beta(\bar{r}_1)\tau_2$ above. We use the analogue of (5.9) for τ_3 and τ_4 and $\alpha(p_3)\tau_3 \geq \frac{7}{12}\tilde{\gamma}(p_3)\tau_3$, which follows from (4.4), to bound

$$-\beta(p_3)\tau_4 \geq \frac{1}{4}\tilde{\gamma}(p_3)\tau_3^2 + \frac{1}{3}\tilde{\eta}(p_3)\tau_4^2 = \frac{1}{4}\tilde{\gamma}(p_2)\tau_3^2 + \frac{1}{3}\tilde{\eta}(p_2)\tau_4^2.$$

Using $\tau_4 \leq \frac{3}{4}\tau_2$ we therefore get

$$\begin{aligned} \tilde{\beta}(\bar{r}_1)\tau_2 &= \tilde{\beta}(p_3)\tau_2 - \tilde{\delta}(p_3)\tau_2\tau_3 - \frac{1}{3}\tilde{\eta}(p_2)\tau_2^2 \\ &\leq -\frac{1}{3}\tilde{\gamma}(p_2)\tau_3^2 - \tilde{\delta}(p_2)\tau_2\tau_3 - \frac{1}{3}\tilde{\eta}(p_2)\tau_2^2 - \frac{4}{9}\tilde{\eta}(p_2)\tau_4^2. \end{aligned}$$

We substitute this into (5.21) and compensate $\varepsilon\tau_2^2$ through $\gamma\tau_3^2$ to obtain

$$\begin{aligned} -(\frac{1}{3}\tilde{\eta}(p_2)\tau_2 + \tilde{\delta}(p_2)\tau_3 + \frac{1}{3}\tilde{\eta}(p_2)\tau_4)\tau_4 &\geq \frac{1}{6}\tilde{\gamma}(p_2)\tau_3^2 - \beta(\bar{r}_1)\tau_2 \\ &\geq \frac{1}{2}\tilde{\gamma}(p_2)\tau_3^2 + \tilde{\delta}(p_2)\tau_2\tau_3 + \frac{1}{3}\tilde{\eta}(p_2)\tau_2^2 + \frac{4}{9}\tilde{\eta}(p_2)\tau_4^2 \\ &\geq \frac{1}{12}\tilde{\gamma}(p_2)\tau_3^2 + \frac{4}{9}\tilde{\eta}(p_2)\tau_4^2 - \varepsilon\tau_2^2 \end{aligned}$$

where the last inequality follows from (5.18). Again we can compensate $-\varepsilon\tau_2^2$ through $\gamma\tau_3^2$ and so the right-hand side is positive. This proves the lemma. \square

Relations (5.19) and (5.20) imply as necessary condition near p that $\tau_4 \leq \varepsilon\tau_3$. This contradicts the fact that τ_3 and τ_4 are comparable. Q.E.D.

The proofs for the remaining generic cases—codimension 3 when $\alpha(p)$ and $\beta(p)$ vanish—are longer and much more technical. Therefore we do not even attempt to give an outline here. Complete and detailed proofs can be found in the author's doctoral dissertation [11].

REFERENCES

- [1] H. J. BAUES, *Commutator calculus and groups of homotopy classes*, London Mathematical Society, Lecture Note Series, Vol. 50, 1981.
- [2] V. G. BOLTYANSKY, *Sufficient conditions for optimality and the justification of the dynamic programming method*, this Journal, 4 (1966), pp. 326–361.
- [3] B. BONNARD, *On singular extremals in the time-minimal control problem in \mathbb{R}^3* , this Journal, 23 (1985), pp. 794–802.
- [4] A. BRESSAN, *The generic local time-optimal stabilizing controls in dimension 3*, this Journal, 24 (1986), pp. 177–190.
- [5] P. BRUNOVSKY, *Existence of regular synthesis for general control problems*, J. Differential Equations, 38 (1980), pp. 317–343.
- [6] A. T. FULLER, *Study of an optimum nonlinear control system*, J. Electronics Control, 15 (1963), pp. 63–71.
- [7] A. J. KRENER, *The higher-order maximum principle and its application to singular extremals*, this Journal, 15 (1977), pp. 256–293.
- [8] I. KUPKA, *The ubiquity of the Fuller phenomenon*, preprint, Institut Fourier, Grenoble, France, to appear.
- [9] PONTRYAGIN et al., *Mathematical Theory of Optimal Processes*, Wiley Interscience, New York, 1962.
- [10] H. SCHÄTTLER, *On the local structure of time-optimal controls in \mathbb{R}^3* , in Proc. 24th IEEE Conference on Decision and Control, Ft. Lauderdale, FL, 1985, pp. 714–720.
- [11] ———, *On the local structure of time-optimal trajectories for a single-input control-linear system in dimension 3*, Ph.D. thesis, Rutgers University, New Brunswick, NJ, October 1986.
- [12] H. J. SUSSMANN, *Time optimal control in the plane*, in Feedback Control of Linear and Nonlinear Systems, LN in Control and Information Sciences, Vol. 39, Springer-Verlag, Berlin, 1982, pp. 244–260.
- [13] ———, *Lie brackets, real analyticity and geometric control*, in Differential Geometric Control Theory, R. Brockett, R. Millman and H. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 1–116.
- [14] ———, *Lie brackets and real analyticity in control theory*, in Mathematical Control Theory, Banach Center Publications, Vol. 14, Warsaw, 1985, pp. 515–542.
- [15] ———, *Envelopes, conjugate points, and optimal bang-bang extremals*, in Proc. 1985 Paris Conference on Nonlinear Systems, M. Fliess and M. Hazewinkel, eds., Reidel Publishing Company, Dordrecht, The Netherlands, to appear.
- [16] ———, *A product expansion for the Chen series*, to appear.
- [17] ———, *The structure of time-optimal trajectories for single-input systems in the plane: the C^∞ nonsingular case*, this Journal, 25 (1987), pp. 433–465.
- [18] ———, *The structure of time-optimal trajectories for single-input systems in the plane: the general real-analytic case*, this Journal, 25 (1987), pp. 868–904.
- [19] ———, *Regular synthesis for time-optimal control of single-input real-analytic systems in the plane*, this Journal, 25 (1987), pp. 1145–1162.

A MAXIMUM PRINCIPLE FOR OPTIMAL PROCESSES WITH DISCONTINUOUS TRAJECTORIES*

R. B. VINTER[†] AND F. M. F. L. PEREIRA[‡]

Abstract. A Maximum Principle is proved which governs solutions to dynamic optimization problems in which the controls driving the system may be impulsive and give rise to discontinuous trajectories. The approach, which involves approximating the problem by a conventional one and using Ekeland's theorem, is new. It permits us to weaken very considerably the hypotheses under which Maximum Principles for such problems have previously been proved.

Key words. optimal control, impulse control, Maximum Principle

AMS(MOS) subject classifications. 49B10, 49B34

1. Introduction. We give necessary conditions for optimality, in the form of a Maximum Principle, for an optimal control problem (we label it (P)) in which the state trajectories are permitted to be functions of bounded variation and the control policies incorporate measures:

$$\text{Minimize} \quad h(x(1)) + \int_0^1 f_0(t, x(t), u(t)) \, dt + \int_{[0,1]} g_0(t, u(t)) \mu(dt)$$

$$\text{subject to} \quad dx(t) = f(t, x(t), u(t)) \, dt + G(t, u(t)) \mu(dt) \quad \text{on } [0, 1],$$

$$x(0) \in C_0, \quad x(1) \in C_1,$$

$$u(t) \in \Omega_t \quad \mu \text{ and } \mathcal{L}\text{-a.e.}$$

and

$$\text{Range } \{\mu\} \subset K.$$

Here,

$$f_0: [0, 1] \times R^n \times R^m \rightarrow R, \quad g_0: [0, 1] \times R^m \rightarrow R^k,$$

$$f: [0, 1] \times R^n \times R^m \rightarrow R^n, \quad G: [0, 1] \times R^m \rightarrow R^{n \times k}$$

are given functions,

C_0, C_1 are closed subsets of R^n ,

Ω is a subset of $[0, 1] \times R^m$ (Ω_t denotes the section $\{u: (u, t) \in \Omega\}$), and

K is a closed convex cone in R^k .

To be more precise, we define a pair of elements (u, μ) to be a control policy if μ is a K valued regular measure on \mathcal{B} (\mathcal{B} denotes the Borel subsets of $[0, 1]$), $u(\cdot): [0, 1] \rightarrow R^m$ is a Borel measurable function such that

$$\mu(A) \in K \quad \text{for all } A \in \mathcal{B},$$

$$u(t) \in \Omega_t,$$

almost everywhere with respect to both μ and Lebesgue measure, and if the function $t \rightarrow G(t, u(t))$ has a μ integrability property made precise in § 2.

* Received by the editors September 16, 1986; accepted for publication March 12, 1987.

[†] Electrical Engineering Department, Imperial College, London, England SW7 2AZ.

[‡] Departamento de Engenharia Electrotecnica da Universidade do Porto, Rua dos Bragas, 4099 Porto, Portugal.

A trajectory, $x: [0, 1] \rightarrow R^n$ (associated with a control policy (u, μ)), is a function of bounded variation such that

$$x(t) = x(0) + \int_0^t f(s, x(s), u(s)) ds + \int_{[0, t]} G(s, u(s)) \mu(ds)$$

for all $t \in (0, 1]$. A trajectory and a control policy with which it is associated, (x, u, μ) , is called a control process.

Problem (P) then is that of minimizing the functional

$$h + \int_0^1 f_0 dt + \int_{[0, 1]} g_0 \mu(dt)$$

over control processes for which $t \rightarrow g_0(t, u(t))$ is μ integrable and which satisfy the endpoint constraints $x(0) \in C_0$ and $x(1) \in C_1$. A minimizing control process is called an optimal process. The components are called "optimal trajectory," etc.

This problem is an extension of a control problem of a conventional kind, denoted (P'), where some components w of the control vector $\text{col}(u, w)$ (i.e., $[u^T \mid w^T]^T$) enter linearly into the cost and dynamics and the values of w are unbounded in certain directions:

$$\text{Minimize } h(x(1), x(0)) + \int_0^1 (f_0(t, x(t), u(t)) + g_0(t, u(t))w(t)) dt$$

$$\text{subject to } \dot{x}(t) = f(t, x(t), u(t)) + G(t, u(t))w(t) \quad \text{a.e. } t \in [0, 1],$$

$$(u(t), w(t)) \in \Omega_t \times K \quad \mathcal{L}\text{-a.e.},$$

$$x(0) \in C_0, \quad x(1) \in C_1.$$

Certain problems arising in midcourse guidance of space vehicles [6], [7], [9], [11], and also in resource economics [2], can be formulated in this way. A noteworthy feature of such problems is that minimizing sequences of trajectories can be expected to have accumulation point functions which are of bounded variation, but not absolutely continuous as in the standard theory. The framework adopted in this paper is one in which the usual class of admissible trajectories is enlarged to include those of bounded variation. Indeed we can view the processes for (P') as a subclass of the processes for (P), via the embedding

$$(x, u, w) \rightarrow (x, u, \mu)$$

where μ is the measure

$$\mu(A) = \int_A w(s) ds.$$

Extra hypotheses can be given under which (P) is a proper extension to (P') in the sense that the existence of solutions to (P) is guaranteed and the infimum cost for (P') coincides with the infimum cost for (P). Results of this nature, but developed in a somewhat different framework, are to be found in [8], [12] and [13]. We do not impose such hypotheses here, but rather strive to provide the weakest hypotheses our methods permit under which necessary conditions can be formulated and proved. This is not to deny the desirability of having a proper extension. Necessary conditions, valid under weak hypotheses which do not guarantee the extension considered is proper, are valuable for several reasons however. Special features of the optimal control problem at hand might permit one to deduce the extension is proper even though

hypotheses of available theorems for properness, directed at a broad class of problems, are not satisfied. Necessary conditions also have a bearing on the question of regularity of solutions: if the necessary conditions lead to a contradiction, then no solutions exist with regularity for the minimizing process to which the necessary conditions apply.

Optimality conditions for problem (P) have already been supplied by Rishel [11]. We now briefly comment on the techniques used by Rishel, in order to call attention to the advantages of the novel methods of this paper.

Rishel's approach is to introduce a new independent variable, with respect to which trajectories become absolutely continuous. This leads to consideration of an auxiliary optimal control problem in which the time variable now has the role of a component of the state variable. The auxiliary problem has the character of a conventional free-time optimal control problem. Under a so-called "constancy" hypothesis, we can associate with an optimal process for the original problem an optimal process for the auxiliary problem. (In the case that the measures are scalar valued and $K = R^+$, a simple version of the constancy hypothesis is

$$(1.1) \quad \overline{\text{co}} \tilde{g}(t, \Omega_t) \subset \bigcup_{0 \leq \alpha \leq 1} \alpha \tilde{g}(t, \Omega_t) \quad \text{for all } t \in [t_0, t_1].$$

Here $\tilde{g} := \text{col}(g_0, g_1)$.)

This permits us to apply the standard Maximum Principle to the associated process and thence to deduce optimality conditions for the original optimal process.

Now, in order to follow through this programme, we require, in addition to other hypotheses,

$$(1.2) \quad \Omega_t = \tilde{\Omega} \quad \text{for all } t \in [t_0, t_1] \quad (\text{where } \tilde{\Omega} \text{ is fixed})$$

and

$$(1.3) \quad f, G, f_0, g_0 \quad \text{are regular in their } t\text{-dependence.}$$

(By "regular" we mean here "at least Lipschitz continuous.") These arise because time becomes a component of the state variable in the auxiliary problem and, except in special circumstances, the conventional Maximum Principle, which we need to apply to the auxiliary problem, is valid only when the data are regular in the state variable and the control constraint set does not depend on the state.

By contrast with what is achievable by Rishel's approach, we prove a Maximum Principle for (P) in which the data are merely required to be measurable in t , time dependency of the control constraint is permitted and the constancy hypothesis (1.1) is dispensed with altogether. The character of our optimality conditions is new too, since Rishel's Maximum Principle involves time derivatives of the data and consequently does not even make sense when the data are not regular in t .

Quite different methods are required then for the proof of the results of the present paper; our approach is to *approximate* the extended problem by a conventional optimal control problem in a manner which does not involve change of the independent variable, and to use limiting arguments based on Ekeland's theorem.

Warga has also provided necessary conditions for optimal control problems with discontinuous trajectories [14], [15]. Warga's approach is essentially to view Rishel's auxiliary problem as a relaxed version of the original optimal control problem and to give necessary conditions on optimal processes for the relaxed problem. Clearly there is no need here for a constancy hypothesis, since we are no longer concerned to translate optimal relaxed processes back into processes for the original problem. Warga then addresses a different problem, and his conditions cannot be compared directly

with ours. Note, however, that Warga invokes hypotheses (1.2) and (1.3) since time in the auxiliary problem is a component of the state.

We mention too that optimality conditions for problems with discontinuous trajectories, when the data is required to be merely measurable in t , are given by Rockafellar [12], [13] (in addition to duality interpretations of the multipliers arising and hypotheses for properness). Rockafellar, however, limits attention to generalized Bolza problems having Lagrangians jointly convex in the (x, \dot{x}) variables, a restriction which is not present here.

Finally, we mention that the results of this paper can be extended in a number of directions, notably to accommodate an affine state inequality constraint of the form

$$A(t)x(t) + b(t) \leq 0 \quad \text{for all } t \in [0, 1].$$

Some refinements appear in [10].

2. Preliminaries. We shall denote the Borel subsets of R^k by \mathcal{B}^k , the Borel subsets of $[0, 1]$ by \mathcal{B} and the Lebesgue subsets of $[0, 1]$ by \mathcal{L} . The symbol \mathcal{L} also refers to Lebesgue measure, as when we write \mathcal{L} -a.e. (meaning “almost everywhere with respect to the Lebesgue measure”), \mathcal{L} -meas (A) (meaning the “Lebesgue measure of A ”), etc.

We refer to signed, regular measures (on some extension of the Borel subsets of $[0, 1]$) briefly as measures. $C^*(R^k)$ denotes the usual linear space comprising k tuples of measures. We write C^* for $C^*(R^1)$. C^\oplus is the class of nonnegative scalar valued measures.

All norms are written $\|\cdot\|$. In R^k the norm is the Euclidean norm. In $C^*(R^k)$ the norm is the sum of the total variations of the component measures.

We denote by B the unit ball in Euclidean space.

As is well known, there is a one-to-one correspondence between elements in $C^*(R^k)$ and k -vector valued functions of bounded variation on $[0, 1]$ which are right continuous on $[0, 1]$. To indicate that a function x is to be associated with a measure $\mu \in C^*(R^k)$ we write

$$dx(t) = \mu(dt).$$

This may be viewed as a shorthand for

$$\int_0^s dx(t) = \int_{[0,s]} \mu(dt) \quad \text{for all } s \in [0, 1]$$

in which the first integral is a Stieltjes integral.

Given $\mu \in C^*(R^k)$, $|\mu|$, the “total variation measure” denotes the element in C^\oplus :

$$|\mu| = \sum_i \mu_i^+ + \sum_i \mu_i^-$$

where $\mu_i = \mu_i^+ - \mu_i^-$ is the Jordan decomposition of the i th component μ_i of μ .

The notation

$$\mu_j \xrightarrow{*} \mu$$

indicates weak* convergence. If the measures are scalar valued this means

$$\int_{[0,1]} h(t) \mu_j(dt) \rightarrow \int_{[0,1]} h(t) \mu(dt) \quad \text{as } j \rightarrow \infty$$

for every continuous function h on $[0, 1]$. When the measures are vector valued we require weak* convergence of the components as just defined.

It is convenient at this stage to comment on our interpretation of integrals of the form

$$\int_A g(t) \cdot \mu(dt) \quad \text{for } A \in \mathcal{B}$$

in which g and μ are both vector valued.

It would seem natural to take it to mean $\sum_i \int_A g_i(t) \mu_i(dt)$ (where g_i and μ_i are components) and to require g_i to be μ_i -integrable for $i=1, \dots, k$. However, transformations we employ in the proof of Theorem 4.1 compel us to adopt a slightly more general definition. Our approach here is to note that the components of μ are absolutely continuous with respect to the total variation measure $|\mu|$. Consequently we can define the Radon-Nikodym derivative w of μ with respect to $|\mu|$. We summarize the relationships involved by writing

$$\mu(dt) = w(t)|\mu|(dt).$$

The k vector valued function g will be said to be μ integrable if $t \rightarrow |g(t) \cdot w(t)|$ is integrable with respect to $|\mu|$. If g is μ integrable we define $\int_A g(t) \cdot \mu(dt)$ to be

$$\int_A g(t) \cdot \mu(dt) := \int_A g(t) \cdot w(t) |\mu|(dt),$$

where the right-hand side is the usual integral. (The extra generality arises because the earlier definition requires each $g_i w_i$ to be $|\mu|$ integrable, whereas (2.1) permits components which are not integrable to cancel out in the inner product $g \cdot w$.) This definition should be borne in mind when, for example, we interpret components of the term

$$\int_{[0,1]} G(t, u(t)) \mu(dt)$$

in the state equation.

We shall use the concept of continuity set. Take $\mu \in C^\oplus$. A Borel set $A \subset [0, 1]$ is a μ continuity set [1] if $\mu(\partial A) = 0$, where ∂A is the relative boundary of A .

Given a locally Lipschitz continuous function $\phi: R^p \rightarrow R^q$, $\partial\phi$ is the generalized Jacobian (or generalized gradient in the case $q=1$), as defined by Clarke [4, § 2.6]. If $C \subset R^k$ is a given closed subset and $s \in C$ then $N_C(s)$ is the normal cone to C at s in the sense of Clarke [4, p. 11]. We denote by d_C the Euclidean distance function from the set C .

Let $D \subset R^k$ be a given subset. Then σ_D is the support function of the set D :

$$\sigma_D(e) = \max \{e \cdot d : d \in D\}.$$

3. Hypotheses. We state here the hypotheses on the data invoked in the Maximum Principle to follow. The function x^* referred to in hypothesis H4 will be the trajectory under consideration (either an optimal trajectory or one associated with a boundary point of the reachable set).

H1 Given $\delta > 0$, there exists $K_\delta \in L^1$ such that

$$\|f(t, x, u) - f(t, y, u)\| \leq K_\delta(t) \|x - y\|$$

when $\|x\|, \|y\| \leq \delta$, $u \in \Omega_t$ and $t \in [0, 1]$.

H2 The function $(t, u) \rightarrow f(t, x, u)$ is $\mathcal{L} \times \mathcal{B}^m$ measurable for each x (where $\mathcal{L} \times \mathcal{B}^m$ denotes the product σ -algebra of \mathcal{L} and \mathcal{B}^m).

H3 The subset $\Omega \subset R^{m+1}$ is Borel measurable.

H4 There exists $\alpha \in L^1$ such that

$$\sup \{ \|f(t, x^*(t), u)\| : u \in \Omega_t \} \leq \alpha(t) \quad \mathcal{L}\text{-a.e.}$$

H5 $h(\cdot)$ is locally Lipschitz continuous.

H6 The multifunction Γ on $[0, 1]$, defined by

$$\Gamma(t) = \overline{\text{co}} \left\{ \frac{G(t, u)w}{1 + \sum_i |\sum_j g_{ij}(t, u)w_j|} : u \in \Omega_t, w \in K \cap \{\xi : |\xi_i| \leq 1\} \right\}$$

is continuous with respect to the Hausdorff metric on the compact, convex subsets of R^n . ($\overline{\text{co}}$ denotes closed convex hull.)

Hypotheses H1–H5 are slightly strengthened forms of standard hypotheses under which the Maximum Principle has been derived in the absence of singular terms (see, e.g., [4, Thm. 5.2.1]). The most notable difference is that hypothesis H4, a uniform integrability condition on the dynamics at the optimal trajectory, is added here. We can drop hypothesis H4 in many cases of interest and, in particular, when the singular and nonsingular terms in the dynamics separate, i.e., when the control variable u comprises two vector components $u = \text{col}(v, w)$,

$$\Omega_t = \{\text{col}(v, w) : v \in V_t, w \in W_t\}$$

and

$$f(t, x, (v, w)) = \tilde{f}(t, x, v), \quad g(t, (v, w)) = \tilde{g}(t, w)$$

for sets V and W and functions \tilde{f} and \tilde{g} . (An argument along the lines of [4, p. 207] is involved.) Of course, this last case subsumes the conventional control problem (no singular term).

The remaining hypothesis, H6, is a kind of continuity condition on a compactification of the singular term in the dynamics. It is automatically satisfied by autonomous systems and also by problems where the multifunction $t \rightarrow G(t, \Omega_t)$ is uniformly bounded on $[0, 1]$ and continuous in the sense of Kuratowski.

4. The main results. Necessary conditions are first given for processes associated with boundary points of a reachable set. The new Maximum Principle for problem (P) will follow directly from these.

Let $\Psi : R^n \rightarrow R^q$ be a locally Lipschitz continuous function. We define the Ψ reachable set, R_Ψ , to be

$$R_\Psi = \{\Psi(x(1)) : (x, u, \mu) \text{ is some control process and } x(0) \in C_0\}.$$

A control process (x, u, μ) such that $x(0) \in c_0$ and $\Psi(x(1))$ is a boundary point of R_Ψ is called a Ψ boundary process.

We denote by H the unmaximized Hamiltonian function:

$$H(t, x, u, p) := p \cdot f(t, x, u).$$

THEOREM 4.1. *Let the data satisfy hypotheses H1–H6, and let (x^*, u^*, μ^*) be a Ψ boundary process.*

Then there exists an absolutely continuous function p and a nonzero vector $d \in R^k$ such that

$$\begin{aligned}
 & -\dot{p}(t) \in \partial_x H(t, x^*(t), u^*(t), p(t)) \quad \mathcal{L}\text{-a.e.}, \\
 & -p(1) \in d \cdot \partial \Psi(x^*(1)), \\
 & p(0) \in N_{C_0}(x^*(0)), \\
 & H(t, x^*(t), u^*(t), p(t)) = \max \{H(t, x^*(t), u, p(t)) : u \in \Omega_t\} \quad \mathcal{L}\text{-a.e.}, \\
 (4.1) \quad & \sup \{\sigma_K(p(t) \cdot G(t, u)) : u \in \Omega_t\} \leq 0 \quad \text{all } t \in [0, 1]
 \end{aligned}$$

and

$$(4.2) \quad \sigma_K(p(t) \cdot G(t, u^*(t))) = 0 \quad \mu^*\text{-a.e.}$$

Here σ_K denotes the support function of the set K defined in § 2 and ∂_x is the generalized gradient in the state variable.

For purposes of comparison with some of the earlier literature (notably [11]) and because later we must refer to this special case, we note that the two conditions (4.1) and (4.2), which give information about the measure μ in Theorem 4.1, take the following form when μ is scalar valued and $K = [0, \infty)$. In this case, (4.1) and (4.2) become

$$\sup_{u \in \Omega_t} \{p(t) \cdot G(t, u)\} \leq 0 \quad \text{for all } t \in [0, 1]$$

and

$$p(t) \cdot G(t, u^*(t)) = 0, \quad \mu^*\text{-a.e.}$$

respectively.

And now a Maximum Principle for problem (P). Here we take \tilde{H} to be the unmaximized Hamiltonian function arising in this problem:

$$\tilde{H}(t, x, u, p, \lambda) := p \cdot f(t, x, u) - \lambda f_0(t, x, u).$$

THEOREM 4.2. *Let hypotheses H1-H6 be satisfied when the functions $\text{col}(f_0, f)$ and $\text{col}(g_0, G)$ are inserted into the conditions in place of f and G . Suppose that (x^*, u^*, μ^*) is an optimal process.*

Then there exists an absolutely continuous function p and a nonnegative number λ , $\|p\|_{L_\infty} + \lambda \neq 0$, such that

$$\begin{aligned}
 & -\dot{p}(t) \in \partial_x \tilde{H}(t, x^*(t), u^*(t), p(t), \lambda) \quad \mathcal{L}\text{-a.e.}, \\
 & -p(1) \in \lambda \partial h(x^*(1)) + N_{C_1}(x^*(1)), \\
 & p(0) \in N_{C_0}(x^*(0)), \\
 & \tilde{H}(t, x^*(t), u^*(t), p(t), \lambda) = \max_{u \in \Omega_t} \{\tilde{H}(t, x^*(t), u, p(t), \lambda)\} \quad \mathcal{L}\text{-a.e.}, \\
 (4.3) \quad & \sup \{\sigma_K(p(t) \cdot G(t, u) - \lambda g_0(t, u)) : u \in \Omega_t\} \leq 0 \quad \text{for all } t \in [0, 1]
 \end{aligned}$$

and

$$(4.4) \quad \sigma_K(p(t) \cdot G(t, u^*(t)) - \lambda g_0(t, u^*(t))) = 0 \quad \mu^*\text{-a.e.}$$

We point out that, in the special case earlier alluded to ($K = [0, \infty)$), conditions (4.3) and (4.4) reduce to

$$(4.5) \quad \sup_{u \in \Omega_t} \{p(t) \cdot G(t, u) - \lambda g_0(t, u)\} \leq 0 \quad \text{for all } t \in [0, 1]$$

and

$$(4.6) \quad p(t) \cdot G(t, u^*(t)) - \lambda g_0(t, u^*(t)) = 0 \quad \mu^* \text{-a.e.}$$

We now prove, as claimed, that Theorem 4.2 is a simple consequence of Theorem 4.1.

Let (x^*, u^*, μ^*) be an optimal process for (P). Consider a new control system (S) in which the state vector $z = \text{col}(y, w, x)$ now comprises a scalar component y and n vector blocks w and x :

$$dy(t) = f_0(t, x(t), u(t)) dt + g_0(t, u(t)) \mu(dt),$$

$$dx(t) = f(t, x(t), u(t)) dt + G(t, u(t)) \mu(dt),$$

$$\frac{dw(t)}{dt} = 0,$$

$$(y(0), w(0), x(0)) \in [0, \infty) \times C_1 \times C_0.$$

Define the mapping $\tilde{\Psi}: R^{2n+1} \rightarrow R^{n+1}$ to be

$$\tilde{\Psi}(z) = \begin{bmatrix} y + h(x) \\ w - x \end{bmatrix}$$

and let $\tilde{R}_{\tilde{\Psi}}$ denote the $\tilde{\Psi}$ reachable set for (S).

Let us examine the control process π :

$$\pi := (z^* = (y^*, w^*, x^*), u^*, \mu^*)$$

for (S) in which (x^*, u^*, μ^*) is the optimal process for (P) under consideration, $w^*(\cdot) \equiv x^*(1)$ and

$$y^*(t) = \int_0^t f_0(s, x^*(s), u^*(s)) ds + \int_{[0,t]} g_0(s, u^*(s)) \mu^*(ds).$$

It is easy to see that $h(z^*(1))$ lies in the boundary of $\tilde{R}_{\tilde{\Psi}}$.

Now apply Theorem 4.1 to π . (The hypotheses hold under which this is permissible.) It is a straightforward task to show that the assertions of Theorem 4.2 follow.

What does Theorem 4.2 tell us about the minimizing measure μ^* ? We see that (4.3) and (4.4) (or (4.5) and (4.6)) locate the support of μ^* . At first sight this information might appear rather meagre. But the fact that the optimal trajectory must satisfy the terminal constraints, and the conditions on the conventional control, embody additional implicit information about μ^* . In fact it is a simple matter to show that for linear convex problems, under a normality hypothesis, the conditions of Theorem 4.2 are also sufficient for optimality of (x^*, u^*, μ^*) . This attests to the strength of the necessary conditions.

5. Approximation of measures. Approximation arguments underlie the proof of Theorem 4.1, in which the measures driving the system equation are replaced by control functions (in the customary sense).

Clearly we need to devise schemes for approximation of measures and to examine the effects of approximation on the state trajectories. Various results required for these purposes are gathered together in this section. Proofs appear in the Appendix.

PROPOSITION 5.1. Consider sequences $\Phi_i: [0, 1] \times R^n \rightarrow R^n$, $\{\gamma_i\}$ and $\{a_i\}$ with $\gamma_i \in C^*(0, 1; R^n)$ and $a_i \in R^n$, $i = 1, 2, \dots$.

Suppose that

$$(5.1) \quad \gamma_i \xrightarrow{*} \gamma_0,$$

$$(5.2) \quad a_i \rightarrow a_0$$

and

$$(5.3) \quad \mathcal{L}\text{-meas}(\{t: \Phi_i(t, x) = \Phi_0(t, x), \text{ all } x \in R^n\}) \rightarrow 1$$

for some point $\gamma_0 \in C^*$, point $a_0 \in R^n$ and function $\Phi_0: [0, 1] \times R^n \rightarrow R^n$.

Let $z_0: [0, 1] \rightarrow R^n$ be a function of bounded variation such that

$$(5.4) \quad \begin{aligned} z_0(t) &= z_0(0) + \int_0^t \Phi_0(s, z_0(s)) \, ds + \int_{[0, t]} \gamma_0(ds) \quad \text{for all } t \in (0, 1], \\ z_0(0) &= a_0. \end{aligned}$$

It is assumed that, given any $\delta > 0$, there exist $\bar{K}_\delta, \bar{\alpha} \in L^1$ such that

$$(5.5) \quad \Phi_i(\cdot, x) \text{ is Lebesgue measurable for each } x \in R^n,$$

$$(5.6) \quad \|\Phi_i(t, x) - \Phi_i(t, y)\| \leq \bar{K}_\delta(t) \|x - y\| \quad \text{when } \|x\|, \|y\| \leq \delta \quad \mathcal{L}\text{-a.e. } t \in [0, 1]$$

and

$$(5.7) \quad \|\Phi_i(t, z_0(t))\| \leq \bar{\alpha}(t) \quad \mathcal{L}\text{-a.e. } t \in [0, 1]$$

for $i = 1, 2, \dots$.

Then, for each i sufficiently large, there exists a function z_i of bounded variation such that

$$(5.8) \quad \begin{aligned} z_i(t) &= z_i(0) + \int_0^t \Phi_i(s, z_i(s)) \, ds + \int_{[0, t]} \gamma_i(ds) \quad \text{for all } t \in (0, 1], \\ z_i(0) &= a_i, \end{aligned}$$

$$(5.9) \quad z_i(t) - \int_{[0, t]} \gamma_i(ds) \rightarrow z_0(t) - \int_{[0, t]} \gamma_0(ds) \quad \text{uniformly in } t \in [0, 1]$$

and

$$(5.10) \quad dz_i \xrightarrow{*} dz_0.$$

We remark that limit (5.10) is to be interpreted in terms of weak* convergence of elements in $C^*(0, 1; R^n)$ associated with the functions of bounded variation z_i , $i = 1, 2, \dots$.

PROPOSITION 5.2. Take $\nu \in C^*(0, 1; R^k)$ and $\{\nu_i\}$ with $\nu_i \in C^*(0, 1; R^k)$, $i = 1, \dots$.

(a) If there exists a Borel measurable subset $S \subset [0, 1]$, having full Lebesgue measure and containing the point $\{1\}$, if the ν_i 's are uniformly bounded in total variation and

$$\int_{[0, t]} \nu_i(ds) \rightarrow \int_{[0, t]} \nu(ds)$$

as $i \rightarrow \infty$, for all $t \in S$, then

$$(5.11) \quad \nu_i \xrightarrow{*} \nu.$$

Conversely,

(b) If (5.11) is true, there exists a subset $S \subset [0, 1]$ which contains $\{1\}$ and is the complement of a countable set and also a subsequence $\{\nu_{i_j}\}$ of $\{\nu_i\}$ such that

$$\int_{[0,t]} \nu_{i_j}(ds) \rightarrow \int_{[0,t]} \nu(ds) \quad \text{for all } t \in S.$$

PROPOSITION 5.3. Let $U \subset [0, 1] \times R^m$ be a Borel measurable subset and let $r: [0, 1] \times R^m \rightarrow R^n$ be a Borel measurable function.

We assume that r is uniformly bounded on U and the multifunction $t \rightarrow \overline{\text{co}} r(t, U_t)$ is continuous with respect to the Hausdorff metric.

Take $\mu \in C^\oplus(0, 1)$ and a Borel measurable function u such that

$$u(t) \in U_t \quad \mathcal{L} \text{ and } \mu\text{-a.e.}$$

Then there exists a sequence of Borel measurable functions $\{u_i: [0, 1] \rightarrow R^m\}$ and a sequence $\{m_i\}$ in $L^1(0, 1; R)$ such that

$$u_i(t) \in U_t \quad \mathcal{L}\text{-a.e.},$$

$$m_i(t) \geq 0 \quad \mathcal{L}\text{-a.e.},$$

$$t \rightarrow r(t, u_i(t))m_i(t) \quad \text{is } \mathcal{L} \text{ integrable}$$

for $i = 1, 2, \dots$,

$$m_i(t) dt \xrightarrow{*} \mu(dt)$$

$$r(t, u_i(t))m_i(t) dt \xrightarrow{*} r(t, u(t))\mu(dt)$$

and

$$\mathcal{L}\text{-meas}(\{t: u(t) \neq u_i(t)\}) \rightarrow 0 \quad \text{as } i \rightarrow \infty.$$

6. Proof of Theorem 4.1. Let us show that, in proving Theorem 4.1, it is permissible to make a few simplifying assumptions at the outset.

H7 $K = R^+$,

G is uniformly bounded on Ω and the multifunction $t \rightarrow \overline{\text{co}} G(t, \Omega_t)$ is continuous with respect to the Hausdorff metric.

LEMMA 6.1. If the assertions of Theorem 4.1 are true under hypotheses H1–H7, then they are true merely under hypotheses H1–H6.

Proof. Suppose hypotheses H1–H7 are in force and (x^*, u^*, μ^*) is a control process such that $x^*(0) \in C_0$ and $\Psi(x^*(1))$ is a boundary point of R_Ψ .

We construct a new dynamic system (S') from the former data as follows:

$$dx(t) = f(t, x(t), u(t)) dt + \tilde{g}(t, u(t), w(t))\nu(dt),$$

$$(u(t), w(t)) \in \Omega_t \times \tilde{K} \quad \mathcal{L} \text{ and } \mu\text{-a.e.},$$

$$x(0) \in C_0,$$

$$\nu(A) \geq 0 \quad \text{for all } A \in \mathcal{B}.$$

Here,

$$\tilde{g}(t, u, w) := \frac{G(t, u)w}{1 + \sum_i |\sum_j g_{ij}(t, u)w_j|},$$

$$\tilde{K} := \{w \in K: \sum_i |w_i| \leq 1\},$$

and u and w are regarded as block components of a composite control vector $\text{col}(u, w)$. The g_{ij} 's are the components of G .

We define w^* to be the Radon-Nikodym derivative of μ^* with respect to $|\mu^*|$. A simple separating hyperplane argument reveals that

$$w^*(t) \in \{v \in K : |v_i| \leq 1, i = 1, 2, \dots, k\} \quad |\mu^*| \text{-a.e.}$$

Now define $\nu^* \in C^\oplus(0, 1)$ according to

$$\nu^*(dt) = (1 + \sum_i |\sum_j g_{ij}(t, u^*(t)) w_j^*(t)|) |\mu^*(dt)|.$$

Note that $1 + \sum_i |\sum_j g_{ij} w_j^*|$ is $|\mu^*|$ integrable in view of the definition of μ integrability that we have adopted; consequently ν^* is well defined.

We claim that $(x^*, \text{col}(u^*, w^*), \nu^*)$ is a Ψ boundary process for (S') .

To prove this, it suffices to show that the set of trajectories is the same for (S) and (S') . This in turn will follow provided we can always arrange that the singular terms in the two extended differential equations, namely $G(t, u) d\mu$ and $\tilde{g}(t, u) d\nu$, are interchangeable in the following sense.

Let u be any given function. We require that

(i) given any $\mu \in C^*$ such that (u, μ) is a control policy for (S) , there are functions $w: [0, 1] \rightarrow R^k$ and $\nu \in C^\oplus$ such that $(\text{col}(u, w), \nu)$ is a control policy for (S') and

$$(6.1) \quad \int_A G(t, u(t)) \mu(dt) = \int_A \tilde{g}(t, u(t), w(t)) \nu(dt) \quad \text{for all } A \in \mathcal{B},$$

and also

(ii) given any function $w: [0, 1] \rightarrow R^k$ and $\nu \in C^\oplus$ such that $(\text{col}(u, w), \nu)$ is a control policy for (S') , then there exists $\mu \in C^*$ such that (u, μ) is a control policy for (S) , and (6.1) is satisfied.

Full details of the proofs of assertions (i) and (ii) are routine and are therefore omitted; we point out here merely that, in proving (i), we can take the function w to be the Radon-Nikodym derivative $d\mu/d|\mu|$ and ν to be that element in C^\oplus defined by

$$\nu(dt) = (1 + \sum_i |\sum_j g_{ij}(t, u(t)) w_j(t)|) |\mu|(dt),$$

and, in proving (ii), we can choose μ to be

$$\mu(dt) = \frac{w(t) \nu(dt)}{1 + \sum_i |\sum_j g_{ij}(t, u(t)) w_j(t)|}.$$

We accept the claim then that $(x^*, \text{col}(u^*, w^*), \nu^*)$ is a Ψ boundary process for (S') .

It will be observed that (S') is a special case of the former system in which the additional hypotheses on the data, listed in the lemma, are valid.

Now apply Theorem 4.1 to $(x^*, \text{col}(u^*, w^*), \nu^*)$, the Ψ boundary process for (S') , as is permitted in view of the additional hypotheses. This yields existence of an absolutely continuous function p and $d \in R^k$, with $\|d\| = 1$, such that

$$-p(t) \in \partial_x H(t, x^*(t), u^*(t), p(t)) \quad \mathcal{L}\text{-a.e. } t \in [0, 1]$$

(where $H(t, x, u, p) = p \cdot f(t, x, u)$),

$$-p(1) = d \cdot Q \quad \text{for some } Q \in \partial \Psi(x^*(1)),$$

$u^*(t)$ maximizes

$$(6.2) \quad \begin{aligned} & u \rightarrow p(t) \cdot f(t, x^*(t), u) \quad \mathcal{L}\text{-a.e.}, \\ & \sup_{u \in \Omega_t, w \in \tilde{K}} \left\{ \frac{p(t) \cdot G(t, u) w}{1 + \sum_i |\sum_j g_{ij}(t, u) w_j|} \right\} \leq 0 \quad \text{for all } t, \end{aligned}$$

and

$$(6.3) \quad \frac{p(t) \cdot G(t, u^*(t)) w^*(t)}{1 + \sum_i |\sum_j g_{ij}(t, u^*(t)) w_j|} = 0 \quad \mu^* \text{-a.e.}$$

However (6.2) implies that

$$\sup_{u \in \Omega_t} \sigma_K(p(t) \cdot G(t, u)) \leq 0 \quad \text{for all } t \in [0, 1],$$

and we deduce from (6.3) that

$$\sigma_K(p(t) \cdot G(t, u^*(t))) = 0 \quad \mu^* \text{-a.e.}$$

We see that the special case of Theorem 4.1 applied to (S') yields the full theorem for (S). The lemma is proved.

It is assumed henceforth that H7 is in force. We write g in place of G to emphasize the fact that, under the additional hypothesis, the function is vector, not matrix, valued.

By assumption $\Psi(x^*(1))$ is a boundary point of R_Ψ . It follows that there is a sequence of vectors $\{\xi_i\}$ such that $\xi_i \notin R_\Psi$, $i = 1, 2, \dots$, and

$$(6.4) \quad \xi_i \rightarrow \Psi(x^*(1)) \quad \text{as } i \rightarrow \infty.$$

In consequence of Propositions 5.1 and 5.3, for $i = 1, 2, \dots$, there exists \mathcal{L} -measurable functions \bar{m}_i and \bar{u}_i such that \bar{m}_i is \mathcal{L} -integrable and $\bar{u}_i(t) \in \Omega_t$, \mathcal{L} -a.e., and a trajectory \bar{x}_i emanating from $x^*(0)$ with the following properties:

$$\frac{d}{dt} \bar{x}_i(t) = f(t, \bar{x}_i(t), \bar{u}_i(t)) + g(t, \bar{u}_i(t)) \bar{m}_i(t) \quad \mathcal{L}\text{-a.e. } t \in [0, 1]$$

for $i = 1, 2, \dots$,

$$(6.5) \quad \bar{m}_i(t) dt \xrightarrow{*} \mu^*(dt),$$

$$(6.6) \quad \mathcal{L}\text{-meas}(\{t: \bar{u}_i(t) \neq u^*(t)\}) \rightarrow 0$$

and

$$d\bar{x}_i(t) \xrightarrow{*} dx^*(t).$$

Now, by assumption, the \bar{x}_i 's have common left point,

$$(6.7) \quad \bar{x}_i(0) = x^*(0).$$

From Proposition 5.2 it follows we can arrange, by restricting attention to appropriate subsequences, that

$$\bar{x}_i(t) \rightarrow x^*(t),$$

on a subset of $[0, 1]$, of full Lebesgue measure, containing the points 0 and 1. Since the functions $\|x_i(t)\|$ are majorized by a common constant function we deduce from the dominated convergence theorem that

$$\int_0^1 \|\bar{x}_i(t) - x^*(t)\|^2 dt \rightarrow 0.$$

For $i = 1, 2, \dots$, define ε_i to be

$$(6.8) \quad \varepsilon_i := \left(\int_0^1 \|\bar{x}_i(t) - x^*(t)\|^2 dt + \|\xi_i - \Psi(\bar{x}_i(1))\| + \|\bar{x}_i(1) - x^*(1)\|^2 \right)^{1/2}.$$

It follows from (6.4) and the convergence properties of $\{\bar{x}_i\}$ just described that $\varepsilon_i \rightarrow 0$.

Let $\{K_i\}$ be the sequence of L^1 functions defined by

$$K_i(t) := i + \max_{j \leq i} \{\bar{m}_j(t)\}.$$

Note that each K_i majorizes \bar{m}_i and the sequence is monotone with pointwise limit $+\infty$. For each i , consider the optimal control problem (P_i) :

$$\begin{aligned} &\text{Minimize} \quad \|\xi_i - \Psi(x(1))\| + \int_0^1 \|x(t) - x^*(t)\|^2 dt + \|x(1) - x^*(1)\|^2 \\ &\text{subject to} \quad \frac{dx(t)}{dt} = f(t, x(t), u(t)) + g(t, u(t))m(t) \quad \mathcal{L}\text{-a.e.}, \\ &\quad x(0) \in C_0, \quad \text{and} \\ &\quad (u(t), m(t)) \in \Omega_i \times [0, K_i(t)] \quad \mathcal{L}\text{-a.e.}, \end{aligned}$$

in which m is treated as a block component in the control variable $\text{col}(u, m)$. The minimization is conducted over processes $\{x, \text{col}(u, m)\}$ for (P_i) , that is to say, over triples each comprising Lebesgue measurable functions u and m and an absolutely continuous function x which satisfy the specified constraints.

Problem (P_i) can be reformulated as

$$\begin{aligned} &\text{Minimize} \quad \Phi_i(a) \\ &\text{over } a \in Q_i. \end{aligned}$$

Here Q_i is the set of elements (s, u, w) in which $s \in C_0$ and $u: [0, 1] \rightarrow R^m$ and $w: [0, 1] \rightarrow R$ are \mathcal{L} -measurable functions such that $u(t) \in \Omega_i$ and $w(t) \in [0, K_i(t)]$ \mathcal{L} -a.e. The function Φ_i on Q_i is taken to be

$$\Phi_i((s, u, m)) = \|\xi_i - \Psi(x(1))\| + \int_0^1 \|x(t) - x^*(t)\|^2 dt + \|x(1) - x^*(1)\|^2$$

in which x is the unique solution to the differential equation corresponding to u and m , and the initial condition $x(0) = s$.

We provide Q_i with the metric ρ :

$$\rho((s, u, m), (\bar{s}, \bar{u}, \bar{m})) = \mathcal{L}\text{-meas}(\{t: u(t) \neq \bar{u}(t)\}) + \int_0^1 |m(t) - \bar{m}(t)| dt + \|s - \bar{s}\|.$$

Q_i thereby becomes a complete metric space (cf. [4, Lemma 1, p. 202]).

We readily deduce from Proposition 5.1 that the mapping Φ is continuous for this choice of metric.

Observe that, for each i ,

$$\Phi_i((\bar{x}_i(0), \bar{u}_i, \bar{m}_i)) \leq \inf_{a \in Q_i} \Phi_i(a) + \varepsilon_i^2$$

by (6.8) and, since $\Phi_i \geq 0$, it follows from Ekeland's theorem [5], that, for each i , there exists $(s_i, u_i, m_i) \in Q_i$ such that

$$(6.9) \quad \Phi_i((s_i, u_i, m_i)) \leq \Phi_i(a) + \varepsilon_i \rho(a, (s_i, u_i, m_i))$$

for all $a \in Q_i$, and

$$(6.10) \quad \rho((s_i, u_i, m_i), (x^*(0), \bar{u}_i, \bar{m}_i)) \leq \varepsilon_i.$$

Let x_i be the trajectory corresponding to (s_i, u_i, m_i) . Bearing in mind that $\varepsilon_i \rightarrow 0$, we deduce from (6.5)–(6.7) and (6.10) that

$$(6.11) \quad \begin{aligned} x_i(0) &\rightarrow x^*(0), \\ \mathcal{L}\text{-meas}(\{t: u_i(t) \neq u^*(t)\}) &\rightarrow 0, \quad \text{and} \\ m_i(t) \, dt &\xrightarrow{*} \mu^*(dt). \end{aligned}$$

A simple contradiction argument based on (6.9) and (6.10) establishes that $\Phi_i((s_i, u_i, m_i)) \rightarrow 0$ as $i \rightarrow \infty$. Examining Φ_i , we see that this implies

$$\int_0^1 \|x_i(t) - x^*(t)\|^2 \, dt \rightarrow 0 \quad \text{and} \quad \|x_i(1) - x^*(1)\|^2 \rightarrow 0.$$

Following extraction of a subsequence, we have then that

$$(6.12) \quad x_i(t) \rightarrow x^*(t) \quad \text{for all } t \in S \cup \{0\} \cup \{1\}$$

for some subset $S \subset (0, 1)$ of full Lebesgue measure.

It is not difficult to show (application of Gronwall's inequality is involved) that the sequence of measures $\{dx_i\}$ associated with the x_i 's is uniformly bounded in total variation. By Proposition 5.2 then

$$dx_i(t) \xrightarrow{*} dx^*(t).$$

In view of (6.11), (6.12) and hypothesis H3 we may apply the dominated convergence theorem to obtain

$$f(t, x_i(t), u_i(t)) \, dt \xrightarrow{*} f(t, x^*(t), u^*(t)) \, dt.$$

Noting that $gm_i \, dt = dx_i - f \, dt$, we deduce that

$$(6.13) \quad g(t, u_i(t))m_i(t) \, dt \xrightarrow{*} g(t, u^*(t))\mu^*(dt).$$

Inequality (6.9) which holds for all $a \in Q_i$ implies that the process (x_i, u_i, m_i) is optimal for the control problem (\bar{P}_i) :

$$\begin{aligned} \text{Minimize} \quad & \|\xi_i - \Psi(x(1))\| + \int_0^1 \|x(t) - x^*(t)\|^2 \, dt + \|x(1) - x^*(1)\|^2 \\ & + \varepsilon_i(\|x(0) - x_i(0)\| + \int_0^1 \|m(t) - m_i(t)\| \, dt \\ & + \int_0^1 \chi_i(t, u(t)) \, dt) \end{aligned}$$

subject to $\dot{x} = f(t, x, u) + g(t, u)m \quad \mathcal{L}\text{-a.e.},$

$x(0) \in C_0,$ and

$(u(t), m(t)) \in \Omega_t \times [0, K_i(t)] \quad \mathcal{L}\text{-a.e.}$

Here

$$\chi_i(t, u) := \begin{cases} 0 & \text{if } u = u_i(t), \\ 1 & \text{otherwise.} \end{cases}$$

Now (\bar{P}_i) is an optimal control problem for which necessary conditions of optimality are available.

As usual let H be the function $H(t, x, u, p) = p \cdot f(t, x, u)$. It is convenient also to introduce the function H_i defined by

$$H_i(t, u, m) := p_i(t) \cdot [f(t, x_i(t), u) + g(t, u)m] - \varepsilon_i[\chi_i(t, u) + |m - m_i(t)|].$$

Application of [4, Thm. 5.2.1] to the optimal process, (x_i, u_i, m_i) , yields the following information. (Note that (\bar{P}_i) is a free right endpoint problem whose extremals are normal.)

There exists an absolutely continuous function $p_i: [0, 1] \rightarrow R^n$ and a number $K > 0$ which does not depend on i , such that

$$(6.14) \quad -\dot{p}_i(t) \in \partial_x H(t, x_i(t), u_i(t), p_i(t)) - 2(x_i(t) - x^*(t)),$$

$$(6.15) \quad H_i(t, u_i(t), m_i(t)) = \max \{H_i(t, u, m): u \in \Omega_i, m \in [0, K_i(t)]\},$$

$$(6.16) \quad p_i(0) \in K \partial d_{C_0}(x_i(0)) + \varepsilon_i K \partial_x \|x - x_i(0)\| \Big|_{x=x_i(0)},$$

$$(6.17) \quad -p_i(1) \in \partial_x \|\xi_i - \Psi(x_i(1))\| + 2(x_i(1) - x^*(1)).$$

In view of the fact that (x_i, u_i, m_i) defines a process for S and $\xi_i \notin R_\Psi$ we have that $\|\xi_i - \Psi(x_i(1))\| \neq 0$.

It follows then from (6.17) and [4, Thm. 2.6.6] that

$$(6.18) \quad -p_i(1) \in d_i \cdot \partial \Psi(x_i(1)) + 2(x_i(1) - x^*(1))$$

for some vector d_i of unit length.

Appealing to Gronwall's inequality, we deduce from (6.14) and (6.17) that the p_i s are a bounded equicontinuous family of functions. Extraction of a subsequence therefore ensures that

$$(6.19) \quad p_i(t) \rightarrow p(t) \quad \text{for all } t \in [0, 1]$$

where p is some absolutely continuous function.

We now show that p has the requisite properties for it to serve in Theorem 4.1.

Limit attention to a subsequence, thereby arranging that the vector d_i in (6.18) converges to a limit d . Clearly $\|d\| = 1$.

The transversality conditions

$$p(0) \in K \partial d_{C_0}(x^*(0)) \quad (\subset N_{C_0}(x^*(0)))$$

and

$$-p(1) \in d \cdot \partial \Psi(x^*(1))$$

follow from (6.16) and (6.18) and the upper semicontinuity of the generalized gradient.

We now derive the costate differential inclusion

$$(6.20) \quad -\dot{p}(t) \in \partial_x H(t, x^*(t), u^*(t), p(t)) \quad \mathcal{L}\text{-a.e.}$$

Notice that, by (6.14), p_i satisfies

$$(6.21) \quad \dot{p}_i(t) \in \Gamma(t, p_i(t)) + r_i(t)B \quad \text{on } A_i$$

where

$$\begin{aligned} \Gamma(t, p) &:= -\partial_x H(t, x^*(t), p, u^*(t)), \\ r_i(t) &= \tilde{k}(t)|x_i(t) - x^*(t)| \end{aligned}$$

and

$$A_i = \{t: u_i(t) \neq u^*(t)\}.$$

Here \tilde{k} is an integrable function which does not depend on i . We now appeal to [4, Thm. 3.1.7], which concerns limiting solutions of differential inclusions such as (6.21). Scrutiny of the proof of [4, Thm. 3.1.7] reveals that the assertions of the theorem remain true if the hypothesis "the r_i 's converge uniformly to zero" is replaced by "the r_i 's are uniformly integrably bounded and converge almost everywhere to zero." The hypotheses are satisfied under which the theorem, strengthened in the manner just indicated, applies. (Note in particular that the r_i 's have the required properties, since the x_i 's are uniformly bounded in L_∞ norm and converge almost everywhere to x^* .) It follows that p satisfies the differential inclusion (6.20).

Next we show that

$$(6.22) \quad p(t) \cdot f(t, x^*(t), u^*(t)) = \max_{u \in \Omega_t} p(t) \cdot f(t, x^*(t), u) \quad \mathcal{L}\text{-a.e.}$$

and

$$(6.23) \quad \sup_{u \in \Omega_t} p(t) \cdot g(t, u) \leq 0 \quad \text{for all } t \in [0, 1].$$

We shall make repeated use of the following simple observation: If $\{\Sigma_i\}$ is a sequence of \mathcal{L} -measurable subsets of $[0, 1]$ and $\mathcal{L}\text{-meas}(\Sigma_i) \rightarrow 0$, then we may replace $\{\Sigma_i\}$ by a subsequence (we do not relabel) with the property

$$\mathcal{L}\text{-meas}(\{t: t \in [0, 1] \setminus \Sigma_j \text{ for all } j \geq i\}) \rightarrow 1$$

as $i \rightarrow \infty$. In fact it suffices to arrange, by extraction of a subsequence if necessary, that $\sum_{j=0}^{\infty} \mathcal{L}\text{-meas}(\Sigma_j) < \infty$.

Let

$$A_i = \{t: p_i(t) \cdot g(t, u_i(t)) > \varepsilon_i\},$$

$$B_i = \left\{t: m_i(t) > \frac{1}{\sqrt{\varepsilon_i}}\right\}$$

and

$$C_i = \left\{t: \sup_{u \in \Omega_t} |p_i(t) \cdot f(t, x_i(t), u)| > \sqrt{K_i(t)}\right\}.$$

The fact that the sequence $\{m_i\}$ is norm bounded in L^1 tells us that $\mathcal{L}\text{-meas}(B_i) \rightarrow 0$ as $i \rightarrow \infty$. We also know that $\mathcal{L}\text{-meas}(A_i) \rightarrow 0$ as $i \rightarrow \infty$ since the maximization of the Hamiltonian condition (6.15) implies that $m_i(t) = K_i(t)$ a.e. on A_i and the $K_i(t)$'s increase uniformly to infinity.

Furthermore $\mathcal{L}\text{-meas}(C_i) \rightarrow 0$ as $i \rightarrow \infty$ since the sequence of functions $\{t \rightarrow \sup_{u \in \Omega_t} |p_i(t) \cdot f(t, x_i(t), u)|\}$ are uniformly integrably bounded.

Now let S_i be the set of points $t \in [0, 1]$ such that

$$(6.24) \quad p_j(t) \cdot g(t, u_j(t)) \leq \varepsilon_j \quad \text{for all } j \geq i,$$

$$(6.25) \quad \sup_{u \in \Omega_t} |p_j(t) \cdot f(t, x_j(t), u)| \leq \sqrt{K_i(t)} \quad \text{for all } j \geq i,$$

$$(6.26) \quad m_j(t) \leq \frac{1}{\sqrt{\varepsilon_j}} \quad \text{for all } j \geq i,$$

$$(6.27) \quad u_j(t) = u^*(t) \quad \text{for all } j \geq i,$$

$$(6.28) \quad H_j(t, u_j(t), m_j(t)) = \max \{H_j(t, u, m): u \in \Omega_t, m \in [0, K_j(t)]\} \quad \text{for all } j \geq i$$

and

$$(6.29) \quad x_j(t) \rightarrow x^*(t) \quad \text{as } j \rightarrow \infty.$$

In view of our earlier observations, we can arrange by extraction of subsequences that $\mathcal{L}\text{-meas}(S_i) \rightarrow 1$ as $i \rightarrow \infty$.

Now take any $t \in \bigcup_i S_i$ and any $\bar{u} \in \Omega_t$. For some i then $t \in S_i$ whence, by (6.27) and (6.28),

$$H_j(t, u^*(t), m_j(t)) \geq H_j(t, \bar{u}, 0) \quad \text{for all } j \geq i.$$

This inequality coupled with (6.24) and (6.26) gives

$$p_j(t) \cdot f(t, x_j(t), u^*(t)) + \frac{\varepsilon_j}{\sqrt{\varepsilon_j}} \geq p_j(t) \cdot f(t, x_j(t), \bar{u}) - \varepsilon_j - \frac{\varepsilon_j}{\sqrt{\varepsilon_j}} \quad \text{for all } j \geq i.$$

In view of (6.19) and (6.29) we may pass to the limit, $j \rightarrow \infty$, and obtain

$$p(t) \cdot f(t, x^*(t), u^*(t)) \geq p(t) \cdot f(t, x^*(t), \bar{u}).$$

We have established (6.22) (on the subset $\bigcup_i S_i$ of full \mathcal{L} -measure). Again take $t \in \bigcup_i S_i$. For every j , let \bar{u}_j be chosen so that

$$(6.30) \quad p_j(t) \cdot g(t, \bar{u}_j) > \sup_{u \in \Omega_t} p_j(t) \cdot g(t, u) - \varepsilon_j.$$

Properties (6.27) and (6.28) give, for some i ,

$$H_j(t, u^*(t), m_j(t)) \geq H_j(t, \bar{u}_j, K_j(t)) \quad \text{for all } j \geq i.$$

It follows then from (6.24)–(6.26) and (6.30) that

$$\sqrt{K_j(t)} + \sqrt{\varepsilon_j} \geq -\sqrt{K_j(t)} + \sup_{u \in \Omega_t} (p_j(t) \cdot g(t, u)) K_j(t) - \varepsilon_j - 2\varepsilon_j K_j(t) \quad \text{for all } j \geq i.$$

But $K_j(t) \rightarrow \infty$ as $j \rightarrow \infty$. Dividing across this inequality by $K_j(t)$ and passing to the limit, $j \rightarrow \infty$, we obtain

$$\sup_{u \in \Omega_t} p(t) \cdot g(t, u) \left(= \lim_{j \rightarrow \infty} \sup_{u \in \Omega_t} p_j(t) \cdot g(t, u) \right) \leq 0.$$

(We have used (6.19) and the continuity of $p \rightarrow \sup_{u \in \Omega_t} p \cdot g(t, u)$.)

We see that (6.23) is true on the subset of full \mathcal{L} -measure, $\bigcup_i S_i$. It is true everywhere then since, under our hypotheses, the function $t \rightarrow \sup_{u \in \Omega_t} p(t) \cdot g(t, u)$ is continuous. It remains to show that

$$p(t) \cdot g(t, u^*(t)) = 0 \quad \mu^*\text{-a.e.}$$

For any i we have

$$H_i(t, u_i(t), m_i(t)) \geq H_i(t, u_i(t), m) \quad \text{for all } m \in [0, K_i(t)] \quad \mathcal{L}\text{-a.e.}$$

It follows that

$$p_i(t) \cdot g(t, u_i(t)) m_i(t) \geq p_i(t) \cdot g(t, u_i(t)) m - \varepsilon_i |m - m_i(t)|$$

for all $m \in [0, K_i(t)] \quad \mathcal{L}\text{-a.e.}$

Now suppose that $m_i(t) > 0$. Examination of the last inequality reveals that this is possible only if

$$(6.31) \quad p_i(t) \cdot g(t, u_i(t)) > -\varepsilon_i.$$

To be precise, we have shown that (6.31) is true \mathcal{L} -a.e. on the set $\{t: m_i(t) > 0\}$.

Define $\{\nu_i\}$ and ν , with ν and $\nu_i \in C^*(0, 1)$, $i = 1, \dots$, by

$$\nu_i(dt) = p_i(t) \cdot g(t, u_i(t)) m_i(t) dt, \quad i = 1, 2, \dots$$

and

$$\nu^*(dt) = p(t) \cdot g(t, u^*(t)) \mu^*(dt).$$

We readily deduce from the properties (6.13) and (6.19) that

$$\nu_i \xrightarrow{*} \nu^*.$$

According to Lemma 5.2, we can arrange by subsequence extraction that

$$(6.32) \quad \int_{[0,t]} \nu_i(ds) \rightarrow \int_{[0,t]} \nu^*(ds)$$

for all t in some dense subset P of $[0, 1]$ which contains $\{1\}$. For any $t \in P$ we have

$$\begin{aligned} \int_{[0,t]} \nu_i(ds) &= \int_0^t p_i(s) \cdot g(s, u_i(s)) m_i(s) ds \\ &= \int_{\{s: m_i(s) > 0\} \cap [0,t]} p_i(s) \cdot g(s, u_i(s)) m_i(s) ds \quad (\text{since } m_i(s) \geq 0 \text{ } \mathcal{L}\text{-a.e.}) \\ &\geq -\varepsilon_i \int_{[0,t]} m_i(s) ds, \end{aligned}$$

by (6.31). Bearing in mind that $\{m_i\}$ is norm bounded in L^1 we deduce, from this inequality and (6.32), that

$$\int_{[0,t]} \nu(dt) \geq 0,$$

i.e.,

$$\int_{[0,t]} p(t) \cdot g(t, u^*(t)) \mu^*(dt) \geq 0.$$

Since the sets $\{[0, t]: t \in P\}$ generate the Borel subsets of $[0, 1]$ we have

$$\int_E p(t) \cdot g(t, u^*(t)) \mu^*(dt) \geq 0$$

for all Borel subsets $E \subset [0, 1]$. It follows that

$$p(t) \cdot g(t, u^*(t)) \geq 0 \quad \mu^*\text{-a.e.}$$

However we have shown that

$$\sup_{u \in \Omega_t} p(t) \cdot g(t, u) \leq 0 \quad \text{for all } t \in [0, 1].$$

Of course

$$u^*(t) \in \Omega_t \quad \mu^*\text{-a.e.}$$

We conclude that

$$p(t) \cdot g(t, u^*(t)) = 0 \quad \mu^*\text{-a.e.}$$

Appendix.

Proof of Proposition 5.1. We shall use the fact that (5.5)–(5.7) remain true when Φ_0 replaces Φ_i . This is clearly a consequence of (5.3).

Let $\tilde{z}_0(t) = z_0(t) - \int_{[0,t]} \gamma_0(ds)$.

Then $\tilde{z}_0(\cdot)$ is an absolutely continuous function which satisfies the differential equation

$$(A.1) \quad \frac{d}{dt} \tilde{z}_0(t) = \Phi_0\left(t, \tilde{z}_0(t) + \int_{[0,t]} \gamma_0(ds)\right), \quad \tilde{z}_0(0) = a_0.$$

Choose $\delta > 0$ such that

$$\max_{t \in [0,1]} \|\tilde{z}_0(t)\| + \|\gamma_i\|_{TV} < \frac{\delta}{2}$$

for all i sufficiently large.

Consider now the functions ρ_i , $i = 1, 2, \dots$,

$$\rho_i(t) := \left| \dot{\tilde{z}}_0(t) - \Phi_i\left(t, \tilde{z}_0(t) + a_i - a_0 + \int_{[0,t]} \gamma_i(ds)\right) \right|.$$

Since $\tilde{z}_0(\cdot)$ solves the differential equation (A.1), ρ_i can be written

$$\rho_i(t) = \left| \Phi_0\left(t, \tilde{z}_0(t) + \int_{[0,t]} \gamma_0(ds)\right) - \Phi_i\left(t, \tilde{z}_0(t) + a_i - a_0 + \int_{[0,t]} \gamma_i(ds)\right) \right|.$$

Introduce the sets S_i , $i = 1, 2, \dots$,

$$S_i = \{t: \Phi_i(t, x) = \Phi_0(t, x), \text{ for all } x \in \mathbb{R}^n\}.$$

For $t \in [0, 1] \setminus S_i$,

$$\begin{aligned} \rho_i(t) &\leq \left| \Phi_0\left(t, \tilde{z}_0(t) + \int_{[0,t]} \gamma_0(ds)\right) \right| + \left| \Phi_i\left(t, \tilde{z}_0(t) + \int_{[0,t]} \gamma_0(ds)\right) \right| + \bar{K}_\delta(t) s_i(t) \\ &\leq 2\bar{\alpha}(t) + \bar{K}_\delta(t) s_i(t). \end{aligned}$$

In these estimates,

$$s_i(t) = |a_i - a_0| + \left| \int_{[0,t]} (\gamma_i(ds) - \gamma_0(ds)) \right|.$$

(We have used bounds (5.6) and (5.7) for Φ_0 in place of Φ_i .)

For $t \in S_i$,

$$\rho_i(t) \leq \bar{K}_\delta(t) s_i(t).$$

Now (5.1) and (5.2) imply that the s_i 's are uniformly bounded in the L^∞ norm and, on a Lebesgue set of full measure, $s_i(t) \rightarrow 0$. Noting also (5.3), we deduce by application of the dominated convergence theorem that

$$\int_0^1 \rho_i(t) dt \rightarrow 0 \quad \text{as } i \rightarrow \infty.$$

By (5.5)–(5.7), the function $(t, x) \rightarrow \Phi_i(t, x + a_i - a_0 + \int_{[0,t]} \gamma_i(ds))$ is measurably Lipschitz on the $(\delta/4)$ tube about $t \rightarrow \tilde{z}_0(t) + a_i - a_0$, for i sufficiently large. (“Measurably Lipschitz” is taken in the sense of [4, Def. 3.1.4].) It is now evident from the

proof of [4, Thm. 3.1.6], that there exists an absolutely continuous function $\tilde{z}_i(\cdot)$ satisfying

$$(A.2) \quad \frac{d}{dt} \tilde{z}_i(t) = \Phi_i \left(t, \tilde{z}_i(t) + \int_{[0,t]} \gamma_i(ds) \right), \quad \tilde{z}_i(0) = a_i$$

for i sufficiently large; furthermore

$$(A.3) \quad \tilde{z}_i(t) - \tilde{z}_0(t) \rightarrow 0 \quad \text{uniformly in } t \in [0, 1]$$

and

$$(A.4) \quad \int_0^1 |\dot{\tilde{z}}_i(t) - \dot{\tilde{z}}_0(t)| dt \rightarrow 0.$$

Now define $z_i(t) := \tilde{z}_i(t) + \int_{[0,t]} d\gamma_i$, for $t > 0$, and $z_i(0) := a_i$. By (A.2), $z_i(\cdot)$ is a normalized function of bounded variation which satisfies (5.8).

Property (A.3), expressed in terms of $z_i(\cdot)$ and $z_0(\cdot)$, is (5.9). Property (A.4) implies that

$$d\tilde{z}_i \xrightarrow{*} d\tilde{z}_0.$$

But then

$$dz_i (= d\tilde{z}_i + \gamma_i(dt)) \xrightarrow{*} (dz_0 + \gamma_0(dt)) = d\tilde{z}_0,$$

since $\gamma_i \xrightarrow{*} \gamma_0$. We have proved (5.10).

Proof of Proposition 5.2. We may limit attention to the scalar case ($k = 1$), since the scalar version of the lemma, applied componentwise, leads to the same conclusions in the vector setting.

(a) Replace $\{\nu_i\}$ by an arbitrary subsequence. (We do not relabel.) Define ν_i^+ , $\nu_i^- \in C^\oplus(0, 1)$ by the Jordan decomposition of ν_i :

$$\nu_i = \nu_i^+ - \nu_i^-$$

for $i = 1, 2, \dots$. The sequences $\{\nu_i^+\}$, $\{\nu_i^-\}$ are uniformly bounded, and we can arrange by extraction of subsequences that

$$\nu_i^+ \xrightarrow{*} \nu^+, \quad \nu_i^- \xrightarrow{*} \nu^-$$

for some $\nu^+, \nu^- \in C^\oplus(0, 1)$. Then

$$\nu_i (= \nu_i^+ - \nu_i^-) \xrightarrow{*} \nu'$$

where $\nu' = \nu^+ - \nu^-$. Since the initial subsequence was arbitrary, part (a) will be proved if we can show that $\nu' = \nu$.

Let $S_1 \subset [0, 1]$ be the set

$$S_1 = \{t \in S : \{t\} \text{ is not an atom of } \nu^+ \text{ or } \nu^-\} \cup \{1\}.$$

(Here S is the set mentioned in the hypotheses.) Now take any $t \in S_1$

$$\int_{[0,t]} \nu_i(ds) \left(= \int_{[0,t]} \nu_i^+(ds) - \int_{[0,t]} \nu_i^-(ds) \right) \rightarrow \int_{[0,t]} \nu'(ds)$$

(see [1, Thm. 2.1]).

Noting that $S_1 \subset S$, we deduce from the hypotheses that

$$(A.5) \quad \int_{[0,t]} \nu(ds) = \int_{[0,t]} \nu'(ds).$$

But the set S_1 is certainly dense in $[0, 1]$, and contains $\{1\}$. It follows from the continuity of $t \rightarrow \int_{[0,t]} \nu(ds)$ from the right on $[0, 1]$ that (A.5) holds for all $t \in [0, 1]$. It is now easy to show, using the regularity of ν and ν' , that

$$\int_A \nu'(ds) = \int_A \nu(ds)$$

for all relatively open subintervals A of $[0, 1]$, and therefore for all Borel sets A , since these sets are generated by the relatively open subintervals. In other words $\nu = \nu'$. Part (a) is proved.

(b) Now suppose (5.11). Define ν_i^+ and ν_i^- by the Jordan decomposition

$$\nu_i = \nu_i^+ - \nu_i^-$$

for $i = 1, 2, \dots$. By uniform boundedness, there exists a subsequence $\{\nu_{i_j}\}$ of $\{\nu_i\}$ such that

$$\nu_{i_j}^+ \xrightarrow{*} \nu^+, \quad \nu_{i_j}^- \xrightarrow{*} \nu^-$$

for some $\nu^+, \nu^- \in C^\oplus$. Then

$$\nu = \text{weak}^* \lim_{j \rightarrow \infty} \{\nu_{i_j}\} = \text{weak}^* \lim_{j \rightarrow \infty} \{\nu_{i_j}^+ - \nu_{i_j}^-\} = \nu^+ - \nu^-.$$

Now take S to be

$$S = \{t \in [0, 1]: \{t\} \text{ is not an atom of } \nu^+ \text{ or } \nu^-\} \cup \{1\}.$$

S is the complement in $[0, 1]$ of a countable set, and contains $\{1\}$. Since $[0, t]$ is a ν^+ - and ν^- -continuity set for each $t \in S$, we conclude that

$$\begin{aligned} \int_{[0,t]} \nu_{i_j}(ds) &= \int_{[0,t]} (\nu_{i_j}^+ - \nu_{i_j}^-)(ds) \\ &\rightarrow \int_{[0,t]} (\nu^+ - \nu^-)(ds) = \int_{[0,t]} \nu(ds) \end{aligned}$$

for each $t \in S$. The subsequence $\{\nu_{i_j}\}$ and the subset S have the required properties then, and the proposition is proved.

Proof of Proposition 5.3. Let us first note that we can without loss of generality suppose that

(A.6) μ is singular with respect to Lebesgue measure.

Indeed if μ does not satisfy this condition, we can decompose it into the sum of absolutely continuous and singular components (with respect to Lebesgue measure) thus

$$\mu(dt) = w(t) dt + \mu^*(dt).$$

Here w is an integrable function and μ^* is a measure singular with respect to Lebesgue measure. Since μ is nonnegative valued, we know that $w(t) \geq 0$, \mathcal{L} -a.e., and μ^* is nonnegative valued. We now apply the proposition (under the additional hypothesis (A.6)) and conclude existence of sequences $\{u_i\}$ and $\{m_i\}$ with the stated properties, corresponding to u and μ^* . It will be seen that the sequences $\{u_i\}$ and $\{m_i + w\}$ fulfill the claims of the proposition in relation to u and μ .

We then impose (A.6). It follows that a sequence $\{A_i\}$ of open subsets of $[0, 1]$ exists such that μ is concentrated on each of the A_i 's and

$$\mathcal{L}\text{-meas}(A_i) \rightarrow 0 \quad \text{as } i \rightarrow \infty.$$

For $j = 0, 1, \dots, i-1$ and $i = 1, 2, \dots$, define the intervals I_{ij} and J_{ij} :

$$I_{ij} = \left[\frac{j}{i}, \frac{j+1}{i} \right), \quad J_{ij} = I_{ij} \cap A_i.$$

Define $\hat{m}_{ij} \in R^+$:

$$\hat{m}_{ij} = \begin{cases} \frac{1}{|J_{ij}|} \int_{J_{ij}} \mu(dt) & \text{when } J_{ij} \neq \emptyset, \\ 0 & \text{when } J_{ij} = \emptyset \end{cases}$$

in which $|J_{ij}|$ denotes \mathcal{L} -measure (J_{ij}) , and $m_{ij}: I_{ij} \rightarrow R^+$:

$$m_{ij}(t) = \begin{cases} \hat{m}_{ij} & \text{if } t \in J_{ij}, \\ 0 & \text{if } t \in I_{ij} \setminus A_i. \end{cases}$$

We also define $\hat{\gamma}_{ij} \in R^n$:

$$\hat{\gamma}_{ij} = \frac{\int_{J_{ij}} r(t, u(t)) \mu(dt)}{\int_{J_{ij}} \mu(dt)} \quad \text{if } \hat{m}_{ij} \neq 0$$

and $\tilde{\gamma}_{ij}: I_{ij} \rightarrow R^n$:

$$\tilde{\gamma}_{ij}(t) = \begin{cases} \hat{\gamma}_{ij} & \text{if } t \in J_{ij} \text{ and } \hat{m}_{ij} \neq 0, \\ r(t, u(t)) & \text{if } t \in I_{ij} \setminus A_i \text{ or } \hat{m}_{ij} = 0. \end{cases}$$

Next define $\gamma_{ij}: I_{ij} \rightarrow R^n$ by

$$\gamma_{ij}(t) = \begin{cases} p_{ij}(t) & \text{if } t \in J_{ij} \text{ and } \hat{m}_{ij} \neq 0, \\ r(t, u(t)) & \text{if } t \in I_{ij} \setminus A_i \text{ or } \hat{m}_{ij} = 0 \end{cases}$$

where the function p_{ij} is uniquely specified by

$$|p_{ij}(t) - \hat{\gamma}_{ij}| = \min_{p \in \overline{\text{co}} r(t, U_i)} |p - \hat{\gamma}_{ij}|.$$

Observe that p_{ij} is continuous on its domain of definition (this follows from the continuity properties of $t \rightarrow \overline{\text{co}} r(t, U_i)$). Consequently γ_{ij} is a Borel measurable function. A simple separating hyperplane argument tells us that $\tilde{\gamma}_{ij}(t) \in \overline{\text{co}} [\bigcup_{s \in I_{ij}} r(s, U_s)]$; appealing to this fact and also again to the continuity properties of $t \rightarrow \overline{\text{co}} r(t, U_i)$, we deduce existence of a sequence $\{\varepsilon_i\}$ of real numbers, $\varepsilon_i \rightarrow 0$, such that

$$(A.7) \quad \|\tilde{\gamma}_{ij}(t) - \gamma_{ij}(t)\| \leq \varepsilon_i \quad \text{for all } i, j \text{ and } t \in I_{ij}.$$

Now define Borel measurable functions $m_i: [0, 1] \rightarrow R$ and $\gamma_i: [0, 1] \rightarrow R^n$ by

$$m_i(t) = \sum_j m_{ij}(t) \chi_{I_{ij}}(t)$$

and

$$\gamma_i(t) = \sum_j \gamma_{ij}(t) \chi_{I_{ij}}(t).$$

Here χ_A denotes the indicator function of the set A :

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

Let us verify that $\{m_i\}$ has the properties asserted in the proposition. Clearly $t \rightarrow r(t, u_i(t))m_i(t)$ is \mathcal{L} integrable. It is immediately apparent that the m_i 's are nonnegative valued L_∞ functions. To ascertain the limit we take arbitrary $\phi \in C(0, 1)$ and calculate

$$\begin{aligned}
 (A.8) \quad \int_0^1 \phi(t) m_i(t) dt &= \sum_{j \in S_i} \int_{J_{ij}} \phi(t) m_i(t) dt \\
 &= \sum_{j \in S_i} \int_{J_{ij}} \left(\int_{J_{ij}} \phi(t) dt / |J_{ij}| \right) \mu(dt) \\
 &= \sum_{j \in S_i} \int_{J_{ij}} \phi(t) \mu(dt) + \sum_{j \in S_i} \int_{J_{ij}} q_i(t) \mu(dt)
 \end{aligned}$$

where the function $q_i(\cdot): [0, 1] \rightarrow R$ is

$$(A.9) \quad q_i(t) = \sum_{j \in S_i} \left[\int_{J_{ij}} \phi(s) ds / |J_{ij}| - \phi(t) \right] \chi_{I_{ij}}(t).$$

Since ϕ is a continuous function, the q_i 's converge uniformly to zero. Consequently the second term in the right-hand side of the second equality in (A.8) converges to zero as $i \rightarrow \infty$. In view of the fact that μ is concentrated on a subset of A_i , for each i , the first term is just the desired limit $\int_{[0,1]} \phi(t) \mu(dt)$. Convergence is proved.

Our next objective is to show that

$$(A.10) \quad \gamma_i(t) m_i(t) dt \xrightarrow{*} r(t, u(t)) \mu(dt).$$

Now take arbitrary $\phi \in C^n(0, 1)$. We have

$$\begin{aligned}
 (A.11) \quad \int_0^1 \phi(t) \gamma_i(t) m_i(t) dt &= \sum_{j \in T_i} \int_{I_{ij}} \phi(t) \tilde{\gamma}_{ij} m_{ij}(t) dt \\
 &\quad + \sum_{j \in T_i} \int_{I_{ij}} \phi(t) \cdot [\gamma_{ij}(t) - \tilde{\gamma}_{ij}(t)] m_{ij}(t) dt
 \end{aligned}$$

where T_i is the set of indices j such that $\int_{J_{ij}} \mu(dt) \neq 0$. By (A.7), the norm of the second term on the right-hand side of (A.11) is bounded by

$$\varepsilon_i \|\phi\|_C \sum_{j \in T_i} \int_{I_{ij}} m_i(t) dt = \varepsilon_i \|\phi\|_C \|\mu\|_{TV},$$

which tends to zero as $i \rightarrow \infty$. Consider now the first term,

$$\sum_{j \in T_i} \int_{I_{ij}} \phi(t) \tilde{\gamma}_{ij}(t) m_{ij}(t) dt.$$

This can be expressed as

$$\begin{aligned}
 &\sum_{j \in T_i} \int_{J_{ij}} \phi(t) dt \left(\int_{J_{ij}} r(t, u(t)) \mu(dt) / \int_{J_{ij}} \mu(dt) \right) \left(\int_{J_{ij}} \mu(dt) / |J_{ij}| \right) \\
 &= \sum_{j \in T_i} \int_{J_{ij}} \left(\int_{J_{ij}} \phi(t) dt / |J_{ij}| \right) r(t, u(t)) \mu(dt) \\
 &= \sum_{j \in T_i} \int_{J_{ij}} \phi(t) r(t, u(t)) \mu(dt) + \sum_{j \in T_i} \int_{J_{ij}} q_i(t) r(t, u(t)) \mu(dt).
 \end{aligned}$$

Here $q_i(\cdot)$ is defined once again by (A.9), though now of course we view ϕ as a vector valued function. As before the q_i 's converge uniformly to zero as $i \rightarrow \infty$. (At this point we make use of the uniform bound on $r(t, U_t)$.) The first term, however, is simply

$$\int_0^1 \phi(t) \cdot r(t, u(t)) \mu(dt),$$

since μ is concentrated on a subset of A_i . We have proved (A.10).

Now define the functions $h_i: [0, 1] \times R^m \rightarrow R^n$ and the multifunctions U^i on $[0, 1]$ to be

$$h_i(t, v) = r(t, v) m_i(t)$$

and

$$U_t^i = \begin{cases} U_t & \text{if } t \in A_i, \\ \{u(t)\} & \text{if } t \in [0, 1] \setminus A_i. \end{cases}$$

Taking note of the manner in which γ_i was constructed, we see that

$$\gamma_i(t) m_i(t) \in \overline{\text{co}} h_i(t, U_t^i) \quad \text{for all } t \in [0, 1].$$

Now let $\{\delta_i\}$ be an arbitrary sequence of positive numbers such that $\delta_i \rightarrow 0$. A routine argument in which we consider increasingly fine uniform partitions, Σ_i , of $[0, 1]$ into intervals and apply Aumann's theorem (see, e.g., [4, Thm. 3.1.3]) to the multifunction

$$t \rightarrow \overline{h_i(t, U_t^i)}$$

on each element of Σ_i leads to the conclusion: there exists a measurable function $d_i(\cdot): [0, 1] \rightarrow R^n$ such that

$$d_i(t) \in \overline{h_i(t, U_t^i)} \quad \mathcal{L}\text{-a.e.}, \quad t \in [0, 1]$$

and

$$\sup_{t \in [0, 1]} \left| \int_0^t [d_i(s) - \gamma_i(s) m_i(s)] ds \right| < \frac{\delta_i}{2}$$

for $i = 1, 2, \dots$. However, [3] tells us that there exists a measurable function u_i such that

$$u_i(t) \in U_t^i \quad \text{a.e. } t \in [0, 1]$$

and

$$\sup_{t \in [0, 1]} \left| \int_0^t [d_i(s) - h_i(s, u_i(s))] ds \right| < \frac{\delta_i}{2}$$

for $i = 1, 2, \dots$. Taking account of the definitions of h_i and U_t^i , we see from these inequalities that

$$(A.12) \quad \sup_{t \in [0, 1]} \left| \int_0^t \gamma_i(s) m_i(s) ds - \int_0^t r(s, u_i(s)) m_i(s) ds \right| < \delta_i,$$

and

$$(A.13) \quad \begin{aligned} u_i(t) &= u(t) \quad \text{a.e. } t \in [0, 1] \setminus A_i, \\ u_i(t) &\in \Omega_t \quad \text{a.e. } t \in A_i. \end{aligned}$$

In view of Proposition 5.2, (A.12) implies

$$[\gamma_i(t) - r(t, u_i(t))]m_i(t) dt \xrightarrow{*} 0.$$

It now follows from (A.10) that

$$(A.14) \quad r(t, u_i(t))m_i(t) dt \xrightarrow{*} r(t, u(t))\mu(dt).$$

Bearing in mind that $\mathcal{L}\text{-meas}(A_i) \rightarrow 0$ as $i \rightarrow \infty$, we see from (A.13) and (A.14) that the functions u_i which we have constructed have the properties asserted in the lemma. The proof is complete.

REFERENCES

- [1] P. BILLINGSLEY, *Convergence of Probability Measures*, Wiley-Interscience, New York, 1968.
- [2] C. W. CLARK, F. H. CLARKE AND G. R. MUNRO, *The optimal exploitation of renewable resource stocks*, *Econometrica*, 47 (1979), pp. 25–47.
- [3] F. H. CLARKE, *The maximum principle under minimal hypotheses*, this Journal, 14 (1976), pp. 1078–1091.
- [4] ———, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
- [5] I. EKELAND, *Nonconvex minimization problems*, *Bull. Amer. Math. Soc.*, 1 (1979), pp. 443–474.
- [6] D. F. LAWDEN, *Optimal Trajectories for Space Navigation*, Butterworth, London, 1963.
- [7] J. P. MAREC, *Optimal Space Trajectories*, Elsevier, Amsterdam-Oxford, 1979.
- [8] J. M. MURRAY, *Existence theorems for optimal control and calculus of variations problems where the states can jump*, this Journal, 24 (1986), pp. 412–438.
- [9] L. W. NEUSTADT, *A general theory of minimum-fuel space trajectories*, this Journal, 3 (1965), pp. 317–356.
- [10] F. M. F. L. PEREIRA, *A maximum principle for impulsive control systems*, Ph.D. thesis, Imperial College, Univ. of London, London, 1986.
- [11] R. W. RISHEL, *An extended Pontryagin principle for control systems whose control laws contain measures*, this Journal, 3 (1965), pp. 191–205.
- [12] R. T. ROCKAFELLAR, *Dual problems for arcs of bounded variation*, in *Calculus of Variations and Control Theory*, D. L. Russell, ed., Academic Press, New York; Springer-Verlag, Berlin-New York, 1976, pp. 155–192.
- [13] ———, *Optimality conditions for convex control problems with nonnegative states and the possibility of jumps*, in *Game Theory and Mathematical Economics*, O. Moeschlin and D. Pallaschke, eds., North-Holland, Amsterdam, 1981, pp. 339–349.
- [14] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [15] ———, *Variational problems with unbounded controls*, this Journal, 3 (1966), pp. 424–438.

SUFFICIENT CONDITIONS FOR THE FUNCTIONAL REPRODUCIBILITY OF TIME-VARYING, INPUT-OUTPUT SYSTEMS*

KEVIN A. GRASSE†

Abstract. The functional reproducibility of an input-output system refers to the capability of realizing a given class of functions as outputs of the system by the choice of appropriate inputs and initial conditions. For linear, time-varying, input-output systems we derive rank conditions that are sufficient for the generation of all C^k outputs by means of C^0 inputs. These rank conditions provide an alternative to the structure algorithm of L. M. Silverman and, in the linear, time-invariant case, reduce to a rank condition of M. K. Sain and J. L. Massey. We also give sufficient conditions for the local functional reproducibility of nonlinear, time-varying, input-output systems.

Key words. input-output system, functional reproducibility, path controllability.

AMS(MOS) subject classifications. 93C05, 93C50

1. Introduction. An important problem in the study of input-output systems is the determination of whether the output of the system can be made to follow a preassigned trajectory over a specified interval of time. The term “functional reproducibility” was introduced by Brockett and Mesarović [3] and, roughly speaking, means the capability of realizing a given class of functions as outputs of the system by the choice of appropriate inputs and initial conditions. Other authors have used the terms “output tracking” [7] and “path controllability” [14] with a similar connotation.

To illustrate the problem to be considered here and to relate our work to the existing literature, we consider the linear, time-varying, input-output system:

$$(1) \quad \dot{x} = A(t)x + B(t)w, \quad y = C(t)x + D(t)w,$$

where $t \in I$ (an interval in \mathbb{R}), $x \in \mathbb{R}^n$, $w \in \mathbb{R}^m$ and $y \in \mathbb{R}^p$; the matrices A , B , C , D have the obvious dimensions and their entries are assumed to be at least continuous. It is known that under a fixed initial condition $x(t_0) = x_0$ one can realize every continuous function $\psi: I \rightarrow \mathbb{R}^p$ as an output of (1) by means of a continuous input $u: I \rightarrow \mathbb{R}^m$ if and only if $\text{rank } D(t) = p$ for every $t \in I$ [1]. This rank condition is quite strong and it is natural to seek weaker conditions that will still ensure the generation of a reasonably large class of outputs.

There are two aspects of the problem that can be modified so as to yield a weaker form of the above rank condition. First, instead of requiring that all continuous functions $\psi: I \rightarrow \mathbb{R}^p$ be generated as outputs of (1), we will only require that all *sufficiently differentiable* (i.e., C^k for $k \geq 1$) functions be generated as outputs. Second, instead of fixing the initial condition $x(t_0) = x_0$ in advance, we will allow the initial condition to “float” and possibly depend on the output. With these considerations in mind, the objective of this paper is to find computable rank conditions involving the system matrices of (1) which guarantee that given an arbitrary C^k function $\psi: I \rightarrow \mathbb{R}^p$ one can find a continuous input $u: I \rightarrow \mathbb{R}^m$ and an initial condition $x(t_0) = x_0$ such that ψ is the output of (1) with this input and initial condition. Our results will also show how the required input and initial condition can be obtained from the desired output by means of a “right-inverse” system of (1).

* Received by the editors November 25, 1985; accepted for publication (in revised form) December 3, 1986.

† Department of Mathematics, The University of Oklahoma, Norman, Oklahoma 73019.

The results of this paper are very much in the spirit of the paper of Sain and Massey [13]. Sain and Massey consider both the invertibility and the functional reproducibility of linear, time-invariant, input-output systems. Their work requires certain results from the theory of sequential circuits and makes extensive use of Laplace-transform techniques, which effectively limits the scope of their results to the time-invariant case. Our proofs are self-contained and avoid Laplace transform methods since we deal with the time-varying case. The general method employed here could be described as "coordinate-free, time-varying, linear algebra." As we will see, the rank conditions derived here in the time-varying case, when specialized to the time-invariant case, reduce to precisely the rank condition of Sain and Massey.

The functional reproducibility of linear, time-invariant, input-output systems has also been treated by Silverman [9] and Silverman and Payne [10]; this work is primarily based on Silverman's introduction of the "structure algorithm." In [8] Silverman proves a necessary and sufficient condition for the functional reproducibility of a linear, time-varying, input-output system, but this result is restricted to the single-input/single-output case. Also in [9] he briefly mentions that the techniques used there to prove results on the invertibility of linear, time-varying, input-output systems can also be applied to produce results on the functional reproducibility of such systems, but he does not explicitly state any such results.

The structure algorithm was extended to nonlinear input-output systems by Hirschorn in [6] and he used this extension to prove a result on the functional reproducibility of nonlinear input-output systems in [7]. Hirschorn's results were refined somewhat by Singh in [11] and [12]. The results of Hirschorn and Singh are stated for the time-invariant case, but they can easily be extended to the time-varying case by using the standard device of incorporating time as the $(n+1)$ st coordinate of an augmented state space. In particular, one can apply these results to obtain necessary and sufficient conditions for the functional reproducibility of linear, time-varying, input-output systems.

In view of the above remarks, it might appear that the problem of functional reproducibility has been completely settled, even for nonlinear time-varying systems. However, the existing results all (with one exception on which we will comment later) have a drawback. In rough terms, these results say that if the rank of a certain matrix, produced by k applications of the structure algorithm, is p (the output space dimension), then a C^k function $\psi: I \rightarrow \mathbb{R}^p$ is an output of the system corresponding to some continuous input if and only if there exists an initial condition $x(t_0) = x_0$ such that a certain functional relationship

$$(*) \quad F_i(t_0, x_0, \psi(t_0), \dots, \psi^{(k-1)}(t_0)) = 0$$

holds for $0 \leq i \leq k-1$. The functions F_i are generated by the structure algorithm and will not be described explicitly here (see, e.g., [7] for details). If one is interested in the capability of producing *all* C^k functions as outputs, then a natural question occurs at this point. Given an arbitrary C^k function $\psi: I \rightarrow \mathbb{R}^p$, does there exist an initial condition $x(t_0) = x_0$ satisfying $(*)$ for $0 \leq i \leq k-1$?

For nonlinear input-output systems the answer is no. As an easy example, consider the input-output system:

$$\dot{x}_1 = e^{x_2}, \quad \dot{x}_2 = u, \quad y = x_1.$$

Taking derivatives of y , we obtain $\dot{y} = \dot{x}_1 = e^{x_2}$, $\ddot{y} = \dot{x}_2 e^{x_2} = u e^{x_2}$, so the structure algorithm terminates after two steps and the rank condition is satisfied by every point of the state

space. Nevertheless, any C^2 function $\psi: \mathbb{R} \rightarrow \mathbb{R}$ for which $\dot{\psi}(t) \leq 0$ for at least one $t \in \mathbb{R}$ cannot be realized as an output.

For linear input-output systems, however, the answer is yes. Our results will show that if certain rank conditions are satisfied (see (31) and (32)), then for every C^k function $\psi: I \rightarrow \mathbb{R}^p$ there exists a continuous input $u: I \rightarrow \mathbb{R}^m$ and an initial condition $x(t_0) = x_0$ that generate ψ as the output of (1). The existence of the appropriate initial condition is nontrivial and is proved in Proposition 2.13. This is a rather delicate point and seems to be overlooked in most of the existing literature. The exception is in [8], where Silverman does address this point for single-input/single-output systems. However, it appears that he did not pursue it in his later work on multi-input/multi-output systems. For more discussion on this matter, we refer the reader to [2].

We feel that the formulation of our results has certain other advantages over the structure algorithm. Our rank conditions are intrinsic to the system matrices of (1) and do not depend on a series of possibly nonunique matrix operations, as is the case with the structure algorithm. Also, the matrices whose rank must be examined in our results are generated by matrix additions, multiplications and differentiations; no inversions or row reductions are required. Consequently, these matrices could be easily generated by a symbolic manipulation program.

A short outline of our paper follows. The main result on the functional reproducibility of linear, time-varying, input-output systems is proved in § 2. In § 3 we briefly discuss how our sufficient condition for the linear case can be applied to the linearization of a nonlinear input-output system to yield a local functional reproducibility result in the nonlinear case. As is pointed out in [15], the existing results of Hirschorn and Singh on nonlinear functional reproducibility are also, in some sense, only local results. Additional results on nonlinear functional reproducibility can be found in [1], [5], [15].

2. The linear time-varying case. Let I be an interval in \mathbb{R} (possibly unbounded) and for nonnegative integers k, l (or $k = \infty$) let $C^{k,l}(I)$ denote the set of all C^k mappings of I into \mathbb{R}^l . We consider the linear, time-varying, input-output system (1), where, as mentioned previously, $t \in I$, $x \in \mathbb{R}^n$, $w \in \mathbb{R}^m$ and $y \in \mathbb{R}^p$. The entries of the time-varying matrix functions A, B, C, D are assumed to be C^∞ functions of t . The C^∞ assumption is invoked to avoid the necessity of counting orders of differentiability and “ C^∞ ” could be replaced by “ C^r for r sufficiently large.” We assume that the inputs of (1) are in $\mathcal{C}^{0,m}(I)$.

DEFINITION 2.1. The system (1) is called *weakly C^k -reproducible* ($k \geq 0$) on I if for every $\psi \in \mathcal{C}^{k,p}(I)$ there exist $(t_0, x_0) \in I \times \mathbb{R}^n$ and $u \in \mathcal{C}^{0,m}(I)$ such that for every $t \in I$

$$\psi(t) = C(t)\phi(t) + D(t)u(t)$$

where ϕ is the solution of the initial-value problem

$$\dot{\phi}(t) = A(t)\phi(t) + B(t)u(t), \quad \phi(t_0) = x_0.$$

In other words (1) is weakly C^k -reproducible if every $\psi \in \mathcal{C}^{k,p}(I)$ can be realized as an output of (1) through the choice of an appropriate initial condition in $I \times \mathbb{R}^n$ and input in $\mathcal{C}^{0,m}(I)$. We call the reproducibility “weak” because the initial condition may depend on the output and cannot be fixed in advance. The reader is referred to [2] for a discussion and comparison of this type of reproducibility as it relates to other notions of functional reproducibility that have appeared in the literature.

Our objective in this section is to give a sufficient condition for system (1) to be weakly C^k -reproducible for $k \geq 1$ (a necessary and sufficient condition for weak C^0 -reproducibility was mentioned in the Introduction). This sufficient condition will

be phrased in terms of rank conditions involving the system matrices A , B , C , D and their derivatives. It will be shown that if the sufficient condition is satisfied, then there exist $C^\infty(m \times p)$ -matrix functions

$$E_0(t), E_1(t), \dots, E_k(t)$$

($t \in I$) with the property that every $\psi \in \mathcal{C}^{k,p}(I)$ determines $(t_0, x_0) \in I \times \mathbb{R}^n$ and $v \in \mathcal{C}^{k,p}(I)$ such that for the initial condition $x(t_0) = x_0$ and the (continuous) input

$$(2) \quad u(t) = \sum_{i=0}^k E_i(t) v^{(i)}(t)$$

where $v^{(0)}(t) = v(t)$ and $v^{(i)}(t)$ = the i th derivative of $v(t)$ for $i \geq 1$, system (1) generates ψ as its output.

Before we specifically determine the matrix functions E_0, E_1, \dots, E_k , let us examine the relationship between outputs ψ of (1) corresponding to inputs u of the form (2), where for the time being E_0, E_1, \dots, E_k are arbitrary $C^\infty(m \times p)$ -matrix functions of $t \in I$ and we leave the initial condition unspecified. Then (in the sequel we suppress t when convenient)

$$(3) \quad \psi = C\phi + \sum_{i=0}^k DE_i v^{(i)},$$

where

$$(4) \quad \dot{\phi} = A\phi + \sum_{i=0}^k BE_i v^{(i)}.$$

We temporarily assume that the function $v \in \mathcal{C}^{k,p}(I)$ is actually C^∞ . Differentiating both sides of (3) and using (4), we obtain

$$(5) \quad \dot{\psi} = (\dot{C} + CA)\phi + (CBE_0 + \overline{DE_0})v^{(0)} + \sum_{i=1}^k (CBE_i + \overline{DE_i} + DE_{i-1})v^{(i)} + DE_k v^{(k+1)}.$$

Define an operator Γ on differentiable ($q \times n$)-matrix functions $M(t)$ ($q \geq 1$) by

$$(6) \quad \Gamma M = \dot{M} + MA.$$

For $l \geq 2$ we define $\Gamma^l M$ inductively by $\Gamma^l M = \Gamma(\Gamma^{l-1} M)$ and by convention $\Gamma^0 M = M$. Setting

$$(7) \quad F_i^0 = DE_i, \quad 0 \leq i \leq k,$$

we can rewrite (3) and (5) as

$$\psi = (\Gamma^0 C)\phi + \sum_{i=0}^k F_i^0 v^{(i)} \quad \text{and} \quad \dot{\psi} = (\Gamma^1 C)\phi + \sum_{i=0}^{k+1} F_i^1 v^{(i)}$$

where $F_0^1 = (\Gamma^0 C)BE_0 + \dot{F}_0^0$, $F_{k+1}^1 = F_k^0 = DE_k$, and

$$F_i^1 = (\Gamma^0 C)BE_i + \overline{DE_i} + DE_{i-1}, \quad 1 \leq i \leq k.$$

This process can be continued to yield for $0 \leq l \leq k$

$$(8) \quad \psi^{(l)} = (\Gamma^l C)\phi + \sum_{i=0}^{k+l} F_i^l v^{(i)}$$

where for $l=0$ the F_i^0 , $0 \leq i \leq k$, are defined by (7) and for $1 \leq l \leq k$ the F_i^l , $0 \leq i \leq k+l$, are defined inductively by

$$(9a) \quad F_0^l = (\Gamma^{l-1}C)BE_0 + \dot{F}_0^l,$$

$$(9b) \quad F_i^l = (\Gamma^{l-1}C)BE_i + \dot{F}_i^{l-1} + F_{i-1}^{l-1}, \quad 1 \leq i \leq k,$$

$$(9c) \quad F_i^l = \dot{F}_i^{l-1} + F_{i-1}^{l-1}, \quad k < i < k+l,$$

$$(9d) \quad F_{k+l}^l = F_{k+l-1}^{l-1} = \cdots = F_k^0 = DE_k.$$

We will derive an alternative formula for the $(p \times p)$ -matrix functions $F_i^l(t)$, $t \in I$, for $0 \leq i, l \leq k$. To this end we define $(m \times p)$ -matrix functions $H_i^r(t)$, $t \in I$, for $0 \leq i, r \leq k$ by

$$(10) \quad H_i^r = \sum_{j=0}^r \binom{r}{j} (E_{i-j})^{(r-j)},$$

where $\binom{r}{j}$ is the usual binomial coefficient and we take $E_q = 0$ for $q < 0$.

LEMMA 2.2. For $1 \leq i \leq k$ and $0 \leq r \leq k-1$ we have

$$\dot{H}_i^r + H_{i-1}^r = H_i^{r+1}.$$

Proof. We simply compute and obtain

$$\begin{aligned} \dot{H}_i^r + H_{i-1}^r &= \sum_{j=0}^r \binom{r}{j} (E_{i-j})^{(r-j+1)} + \sum_{j=0}^r \binom{r}{j} (E_{i-1-j})^{(r-j)} \\ &= \sum_{j=0}^r \binom{r}{j} (E_{i-j})^{(r-j+1)} + \sum_{j=1}^{r+1} \binom{r}{j-1} (E_{i-j})^{(r-j+1)} \\ &= \sum_{j=0}^{r+1} \binom{r+1}{j} (E_{i-j})^{(r+1-j)} \quad \left(\text{since } \binom{r}{j} + \binom{r}{j-1} = \binom{r+1}{j}, 1 \leq j \leq r \right) \\ &= H_i^{r+1}. \end{aligned} \quad \square$$

Remark 2.3. For later reference we observe that

$$(11) \quad H_i^0 = E_i, \quad 0 \leq i \leq k,$$

and $H_0^r = (E_0)^{(r)}$, $0 \leq r \leq k$, so that

$$(12) \quad \dot{H}_0^r = H_0^{r+1}, \quad 0 \leq r \leq k-1.$$

We also define $(p \times m)$ -matrix functions $G_r^l(t)$, $t \in I$, for $r, l \geq 0$ by

$$(13) \quad G_0^0 = D, \quad G_r^0 = 0, \quad r \geq 1,$$

for $l=0$ and for $l \geq 1$ inductively by

$$(14) \quad G_0^l = (\Gamma^{l-1}C)B + \dot{G}_0^{l-1}, \quad G_r^l = G_{r-1}^{l-1} + \dot{G}_r^{l-1}, \quad r \geq 1.$$

It is worthwhile to observe that

$$(15) \quad G_i^l = D, \quad l \geq 0,$$

and

$$(16) \quad G_r^l = 0, \quad l < r.$$

PROPOSITION 2.4. For $0 \leq i, l \leq k$ we have $F_i^l = \sum_{r=0}^l G_r^l H_i^r$.

Proof. We proceed by induction on l . For $l=0$ and $0 \leq i \leq k$ (7), (11) and (13) imply that $G_0^0 H_i^0 = DE_i = F_i^0$, so the formula holds when $l=0$. Let $1 \leq l \leq k$ and assume that the formula holds for $l-1$. Then for $1 \leq i \leq k$ (9b) yields

$$\begin{aligned}
 F_i^l &= (\Gamma^{l-1} C) B E_i + \dot{F}_i^{l-1} + F_{i-1}^{l-1} \\
 &= (\Gamma^{l-1} C) B H_i^0 + \frac{d}{dt} \left(\sum_{r=0}^{l-1} G_r^{l-1} H_i^r \right) + \sum_{r=0}^{l-1} G_r^{l-1} H_{i-1}^r \\
 &\quad \text{(by (11) and the induction assumption)} \\
 &= (\Gamma^{l-1} C) B H_i^0 + \sum_{r=0}^{l-1} (\dot{G}_r^{l-1} H_i^r + G_r^{l-1} \dot{H}_i^r) + \sum_{r=0}^{l-1} G_r^{l-1} H_{i-1}^r \\
 &= (\Gamma^{l-1} C) B H_i^0 + \sum_{r=0}^{l-1} \dot{G}_r^{l-1} H_i^r + \sum_{r=0}^{l-1} G_r^{l-1} (\dot{H}_i^r + H_{i-1}^r) \\
 &= (\Gamma^{l-1} C) B H_i^0 + \sum_{r=0}^{l-1} \dot{G}_r^{l-1} H_i^r + \sum_{r=0}^{l-1} G_r^{l-1} H_i^{r+1} \quad \text{(by Lemma 2.2)} \\
 &= (\Gamma^{l-1} C) B H_i^0 + \sum_{r=0}^{l-1} \dot{G}_r^{l-1} H_i^r + \sum_{r=1}^l G_{r-1}^{l-1} H_i^r \\
 &= ((\Gamma^{l-1} C) B + \dot{G}_0^{l-1}) H_i^0 + \sum_{r=1}^{l-1} (G_{r-1}^{l-1} + \dot{G}_r^{l-1}) H_i^r + G_{l-1}^{l-1} H_i^l.
 \end{aligned}$$

By (14) and (15) we infer that

$$F_i^l = G_0^l H_i^0 + \sum_{r=1}^{l-1} G_r^l H_i^r + G_l^l H_i^l = \sum_{t=0}^l G_t^l H_i^t,$$

which is the desired formula for $1 \leq i \leq k$. A similar computation, starting with (9a), proves the induction step when $i=0$. \square

Proposition 2.4 can be summarized by the following block-matrix equation (note also relation (16)):

$$(17) \quad \begin{bmatrix} F_0^0 & F_1^0 & \cdots & F_k^0 \\ F_0^1 & F_1^1 & \cdots & F_k^1 \\ \vdots & \vdots & & \vdots \\ F_0^k & F_1^k & \cdots & F_k^k \end{bmatrix} = \begin{bmatrix} G_0^0 & 0 & 0 & \cdots & 0 \\ G_0^1 & G_1^1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ G_0^k & G_1^k & G_2^k & \cdots & G_k^k \end{bmatrix} \begin{bmatrix} H_0^0 & H_1^0 & \cdots & H_k^0 \\ H_0^1 & H_1^1 & \cdots & H_k^1 \\ \vdots & \vdots & & \vdots \\ H_0^k & H_1^k & \cdots & H_k^k \end{bmatrix}.$$

Up to this point, the E_0, E_1, \dots, E_k have been arbitrary C^∞ ($m \times p$)-matrix functions of $t \in I$. We will now specify how the E_0, E_1, \dots, E_k should be chosen so as to obtain the desired property of weak C^k reproducibility. In the following proposition I_p denotes the $p \times p$ identity matrix and 0_p denotes the $p \times p$ zero matrix.

PROPOSITION 2.5. *Suppose that there exist C^∞ ($m \times p$)-matrix functions K_0, K_1, \dots, K_k of $t \in I$ such that*

$$(18) \quad \begin{bmatrix} G_0^0 & 0 & 0 & \cdots & 0 \\ G_0^1 & G_1^1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ G_0^k & G_1^k & G_2^k & \cdots & G_k^k \end{bmatrix} \begin{bmatrix} K_0 \\ K_1 \\ \vdots \\ K_k \end{bmatrix} = \begin{bmatrix} 0_p \\ 0_p \\ \vdots \\ I_p \end{bmatrix}$$

at every $t \in I$. Then there exist C^∞ ($m \times p$)-matrix functions E_0, E_1, \dots, E_k of $t \in I$ such that for the $(m \times p)$ -matrix functions H_i^r defined by (10) we have

$$(19) \quad H_k^r = K_r, \quad 0 \leq r \leq k.$$

Furthermore, for the $(p \times p)$ -matrix functions F_i^l defined by (7) and (9) we have

$$(20) \quad F_k^0 = F_k^1 = \cdots = F_k^{k-1} = 0_p \quad \text{and} \quad F_k^k = I_p$$

at every $t \in I$.

Proof. By the definition of the matrix functions H_i^r in terms of the matrix functions E_0, E_1, \dots, E_k (see (10)), in order to satisfy (19) we must choose E_0, E_1, \dots, E_k so that

$$(21) \quad K_r = \sum_{j=0}^r \binom{r}{j} (E_{k-j})^{(r-j)}$$

for $0 \leq r \leq k$. If we set $E_k = K_0$, then (21) is satisfied when $r = 0$ (recall the convention that $E_q = 0$ if $q < 0$). Let $1 \leq s \leq k$ and assume that E_k, \dots, E_{k-s+1} have been chosen so that (21) is satisfied for $0 \leq r \leq s-1$. Setting

$$E_{k-s} = K_s - \sum_{j=0}^{s-1} \binom{s}{j} (E_{k-j})^{(s-j)},$$

we see that (21) is satisfied for $r = s$. By induction we can choose E_0, E_1, \dots, E_k so as to satisfy (21) (and hence (19)) for $0 \leq r \leq k$. The relations (20) follow by using (18) and (19) and equating the last columns of both sides of (17). \square

COROLLARY 2.6. *Suppose that there exist C^∞ $(m \times p)$ -matrix functions K_0, K_1, \dots, K_k of $t \in I$ that satisfy (18) and let E_0, E_1, \dots, E_k be as given in Proposition 2.4. Then for every input u of the form (2), where $v \in \mathcal{C}^{k,p}(I)$, and every initial condition $(t_0, x_0) \in I \times \mathbb{R}^n$, the corresponding output ψ of system (1) is in $\mathcal{C}^{k,p}(I)$ and we have*

$$(22) \quad \psi^{(l)} = (\Gamma^l C) \phi + \sum_{i=0}^{k-1} F_i^l v^{(i)}, \quad 0 \leq l \leq k-1,$$

and

$$(23) \quad \psi^{(k)} = (\Gamma^k C) \phi + \sum_{i=0}^{k-1} F_i^k v^{(i)} + v^{(k)}$$

where ϕ satisfies (4).

Proof. This follows from (8) with the use of (20), (7) and (9). \square

Equation (23) is the key factor in determining the required input u from the desired output ψ . Observe that, for the matrix functions E_0, E_1, \dots, E_k given by Proposition 2.5, we do not have to assume that v has more than k derivatives to compute $\psi^{(1)}, \dots, \psi^{(k)}$, since the k th derivative of v does not appear until we compute $\psi^{(k)}$. The next task is to find a sufficient condition for (18) to hold. This will be done subsequent to the next two lemmas, the first of which is a modification of a theorem of Dolezal [4].

LEMMA 2.7. *Let N be a positive integer, let $\{\omega_1, \dots, \omega_q\}$ be a finite subset of $\mathcal{C}^{r,N}(I)$, $0 \leq r \leq \infty$, let $\Omega(t) \subseteq \mathbb{R}^N$ for $t \in I$ denote the subspace*

$$\Omega(t) = \text{span} [\omega_1(t), \dots, \omega_q(t)],$$

and assume that for every $t \in I$ we have $\dim \Omega(t) = \rho$, where ρ is a constant satisfying $\rho < N$. Let $\eta \in \mathcal{C}^{r,N}(I)$ be such that $\eta(t) \notin \Omega(t)$ for every $t \in I$. Then there exists $e \in \mathcal{C}^{r,N}(I)$ such that for every $t \in I$

$$\langle \eta(t), e(t) \rangle = 1$$

and

$$\langle \omega_i(t), e(t) \rangle = 0, \quad 1 \leq i \leq q,$$

where \langle, \rangle is the standard inner product on \mathbb{R}^N .

Proof. For each $t \in I$ let $\eta_0(t)$ denote the orthogonal projection of $\eta(t)$ onto the subspace $\Omega(t)$ and observe that $\|\eta(t) - \eta_0(t)\| > 0$. We will first show that η_0 is C^r and to do this it suffices to show that for every $\bar{t} \in I$ there is a relatively open neighborhood $U_{\bar{t}}$ of \bar{t} in I such that $\eta_0|_{U_{\bar{t}}}$ is C^r . Indeed, given $\bar{t} \in I$ we have $\dim \Omega(\bar{t}) = \rho$ by assumption, so there exists a subset $\{k_1, \dots, k_\rho\}$ of $\{1, \dots, q\}$ such that $\{\omega_{k_i}(\bar{t}) | 1 \leq i \leq \rho\}$ is a basis for $\Omega(\bar{t})$. Since the ω_{k_i} 's are at least continuous, the function

$$\text{rank} [\omega_{k_1}(t), \dots, \omega_{k_\rho}(t)]$$

is locally nondecreasing, so there exists a relatively open neighborhood $U_{\bar{t}}$ of \bar{t} in I such that

$$t \in U_{\bar{t}} \Rightarrow \rho \leq \text{rank} [\omega_{k_1}(t), \dots, \omega_{k_\rho}(t)].$$

On the other hand, for every $t \in I$ we have

$$\text{rank} [\omega_{k_1}(t), \dots, \omega_{k_\rho}(t)] \leq \dim \Omega(t) = \rho,$$

so we infer that

$$t \in U_{\bar{t}} \Rightarrow \text{rank} [\omega_{k_1}(t), \dots, \omega_{k_\rho}(t)] = \rho.$$

Consequently, $\{\omega_{k_i}(t) | 1 \leq i \leq \rho\}$ is a basis of $\Omega(t)$ for every $t \in U_{\bar{t}}$. By definition $\eta_0(t) \in \Omega(t)$ for every $t \in I$, so for $t \in U_{\bar{t}}$ there exist real-valued functions $\alpha_1(t), \dots, \alpha_\rho(t)$ such that

$$(24) \quad \eta_0(t) = \sum_{j=1}^{\rho} \alpha_j(t) \omega_{k_j}(t).$$

The relations

$$\langle \eta(t) - \eta_0(t), \omega_{k_i}(t) \rangle = 0, \quad 1 \leq i \leq \rho$$

imply that for each $t \in U_{\bar{t}}$ the ρ -tuple $(\alpha_1(t), \dots, \alpha_\rho(t))$ is a solution of the linear system of equations

$$(25) \quad \sum_{j=1}^{\rho} \langle \omega_{k_j}(t), \omega_{k_i}(t) \rangle \alpha_j(t) = \langle \eta(t), \omega_{k_i}(t) \rangle, \quad 1 \leq i \leq \rho.$$

The linear independence of $\{\omega_{k_i}(t) | 1 \leq i \leq \rho\}$ for $t \in U_{\bar{t}}$ implies that the $\rho \times \rho$ coefficient matrix of (25) is nonsingular. Therefore the linear system (25) has a unique solution $(\alpha_1(t), \dots, \alpha_\rho(t))$ and by Cramer's rule this solution is a C^r function of $t \in U_{\bar{t}}$, since the ω_{k_i} and η are C^r . From (24) it is clear that $\eta_0|_{U_{\bar{t}}}$ is C^r and, as was mentioned earlier, this is enough to conclude that η_0 is C^r on I .

From the relations

$$\|\eta(t) - \eta_0(t)\| > 0, \quad \langle \eta(t) - \eta_0(t), \eta_0(t) \rangle = 0,$$

we obtain

$$0 < \|\eta(t) - \eta_0(t)\|^2 = \langle \eta(t) - \eta_0(t), \eta(t) \rangle,$$

for every $t \in I$, so the function $e: I \rightarrow \mathbb{R}^N$ given by

$$e(t) = \langle \eta(t) - \eta_0(t), \eta(t) \rangle^{-1} (\eta(t) - \eta_0(t))$$

is well defined, of class C^r , and clearly satisfies our requirements. \square

LEMMA 2.8. Let p, q, N be positive integers, let $W(t)$ be a time-varying $(q \times N)$ -matrix function of $t \in I$, let $Z(t)$ be a time-varying $(p \times N)$ -matrix function of $t \in I$, and assume that the entries of $W(t)$ and $Z(t)$ are of class C^r , $0 \leq r \leq \infty$. Suppose that

$$(26) \quad \text{rank } W(t) = \text{const.} \quad \text{for } t \in I$$

and

$$(27) \quad \text{rank} \begin{bmatrix} W(t) \\ Z(t) \end{bmatrix} - \text{rank } W(t) = p \quad \text{for } t \in I.$$

Then there exists an $(N \times p)$ -matrix function $K(t)$ of $t \in I$ with C^r entries such that

$$(28) \quad \begin{bmatrix} W(t) \\ Z(t) \end{bmatrix} K(t) = \begin{bmatrix} 0_{q \times p} \\ I_p \end{bmatrix} \quad \text{for } t \in I.$$

Proof. Let $\omega_1(t), \dots, \omega_q(t)$ denote the N -dimensional row vectors of $W(t)$ and let $\eta_1(t), \dots, \eta_p(t)$ denote the N -dimensional row vectors of $Z(t)$. For each $t \in I$ define subspaces of \mathbb{R}^N by

$$\Omega_0(t) = \text{span} [\omega_1(t), \dots, \omega_q(t)]$$

and

$$\Omega_i(t) = \text{span} [\{\omega_1(t), \dots, \omega_q(t), \eta_1(t), \dots, \eta_p(t)\} \setminus \{\eta_i(t)\}], \quad 1 \leq i \leq p.$$

From (26) and (27) we see that for $t \in I$

$$\eta_i(t) \notin \Omega_i(t), \quad 1 \leq i \leq p,$$

and

$$\dim \Omega_i(t) = \dim \Omega_0(t) + (p - 1) = \text{const.}, \quad 1 \leq i \leq p.$$

Thus for each $i = 1, \dots, p$ we can apply Lemma 2.7 to obtain a function $e_i \in \mathcal{C}^{r,N}(I)$ such that for $t \in I$

$$(29) \quad \begin{aligned} \langle \omega_j(t), e_i(t) \rangle &= 0, & 1 \leq j \leq q, \\ \langle \eta_j(t), e_i(t) \rangle &= 0, & 1 \leq j \leq p, \quad j \neq i, \\ \langle \eta_i(t), e_i(t) \rangle &= 1. \end{aligned}$$

Form an $(N \times p)$ -matrix function $K(t)$ in such a way that its i th column consists of the N components of the vector $e_i(t)$, $1 \leq i \leq p$. Then the relations (29) show that $K(t)$ satisfies (28). \square

We will associate to the input-output system (1) a family $\{M_k(t) | k \geq 0\}$ of C^∞ block-lower-triangular matrix functions of $t \in I$ defined by

$$(30) \quad M_k = \begin{bmatrix} G_0^0 & 0 & 0 & \cdots & 0 \\ G_0^1 & G_1^1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ G_0^k & G_1^k & G_2^k & \cdots & G_k^k \end{bmatrix}$$

where the matrices G_r^l are as defined in (13) and (14). Observe that the dimensions of M_k are $(k+1)p$ by $(k+1)m$. It is perhaps instructive to write out one of

the M_k 's explicitly directly in terms of the system matrices A, B, C, D ; for example,

$$M_3 = \begin{bmatrix} D & 0 & 0 & 0 \\ CB + \dot{D} & D & 0 & 0 \\ (\Gamma C)B + \overline{CB} + \ddot{D} & CB + 2\dot{D} & D & 0 \\ (\Gamma^2 C)B + (\overline{\Gamma C})\overline{B} + \overline{\overline{CB}} + \ddot{D} & (\Gamma C)B + 2\overline{CB} + 3\ddot{D} & CB + 3\dot{D} & D \end{bmatrix}.$$

PROPOSITION 2.9. Let k be a positive integer and suppose that

$$(31) \quad \text{rank } M_{k-1}(t) = \text{const. for every } t \in I$$

and

$$(32) \quad \text{rank } M_k(t) - \text{rank } M_{k-1}(t) = p \quad \text{for every } t \in I.$$

Then there exist $C^\infty(m \times p)$ -matrix functions K_0, K_1, \dots, K_k of $t \in I$ such that (18) holds everywhere on I .

Proof. From the form of the matrix M_k it is clear that the rank of the first kp rows of M_k equals the rank of M_{k-1} , which is constant on I by (31). Assumption (32) allows us to use Lemma 2.8 to obtain a $C^\infty((k+1)m \times p)$ -matrix function $K(t)$ of $t \in I$ such that

$$M_k(t)K(t) = \begin{bmatrix} 0_{kp \times p} \\ I_p \end{bmatrix}.$$

If we partition the matrix $K(t)$ into a column of $k+1$ $m \times p$ matrices as

$$K(t) = \begin{bmatrix} K_0(t) \\ \vdots \\ K_k(t) \end{bmatrix},$$

then the matrix functions K_0, K_1, \dots, K_k are seen to satisfy our requirements. \square

We will fit together the results obtained thus far in the following theorem.

THEOREM 2.10. Consider the input-output system (1) with C^∞ system matrices A, B, C, D and for $k \geq 0$ let M_k be as defined in (30), where the block entries G_r^l are as defined in (13) and (14). Suppose that for some $k \geq 1$ the rank conditions (31) and (32) hold. Then for every $\psi \in \mathcal{C}^{k,p}(I)$ and for every initial condition $(t_0, x_0) \in I \times \mathbb{R}^n$ there exists $u \in \mathcal{C}^{0,m}(I)$ such that the output $\tilde{\psi}$ of (1) corresponding to the input u and initial condition (t_0, x_0) satisfies $\tilde{\psi} \in \mathcal{C}^{k,p}(I)$ and $\tilde{\psi}^{(k)} = \psi^{(k)}$.

Proof. By Proposition 2.9 there exist $C^\infty(m \times p)$ -matrix functions K_0, K_1, \dots, K_k of $t \in I$ such that (18) holds everywhere on I . Then from Proposition 2.5 we obtain $C^\infty(m \times p)$ -matrix functions E_0, E_1, \dots, E_k such that the matrix functions F_i^l defined by (7) and (9) satisfy

$$F_k^0 = \dots = F_k^{k-1} = 0_p \quad \text{and} \quad F_k^k = I_p.$$

Let $\psi \in \mathcal{C}^{k,p}(I)$ and $(t_0, x_0) \in I \times \mathbb{R}^n$ be given and consider the linear time-varying, nonhomogeneous system of ODEs of order $n + kp$ in the vector variables $x \in \mathbb{R}^n$ and $z_0, \dots, z_{k-1} \in \mathbb{R}^p$ given by

$$(33) \quad \begin{aligned} \dot{x} &= A(t)x + \sum_{i=0}^{k-1} B(t)E_i(t)z_i + B(t)E_k(t) \cdot \left\{ \psi^{(k)}(t) - (\Gamma^k C)(t)x - \sum_{i=0}^{k-1} F_i^k(t)z_i \right\}, \\ \dot{z}_0 &= z_1, \\ &\vdots \\ \dot{z}_{k-2} &= z_{k-1}, \\ \dot{z}_{k-1} &= \psi^{(k)}(t) - (\Gamma^k C)(t)x - \sum_{i=0}^{k-1} F_i^k(t)z_i. \end{aligned}$$

Choose an initial condition of the form $(t_0, x_0, w_0, \dots, w_{k-1}) \in I \times \mathbb{R}^{n+kp}$, where the $w_i \in \mathbb{R}^p$ are arbitrary, $0 \leq i \leq k-1$, and for $t \in I$ let $(\phi(t), \zeta_0(t), \dots, \zeta_{k-1}(t))$ denote the solution of (33) subject to this initial condition. If we set $v(t) = \zeta_0(t)$, then (33) implies that

$$\begin{aligned} v^{(1)}(t) &= \dot{\zeta}_0(t) = \zeta_1(t), \\ &\vdots \\ v^{(k-1)}(t) &= \dot{\zeta}_{k-2}(t) = \zeta_{k-1}(t), \end{aligned}$$

and

$$(34) \quad v^{(k)}(t) = \psi^{(k)}(t) - (\Gamma^k C)(t)\phi(t) - \sum_{i=0}^{k-1} F_i^k(t)v^{(i)}(t);$$

in particular, $v \in \mathcal{C}^{k,p}(I)$. For $t \in I$ let

$$(35) \quad u(t) = \sum_{i=0}^k E_i(t)v^{(i)}(t).$$

Then $u \in \mathcal{C}^{0,m}(I)$, ϕ satisfies $\phi(t_0) = x_0$, and

$$\dot{\phi}(t) = A(t)\phi(t) + B(t)u(t)$$

for every $t \in I$. Setting $\tilde{\psi}(t) = C(t)\phi(t) + D(t)u(t)$, we see by (23) that for every $t \in I$

$$\begin{aligned} \tilde{\psi}^{(k)}(t) &= (\Gamma^k C)(t)\phi(t) + \sum_{i=0}^{k-1} F_i^k(t)v^{(i)}(t) + v^{(k)}(t) \\ &= \psi^{(k)}(t) \end{aligned}$$

where the last equality follows from (34). This completes the proof. \square

A concise restatement of Theorem 2.10 is that, in the presence of the rank conditions (31) and (32), given any $\psi \in \mathcal{C}^{k,p}(I)$ we can generate an output $\tilde{\psi}$ of (1) such that $\tilde{\psi}^{(k)} = \psi^{(k)}$. The initial condition is arbitrary and the input is of the form (35). The next step is to show that if the initial condition is chosen properly, then we can obtain $\tilde{\psi} = \psi$. We first observe that by (22) (Corollary 2.6) for every initial condition $(t_0, x_0) \in I \times \mathbb{R}^n$ we have

$$(36) \quad \tilde{\psi}^{(l)}(t_0) = (\Gamma^l C)(t_0)x_0 + \sum_{i=0}^{k-1} F_i^l(t_0)w_i, \quad 0 \leq l \leq k-1$$

where $w_i = \zeta_i(t_0) = v^{(i)}(t_0)$ for $0 \leq i \leq k-1$. Suppose that for a given output $\psi \in \mathcal{C}^{k,p}(I)$ we can find an initial condition $(t_0, x_0, w_0, \dots, w_{k-1})$ for the system (33) that satisfies

$$(37) \quad \psi^{(l)}(t_0) = (\Gamma^l C)(t_0)x_0 + \sum_{i=0}^{k-1} F_i^l(t_0)w_i, \quad 0 \leq l \leq k-1.$$

In this case, comparing (36) and (37), we see that the output $\tilde{\psi}$ of (1) generated by the initial condition (t_0, x_0) and the input (35) satisfies

$$\tilde{\psi}^{(l)}(t_0) = \psi^{(l)}(t_0), \quad 0 \leq l \leq k-1,$$

and

$$\tilde{\psi}^{(k)}(t) = \psi^{(k)}(t), \quad t \in I,$$

which clearly implies that $\tilde{\psi} = \psi$. Thus to generate every $\psi \in \mathcal{C}^{k,p}(I)$ as an output of

(1) it is sufficient that the rank conditions (31), (32) hold and that for every $\psi \in \mathcal{C}^{k,p}(I)$ there exist $(t_0, x_0, w_0, \dots, w_{k-1}) \in I \times \mathbb{R}^{n+kp}$ satisfying (37).

We can rewrite the system of equations (37) in matrix form:

$$(38) \quad \begin{bmatrix} \psi^{(0)}(t_0) \\ \psi^{(1)}(t_0) \\ \vdots \\ \psi^{(k-1)}(t_0) \end{bmatrix} = \begin{bmatrix} (\Gamma^0 C)(t_0) & F_0^0(t_0) & \cdots & F_{k-1}^0(t_0) \\ (\Gamma^1 C)(t_0) & F_0^1(t_0) & \cdots & F_{k-1}^1(t_0) \\ \vdots & \vdots & \ddots & \vdots \\ (\Gamma^{k-1} C)(t_0) & F_0^{k-1}(t_0) & \cdots & F_{k-1}^{k-1}(t_0) \end{bmatrix} \begin{bmatrix} x_0 \\ w_0 \\ \vdots \\ w_{k-1} \end{bmatrix}.$$

It is clear that (38) will have a solution for every $\psi \in \mathcal{C}^{k,p}(I)$ if and only if the matrix

$$(39) \quad S_k(t) = \begin{bmatrix} (\Gamma^0 C)(t) & F_0^0(t) & \cdots & F_{k-1}^0(t) \\ (\Gamma^1 C)(t) & F_0^1(t) & \cdots & F_{k-1}^1(t) \\ \vdots & \vdots & \ddots & \vdots \\ (\Gamma^{k-1} C)(t) & F_0^{k-1}(t) & \cdots & F_{k-1}^{k-1}(t) \end{bmatrix}$$

has full rank kp when $t = t_0$. It turns out that the rank conditions (31) and (32) imply that $\text{rank } S_k(t) = kp$ for every t in an open dense subset of I (later we will improve this to $\text{rank } S_k(t) = kp$ for every t in I). The proof of this fact requires two lemmas.

LEMMA 2.11. *Let η be a C^∞ p -dimensional (column) vector function of $t \in I$ and let C be a C^∞ $(p \times n)$ -matrix function of $t \in I$. Then for every nonnegative integer r the operator Γ defined in (6) satisfies*

$$\Gamma^r(\eta^T C) = \sum_{q=0}^r \binom{r}{q} (\eta^T)^{(q)} \Gamma^{r-q} C$$

where the superscript T denotes transpose.

Proof. The proof is a straightforward induction on r . \square

LEMMA 2.12. *Let E_0, E_1, \dots, E_k ($k \geq 1$) be arbitrary C^∞ $(m \times p)$ -matrix functions of $t \in I$ and let the C^∞ $(p \times p)$ -matrix functions F_i^l of $t \in I$ be as defined in (7) and (9). Let $1 \leq l \leq k$, let J be an open subinterval of I , and let $\eta_0, \dots, \eta_{l-1}$ be C^∞ p -dimensional (column) vector functions of $t \in I$ such that*

$$(40) \quad [\eta_0^T, \dots, \eta_{l-1}^T] \begin{bmatrix} \Gamma^0 C & F_0^0 & \cdots & F_{k-1}^0 \\ \Gamma^1 C & F_0^1 & \cdots & F_{k-1}^1 \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma^{l-1} C & F_0^{l-1} & \cdots & F_{k-1}^{l-1} \end{bmatrix} = 0_{1 \times (n+kp)}$$

for every $t \in J$. Then for every $t \in J$ and $0 \leq i \leq k-1$ we have

$$(41) \quad \sum_{j=0}^{l-1} \sum_{q=0}^{k-l} \binom{k-l}{q} (\eta_j^T)^{(q)} F_i^{j+k-l-q} = 0.$$

Proof. We will prove by induction on $r \in \{0, \dots, k-l\}$ that for $t \in J$ and $0 \leq i \leq k-1$ we have

$$(42) \quad \sum_{j=0}^{l-1} \sum_{q=0}^r \binom{r}{q} (\eta_j^T)^{(q)} F_i^{j+r-q} = 0.$$

Formula (41) will then follow from (42) by setting $r = k-l$.

For $r = 0$ we have

$$\sum_{j=0}^{l-1} \sum_{q=0}^r \binom{r}{q} (\eta_j^T)^{(q)} F_i^{j+r-q} = \sum_{j=0}^{l-1} \eta_j^T F_i^j$$

and (40) implies that this expression vanishes for $t \in J$ and $0 \leq i \leq k-1$. Thus (42) holds when $r = 0$.

Let $0 \leq r \leq k-l-1$ and assume that (42) holds for r . Then for $1 \leq i \leq k-1$ we can replace i by $i-1$ in (42) to obtain

$$(43) \quad \sum_{j=0}^{l-1} \sum_{q=0}^r \binom{r}{q} (\eta_j^T)^{(q)} F_{i-1}^{j+r-q} = 0,$$

and we can differentiate (42) to obtain

$$(44) \quad \sum_{j=0}^{l-1} \sum_{q=0}^r \binom{r}{q} \{(\eta_j^T)^{(q+1)} F_i^{j+r-q} + (\eta_j^T)^{(q)} \dot{F}_i^{j+r-q}\} = 0.$$

Equation (40) implies that

$$\sum_{j=0}^{l-1} \eta_j^T \Gamma^j C = 0$$

and if we apply Γ^r to this equation and use Lemma 2.11 we get

$$0 = \sum_{j=0}^{l-1} \Gamma^r (\eta_j^T \Gamma^j C) = \sum_{j=0}^{l-1} \sum_{q=0}^r \binom{r}{q} (\eta_j^T)^{(q)} \Gamma^{r-q} (\Gamma^j C).$$

Multiplication of this last equation by BE_i on the right yields

$$(45) \quad \sum_{j=0}^{l-1} \sum_{q=0}^r \binom{r}{q} (\eta_j^T)^{(q)} (\Gamma^{j+r-q} C) BE_i = 0.$$

We can now add (43)–(45) to obtain for $1 \leq i \leq k-1$ and $t \in J$

$$\begin{aligned} 0 &= \sum_{j=0}^{l-1} \sum_{q=0}^r \binom{r}{q} [(\eta_j^T)^{(q+1)} F_i^{j+r-q} + (\eta_j^T)^{(q)} \{(\Gamma^{j+r-q} C) BE_i + \dot{F}_i^{j+r-q} + F_{i-1}^{j+r-q}\}] \\ &= \sum_{j=0}^{l-1} \sum_{q=0}^r \binom{r}{q} [(\eta_j^T)^{(q+1)} F_i^{j+r-1} + (\eta_j^T)^{(q)} F_i^{j+(r+1)-q}] \quad (\text{by (9b)}) \\ &= \sum_{j=0}^{l-1} \sum_{q=0}^{r+1} \binom{r+1}{q} (\eta_j^T)^{(q)} F_i^{j+(r+1)-q} \end{aligned}$$

This shows that (42) holds for $1 \leq i \leq k-1$ with r replaced by $r+1$. A similar argument works when $i=0$ (using (9a) instead of (9b)), so this completes the induction and the proof. \square

PROPOSITION 2.13. *Suppose that for some $k \geq 1$ the matrix function M_k defined in (30) satisfies the rank conditions (31) and (32). Let E_0, E_1, \dots, E_k be $C^\infty(m \times p)$ -matrix functions of $t \in I$ such that the $C^\infty(p \times p)$ -matrix functions F_i^t of $t \in I$ defined by (7) and (9) satisfy (20). Then the matrix $S_k(t)$ defined in (39) has rank kp for all t in an open dense subset of I .*

Proof. The existence of the matrix functions E_0, E_1, \dots, E_k follows from Propositions 2.9 and 2.5. Since the entries of the matrix $S_k(t)$ are C^∞ functions of t , the set

$$V = \{t \in I \mid \text{rank } S_k(t) = kp\}$$

is open relative to I . Thus it suffices to show that V is dense in I .

If V is not dense in I , then there exists a nonempty open subinterval J_0 of I such that

$$t \in J_0 \Rightarrow \text{rank } S_k(t) < kp.$$

Let

$$\rho = \max \{ \text{rank } S_k(t) \mid t \in J_0 \} < kp;$$

then there exists a nonempty open subinterval J of J_0 such that

$$t \in J \Rightarrow \text{rank } S_k(t) = \rho.$$

Since $\rho < kp$, Dolezal's theorem [4] (or, alternatively, an argument similar to that in the proof of Lemma 2.7) implies the existence of a nowhere zero C^∞ function $\eta: J \rightarrow \mathbb{R}^{kp}$ such that

$$(46) \quad \eta(t)^T S_k(t) = 0_{1 \times (n+kp)} \quad \text{for every } t \in J.$$

Partitioning the kp -dimensional (column) vector function η into k p -dimensional column vector functions $\eta_0, \dots, \eta_{k-1}$ we can rewrite (46) as

$$(47) \quad [\eta_0^T, \dots, \eta_{k-1}^T] \begin{bmatrix} \Gamma^0 C & F_0^0 & \dots & F_{k-1}^0 \\ \Gamma^1 C & F_0^1 & \dots & F_{k-1}^1 \\ \vdots & \vdots & & \vdots \\ \Gamma^{k-1} C & F_0^{k-1} & \dots & F_{k-1}^{k-1} \end{bmatrix} = 0_{1 \times (n+kp)}$$

for every $t \in J$. The contradiction will be obtained by showing that each of the vector functions $\eta_{k-1}, \dots, \eta_0$ must vanish identically on J if (47) is to hold.

We first show that η_{k-1} must vanish on J . From (47) we obtain the equations

$$\begin{aligned} \eta_0^T \Gamma^0 C + \dots + \eta_{k-2}^T \Gamma^{k-2} C + \eta_{k-1}^T \Gamma^{k-1} C &= 0, \\ \eta_0^T F_{k-1}^0 + \dots + \eta_{k-2}^T F_{k-1}^{k-2} + \eta_{k-1}^T F_{k-1}^{k-1} &= 0. \end{aligned}$$

Multiply the first by BE_k on the left and add to the second to get

$$(48) \quad \begin{aligned} \eta_0^T ((\Gamma^0 C)BE_k + F_{k-1}^0) + \dots + \eta_{k-2}^T ((\Gamma^{k-2} C)BE_k + F_{k-1}^{k-2}) \\ + \eta_{k-1}^T ((\Gamma^{k-1} C)BE_k + F_{k-1}^{k-1}) = 0. \end{aligned}$$

From (9b) and (20) we see that for $1 \leq j \leq k$

$$\begin{aligned} F_k^j &= (\Gamma^{j-1} C)BE_k + \dot{F}_{k-1}^{j-1} + F_{k-1}^{j-1} \\ &= (\Gamma^{j-1} C)BE_k + F_{k-1}^{j-1} = \begin{cases} 0_p, & 1 \leq j \leq k-1, \\ I_p, & j = k, \end{cases} \end{aligned}$$

and using these relations in (48) we obtain $\eta_{k-1} = 0$ on J .

Next we suppose we have shown that $\eta_{k-1} = \dots = \eta_l = 0$ on J , where $1 \leq l \leq k-1$; from this we want to infer that $\eta_{l-1} = 0$. The induction assumption implies that (47) reduces to (40). Thus we have $\sum_{j=0}^{l-1} \eta_j^T \Gamma^j C = 0$, which by Lemma 2.11 implies

$$0 = \Gamma^{k-l} \left(\sum_{j=0}^{l-1} \eta_j^T \Gamma^j C \right) = \sum_{j=0}^{l-1} \sum_{q=0}^{k-l} \binom{k-l}{q} (\eta_j^T)^{(q)} \Gamma^{k-l-q} (\Gamma^j C).$$

Multiplying this equation on the right by BE_k , we obtain

$$(49) \quad \sum_{j=0}^{l-1} \sum_{q=0}^{k-l} \binom{k-l}{q} (\eta_j^T)^{(q)} (\Gamma^{j+k-l-q} C) BE_k = 0.$$

Equation (40) also implies (41) (Lemma 2.12), so we can add (41), with $i = k-1$, to (49) and get

$$(50) \quad \sum_{j=0}^{l-1} \sum_{q=0}^{k-l} \binom{k-l}{q} (\eta_j^T)^{(q)} ((\Gamma^{j+k-l-q} C) BE_k + F_{k-1}^{j+k-l-q}) = 0.$$

Again from (9b) and (20) we see that

$$(\Gamma^{j+k-l-q}C)BE_k + F_{k-1}^{j+k-l-q} = F_k^{j+k-l-q+1} = \begin{cases} 0_p, & 0 \leq j+k-l-q+1 \leq k-1, \\ I_p, & j+k-l-q+1 = k. \end{cases}$$

But given the inequalities $0 \leq j \leq l-1$ and $0 \leq q \leq k-l$, it is clear that $j+k-l-q+1 = k \Leftrightarrow j = l-1$ and $q = 0$, so (50) reduces to $(\begin{smallmatrix} k-l \\ 0 \end{smallmatrix})(\eta_{l-1}^T)^{(0)} = 0$, or $\eta_{l-1} = 0$ on J . By induction, we infer that $\eta_{k-1} = \dots = \eta_0 = 0$ on J . However, this contradicts the fact that the function $\eta: J \rightarrow \mathbb{R}^{kp}$ is nowhere zero, and the proof is complete. \square

We can now prove the principal result of this paper. Some of the earlier notation will be repeated here so as to make the statement of the theorem reasonably self-contained.

THEOREM 2.14. *Consider the input-output system (1) with C^∞ system matrices A , B , C , D and for $k \geq 0$ let M_k be the time-varying matrix function defined by*

$$M_k = \begin{bmatrix} G_0^0 & 0 & 0 & \cdots & 0 \\ G_0^1 & G_1^1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ G_0^k & G_1^k & G_2^k & \cdots & G_k^k \end{bmatrix}$$

where

$$G_0^0 = D, \quad G_r^0 = 0, \quad r \geq 1$$

and for $l \geq 1$

$$G_0^l = (\Gamma^{l-1}C)B + \dot{G}_0^{l-1}, \quad G_r^l = G_{r-1}^{l-1} + \dot{G}_r^{l-1}, \quad r \geq 1.$$

Let $k \geq 1$ and suppose that

$$\text{rank } M_{k-1}(t) = \text{const.} \quad \text{for every } t \in I$$

and

$$\text{rank } M_k(t) - \text{rank } M_{k-1}(t) = p \quad \text{for every } t \in I.$$

Then for every $\psi \in \mathcal{C}^{k,p}(I)$ there exists $u \in \mathcal{C}^{0,m}(I)$ and $(t_0, x_0) \in I \times \mathbb{R}^n$ such that ψ is the output of (1) corresponding to the input u and the initial condition $x(t_0) = x_0$; that is, system (1) is weakly C^k -reproducible on I .

Proof. Under the assumptions of this theorem, we saw in Theorem 2.10 that for every $\psi \in \mathcal{C}^{k,p}(I)$ and for every $(t_0, x_0) \in I \times \mathbb{R}^n$ there exists $v \in \mathcal{C}^{k,p}(I)$ such that for $u \in \mathcal{C}^{0,m}(I)$ defined by

$$(51) \quad u(t) = \sum_{i=0}^k E_i(t)v^{(i)}(t)$$

the output $\tilde{\psi}$ of (1) corresponding to the input u and initial condition $x(t_0) = x_0$ satisfies $\psi \in \mathcal{C}^{k,p}(I)$ and $\tilde{\psi}^{(k)} = \psi^{(k)}$. The E_0, E_1, \dots, E_k are C^∞ ($m \times p$)-matrix functions of $t \in I$ that are independent of the desired output ψ and are chosen so that the $(p \times p)$ -matrix functions F_i^l defined in (7) and (9) satisfy (20). The function v is in $\mathcal{C}^{k,p}(I)$ and $v(t) = \zeta_0(t)$, where $(\phi(t), \zeta_0(t), \dots, \zeta_{k-1}(t))$ is a solution of the linear system of ODEs (33) subject to the initial conditions $\phi(t_0) = x_0$ and $\zeta_i(t_0)$ arbitrary, $0 \leq i \leq k-1$. From the form of the system (33) it is clear that $v^{(i)}(t) = \zeta_i(t)$ for $0 \leq i \leq k-1$.

Equation (22) of Corollary 2.6 shows that for every initial condition $\phi(t_0) = x_0$ the output $\tilde{\psi}$ of (1) corresponding to the input (51) satisfies

$$(52) \quad \begin{aligned} \tilde{\psi}^{(l)}(t_0) &= (\Gamma^l C)(t_0)\phi(t_0) + \sum_{i=0}^{k-1} F_i^l(t_0)v^{(i)}(t_0), \quad 0 \leq l \leq k-1 \\ \text{or} \\ \tilde{\psi}^{(l)}(t_0) &= (\Gamma^l C)(t_0)x_0 + \sum_{i=0}^{k-1} F_i^l(t_0)\zeta_i(t_0), \quad 0 \leq l \leq k-1. \end{aligned}$$

Now by Proposition 2.13 there exists $t_0 \in I$ such that the matrix $S_k(t_0)$ defined in (39) has rank kp . Hence given any $\psi \in \mathcal{C}^{k,p}(I)$ there exists $(x_0, w_0, \dots, w_{k-1}) \in \mathbb{R}^{n+kp}$ such that

$$(53) \quad \text{col} [\psi(t_0), \psi^{(1)}(t_0), \dots, \psi^{(k-1)}(t_0)] = S_k(t_0) \text{col} [x_0, w_0, \dots, w_{k-1}].$$

If we solve system (33) subject to the initial condition $(t_0, x_0, w_0, \dots, w_{k-1})$ given by (53), if we denote the solution by $(\phi(t), \zeta_0(t), \dots, \zeta_{k-1}(t))$, and if we define $u \in \mathcal{C}^{0,m}(I)$ by (51), where $v(t) = \zeta_0(t)$, then the output $\tilde{\psi}$ of (1) corresponding to the input u and the initial condition (t_0, x_0) satisfies $\tilde{\psi}^{(k)} = \psi^{(k)}$ by Theorem 2.10 and

$$\tilde{\psi}^{(l)}(t_0) = \psi^{(l)}(t_0), \quad 0 \leq l \leq k-1,$$

by (52) and (53). We infer that $\tilde{\psi} = \psi$ and, since $\psi \in \mathcal{C}^{k,p}(I)$ was arbitrary, system (1) is weakly C^k -reproducible on I . \square

Remark 2.15. Contained in the proof of Theorem 2.14 is a procedure for obtaining the required initial condition $(t_0, x_0) \in I \times \mathbb{R}^n$ and input $u \in \mathcal{C}^{0,m}(I)$ from the desired output $\psi \in \mathcal{C}^{k,p}(I)$. Choose $t_0 \in I$ so that $\text{rank } S_k(t_0) = kp$ and let $R_k(t_0)$ be a right inverse for $S_k(t_0)$; that is,

$$S_k(t_0)R_k(t_0) = I_{kp}.$$

For example we could take

$$R_k(t_0) = S_k(t_0)^T (S_k(t_0)S_k(t_0)^T)^{-1}.$$

If we define

$$\text{col} [x_0, w_0, \dots, w_{k-1}] = R_k(t_0) \text{col} [\psi(t_0), \psi^{(1)}(t_0), \dots, \psi^{(k-1)}(t_0)],$$

then the required initial condition for (1) is (t_0, x_0) ; the required input u is the output γ of the system (cf. (33)) with input ω and state $(x, z_0, \dots, z_{k-1}) \in \mathbb{R}^{n+kp}$ given by

$$\dot{x} = (A(t) - B(t)E_k(t)(\Gamma^k C)(t))x + \sum_{i=0}^{k-1} B(t)(E_i(t) - E_k(t)F_i^k(t))z_i + B(t)E_k(t)\omega,$$

$$\dot{z}_0 = z_1,$$

$$\vdots$$

$$\dot{z}_{k-2} = z_{k-1},$$

$$\dot{z}_{k-1} = -(\Gamma^k C)(t)x - \sum_{i=0}^{k-1} F_i^k(t)z_i + \omega,$$

$$\gamma = -E_k(t)(\Gamma^k C)(t)x + \sum_{i=0}^{k-1} (E_i(t) - E_k(t)F_i^k(t))z_i + E_k(t)\omega$$

where we use the input $\omega(t) = \gamma^{(k)}(t)$ and the initial condition $(t_0, x_0, w_0, \dots, w_{k-1})$.

Remark 2.16. In Proposition 2.13 we proved that the rank conditions (31) and (32) imply that the matrix $S_k(t)$ has rank kp for all t in an open dense subset of I .

We can now see that we must have $\text{rank } S_k(t) = kp$ for every $t \in I$. For a given $\psi \in \mathcal{C}^{k,p}(I)$ let $u = \sum_{i=0}^k E_i v^{(i)}$ be the input that generates ψ as described above. Then u and ψ are connected by (22), which we can write in matrix form as

$$\text{col} [\psi(t), \psi^{(1)}(t), \dots, \psi^{(k-1)}(t)] = S_k(t) \text{col} [\phi(t), v(t), v^{(1)}(t), \dots, v^{(k-1)}(t)]$$

for every $t \in I$. Since $\psi \in \mathcal{C}^{k,p}(I)$ is arbitrary, this forces $S_k(t)$ to have rank kp at every $t \in I$.

Remark 2.17. We briefly interpret our results for a time-invariant input-output system

$$\dot{x} = Ax + Bw, \quad y = Cx + Dw.$$

Because the system matrices are constant, we see that $\Gamma C = \dot{C} + CA = CA$, $\Gamma^2 C = CA^2$, etc. It follows that the matrices G_r^l defined in (13) and (14) are given by

$$G_r^l = \begin{cases} 0, & r > l \geq 0, \\ D, & r = l \geq 0, \\ CA^{l-r-1}B, & l > r \geq 0, \end{cases}$$

so the matrix M_k defined in (30) takes the form

$$M_k = \begin{bmatrix} D & 0 & 0 & \cdots & 0 & 0 \\ CB & D & 0 & \cdots & 0 & 0 \\ CAB & CB & D & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ CA^{k-1}B & CA^{k-2}B & CA^{k-3}B & \cdots & CB & D \end{bmatrix}.$$

Hence in the time-invariant case the rank condition (31) is automatically satisfied and the rank condition (32) is precisely that of Sain and Massey [13].

Remark 2.18. We conjecture that the rank conditions in Theorem 2.14 are also necessary for the weak C^k -reproducibility of (1). For $k=1$ this is proved in [1], and the type of argument used there could probably be extended to the general case, though we have not checked the details.

Example 2.19. Consider the linear time-varying input-output system with $n=4$ and $m=p=2$:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \\ e^t & 0 & 0 & 0 \\ 0 & e^t & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \text{cost} & 0 \\ \text{sint} & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix},$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \text{cost} & \text{sint} & 0 & 0 \\ -\text{sint} & \text{cost} & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} t^2 & t \\ t & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

The first few matrices $M_k(t)$ (as defined in (30)) and their ranks are as follows:

$$M_0(t) = \begin{bmatrix} t^2 & t \\ t & 1 \end{bmatrix}, \quad \text{rank } M_0(t) = 1 \quad \text{for every } t \in \mathbb{R},$$

$$M_1(t) = \begin{bmatrix} t^2 & t & 0 & 0 \\ t & 1 & 0 & 0 \\ 2t & 1 & t^2 & t \\ 1 & 0 & t & 1 \end{bmatrix}, \quad \text{rank } M_1(t) = 2 \quad \text{for every } t \in \mathbb{R},$$

$$M_2(t) = \begin{bmatrix} t^2 & t & 0 & 0 & 0 & 0 \\ t & 1 & 0 & 0 & 0 & 0 \\ 2t & 1 & t^2 & t & 0 & 0 \\ 1 & 0 & t & 1 & 0 & 0 \\ 0 & 0 & 4t & 2 & t^2 & t \\ 0 & 0 & 2 & 0 & t & 1 \end{bmatrix}, \quad \text{rank } M_2(t) = 4 \quad \text{for every } t \in \mathbb{R}.$$

Since $\text{rank } M_1(t)$ is constant for every $t \in \mathbb{R}$ and $\text{rank } M_2(t) - \text{rank } M_1(t) = 2$ for every $t \in \mathbb{R}$, Theorem 2.14 implies that the system is weakly C^2 -reproducible on \mathbb{R} . The matrix functions $K_0(t)$, $K_1(t)$, $K_2(t)$ given by Proposition 2.9 can be chosen as

$$K_0(t) = \begin{bmatrix} -1/2 & t/2 \\ t/2 & -t^2/2 \end{bmatrix}, \quad K_1(t) = \begin{bmatrix} 0 & 0 \\ 1/2 & -t/2 \end{bmatrix}, \quad K_2(t) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

(note that these are not uniquely determined). The matrices $E_0(t)$, $E_1(t)$, $E_2(t)$ that define the input $u = \sum_{i=0}^2 E_i v^{(i)}$ are computed by the procedure in Proposition 2.5. We get

$$\begin{aligned} E_2(t) &= K_0(t) = \begin{bmatrix} -1/2 & t/2 \\ t/2 & -t^2/2 \end{bmatrix}, \\ E_1(t) &= K_1(t) - \dot{E}_2(t) = \begin{bmatrix} 0 & -1/2 \\ 0 & t/2 \end{bmatrix}, \\ E_0(t) &= K_2(t) - \ddot{E}_2(t) - 2\dot{E}_1(t) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \end{aligned}$$

The right-inverse system (cf. Remark 2.15) can be explicitly determined once we know the matrix functions F_i^l . These in turn can be determined by Proposition 2.4, where the matrix functions H_i^r are defined in terms of the E_i 's by (10).

3. The nonlinear case. In this section we state a result on the local weak C^k -reproducibility of the nonlinear input-output system

$$(54) \quad \dot{x} = f(t, x, w), \quad y = g(t, x, w)$$

where $f: I \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ and $g: I \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^p$ and C^∞ mappings. It is to be expected that a global reproducibility result for linear input-output systems should give rise to a local reproducibility result for nonlinear input-output systems by applying the global result to the linearization of the nonlinear system. This is indeed the case, but one has to exercise some care in the definition of "local" weak C^k -reproducibility.

NOTATION 3.1. Let I_0 be a compact subinterval of I and fix a point t_0 in I_0 . For $\psi \in \mathcal{C}^{0,l}(I_0)$ we set

$$\|\psi\|_0 = \max \{ |\psi(t)| : t \in I_0 \}$$

where $|\cdot|$ is the usual norm on \mathbb{R}^l , and for $k \geq 1$ and $\psi \in \mathcal{C}^{k,l}(I_0)$ we set

$$\|\psi\|_k = \sum_{i=0}^{k-1} |\psi^{(i)}(t_0)| + \|\psi^{(k)}\|_0.$$

DEFINITION 3.2. Let $u_0 \in \mathcal{C}^{\infty,m}(I)$, let $(t_0, x_0) \in I \times \mathbb{R}^n$, and let $I_0 \subseteq I$ be a compact subinterval containing t_0 on which the solution ϕ_0 of the initial-value problem

$$\dot{\phi}_0(t) = f(t, \phi_0(t), u_0(t)), \quad \phi_0(t_0) = x_0,$$

is defined. Let $\psi_0 \in \mathcal{C}^{\infty,p}(I_0)$ be the output of the input-output system (54) corresponding to the initial condition (t_0, x_0) and input u_0 defined by

$$\psi_0(t) = g(t, \phi_0(t), u_0(t)), \quad t \in I_0.$$

We say that the input-output system (54) is *locally weakly C^k -reproducible* at ψ_0 if for every $\varepsilon > 0$ there exists a $\delta > 0$ having the property that for every $\psi \in \mathcal{C}^{k,p}(I_0)$ with $\|\psi - \psi_0\|_k < \delta$ there exist $x \in \mathbb{R}^n$ with $|x - x_0| < \varepsilon$ and $u \in \mathcal{C}^{0,m}(I_0)$ with $\|u - u_0\|_0 < \varepsilon$ such that

$$\psi(t) = g(t, \phi(t), u(t)),$$

where $\phi(t_0) = x$ and

$$\dot{\phi}(t) = f(t, \phi(t), u(t)) \quad \text{for every } t \in I_0.$$

We omit the proof of the following theorem as it is fairly routine and contains few surprises.

THEOREM 3.3. *Consider the nonlinear input-output system (54), let $u_0 \in \mathcal{C}^{\infty,m}(I)$, let $(t_0, x_0) \in I \times \mathbb{R}^n$, and let $I_0 \subseteq I$ be a compact subinterval containing t_0 on which the solution ϕ_0 of the initial-value problem*

$$\dot{\phi}_0(t) = f(t, \phi_0(t), u_0(t)), \quad \phi_0(t_0) = x_0,$$

is defined. Let $\psi_0 \in \mathcal{C}^{\infty,m}(I_0)$ denote the corresponding output defined by $\psi_0(t) = g(t, \phi_0(t), u_0(t))$. If the linearized input-output system

$$(55) \quad \begin{aligned} \dot{x} &= D_2 f(t, \phi_0(t), u_0(t))x + D_3 f(t, \phi_0(t), u_0(t))w, \\ y &= D_2 g(t, \phi_0(t), u_0(t))x + D_3 g(t, \phi_0(t), u_0(t))w, \end{aligned}$$

satisfies the rank conditions (31) and (32), then the nonlinear input-output system (54) is locally weakly C^k -reproducible at ψ_0 . \square

Remark 3.4. In the course of proving Theorem 3.3, one can also show that given $\psi \in \mathcal{C}^{k,p}(I_0)$ sufficiently close to ψ_0 in the $\|\cdot\|_k$ -norm, the input $u \in \mathcal{C}^{0,m}(I_0)$ that generates ψ is of the form

$$u(t) = u_0(t) + \sum_{i=0}^{k-1} E_i(t) \zeta_i(t) + h(t, \mu(t), \zeta_0(t), \dots, \zeta_{k-1}(t), \psi^{(k)}(t) - \psi_0^{(k)}(t))$$

where the matrix functions E_0, E_1, \dots, E_k are those associated to (55) by Proposition 2.5, the functions $\mu, \zeta_0, \dots, \zeta_{k-1}$ are solutions of the nonlinear system of ODEs

$$\begin{aligned} \dot{x} &= \tilde{f} \left(t, x, \sum_{i=0}^{k-1} E_i(t) z_i + E_k(t) h(t, x, z_0, \dots, z_{k-1}, \psi^{(k)}(t) - \psi_0^{(k)}(t)) \right), \\ \dot{z}_0 &= z_1, \\ &\vdots \\ \dot{z}_{k-2} &= z_{k-1}, \\ \dot{z}_{k-1} &= h(t, x, z_0, \dots, z_{k-1}, \psi^{(k)}(t) - \psi_0^{(k)}(t)), \end{aligned}$$

the function \tilde{f} is defined by

$$\tilde{f}(t, x, w) = f(t, \phi_0(t) + x, u_0(t) + w) - f(t, \phi_0(t), u_0(t)),$$

and h is some function defined and C^∞ in a neighborhood of $I_0 \times \{0\} \times \{0\} \times \dots \times \{0\}$ in $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^{(k+1)p}$. In other words, we can generate the required input from the desired output by a local right-inverse system of (54).

Acknowledgments. The author wishes to thank Professors Felix Albrecht and Nelson Wax for many hours of helpful conversations during the preparation of this work. He also thanks the referees for their useful comments on the first draft of this paper.

REFERENCES

- [1] F. ALBRECHT, K. A. GRASSE AND N. WAX, *Reproducibility of linear and nonlinear input-output systems*, J. Math. Anal. Appl., 79 (1981), pp. 178–202.
- [2] ———, *Path controllability of linear input-output systems*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 569–571.
- [3] R. W. BROCKETT AND M. D. MESAROVIĆ, *The reproducibility of multivariable systems*, J. Math. Anal. Appl., 11 (1965), pp. 548–563.
- [4] V. DOLEZAL, *The existence of a continuous basis of a certain linear subspace of E , which depends on a parameter*, Casopis Pest. Mat., 89 (1964), pp. 466–468.
- [5] R. M. HIRSCHORN, *Invertibility of nonlinear control systems*, this Journal, 17 (1979), pp. 289–297.
- [6] ———, *Invertibility of multivariable nonlinear control systems*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 855–865.
- [7] ———, *Output tracking in multivariable nonlinear systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 593–595.
- [8] L. M. SILVERMAN, *Properties and application of inverse systems*, IEEE Trans. Automat. Control, AC-13 (1968), pp. 436–437.
- [9] ———, *Inversion of multivariable linear systems*, IEEE Trans. Automat. Control, AC-14 (1969), pp. 270–276.
- [10] L. M. SILVERMAN AND H. J. PAYNE, *Input-output structure of linear systems with application to the decoupling problem*, this Journal, 9 (1971), pp. 199–233.
- [11] S. N. SINGH, *Functional reproducibility of multivariable nonlinear systems*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 270–272.
- [12] ———, *Generalized functional reproducibility condition for nonlinear systems*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 958–960.
- [13] M. K. SAIN AND J. L. MASSEY, *Invertibility of linear time-invariant dynamical systems*, IEEE Trans. Automat. Control, AC-14 (1969), pp. 141–149.
- [14] H.-W. WOHLTMANN, *Target path controllability of linear time-varying dynamical systems*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 84–87.
- [15] H. NIJMEIJER, *Right-invertibility for a class of nonlinear control systems: A geometric approach*, Systems Control Lett., 7 (1986), pp. 125–132.

SUR LA MODIFICATION DE LA STRUCTURE A L'INFINI PAR UN RETOUR D'ETAT STATIQUE*

JEAN JACQUES LOISEAU†

Résumé (Abstract). Le bouclage d'un système linéaire par un retour d'état statique non régulier, c'est-à-dire du type $u = Fx + Gv$ avec G non inversible, modifie en général sa structure, et en particulier sa structure à l'infini qui joue un rôle-clé dans les problèmes de commande. L'objet de cet article est d'étudier ces modifications. Nous donnons une condition structurelle nécessaire et suffisante pour qu'une liste donnée soit la liste des ordres des zéros infinis du système bouclé par un retour d'état statique. Nous calculons alors explicitement un retour d'état (F, G) qui produit cette modification de structure à l'infini.

Mots-clés (Key words). système linéaire, commande, retour d'état statique non régulier, structure à l'infini

AMS(MOS) subject classifications. 15A03, 15A21, 3A30, 93B

1. Introduction. Durant la dernière décade, de nombreux auteurs utilisant différentes approches ont contribué à l'étude de la structure à l'infini des systèmes linéaires. On peut citer parmi eux Rosenbrock [1], Pugh et Ratcliffe [2] qui reprennent les travaux de MacMillan [3] et définissent cette structure à partir de la forme de Smith et MacMillan de la matrice de transfert du système.

Rosenbrock [1] définit aussi les zéros à l'infini à partir de la matrice système, en exploitant la décomposition de Kronecker (voir à ce sujet Gantmacher [4]) des faisceaux singuliers de matrices. Jaffe et Karcanias [5] reprennent cette méthodologie et effectuent le lien avec le concept géométrique de sous-espace presque de commandabilité défini par Willems [6], [7].

Une approche purement géométrique a été développée par Commault et Dion [8], qui ont montré que la liste ordonnée des ordres des zéros à l'infini d'un système linéaire (C, A, B) est la liste I_4 d'invariants structurels du système définis par Morse [9].

La motivation de ces travaux est l'importance prépondérante de la structure à l'infini pour la résolution des problèmes classiques de commande des systèmes linéaires, et en particulier pour le problème du découplage (voir Dion [10], et Descusse, Lafay et Malabre [11]).

Les études récentes concernant le découplage dans le cas non régulier, c'est-à-dire à l'aide d'une loi de commande du type $u = Fx + Gv$ lorsque G n'est pas inversible (voir Descusse, Lafay et Malabre [12], [13]) ont mis en évidence le rôle de la liste I_2 définie par Morse [9] dans le problème de la modification de la structure à l'infini du système qui est bouclé par un tel retour d'état statique non régulier. L'emploi de commandes de ce type est très important puisqu'il existe des systèmes qui ne peuvent être découplés qu'en employant des commandes non régulières, mais complique singulièrement le problème comme le montre l'étude du découplage des systèmes ayant deux sorties (voir Descusse, Lafay et Malabre [13]).

La difficulté inhérente à l'emploi de telles commandes non régulières est que nous ne sommes plus en présence d'un groupe de transformations du système, contrairement au cas G inversible. La bibliographie traitant de ce type de loi de commande est de ce fait peu développée. La principale contribution dans ce domaine est due à Heymann [14] qui a complètement caractérisé les modifications possibles de la liste des indices

* Received by the editors February 5, 1986; accepted for publication (in revised form) December 18, 1986.

† Laboratoire d'Automatique, Ecole Nationale Supérieure de Mécanique, Unité Associée au Centre National de la Recherche Scientifique, UA CNRS 4/823, 1 rue de la Noë, 44072 Nantes, Cédex, France. Present address: BE ER, ECAN d'Indrat, 44620 La Montagne, France.

de commandabilité d'une paire (A, B) sous l'action de retours d'états statiques (F, G) non réguliers.

L'objet du présent papier est de caractériser l'ensemble des modifications possibles de la structure à l'infini d'un système (C, A, B) sous l'action de retours d'états (F, G) non réguliers. Le problème est posé de la manière suivante: étant donné un système (C, A, B) et une liste d'entiers positifs $\{n_i\}$, existe-t-il une loi (F, G) telle que $\{n_i\}$ soit la liste des ordres des zéros à l'infini du système bouclé $(C, A + BF, BG)$?

Après avoir au § 2 introduit les notations et rappelé les concepts utilisés, nous répondons à cette question au § 3 sous la forme d'une condition structurelle nécessaire et suffisante (Théorème 3.1) dont la nécessité et la suffisance sont établies dans les paragraphes suivants. Au § 4 sont introduits des sous-espaces qui sont le support géométrique de cette condition structurelle et nous établissons la nécessité de cette condition.

Les §§ 5-7 sont consacrés à la démonstration de la suffisance de notre condition. Nous y exhibons explicitement un retour d'état (F, G) tel que la liste des ordres des zéros à l'infini du système bouclé $(C, A + BF, BG)$ soit une liste $\{n_i\}$ fixée qui vérifie les conditions structurelles du Théorème 3.1. Un cas particulier est examiné au § 5 et généralisé au § 6 où nous résolvons ce problème du placement des ordres des zéros infinis pour le cas où le rang du système (qui est le nombre de ses zéros à l'infini) n'est pas modifié. L'examen d'un second cas particulier au § 7 nous permet de conclure quant à la suffisance de notre condition structurelle.

Enfin, le § 8 est consacré à quelques remarques finales.

2. Préliminaires.

2.1. Notation. Le corps des nombres réels est noté \mathcal{R} . Les minuscules x, u, \dots , désignent des vecteurs, les majuscules romanes A, B, \dots , des applications linéaires ou leur matrice représentative dans une base donnée, les majuscules curvilignes $\mathcal{X}, \mathcal{V}, \dots$, des espaces vectoriels.

$\text{Im } M$ et $\text{Ker } M$ désignent l'image et le noyau de l'application M . La restriction de M à un sous-espace \mathcal{V} est notée M/\mathcal{V} et son image $M\mathcal{V}$. Si $\text{Im } M \subset \mathcal{W}$, nous désignerons par \mathcal{W}/M la restriction de M au codomaine \mathcal{W} .

La dimension d'un sous-espace \mathcal{V} est notée $\dim \mathcal{V}$. L'espace quotient de \mathcal{W} par $\mathcal{V} \subset \mathcal{W}$ sera noté \mathcal{W}/\mathcal{V} ou $\frac{\mathcal{W}}{\mathcal{V}}$ suivant le cas. L'espace engendré par les vecteurs x_1, x_2, \dots, x_i est noté $\text{span } \{x_1, \dots, x_i\}$.

Une suite $x_1, x_2, \dots, x_i \dots$ sera dite finie s'il existe un entier k tel que $x_i = 0$ pour tout i supérieur à k . Le plus petit de ces entiers, soit l , est appelé longueur de la suite, et pour spécifier cette longueur nous écrirons quelquefois la suite $\{x_i\}_l$. De plus, nous identifierons cette suite avec la liste $\{x_1, x_2, \dots, x_l\}$. Le nombre d'éléments d'un ensemble fini $\{\cdot\}$ est noté $\text{card } \{\cdot\}$.

Un nombre nul, un vecteur nul, un espace vectoriel nul ou une matrice nulle seront tous écrits 0.

2.2. Rappel des principaux concepts utilisés. Dans tout ce qui suit, nous appellerons système (C, A, B) le système linéaire stationnaire décrit par les équations suivantes:

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

$$y = Cx,$$

où $x \in \mathcal{X} \simeq \mathcal{R}^n$, $u \in \mathcal{U} \simeq \mathcal{R}^m$, $y \in \mathcal{Y} \simeq \mathcal{R}^p$. Nous pourrions toujours supposer sans restriction de généralité que B est injective et C surjective. Nous poserons $\mathcal{K} = \text{Ker } C$ et $\mathcal{B} = \text{Im } B$.

Rappelons les algorithmes de base de l'approche géométrique (Wonham [15]):

$$(2.1) \quad \begin{aligned} \mathcal{V}_{\mathcal{K}}^0 &= \mathcal{K}, \\ \mathcal{V}_{\mathcal{K}}^i &= \mathcal{K} \cap A^{-1}(\mathcal{V}_{\mathcal{K}}^{i-1} + \mathcal{B}) \end{aligned}$$

pour $i = 1, 2, \dots$, qui est non croissant et converge vers $\mathcal{V}_{\mathcal{K}}^*$, le plus grand sous-espace (A, B) invariant inclus dans \mathcal{K} ;

$$(2.2) \quad \begin{aligned} \mathcal{S}_{\mathcal{B}}^0 &= 0, \\ \mathcal{S}_{\mathcal{B}}^i &= \mathcal{B} + A(\mathcal{S}_{\mathcal{B}}^{i-1} \cap \mathcal{K}) \end{aligned}$$

pour $i = 1, 2, \dots$, qui est non décroissant et converge vers $\mathcal{S}_{\mathcal{B}}^*$, le plus petit sous espace (C, A) invariant contenant \mathcal{B} .

Nous poserons $\mathcal{R}_{\mathcal{K}}^* = \mathcal{V}_{\mathcal{K}}^* \cap \mathcal{S}_{\mathcal{B}}^*$. $\mathcal{R}_{\mathcal{K}}^*$ est le plus grand sous espace de commandabilité de (A, B) inclus dans \mathcal{K} [15].

On définit quatre listes d'entiers à partir des étapes de ces algorithmes:

$$(2.3) \quad p_i = \dim \left(\frac{\mathcal{S}_{\mathcal{B}}^i + \mathcal{V}_{\mathcal{K}}^*}{\mathcal{S}_{\mathcal{B}}^{i-1} + \mathcal{V}_{\mathcal{K}}^*} \right) \quad \text{pour } i = 1, 2, \dots,$$

$$(2.4) \quad n_i = \text{card} \{j \mid p_j \geq i\} \quad \text{pour } i = 1, 2, \dots,$$

$$(2.5) \quad \alpha_i = \dim \left(\frac{\mathcal{S}_{\mathcal{B}}^i \cap \mathcal{V}_{\mathcal{K}}^*}{\mathcal{S}_{\mathcal{B}}^{i-1} \cap \mathcal{V}_{\mathcal{K}}^*} \right) \quad \text{pour } i = 1, 2, \dots,$$

$$(2.6) \quad \sigma_i = \text{card} \{j \mid \alpha_j \geq i\} \quad \text{pour } i = 1, 2, \dots.$$

Ces quatre suites d'entiers positifs sont non croissantes et finies. $\{n_1, \dots, n_{p_1}\}$ est la liste des ordres des zéros à l'infini du système (C, A, B) , qui coïncide avec la liste I_4 d'invariants structurels décrits par Morse [9] (voir Commault et Dion [8]). $\{\sigma_1, \dots, \sigma_{\alpha_1}\}$ est la liste des indices de commandabilité de la paire $(\mathcal{R}_{\mathcal{K}}^* | A + BF | \mathcal{R}_{\mathcal{K}}^*, B_{\mathcal{R}})$, où $\text{Im } B_{\mathcal{R}} = \mathcal{B} \cap \mathcal{R}_{\mathcal{K}}^*$ et $(A + BF)\mathcal{R}_{\mathcal{K}}^* \subset \mathcal{R}_{\mathcal{K}}^*$ (Morse [9]), qui est appelée liste I_2 de Morse.

Rappelons qu'une suite finie et non croissante d'entiers positifs, par exemple $\{p_i\}_{n_1}$, définie en (2.3), est en bijection avec la suite $\{n_i\}_{p_1}$ qui en est déduite par comptage (2.4)

$$n_i = \text{card} \{j \mid p_j \geq i\} \quad \text{pour } i = 1, 2, \dots.$$

Nous avons en effet (Brunovsky [16])

$$(2.7) \quad p_i = \text{card} \{j \mid n_j \geq i\} \quad \text{pour } i = 1, 2, \dots.$$

Les listes $\{\alpha_i\}_{\sigma_1}$ et $\{\sigma_i\}_{\alpha_1}$ satisfont la même bijection.

Cet isomorphisme peut être visualisé sur le schéma suivant (voir Fig. 1):

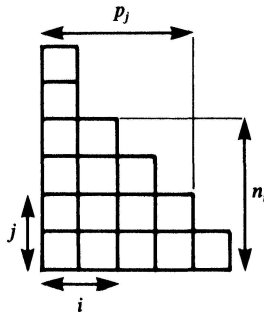


FIG. 1. Si n_i est le nombre de blocs dans la colonne i , alors p_j est le nombre de blocs dans la ligne j .

Nous allons maintenant détailler la structure fine de $\mathcal{S}_{\mathcal{B}}^*$ qui sera largement utilisée dans la suite de cet article.

Les sous-espaces de commandabilité ont été introduits et leurs propriétés étudiées par Morse et Wonham et 1970 (ces résultats sont détaillés dans Wonham [15]). Ainsi, il existe $F_0: \mathcal{X} \rightarrow \mathcal{U}$, et des vecteurs b_1, \dots, b_{α_1} de $\mathcal{B} \cap \mathcal{R}_{\mathcal{K}}^*$ qui nous permettent d'écrire (avec $A_0 = A + BF_0$)

$$(2.8) \quad \mathcal{R}_{\mathcal{K}}^* = \mathcal{R}_1 \oplus \mathcal{R}_2 \oplus \dots \oplus \mathcal{R}_{\alpha_1}$$

avec

$$\mathcal{R}_i = \text{span} \{b_i, A_0 b_i, \dots, A_0^{\sigma_i-1} b_i\} \quad \text{pour } 1 \leq i \leq \alpha_1,$$

$$A_0^{j-1} b_i \in \mathcal{K} \quad \text{pour } 1 \leq j \leq \sigma_i,$$

$$A_0^{\sigma_i} b_i = 0 \quad \text{pour } 1 \leq i \leq \alpha_1.$$

Dion et Commault [8] ont donné une décomposition similaire de la partie irréductible du système qui correspond au quotient $\mathcal{S}_{\mathcal{B}}^* / \mathcal{R}_{\mathcal{K}}^*$. Il existe des vecteurs $\ell_1, \dots, \ell_{p_1}$ de \mathcal{B} tels que

$$(2.9) \quad \mathcal{S}_{\mathcal{B}}^* = \mathcal{R}_{\mathcal{K}}^* \oplus \mathcal{L}_1 \oplus \mathcal{L}_2 \oplus \dots \oplus \mathcal{L}_{p_1}$$

avec

$$\mathcal{L}_i = \text{span} \{\ell_i, A_0 \ell_i, \dots, A_0^{n_i-1} \ell_i\} \quad \text{pour } 1 \leq i \leq p_1,$$

$$A_0^{j-1} \ell_i \in \mathcal{K} \quad \text{pour } 1 \leq j \leq n_i - 1,$$

$$A_0^{n_i-1} \ell_i \notin \mathcal{K}.$$

Dans ces conditions, les vecteurs $C \cdot A_0^{n_i-1} \ell_i$ sont indépendants et engendrent $C \cdot \mathcal{S}_{\mathcal{B}}^*$.

Cette première décomposition est relative aux listes $\{\sigma_i\}_{\alpha_1}$ et $\{n_i\}_{p_1}$. Il existe une deuxième manière de décomposer $\mathcal{S}_{\mathcal{B}}^*$, et qui est relative aux listes $\{\alpha_i\}_{\sigma_1}$ et $\{p_i\}_{n_1}$ que l'on déduit par comptage. Willems [6] a montré qu'il existe un $F: \mathcal{X} \rightarrow \mathcal{U}$, et une chaîne $\{\mathcal{B}_i\}_n$ de sous espaces de \mathcal{B} tels que

$$(2.10) \quad \mathcal{S}_{\mathcal{B}}^* = \mathcal{B} + A_F \mathcal{B}_1 + \dots + A_F^{n-1} \mathcal{B}_{n-1}$$

avec $\mathcal{B} \supset \mathcal{B}_1 \supset \mathcal{B}_2 \supset \dots \supset \mathcal{B}_n$.

La matrice F et la chaîne de \mathcal{B} ne sont pas uniques. Malabre [17] a montré que l'on peut toujours choisir $\{\mathcal{B}_i\}$ de la manière suivante:

$$(2.11) \quad \mathcal{B}_i = \mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^i \quad \text{pour } 1 \leq i \leq n.$$

D'autre part, il est clair que l'on peut toujours choisir F de la manière suivante [18]:

$$(2.12) \quad F = F_0.$$

Ce choix particulier de F et de la chaîne de \mathcal{B} , du fait des conditions $A_0^{\sigma_i} b_i = 0$, permet d'écrire $\mathcal{S}_{\mathcal{B}}^*$ sous la forme d'une somme directe:

$$\mathcal{S}_{\mathcal{B}}^* = \mathcal{B} \oplus A_0(\mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^1) \oplus \dots \oplus A_0^{n-1}(\mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^{n-1}),$$

avec

$$\mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^i = \text{span} \{b_1, \dots, b_{\alpha_i}; \ell_1, \dots, \ell_{p_{i+1}}\}.$$

Nous obtenons aussi la caractérisation suivante:

$$(2.13) \quad \begin{aligned} A_0^{i-1}(\mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^{i-1}) &= A_0^{i-1} \text{span} \{b_1, \dots, b_{\alpha_i}; \ell_1, \dots, \ell_{p_i}\} \\ &= \text{span} \{A_0^{i-1} b_1, \dots, A_0^{i-1} b_{\alpha_i}; A_0^{i-1} \ell_1, \dots, A_0^{i-1} \ell_{p_i}\}. \end{aligned}$$

Et donc

$$(2.14) \quad \dim A_0^{i-1}(\mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^{j-1}) = \alpha_i + p_j \quad \text{pour } i \geq 1 \text{ et } j \geq 1.$$

Notons encore que ce choix particulier de F_0 et de la chaîne $(\mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^i)$ permet d'écrire

$$(2.15) \quad \mathcal{S}_{\mathcal{B}}^i = \mathcal{B} \oplus A_0(\mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^1) \oplus \cdots \oplus A_0^{i-1}(\mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^{i-1}) \quad \text{pour } i \geq 1,$$

$$(2.16) \quad \mathcal{S}_{\mathcal{B}}^* \cap \mathcal{V}_{\mathcal{K}}^i = (\mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^i) \oplus A_0(\mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^{i+1}) \oplus \cdots \oplus A_0^{n-i}(\mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^n) \quad \text{pour } i \geq 1.$$

3. Caractérisation des structures à l'infini atteignables par retour d'état. Etant donné un système (C, A, B) nous dirons qu'une suite finie non croissante d'entiers positifs est une structure à l'infini atteignable par retour d'état si il existe des matrices F et G telles que cette liste soit la liste ordonnée des ordres des zéros à l'infini (liste I_4) du système bouclé $(C, A + BF, BG)$.

Le Théorème 3.1 caractérise l'ensemble des structures à l'infini atteignables par retour d'état pour un système donné.

THÉOREME 3.1. *Soit un système (C, A, B) avec $\{\sigma_i\}_{\alpha_1}$ pour liste I_2 et $\{n_i\}_{p_1}$ pour liste I_4 , et considérons les listes qui en sont déduites par comptage:*

$$\alpha_i = \text{card} \{j | \sigma_j \geq i\} \quad \text{pour } i \geq 1,$$

$$p_i = \text{card} \{j | n_j \geq i\} \quad \text{pour } i \geq 1.$$

Soit $\{n'_i\}_{p'_1}$ une suite non croissante finie d'entiers positifs, et $\{p'_i\}_{n'_1}$ la liste qui en est déduite par comptage:

$$p'_i = \text{card} \{j | n'_j \geq i\} \quad \text{pour } i \geq 1.$$

Considérons alors la liste des différences $\{p'_i - p_i\}$. Cette liste n'est pas ordonnée et comporte même en général des termes négatifs. Soit $\{\Delta_i\}$ la liste obtenue en ne gardant que les différences non négatives et en renumérotant pour obtenir une suite non croissante.

Alors il existe une loi de retour d'état statique (F, G) telle que $\{n'_i\}_{p'_1}$ soit la liste I_4 du système bouclé $(C, A + BF, BG)$ si et seulement si les deux conditions suivantes sont vérifiées:

$$(1) \quad p_1 - p_i \geq p'_1 - p'_i \quad \text{pour } i \geq 1,$$

$$(2) \quad \sum_{j=1}^i \sigma_j \geq \sum_{j=1}^i \Delta_j \quad \text{pour } i \geq 1.$$

La démonstration de ce résultat sera effectuée dans le § 4 pour la nécessité, §§ 5-7 pour la suffisance.

L'énoncé du Théorème 3.1 appelle quelques précisions indispensables. En ce qui concerne le type de transformation utilisée pour modifier la structure à l'infini, la constatation suivante s'impose.

PROPOSITION 3.2. *Etant donné un système (C, A, B) et une suite $\{n'_i\}$ non croissante finie d'entiers positifs, les affirmations suivantes sont équivalentes:*

(i) *Il existe une loi de retour d'état statique (F, G) telle que $\{n'_i\}_{p'_1}$ soit la liste I_4 du système bouclé $(C, A + BF, BG)$.*

(ii) *Il existe une loi de commande (T, F, G) , où T est un isomorphisme, telle que $\{n'_i\}_{p'_1}$ soit la liste I_4 du système bouclé $(CT^{-1}, T(A + BF)T^{-1}, TBG)$.*

(iii) *Il existe une loi de commande (T, F, G, R, H) où T et H sont des isomorphismes, telle que $\{n'_i\}_{p'_1}$ soit la liste I_4 du système bouclé $(HCT^{-1}, T(A + BF + RC)T^{-1}, TBG)$.*

Démonstration. Les trois systèmes $(C, A + BF, BG)$, $(CT^{-1}, T(A + BF)T^{-1}, TBG)$ et $(HCT^{-1}, T(A + BF + RC)T^{-1}, TBG)$ sont équivalents au sens de Morse [9] puisque H et T sont des isomorphismes: ils ont donc en particulier la même structure à l'infini. \square

Le choix d'une base particulière pour \mathcal{X} , qui se traduit concrètement par l'introduction de la transformation supplémentaire T , ne modifie donc pas le problème. Cela va nous permettre d'utiliser sans restriction de généralité la base décrite en (2.9) et (2.10).

La première condition du Théorème 3.1 peut s'écrire directement en termes des listes $\{n_i\}$ et $\{n'_i\}$. Il apparaît alors que la structure à l'infini ne peut être qu'augmentée, ce qui est assez naturel lorsqu'on voit la structure à l'infini comme l'équivalent des retards pour les systèmes discrets.

PROPOSITION 3.3. *Etant données les listes $\{n_i\}_{p'_1}\{p_i\}_{n'_1}\{n'_i\}_{p'_1}$ et $\{p'_i\}_{n'_1}$ du Théorème 3.1, les conditions suivantes sont équivalentes:*

- (i) $p_1 - p_i \geq p'_1 - p'_i$ pour $i \geq 1$,
- (ii) $p_1 \geq p'_1$ et $n_{p_1} - j \leq n'_{p'_1} - j$ pour $0 \leq j \leq p'_1 - 1$.

Démonstration. Supposons que (ii) ne soit pas vérifiée, c'est-à-dire qu'il existe un entier k , $0 \leq k \leq p'_1 - 1$, tel que

$$n_{p_1-k} > n'_{p'_1-k}.$$

Cela implique

$$\text{card} \{j | n_j < n_{p_1-k}\} < \text{card} \{j | n'_j < n_{p_1-k}\},$$

soit:

$$p_1 - p_{n_{p_1-k}} < p'_1 - p'_{n_{p_1-k}}.$$

Pour $i = n_{p_1-k}$, la relation (i) ne serait pas vérifiée. Cela montre que (i) \rightarrow (ii).

Réciproquement, si (ii) est vérifiée, alors il est clair que pour tout entier i , on a

$$\text{card} \{j | n_j < i\} \geq \text{card} \{j | n'_j < i\},$$

soit

$$p_1 - p_i \geq p'_1 - p'_i \quad \text{pour } i \geq 1.$$

Cela montre que (ii) \rightarrow (i), et achève la démonstration. \square

De la même façon, la condition (2) du Théorème 3.1 peut s'énoncer de plusieurs manières différentes.

PROPOSITION 3.4. *Considérons les listes $\{\alpha_i\}$, $\{p_i\}$, $\{p'_i\}$ et $\{\Delta_i\}$ du Théorème 3.1. Les affirmations suivantes sont équivalentes:*

- (i) $\sum_{j=1}^i \alpha_j \geq \sum_{j=1}^i \Delta_j$ pour $i \geq 1$.
- (ii) Pour tout choix $\{k_1, \dots, k_i\}$ d'entiers positifs distincts, on a

$$\sum_{j=1}^i \alpha_j + \sum_{j=1}^i p_{k_j} \geq \sum_{j=1}^i p'_{k_j} \quad \text{pour } i \geq 1.$$

Démonstration. Les termes non nuls dans la liste des différences $\{p'_i - p_i\}$ sont en nombre fini l puisque les deux listes $\{p_i\}$ et $\{p'_i\}$ sont finies.

Considérons une permutation $\{k_i\}$ des indices qui regroupe ces termes non nuls en début de liste et qui les ordonne de la façon suivante:

$$p'_{k_1} - p_{k_1} \geq p'_{k_2} - p_{k_2} \geq \cdots \geq p'_{k_l} - p_{k_l},$$

$$p'_{k_j} - p_{k_j} \neq 0 \quad \text{pour } 1 \leq j \leq l,$$

$$p'_{k_j} - p_{k_j} = 0 \quad \text{pour } j > l.$$

La liste $\{\Delta_i\}$ est alors obtenue en ne gardant que les termes non négatifs:

$$\Delta_i = \max \{0, p'_{k_i} - p_{k_i}\} \quad \text{pour } i \geq 1.$$

Notons que cette liste $\{\Delta_i\}$ est unique, alors que la permutation $\{k_i\}$ ne l'est pas en général.

La somme des i premiers termes de la liste $\{\Delta_i\}$ est un majorant de l'ensemble des sommes de i termes choisis sans répétition d'indice dans la liste $\{p'_i - p_i\}$. De là se déduit la partie (i) \rightarrow (ii) de la Proposition 3.4.

La partie (ii) \rightarrow (i) s'établit en considérant que les premiers termes de la liste $\{\Delta_i\}$ sont choisis parmi les différences $\{p'_i - p_i\}$. \square

La démonstration du Théorème 3.1 va être effectuée dans les paragraphes qui suivent. La nécessité des conditions (1) et (2) du Théorème 3.1 sera montrée dans le § 4 en utilisant directement les énoncés (i) de la Proposition 3.3 et (ii) de la Proposition 3.4. Pour la démonstration de la suffisance, qui est développée dans les §§ 5-7, nous utiliserons plus particulièrement les énoncés (ii) de la Proposition 3.3 et (i) de la Proposition 3.4.

4. Le support géométrique des conditions structurelles.

THÉORÈME 4.1. Soit un système (C, A, B) , et considérons les étapes $\mathcal{V}_{\mathcal{H}}^i$ et $\mathcal{S}_{\mathcal{B}}^i$ des algorithmes définis par (2.1) et (2.2), et les listes $\{\alpha_i\}$ et $\{p_i\}$ définies par (2.3) et (2.5).

Alors nous avons

$$(i) \quad \dim \left(\frac{\mathcal{S}_{\mathcal{B}}^i + \mathcal{H}}{\mathcal{H}} \right) = p_1 - p_{i+1} \quad \text{pour } i \geq 0.$$

De plus, $\{k_1, \dots, k_i\}$ étant une liste strictement croissante d'entiers positifs distincts, posons

$$(ii) \quad \begin{aligned} T_1(k_1, \dots, k_i) &= \mathcal{B} \cap \mathcal{V}_{\mathcal{H}}^{k_i-1}, \\ T_{j+1}(k_1, \dots, k_i) &= (A\mathcal{T}_j(k_1, \dots, k_i) + \mathcal{B}) \cap \mathcal{V}_{\mathcal{H}}^{k_i-j-1} \quad \text{pour } 1 \leq j \leq i-1. \end{aligned}$$

Nous avons alors

$$(4.1) \quad \dim T_j(k_1, \dots, k_i) = \alpha_1 + \dots + \alpha_j + p_{k_i} + p_{k_{i-1}} + \dots + p_{k_{i-j+1}} \quad \text{pour } 1 \leq j \leq i.$$

Et donc en particulier

$$(4.2) \quad \dim \mathcal{T}_i(k_1, \dots, k_i) = \alpha_1 + \dots + \alpha_i + p_{k_1} + \dots + p_{k_i}.$$

Démonstration. (i) Malabre a donné dans [19] quelques déterminations géométriques de la structure à l'infini, et en particulier la suivante:

$$p_{i+1} = \dim \left(\frac{\mathcal{S}_{\mathcal{B}}^* + \mathcal{H}}{\mathcal{S}_{\mathcal{B}}^i + \mathcal{H}} \right) \quad \text{pour } i \geq 0.$$

L'affirmation (i) s'en déduit aisément.

(ii) Nous pouvons écrire les sous espaces \mathcal{T}_j sous forme de sommes directes en utilisant la description donnée dans le § 2. En utilisant le retour d'état particulier introduit en (2.12), on montre aisément que:

$$(4.3) \quad \mathcal{T}_j(k_1, \dots, k_i) = A_0^{j-1}(\mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^{k_i-1}) \oplus A_0^{j-2}(\mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^{k_{i-1}-1}) \oplus \dots \oplus (\mathcal{B} \cap \mathcal{V}_{\mathcal{K}}^{k_{i-j+1}-1})$$

pour $1 \leq j \leq i$.

La dimension de ces sous-espaces se déduit alors de (2.15), et l'on obtient ainsi (4.1) et (4.2).

Démonstration de la nécessité des conditions (1) et (2) du Théorème 3.1. Le Théorème 4.1 permet d'établir rapidement la nécessité des conditions structurelles (1) et (2) du Théorème 3.1. On peut facilement établir l'inclusion suivante entre les sous-espaces $\mathcal{T}_j(k_1, \dots, k_i)$ relatifs au système (C, A, B) et $\mathcal{T}'_j(k_1, \dots, k_i)$ relatifs au système bouclé $(C, A + BF, BG)$:

$$\mathcal{T}'_j(k_1, \dots, k_i) \subset \mathcal{T}_j(k_1, \dots, k_i).$$

Cette inclusion conduit à

$$(4.4) \quad \sum_{j=1}^i \alpha_j + \sum_{j=1}^i p_{k_j} \geq \sum_{j=1}^i \alpha'_j + \sum_{j=1}^i p'_{k_j}.$$

Du fait de la Proposition 3.4, cela établit la nécessité de la condition (2) du Théorème 3.1.

La nécessité de la condition (1) s'établit d'une manière similaire, en considérant les espaces

$$\left(\frac{\mathcal{P}_{\mathcal{B}}^i + \mathcal{K}}{\mathcal{K}} \right) \supset \left(\frac{\mathcal{P}_{\mathcal{B}'}^i + \mathcal{K}}{\mathcal{K}} \right). \quad \square$$

La liste I_2 du système bouclé intervient dans (4.2) qui est de ce fait plus forte que la condition (2) du Théorème 3.1. En fait, (4.2) doit être vue comme une condition nécessaire pour l'assignation simultanée des listes I_2 et I_4 du système [20].

Les sous-espaces $\mathcal{T}_j(k_1, \dots, k_i)$ qui sont décrits dans le présent paragraphe sont le support géométrique naturel des conditions (2) du Théorème 3.1. Comme le montre (4.1), ces sous-espaces sont des sous-espaces presque de commandabilité de (A, B) (voir Willems [6], [7]).

Nous allons maintenant montrer dans les §§ 5-7 la suffisance des conditions structurelles (1) et (2) du Théorème 3.1. Nous allons pour cela construire un retour d'état (F, G) qui permet d'atteindre la structure à l'infini $\{n'_i\}$ choisie à l'avance et vérifiant les conditions (1) et (2) du Théorème 3.1.

5. Le cas particulier fondamental. Deux points importants ont été soulevés par les études récentes concernant le découplage par retour d'état non régulier [12], [13]. D'abord le fait que la liste I_4 de Morse (structure à l'infini) ne puisse qu'être augmentée; ensuite le fait que les augmentations possibles de la liste I_4 de Morse ne proviennent que de la liste I_2 de Morse du système: c'est bien ce que traduisent les conditions (1) et (2) du Théorème 3.1.

Notons toutefois que le problème du découplage par ligne qui est considéré dans [12] et [13] impose l'inversibilité à droite du système ($p_1 = p$), et la conservation de cette inversibilité à droite à travers la transformation du système par retour d'état ($p'_1 = p_1$).

Nous allons dans les §§ 5 et 6 étudier dans une première étape le problème du placement des ordres des zéros à l'infini avec cette condition supplémentaire, imposée pour le problème du découplage par ligne, de conservation du rang du système $p'_1 = p_1$.

Le cas particulier fondamental, qui est l'utilisation d'un seul des termes de la liste I_2 de Morse pour augmenter la liste I_4 , va être décrit dans le présent § 5. Le § 6 sera consacré à la généralisation de ce premier résultat (utilisation de tous les termes de la liste I_2).

THÉORÈME 5.1. *Soit un système (C, A, B) , avec $\{\sigma_i\}_{\alpha_1}$ pour liste I_2 et $\{n_i\}_{p_1}$ pour liste I_4 de Morse, et soient $\{\alpha_i\}_{\sigma_1}$ et $\{p_i\}_{n_1}$ les listes qui en sont déduites par comptage. Soit $\{n_i\}_{p_1^1}$ une liste finie et non croissante d'entiers positifs, et définissons $\{p_i^1\}$:*

$$p_i^1 = \text{card} \{j | n_j^1 \geq i\} \quad \text{pour } i \geq 1.$$

Supposons que l'égalité suivante soit vérifiée:

$$p_1^1 = p_1.$$

Alors il existe un retour d'état statique (F, G) tel que $\{n_1^1, \dots, n_{p_1^1}^1\}$ soit la liste I_4 de $(C, A + BF, BG)$ et $\{\sigma_2, \dots, \sigma_{\alpha_1}\}$ en soit la liste I_2 si et seulement si les deux conditions suivantes sont satisfaites:

- (1) $0 \leq p_i^1 - p_i \leq 1 \quad \text{pour } i \geq 1,$
- (2) $\text{card} \{i | p_i^1 - p_i = 1\} \leq \sigma_1.$

Démonstration.

Nécessité. La nécessité est une simple conséquence des hypothèses et de la nécessité des conditions (1) et (2) du Théorème 3.1. $\{\sigma_i^1\}$ désignant la liste I_2 de $(C, A + BF, BG)$ et $\{\alpha_i^1\}$ la liste qui en est déduite par comptage, nous avons par hypothèse

$$\sigma_i^1 = \sigma_{i+1} \quad \text{pour } i \geq 1,$$

et donc

$$(5.1) \quad \alpha_i^1 = \begin{cases} \alpha_i - 1 & \text{pour } 1 \leq i \leq \sigma_1, \\ 0 & \text{pour } i \geq \sigma_1 + 1. \end{cases}$$

D'autre part, les conditions nécessaires du Théorème 3.1 donnent en particulier

$$(5.2) \quad p_1 - p_i \geq p_1^1 - p_i^1 \quad \text{pour } i \geq 1,$$

$$(5.3) \quad \alpha_1 + p_i \geq \alpha_1^1 - p_i^1 \quad \text{pour } i \geq 1,$$

$$(5.4) \quad \sum_{j=1}^n \alpha_j + \sum_{j=1}^n p_j \geq \sum_{j=1}^n \alpha_j^1 + \sum_{j=1}^n p_j^1.$$

La condition (1) du Théorème 5.1 découle alors de (5.1)–(5.3); la condition (2) découle de (5.4), de (1) et du fait que

$$\sum_{j=1}^n \alpha_j - \sum_{j=1}^n \alpha_j^1 = \sum_{j=1}^n \sigma_j - \sum_{j=1}^n \sigma_j^1 := \sigma_1$$

et que

$$\begin{aligned} \sum_{j=1}^n p_j^1 - \sum_{j=1}^n p_j &= \sum_{j=1}^n (p_j^1 - p_j) \\ &= \text{card} \{j | p_j^1 - p_j = 1\}. \end{aligned}$$

Nous aurons besoin pour montrer la réciproque de traduire les conditions (ii) en termes des listes $\{n_i\}$ et $\{n_i^1\}$.

LEMME 5.2. Soient $\{n_i\}_{p_1}$ et $\{n_i^1\}_{p_1}$ deux listes finies et non croissantes d'entiers positifs et soient $\{p_i\}_{n_1}$ et $\{p_i^1\}_{n_1}$ les listes qui en sont déduites par comptage. Alors:

(1) Les conditions suivantes sont équivalentes:

$$(5.5) \quad 0 \leq p_i^1 - p_i \quad \text{pour } i \geq 1,$$

$$(5.6) \quad 0 \leq n_i^1 - n_i \quad \text{pour } i \geq 1.$$

(2) Les conditions suivantes sont équivalentes:

$$(5.7) \quad p_i^1 - p_i \leq 1 \quad \text{pour } i \geq 1,$$

$$(5.8) \quad n_{i+1}^1 \leq n_i \quad \text{pour } i \geq 1.$$

(3) Les conditions suivantes sont équivalentes:

$$(5.9) \quad \sum_{i=1}^n p_i^1 - \sum_{i=1}^n p_i \leq \sigma_1,$$

$$(5.10) \quad n_1^1 \leq n_1 + \sigma_1 - \sum_{i=2}^n (n_i^1 - n_i).$$

Démonstration du Lemme 5.2. Les conditions (1) et (2) sont des conséquences immédiates de la bijection (2.7) qui existe entre les listes $\{n_i\}$ et $\{p_i\}$ d'une part, et entre $\{n_i^1\}$ et $\{p_i^1\}$ d'autre part.

La condition (3) se déduit par équivalences:

$$\begin{aligned} & \sum_{i=1}^n p_i^1 - \sum_{i=1}^n p_i \leq \sigma_1 \\ \Leftrightarrow & \sum_{i=1}^n n_i^1 - \sum_{i=1}^n n_i \leq \sigma_1 \\ \Leftrightarrow & n_1^1 \leq n_1 + \sigma_1 - \sum_{i=2}^n (n_i^1 - n_i); \end{aligned}$$

ceci achève la démonstration du Lemme 5.2. \square

Suffisance du Théorème 5.1. Nous allons maintenant établir la réciproque du Théorème 5.1 en construisant des matrices F et G qui réalisent les modifications voulues des listes I_2 et I_4 .

Supposons dans ce but que les conditions (1) et (2) du Théorème 5.1 soient vérifiées. Cela s'écrit en utilisant le Lemme 5.2:

$$(5.11) \quad n_1 \leq n_1^1 \leq n_1 + \sigma_1 - \sum_{i=2}^n (n_i^1 - n_i),$$

$$(5.12) \quad n_i \leq n_i^1 \leq n_{i-1} \quad \text{pour } 2 \leq i \leq p_1.$$

Nous pouvons toujours choisir une base de \mathcal{X} sous la forme suivante (voir (2.9) et (2.8)):

$$\{b_1, A_0 b_1, \dots, A_0^{\sigma_1-1} b_1; \dots; b_{\alpha_1}, \dots, A_0^{\sigma_{\alpha_1}-1} b_{\alpha_1}; \ell_1, \dots, A_0^{n_1-1} \ell_1; \\ \dots; \ell_{p_1}, \dots, A_0^{n_{p_1}-1} \ell_{p_1}; x_1, \dots, x_\lambda\}.$$

Définissons maintenant la famille des vecteurs x_{ij} de la façon suivante:

$$(5.13) \quad \begin{aligned} x_{ij} &= A_0^{j-1} \ell_{i-1} && \text{pour } 1 \leq i \leq p_1 \text{ et } 1 \leq j \leq n_i^1 - n_i, \\ x_{ij} &= A_0^{j-1} \ell_{i-1} + A_0^{n_i - n_i^1 + j - 1} \ell_i && \text{pour } 1 \leq i \leq p_1 \text{ et } n_i^1 - n_i < j \leq n_i^1, \\ x_{0j} &= A_0^{\sigma+j-1} \ell_0 && \text{pour } 1 \leq j \leq \sigma_1 - \sigma \end{aligned}$$

avec

$$\begin{aligned}\ell_0 &:= b_1, \\ \sigma &:= \sum_{i=1}^{p_1} (n_i^1 - n_i).\end{aligned}$$

Il est alors facile d'établir l'égalité suivante:

$$\text{span} \{x_{ij}\} = \text{span} \{b_1, \dots, A_0^{\sigma_1-1} b_1; \ell_1, \dots, A_0^{n_1-1} \ell_1; \dots; \ell_{p_1}, \dots, A_0^{n_{p_1}-1} \ell_{p_1}\}.$$

L'inclusion provient directement de la définition des x_{ij} et les deux espaces ont la même dimension du fait de l'indépendance des x_{ij} .

Il en résulte que la famille suivante est une base de \mathcal{X} .

$$\begin{aligned}\{b_2, \dots, A_0^{\sigma_2-1} b_2; \dots; b_{\alpha_1}, \dots, A_0^{\sigma_{\alpha_1}-1} b_{\alpha_1}; x_{01}, \dots, x_{0, \sigma_1-\sigma}; \\ x_{11}, \dots, x_{1, n_1^1}; \dots; x_{p_1, 1}, \dots, x_{p_1, n_{p_1}^1}; x_1, \dots, x_\lambda\}.\end{aligned}$$

Définissons alors F et G sur cette nouvelle base de \mathcal{X} . Posons tout d'abord

$$\ell'_i = x_{i1} \quad \text{pour } 1 \leq i \leq p_1;$$

puis

$$(5.14) \quad \text{Im } BG = \text{span} \{b_2, \dots, b_{\alpha_1}; \ell'_1, \dots, \ell'_{p_1}\}.$$

Enfin définissons F de la manière suivante:

$$F := F_0 + F'$$

avec

$$\begin{aligned}BF'x_{ij} &= \begin{cases} 0 & \text{si } j \neq n_i^1 - n_i, \\ \ell_i & \text{si } j = n_i^1 - n_i \text{ et } 1 \leq i \leq p_1, \end{cases} \\ BF'x_{0j} &= 0 \quad \text{pour } 1 \leq j \leq \sigma_1 - \sigma, \\ BF'A_0^{j-1} b_i &= 0 \quad \text{pour } 2 \leq i \leq \alpha_1 \text{ et } 1 \leq j \leq \sigma_i, \\ BF'x_j &= 0 \quad \text{pour } 1 \leq j \leq \lambda. \end{aligned}$$

Nous obtenons alors

$$\begin{aligned}x_{ij+1} &= (A_0 + BF')x_{ij} \\ &= (A + BF)x_{ij} \quad \text{pour } 1 \leq i \leq p_1 \text{ et } 1 \leq j \leq n_i^1 - 1, \\ x_{0j+1} &= (A + BF)x_{0j} \quad \text{pour } 1 \leq j \leq \sigma_1 - \sigma - 1.\end{aligned}$$

De plus, nous pouvons vérifier que

$$\begin{aligned}x_{i,j} &= (A + BF)^{j-1} \ell'_i \in \mathcal{H} \quad \text{pour } 1 \leq i \leq p_1 \text{ et } 1 \leq j \leq n_i^1 - 1, \\ x_{i, n_i^1} &\notin \mathcal{H} \quad \text{pour } 1 \leq i \leq p_1.\end{aligned}$$

Le triplet $(C, A + BF, BG)$ est donc décomposé par cette nouvelle base de \mathcal{X} comme il est décrit dans le § 2. Les listes de Morse correspondantes sont donc

$$I_2 = \{\sigma_2, \dots, \sigma_{\alpha_1}\}, \quad I_4 = \{n_1^1, \dots, n_{p_1}^1\}.$$

Ceci achève la démonstration du Théorème 5.1. \square

Remarque 5.3. Le retour d'état que nous venons d'exhiber pour la démonstration du Théorème 5.1 modifie aussi la liste des zéros invariants du système, tout au moins

dans le cas où $\sigma_1 > \sigma > 0$. En effet, il apparaît dans ce cas dans la structure du système bouclé un zéro invariant fini de valeur 0 et de multiplicité $\sigma_1 - \sigma$, qui correspond aux vecteurs x_{0j} :

$$\{A_0^\sigma b_1, A_0^{\sigma+1} b_1, \dots, A_0^{\sigma_1-1} b_1\}.$$

En fait il est toujours possible d'assigner la valeur de ce zéro invariant en bouclant le système, avant d'appliquer la procédure décrite précédemment, par un retour d'état F'_0 tel que la restriction $\mathcal{R}_1|(A_0 + BF'_0)|\mathcal{R}_1$ ait α arbitraire pour valeur propre de multiplicité σ_1 .

Un tel retour d'état existe toujours du fait que \mathcal{R}_1 est un sous espace de commandabilité (voir Wonham [15]). Lorsqu'on applique alors la démonstration du Théorème 5.1, les vecteurs x_{0j} correspondent à un zéro invariant α de multiplicité $(\sigma_1 - \sigma)$.

6. Première généralisation: conservation du rang du système. Nous sommes maintenant en mesure d'établir la suffisance des conditions (1) et (2) du Théorème 3.1 sous l'hypothèse de conservation du rang: $p'_1 = p_1$, hypothèse qui sera abandonnée dans le § 7.

Nous n'allons pas construire explicitement le retour d'état (F, G) qui fait que la liste $\{n'_i\}_{p_1}$ choisie à l'avance, et que l'on suppose vérifier les conditions nécessaires du Théorème 3.1, soit la liste des ordres des zéros à l'infini du système bouclé $(C, A + BF, BG)$; mais nous allons utiliser le Théorème 5.1 pour définir (F, G) d'une façon algorithmique.

L'idée de cette construction est de définir une suite de listes $\{p_1^\mu, p_2^\mu, \dots\}$ qui vérifient les propriétés suivantes:

—La suite est initialisée à $\{p_i\}$:

$$p_i^0 = p_i \quad \text{pour } i \geq 1.$$

—Pour $\mu \geq 1$, les conditions du Théorème 5.1 sont vérifiées de l'étape $\mu - 1$ à l'étape μ :

$$\begin{aligned} p_1^\mu &= p_1^{\mu-1}, \\ 0 &\leq p_i^\mu - p_i^{\mu-1} \leq 1 \quad \text{pour } i \geq 1, \\ \text{card } \{i | p_i^\mu - p_i^{\mu-1} = 1\} &\leq \sigma_\mu. \end{aligned}$$

—La suite converge vers $\{p'_i\}$; pour $\mu \geq \alpha_1$, on a

$$p_i^\mu = p'_i \quad \text{pour } i \geq 1.$$

En effet, si de telles listes peuvent être définies, alors le problème peut être résolu de la façon suivante:

—Posons:

$$(C, A_0, B_0) = (C, A, B).$$

—Pour $\mu \geq 1$, nous pouvons calculer à l'aide du Théorème 5.1 un retour d'état (F_μ, G_μ) tel que les listes I_4 et I_2 de Morse du système bouclé:

$$(C, A_\mu, B_\mu) = (C, A_{\mu-1} + B_{\mu-1}F_\mu, B_{\mu-1}G_\mu),$$

soient respectivement, $\{n_i^\mu\}$ et $\{\sigma_i^\mu\}$:

$$\begin{aligned} n_i^\mu &= \text{card } \{j | p_j^\mu \geq i\} \quad \text{pour } i \geq 1, \\ \sigma_i^\mu &= \sigma_{i+1}^{\mu-1} = \sigma_{i+\mu} \quad \text{pour } i \geq 1. \end{aligned}$$

— $\{n'_i\}$ sera alors la liste I_4 de Morse du système

$$(C, A_{\alpha_1}, B_{\alpha_1}) = (C, A + BF, BG)$$

avec

$$F = F_1 + G_1 F_2 + \cdots + G_1 G_2 \cdots G_{\alpha_1-1} F_{\alpha_1},$$

$$G = G_1 G_2 \cdots G_{\alpha_1}.$$

La suite de ce paragraphe va être consacrée à la description de telles listes. Définissons tout d'abord les listes $\{p_i^\mu\}$ de la manière suivante.

DÉFINITION 6.1. Posons pour $\mu = 0$:

$$(6.1) \quad p_i^0 = p_i \quad \text{pour } i \geq 1.$$

Pour $\mu \geq 1$, considérons la liste $\{\Delta_i^{\mu-1}\}$ des différences $(p'_i - p_i^{\mu-1})$ rangées dans un ordre non croissant. Cet ordre de rangement n'est pas unique à cause des possibles répétitions de différences égales pour des indices différents. Soit $\{k_i^{\mu-1}\}$ l'unique permutation des indices définie par

$$(6.2) \quad \Delta_i^{\mu-1} = p'_{k_i^{\mu-1}} - p_{k_i^{\mu-1}}^{\mu-1} \quad \text{pour } i \geq 1$$

et

$$(6.3) \quad \begin{aligned} &\text{Si } \Delta_i^{\mu-1} = \Delta_{i+1}^{\mu-1} \quad \text{et} \quad \Delta_i^{\mu-1} \neq 0, \quad \text{alors} \quad k_i^{\mu-1} > k_{i+1}^{\mu-1}, \\ &\text{Si } \Delta_i^{\mu-1} = \Delta_{i+1}^{\mu-1} = 0, \quad \text{alors} \quad k_i^{\mu-1} < k_{i+1}^{\mu-1} \quad \text{pour } i \geq 1. \end{aligned}$$

Posons alors

$$\begin{aligned} l_1^{\mu-1} &= \text{card} \{j | \Delta_j^{\mu-1} > \Delta_{\sigma_\mu}^{\mu-1}\}, \\ l_2^{\mu-1} &= \text{card} \{j | \Delta_j^{\mu-1} \geq \Delta_{\sigma_\mu}^{\mu-1} \text{ et } \Delta_j^{\mu-1} \geq 1\}, \end{aligned}$$

et définissons

$$(6.4) \quad \Delta_i^\mu = \begin{cases} \Delta_i^{\mu-1} - 1 & \text{pour } 1 \leq i \leq l_1^{\mu-1}, \\ \Delta_i^{\mu-1} & \text{pour } l_1^{\mu-1} + 1 \leq i \leq l_2^{\mu-1} + l_1^{\mu-1} - \sigma_\mu, \\ \Delta_i^{\mu-1} - 1 & \text{pour } l_2^{\mu-1} + l_1^{\mu-1} - \sigma_\mu + 1 \leq i \leq l_2^{\mu-1}, \\ \Delta_i^{\mu-1} & \text{pour } l_2^{\mu-1} + 1 \leq i. \end{cases}$$

Finalement définissons $\{p_i^\mu\}$ et $\{n_i^\mu\}$ de la façon suivante:

$$p_{k_i^{\mu-1}}^\mu = p'_{k_i^{\mu-1}} - \Delta_i^\mu \quad \text{pour } i \geq 1,$$

$$n_i^\mu = \text{card} \{j | p_j^\mu \geq i\} \quad \text{pour } i \geq 1. \quad \square$$

Rappelons que $\{\Delta_i\}$ est une liste finie et non croissante d'entiers positifs. De plus $\{\Delta_i^\mu\}$ possèdera cette propriété si $\{\Delta_i^{\mu-1}\}$ la possède, comme le montre la lecture du tableau 6.1 qui traduit les définitions (6.4),

TABLEAU 6.1

i	1	\cdots	$l_1^{\mu-1}$	$l_1^{\mu-1} + 1$	\cdots	$\beta^{\mu-1}$	$\beta^{\mu-1} + 1$	\cdots	$l_2^{\mu-1}$	$l_2^{\mu-1} + 1$	\cdots	$\delta_1^{\mu-1}$	$\delta_1^{\mu-1} + 1$	\cdots
$\Delta_i^{\mu-1}$	$\Delta_i^{\mu-1} > \Delta_{\sigma_\mu}^{\mu-1}$			$\Delta_i^{\mu-1} = \Delta_{\sigma_\mu}^{\mu-1}$					$\Delta_i^{\mu-1} < \Delta_{\sigma_\mu}^{\mu-1}$			0		
$\Delta_i^{\mu-1} - \Delta_i^\mu$	1			0			1			0			0	

où

$$\beta^{\mu-1} := l_1^{\mu-1} + l_2^{\mu-1} - \sigma_\mu,$$

$$\delta_1^{\mu-1} := \text{card} \{j | \Delta_j^{\mu-1} \geq 1\}.$$

Les listes $\{\Delta_i^\mu\}$ sont donc toutes des listes finies et non croissantes d'entiers positifs. Cela nous permet d'affirmer que $\{k_i^\mu\}$ est toujours une permutation des entiers; pour tout entier j positif, il existe i tel que $k_i^\mu = j$. Par conséquent (6.5) définit complètement la liste $\{p_i^\mu\}$.

La liste I_2 de Morse des systèmes (C, A_μ, B_μ) que nous cherchons à définir doit être $\{\sigma_{\mu+1}, \sigma_{\mu+2}, \dots\}$.

Introduisons donc la notation suivante:

$$\alpha_i^\mu = \text{card} \{j | \sigma_{\mu+j} \geq i\} \quad \text{pour } i \geq 1.$$

Nous allons maintenant vérifier que ces listes $\{p_i^\mu\}$ possèdent les propriétés que nous en attendons.

PROPOSITION 6.2. *Pour $\mu \geq 1$, les propriétés suivantes sont vérifiées:*

- (i) $\{p_i^\mu\}$ est une liste finie et non croissante d'entiers positifs,
- (ii) $p_1^\mu = p'_1 = p_1$,
- (iii) $0 \leq p_i^\mu - p_i^{\mu-1} \leq 1$ pour $i \geq 1$,
- (iv) $\text{card} \{j | p_j^\mu - p_j^{\mu-1} = 1\} \leq \sigma_\mu$,
- (v) $\sum_{j=1}^i \alpha_j^\mu \geq \sum_{j=1}^i \Delta_j^\mu$ pour $i \geq 1$.

Démonstration. La démonstration de (v) est un peu technique et pour cela reportée au § Annexe. $\{p_i^\mu\}$ est bien une liste finie d'entiers positifs, du fait de l'encadrement suivant:

$$p_i^0 \leq p_i^\mu \leq p'_i.$$

De plus, l'examen du tableau (6.6) permet d'affirmer que $\{p_i^\mu\}$ est non croissante si $\{p_i^{\mu-1}\}$ l'est. Cela établit (i).

L'affirmation (ii) se montre aisément à partir du fait que

$$p_1^0 = p'_1 = p_1.$$

En observant les définitions (6.4), nous pouvons constater que $(\Delta_i^{\mu-1} - \Delta_i^\mu)$ ne peut prendre que les valeurs 0 ou 1. De là, et de la définition (6.5) découle l'affirmation (iii).

Enfin, l'affirmation (iv) provient des définitions (6.4) et (6.5) et de l'égalité suivante:

$$\text{card} \{i | p_i^\mu - p_i^{\mu-1} = 1\}. \quad \square$$

Les quatre premières affirmations de la Proposition 6.2 permettent l'application à chaque étape du Théorème 5.1. Il existe donc des retours d'état (F_μ, G_μ) tels que $\{n_i^\mu\}$ et $\{\sigma_i^\mu\}$ soient respectivement les listes I_4 et I_2 de Morse du système bouclé (C, A_μ, B_μ) .

À partir de l'étape $\mu = \alpha_1$, les listes $\{\sigma_i^\mu\}$ et $\{\alpha_i^\mu\}$ sont nulles. Puisque $\{\Delta_i^\mu\}$ est une liste positive, la condition (v) s'écrit:

$$\sum_{j=1}^i \Delta_j^\mu = 0 \quad \text{pour } i \geq 1.$$

La liste $\{\Delta_i^\mu\}$ est donc elle-même nulle, ce qui traduit l'égalité des listes $\{p_i^\mu\}$ et $\{p'_i\}$. Autrement dit, $\{n_i\}$ est la liste des ordres des zéros infinis de $(C, A_{\alpha_1}, B_{\alpha_1})$.

Ainsi s'achève la démonstration de la suffisance des conditions (1) et (2) du Théorème 3.1 pour le cas où $p'_1 = p_1$. Nous allons maintenant résoudre le problème

du placement des ordres des zéros infinis par retour d'état statique dans le cas général: $p'_1 \leq p_1$.

7. Le placement des ordres des zéros à l'infini dans le cas général. Le découplage ligne par ligne [12], [13] a motivé l'étude du cas où le rang du système est conservé: $p'_1 = p_1$. En revanche, le découplage par blocs d'un système n'implique pas son inversibilité à droite, ni même la conservation du rang du système à travers l'action du retour d'état qui modifie sa structure à l'infini. C'est une motivation pour étudier le cas général où le nombre de zéros à l'infini du système bouclé peut être inférieur au nombre des zéros à l'infini du système initial: $p'_1 \leq p_1$.

Le Théorème 7.1 va permettre cette généralisation.

THÉOREME 7.1. Soit un système (C, A, B) et soient $\{n_i\}_{p_1}$ sa liste I_4 et $\{p_i\}_{n_1}$ la liste qui en est déduite par comptage:

$$p_i = \text{card} \{j | n_j \geq i\} \quad \text{pour } i \geq 1.$$

Soient $\{n'_i\}_{p'_1}$ une liste finie et non croissante d'entiers positifs, et $\{p'_i\}_{n'_1}$ la liste qui en est déduite par comptage:

$$p'_i = \text{card} \{j | n'_j \geq i\} \quad \text{pour } i \geq 1.$$

Alors il existe un retour d'état (F, G) tel que la liste I_4 de $(C, A + BF, BG)$ soit $\{n'_i\}_{p'_1}$, et sa liste I_2 soit la liste I_2 de (C, A, B) si et seulement si les inégalités suivantes sont vérifiées:

$$(1) \quad p_1 - p_i \geq p'_1 - p'_i \quad \text{pour } i \geq 1,$$

$$(2) \quad p_i \geq p'_i \quad \text{pour } i \geq 1.$$

Démonstration. Nous n'allons pas détailler cette démonstration qui est très similaire à la démonstration du Théorème 5.1.

Les conditions (1) et (2) sont manifestement un cas particulier des conditions nécessaires (1) et (2) du Théorème 3.1, et donc sont nécessaires.

Montrons maintenant la suffisance de ces conditions, et commençons pour cela par les traduire en termes des listes $\{n_i\}$ et $\{n'_i\}$. Nous obtenons à partir de la Proposition 3.3

$$n_{p_1 - p'_i + i} \leq n'_i \leq n_i \quad \text{pour } 1 \leq i \leq p'_1.$$

La construction de F et G s'effectue alors avec la même technique que pour le Théorème 5.1. Les vecteurs $\{\ell_i, A_0 \ell_i, \dots, A_0^{n_i - 1} \ell_i\}$ étant associés à chaque zéro à l'infini d'ordre n_i , posons pour $1 \leq i \leq p'_1$

$$\ell'_i = \begin{cases} \ell_i & \text{si } n_{p_1 - p'_i + i} < n'_i, \\ \ell_i + \ell_{p_1 - p'_i + i} & \text{si } n_{p_1 - p'_i + i} = n'_i, \end{cases}$$

puis

$$x_{ij} = \begin{cases} A_0^{j-1} \ell_i & \text{pour } 1 \leq j \leq n'_i - n_{p_1 - p'_i + i}, \\ A_0^{j-1} \ell_i + A_0^{j + n'_{p_1 - p'_i + i} - n'_i - 1} \ell_{p_1 - p'_i + i} & \text{pour } n'_i - n_{p_1 - p'_i + i} \leq j \leq n'_i. \end{cases}$$

On peut vérifier que les vecteurs x_{ij} sont tous indépendants. Définissons alors F et G comme suit:

$$(7.1) \quad \text{Im } BG = \{\ell'_1, \dots, \ell'_{p'_1}; b_1, \dots, b_{\alpha_1}\};$$

puis pour $1 \leq i \leq p'_1$:

$$(7.2) \quad BF x_{ij} = \begin{cases} \delta_{p_1-p'_1+i} & \text{pour } j = n'_i - n_{p_1-p'_1+i}, \\ 0 & \text{pour } j \neq n'_i - n_{p_1-p'_1+i}. \end{cases}$$

Le résultat cherché est alors obtenu. \square

Le Théorème 7.1 va nous permettre de compléter la démonstration du Théorème 3.1 pour le cas général. Il nous suffit pour cela de réinitialiser l'algorithme présenté dans le § 6.

PROPOSITION 7.2. *Considérons les listes $\{p_i\}$, $\{p'_i\}$ et $\{\alpha_i\}$ respectivement déduites par comptage de la liste $\{n_i\}$ des ordres des zéros à l'infini, de la liste $\{n'_i\}$ à atteindre, et de la liste $\{\sigma_i\}$ qui est la liste I_2 de Morse de (C, A, B) .*

Soit $\{\Delta_i\}$ la liste des différences $(p'_i - p_i)$ rangées par ordre non croissant, et supposons que les conditions du Théorème 3.1 sont vérifiées:

- (1) $p_1 - p_i \geq p'_1 - p'_i$ pour $i \geq 1$,
- (2) $\sum_{j=1}^i \alpha_j \geq \sum_{j=1}^i \Delta_j$ pour $i \geq 1$.

Définissons la liste $\{p_i^0\}$ de la manière suivante:

$$(7.3) \quad p_i^0 = \min \{p_i, p'_i\} \quad \text{pour } i \geq 1,$$

et appelons $\{\Delta_i^0\}$ la liste des différences $(p'_i - p_i^0)$ rangées par ordre non croissant.

Alors les propriétés suivantes sont vérifiées:

- (i) $p_i \geq p_i^0$ pour $i \geq 1$,
- (ii) $p_1 - p_i \geq p'_1 - p_i^0$ pour $i \geq 1$,
- (iii) $p'_1 - p_i^0 \geq p'_1 - p'_i$ pour $i \geq 1$,
- (iv) $\sum_{j=1}^i \alpha_j \geq \sum_{j=1}^i \Delta_j^0$ pour $i \geq 1$,
- (v) $p_1^0 = p'_1$.

Démonstration. La condition (i) provient directement de la définition (7.3). De la même définition et de l'hypothèse (1) découle (v).

Les conditions (ii) et (iii) découlent également de l'hypothèse (1).

Afin d'établir (iv), notons l'égalité suivante qui provient de la définition (7.3):

$$p'_i - p_i^0 = |p'_i - p_i| \quad \text{pour } i \geq 1.$$

Donc la liste $\{\Delta_i^0\}$ n'est autre que la liste $\{\Delta_i\}$, et (iv) traduit directement l'hypothèse (2).

Les conditions (i) et (ii) étant satisfaites, nous pouvons calculer un retour d'état (F_0, G_0) à la manière du Théorème 7.1. La liste des zéros à l'infini du système bouclé $(C, A + BF_0, BG_0)$ est alors $\{n_i^0\}$:

$$n_i^0 = \text{card} \{j | p_j^0 \geq i\} \quad \text{pour } i \geq 1.$$

Les conditions (iii)–(v) étant satisfaites, il existe un retour d'état (F, G) du type détaillé dans le § 6, tel que $\{n'_i\}$ soit la structure à l'infini du système bouclé $(C, A + BF_0 + BG_0F, BG_0G)$. Cela achève la démonstration du Théorème 3.1. \square

Afin d'illustrer cet article et de récapituler la méthode à suivre pour résoudre un problème de placement de structure à l'infini, nous concluons sur un exemple très simple (et très académique).

Exemple 7.3.

$$A = \begin{array}{|c|c|c|c|} \hline 0 & 1 & & \\ \hline 0 & 0 & & \\ \hline & & 0 & 1 \\ & & 0 & 0 \\ \hline & & & 0 \\ \hline & & & 0 \\ \hline \end{array}, \quad B = \begin{array}{|c|c|c|c|} \hline 0 & & & \\ \hline 1 & & & \\ \hline & 0 & & \\ & 1 & & \\ \hline & & 1 & \\ \hline & & & 1 \\ \hline \end{array},$$

$$C = \begin{array}{|c|c|c|c|} \hline & & 1 & \\ \hline & & & 1 \\ \hline \end{array} \quad (\text{les blocs non spécifiés sont nuls}).$$

Ce système est sous la forme normale qui correspond à

$$\begin{aligned} \{n_i\} &= \{1, 1\}, & \{p_i\} &= \{2\}, \\ \{\sigma_i\} &= \{2, 2\}, & \{\alpha_i\} &= \{2, 2\} \end{aligned}$$

(les termes nuls ne sont pas écrits).

On se propose d'atteindre la structure à l'infini suivante:

$$\{n'_i\} = \{4, 2\}, \quad \{p'_i\} = \{2, 2, 1, 1\}.$$

Or on a

$$\{\Delta_i\} = \{2, 1, 1\}.$$

Les conditions nécessaires du Théorème 3.1 sont vérifiées donc la liste $\{n'_i\}$ est atteignable.

Du fait que $p_1 = p'_1 = 2$, le Théorème 7.1 n'a pas à être utilisé. Nous posons donc

$$(C_0, A_0, B_0) = (C, A, B).$$

Appliquons alors la définition 6.1:

μ	$\{k_i^\mu\}$	l_1^μ	l_2^μ	$\{\Delta_i^\mu\}$	$\{p_i^\mu\}$
1	$\{2, 4, 3\}$	1	3	$\{1, 1\}$	$\{2, 1, 1\}$
2	$\{4, 2\}$	0	2	0	$\{2, 2, 1, 1\}$

Calculons maintenant (F_1, G_1) et (F_2, G_2) comme dans le § 5:

$$G_1 = \begin{array}{|c|c|c|} \hline & 1 & \\ \hline 1 & & \\ \hline & & 1 \\ & & 1 \\ \hline \end{array}, \quad F_1 = \begin{array}{|c|c|c|c|} \hline & & & \\ \hline & & & \\ \hline 1 & 0 & & \\ \hline & & & \\ \hline \end{array},$$

$$G_2 = \begin{array}{|c|c|} \hline 1 & \\ \hline & 1 \\ \hline & \\ \hline \end{array}, \quad F_2 = \begin{array}{|c|c|c|c|} \hline & & & \\ \hline & 0 & 1 & \\ \hline 0 & 1 & -1 & 0 \\ \hline \end{array}.$$

Tous calculs faits, nous obtenons:

$$A + BF_1 + BG_1F_2 = \begin{array}{|c|c|c|c|c|} \hline 0 & 1 & 0 & 0 & \\ \hline 0 & 0 & 0 & 1 & \\ \hline & & 0 & 1 & \\ & & 0 & 0 & \\ \hline 1 & 1 & -1 & 0 & 0 \\ \hline 0 & 1 & -1 & 0 & 0 \\ \hline \end{array}, \quad BG_1G_2 = \begin{array}{|c|c|} \hline & 0 \\ & 1 \\ \hline 0 & \\ 1 & \\ \hline & \\ & \\ \hline \end{array}.$$

8. Remarques finales. Notons tout d'abord l'intérêt pratique du Théorème 3.1. Les conditions structurelles du Théorème 3.1, qui caractérisent l'ensemble des structures à l'infini atteignables par retour d'état statique, sont en effet très simples à vérifier. Cette vérification s'effectue en fait en quatre étapes:

- (i) Calcul des étapes successives des algorithmes (2.1) et (2.2).
- (ii) Calcul des listes I_2 et I_4 de Morse du système par les formules (2.3), (2.4), (2.5) et (2.6).
- (iii) Calcul de la liste ordonnée des différences $\{\Delta_i\}$.
- (iv) Vérification des inégalités (1) et (2) du Théorème 3.1.

Le nombre des inégalités à vérifier est en fait $2 \max \{n_1, n'_1\}$.

Le problème du placement par retour d'état de la structure à l'infini est également résolu de manière pratique. La construction d'une solution (F, G) s'effectue en trois étapes:

- (i) Calcul d'une base de \mathcal{X} qui décompose $\mathcal{R}_{\mathcal{X}}^*$ comme en (2.9) et $\mathcal{S}_{\mathcal{B}}^*$ comme en (2.10). Cela s'effectue de manière standard à partir des étapes des algorithmes (2.1) et (2.2). La procédure est détaillée dans Morse [9].
- (ii) Réduction si nécessaire du rang du système à l'aide des formules (7.1) et (7.2) du § 7.
- (iii) Utilisation successive des termes de la liste I_2 à l'aide de la définition 6.1 et des formules (5.14) et (5.15).

Le calcul, assez lourd dans le cas général, d'un retour d'état qui décompose $\mathcal{R}_{\mathcal{X}}^*$ et $\mathcal{S}_{\mathcal{B}}^*$ ne s'effectue qu'une seule fois, au début de la procédure. A partir de cette première étape, les calculs demandés sont très simples, et particulièrement aptes à être traités sur ordinateur. L'algorithme peut être modifié afin de réduire le nombre des calculs, par exemple en appliquant la procédure de modification aux seuls vecteurs correspondant à des zéros à l'infini dont l'ordre n'est pas dans la structure à atteindre. Ainsi, avec $\{n_i\} = \{3, 3, 2, 1\}$ et $\{n'_i\} = \{4, 4, 3, 2\}$, il suffit d'appliquer la procédure qui correspond aux listes $\{n_i\} = \{3, 1\}$ et $\{n'_i\} = \{4, 4\}$.

La solution (F, G) au problème de placement des ordres des zéros à l'infini qui est déterminée de cette manière est en général différente de la solution obtenue par l'algorithme de départ. En fait, il est bien clair que la solution, si elle existe, n'est pas unique. La détermination de l'ensemble des solutions (F, G) d'un problème de placement de structure à l'infini est pour l'instant un problème ouvert. Il semble que la solution de ce problème, ou plus exactement la détermination de l'ensemble des sous-espaces $\text{Im } BG$ correspondant à une solution, soit une clé importante pour la solution de problèmes de commande tels que le découplage.

L'étude effectuée dans le présent article s'inscrit dans le cadre plus général de l'étude des transformations (sous l'action de retours d'états statiques) des quatre listes

d'invariants structurels de Morse [20]. Une référence importante dans ce domaine est celle d'Heymann [14] qui a caractérisé pour une paire (A, B) donnée l'ensemble des listes d'indices de commandabilité atteignables par retour d'état. Les travaux d'Heymann s'appliquent aux transformations de la liste I_2 de Morse, qui n'est autre que la liste des indices de commandabilité de la paire restreinte $(\mathcal{R}_{\mathcal{H}}^*|A+BF|\mathcal{R}_{\mathcal{H}}^*, B_{\mathcal{R}})$ où $\text{Im } B_{\mathcal{R}} = \mathcal{B} \cap \mathcal{R}_{\mathcal{H}}^*$, et $(A+BF)\mathcal{R}_{\mathcal{H}}^* \subset \mathcal{R}_{\mathcal{H}}^*$.

Etant donné un système (C, A, B) dont la liste I_2 de Morse est $\{\sigma_i\}$, la condition nécessaire et suffisante donnée par Heymann pour qu'une liste finie d'entiers positifs $\{\sigma'_i\}$ soit la liste I_2 d'un système bouclé $(C, A+BF, BF)$ est la suivante:

$$(8.1) \quad \sum_{j \in \mathcal{J}'(i)} \sigma'_j \leq \sum_{j \in \mathcal{J}(i)} \sigma_j \quad \text{pour } i \geq 1,$$

avec

$$\mathcal{J}(i) = \{j | \sigma_j \leq i\}.$$

La condition (8.1) exprime en fait les dimensions possibles des sous-espaces de commandabilité du système considéré inclus dans \mathcal{H} . Ces dimensions ont d'abord été étudiées à l'aide de considérations polynomiales par Warren et Eckberg [21].

Les liens qui relient la décomposition de Morse des systèmes et la décomposition de Kronecker des faisceaux de matrices sont bien connus (voir Morse [9] et Wonham [15] pour l'approche géométrique, Thorp [22] et Rosenbrock [1] pour l'approche des systèmes par la théorie de Kronecker, Jaffe et Karcianas [5] et Loiseau [18] pour la synthèse des deux approches). Utilisant ce lien, nous avons développé dans [20] une approche par les faisceaux de matrices du problème du placement par retour d'état statique des listes de Morse. Cette étude nous suggère que la condition (2) du Théorème 3.1 exprime les dimensions possibles des sous espaces presque de commandabilité inclus dans $(\mathcal{H} + A\mathcal{H} + \mathcal{B})$, alors que la condition (1) est la condition sur la dimension, ou plutôt sur la structure, d'un tel sous-espace pour qu'il soit le plus petit (C, A_F) invariant contenant un sous-espace de \mathcal{B} du type $\text{Im } BG$.

Annexe. Afin de montrer que l'inégalité (v) de la Proposition 6.2 est toujours vérifiée, nous allons utiliser le lemme suivant.

LEMME A.1. Soient $\{\alpha_i\}$ et $\{\Delta_i\}$ deux listes finies et non croissantes d'entiers positifs, et posons:

$$\sigma_i = \text{card } \{j | \alpha_j \geq i\} \quad \text{pour } i \geq 1,$$

$$\delta_i = \text{card } \{j | \Delta_j \geq i\} \quad \text{pour } i \geq 1.$$

Alors les affirmations suivantes sont équivalentes:

$$(i) \quad \sum_{j=1}^i \alpha_j \geq \sum_{j=1}^i \Delta_j \quad \text{pour } i \geq 1,$$

$$(ii) \quad \sum_{j=i}^{\infty} \sigma_j \geq \sum_{j=i}^{\infty} \delta_j \quad \text{pour } i \geq 1.$$

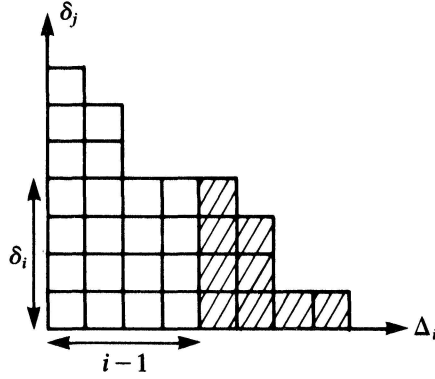
Démonstration du Lemme A.1. Montrons tout d'abord que (i) \rightarrow (ii), et supposons pour cela que (i) soit vérifiée.

Remarquons tout d'abord les égalités suivantes:

$$(A.1) \quad \sum_{j=i}^{\infty} \delta_j = \sum_{j=1}^{\delta_i} \Delta_j - (i-1)\delta_i,$$

$$(A.2) \quad \sum_{j=1}^{\infty} \sigma_j = \sum_{j=1}^{\sigma_i} \alpha_j - (i-1)\sigma_i.$$

La meilleure preuve de ce résultat est le schéma suivant:



Deux cas peuvent se présenter:

(1) Supposons que $\delta_i \leq \sigma_i$.

Nous avons par hypothèse

$$\sum_{j=1}^{\delta_i} \Delta_j \leq \sum_{j=1}^{\delta_i} \alpha_j,$$

et donc

$$\sum_{j=1}^{\delta_i} \Delta_j - (i-1)\delta_i \leq \sum_{j=1}^{\sigma_i} \alpha_j - (i-1)\sigma_i - \sum_{j=\delta_i+1}^{\sigma_i} \alpha_j + (i-1)\sigma_i - (i-1)\delta_i.$$

Mais puisque $\{\alpha_i\}$ est non croissante, nous avons

$$\sum_{j=\delta_i+1}^{\sigma_i} \alpha_j - (i-1)(\sigma_i - \delta_i) \geq (\sigma_i - \delta_i)\alpha_{\sigma_i} - (i-1)(\sigma_i - \delta_i),$$

et puisque

$$\alpha_{\sigma_i} = \text{card} \{j | \sigma_j \geq \sigma_i\} \geq i,$$

nous obtenons finalement

$$\sum_{j=\delta_i+1}^{\sigma_i} \alpha_j - (i-1)(\sigma_i - \delta_i) \geq 0,$$

et donc nous avons

$$\sum_{j=1}^{\delta_i} \Delta_j - (i-1)\delta_i \leq \sum_{j=1}^{\sigma_i} \alpha_j - (i-1)\sigma_i.$$

(2) Supposons que $\sigma_i \leq \delta_i$.

Nous avons par hypothèse

$$\sum_{j=1}^{\delta_i} \Delta_j \leq \sum_{j=1}^{\delta_i} \alpha_j,$$

ce qui s'écrit de la manière suivante:

$$\sum_{j=1}^{\delta_i} \Delta_j - (i-1)\delta_i \leq \sum_{j=1}^{\sigma_i} \alpha_j - (i-1)\sigma_i + \sum_{j=\sigma_i+1}^{\delta_i} \alpha_j - (i-1)\delta_i + (i-1)\sigma_i;$$

or

$$\sum_{j=\sigma_i+1}^{\delta_i} \alpha_j \leq (\delta_i - \sigma_i) \alpha_{\sigma_i+1} \leq (i-1)(\delta_i - \sigma_i).$$

Par conséquent, nous obtenons

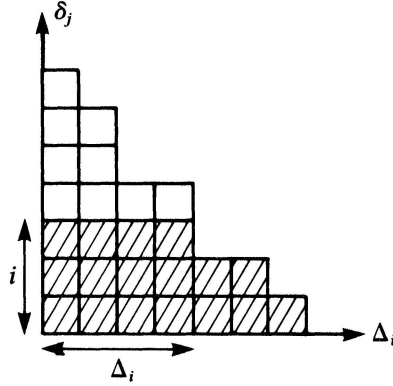
$$\sum_{j=1}^{\delta_i} \Delta_j - (i-1)\delta_i \leq \sum_{j=1}^{\sigma_i} \alpha_j - (i-1)\sigma_i.$$

Ceci, compte tenu des relations (A.1) et (A.2), achève de prouver la partie (i) \rightarrow (ii) du Lemme A.1. La réciproque s'établit de manière identique grâce aux égalités suivantes:

$$(A.3) \quad \sum_{j=1}^i \Delta_j = \sum_{j=\Delta_i+1}^{\infty} \delta_j + i\Delta_i,$$

$$(A.4) \quad \sum_{j=1}^i \alpha_j = \sum_{j=\alpha_i+1}^{\infty} \sigma_j + i\alpha_i,$$

qui peuvent être visualisées par le schéma suivant:



Nous allons maintenant achever la démonstration de la partie (v) de la Proposition 6.2, et montrer que la relation suivante est vérifiée à chaque étape μ :

$$(A.5) \quad \sum_{j=i}^{\infty} \sigma_j^{\mu} \geq \sum_{j=i}^{\infty} \delta_j^{\mu} \quad \text{pour } i \geq 1$$

avec

$$\begin{aligned} \sigma_i^{\mu} &= \sigma_{\mu+i} && \text{pour } i \geq 1, \\ \delta_j^{\mu} &= \text{card} \{j | \Delta_j^{\mu} \geq i\} && \text{pour } i \geq 1. \end{aligned}$$

L'algorithme introduit par la proposition (7.2) est facilement décrit en terme des listes $\{\delta_i^{\mu}\}$:

$$\delta_i^{\mu} = \begin{cases} \delta_i^{\mu-1} & \text{si } \sigma_1^{\mu-1} \leq \delta_{i+1}^{\mu-1}, \\ \delta_i^{\mu-1} + \delta_{i+1}^{\mu-1} - \sigma_1^{\mu-1} & \text{si } \delta_{i+1}^{\mu-1} \leq \sigma_1^{\mu-1} \leq \delta_i^{\mu-1}, \\ \delta_{i+1}^{\mu-1} & \text{si } \delta_i^{\mu-1} \leq \sigma_1^{\mu-1}. \end{cases}$$

L'inégalité que nous voulons établir est vérifiée pour $\mu = 0$, comme l'indiquent la définition 6.1 et le Lemme A.1. Nous allons voir par récurrence qu'elle est vérifiée pour toutes les valeurs de μ .

Supposons donc que l'on ait pour une valeur fixée de μ

$$\sum_{j=i}^{\infty} \delta_j^{\mu-1} \leq \sum_{j=i}^{\infty} \sigma_j^{\mu-1} \quad \text{pour } i \geq 1.$$

Deux cas sont à envisager:

(1) Supposons que $\sigma_1^{\mu-1} \geq \delta_1^{\mu-1}$.

On obtient alors

$$\delta_i^{\mu} = \delta_{i+1}^{\mu-1} \quad \text{pour } i \geq 1.$$

Par conséquent nous pouvons écrire

$$\sum_{j=i}^{\infty} \delta_j^{\mu} = \sum_{j=i+1}^{\infty} \delta_j^{\mu-1} \leq \sum_{j=i+1}^{\infty} \sigma_j^{\mu-1} = \sum_{j=1}^{\infty} \sigma_j^{\mu}.$$

L'inégalité est alors satisfaite pour tout i .

(2) Supposons que $\delta_1^{\mu-1} \geq \sigma_1^{\mu-1}$.

Dans ce cas, il existe un entier l tel que

$$\delta_{l+1}^{\mu-1} \leq \sigma_1^{\mu-1} \leq \delta_l^{\mu-1}.$$

La liste $\{\delta_i^{\mu}\}$ s'écrit donc de la manière suivante:

$$\{\delta_i^{\mu}\} = \{\delta_1^{\mu-1}, \dots, \delta_l^{\mu-1}, \delta_{l+1}^{\mu-1} + \delta_l^{\mu-1} - \sigma_1^{\mu-1}, \delta_{l+2}^{\mu-1}, \dots\}.$$

Suivant la valeur de i , trois cas sont à considérer:

(i) Pour $i \geq l+1$, on a

$$\sum_{j=i}^{\infty} \delta_j^{\mu} = \sum_{j=i+1}^{\infty} \delta_j^{\mu-1} \leq \sum_{j=i+1}^{\infty} \sigma_j^{\mu-1} = \sum_{j=i}^{\infty} \sigma_j^{\mu};$$

(ii) Pour $i \leq l$, on a

$$\sum_{j=i}^{\infty} \delta_j^{\mu} = \sum_{j=i}^{\infty} \delta_j^{\mu-1} - \sigma_1^{\mu-1} \leq \sum_{j=i}^{\infty} \sigma_j^{\mu-1} - \sigma_1^{\mu-1};$$

or

$$(iii) \quad \sum_{j=i}^{\infty} \sigma_j^{\mu-1} - \sigma_1^{\mu-1} = \sum_{j=i+1}^{\infty} \sigma_j^{\mu-1} + \sigma_i^{\mu-1} - \sigma_1^{\mu-1} \leq \sum_{j=i+1}^{\infty} \sigma_j^{\mu-1} = \sum_{j=i}^{\infty} \sigma_j^{\mu}.$$

L'inégalité (A.18) est donc toujours vérifiée à l'étape μ si elle l'était à l'étape $\mu-1$. Cela prouve la partie (v) de la Proposition 6.2 compte tenu du fait que cette inégalité est vérifiée pour $\mu=0$.

REFERENCES

- [1] H. H. ROSENBROCK, *State space and multivariable theory*, John Wiley, New York, 1970.
- [2] A. C. PUGH AND P. A. RATCLIFFE, *On the zeros and poles of a rational matrix*, Internat. J. Control, 30 (1979), pp. 213-226.
- [3] B. MACMILLAN, *Introduction to formal realizability theory*, Bell. System Tech. J., 31 (1952), pp. 541-600.
- [4] F. R. GANTMACHER, *Théorie des matrices*, Dunod, Paris, 1966.
- [5] S. JAFFE AND N. KARCANIAS, *Matrix pencil characterization of almost (A, B)-invariant subspaces: A classification of geometric concepts*, Internat. J. Control, 33 (1981), pp. 51-93.
- [6] J. C. WILLEMS, *Almost A (mod B)-invariant subspaces*, Astérisque, 75-76 (1980), pp. 239-248.
- [7] ———, *Almost invariant subspaces: An approach to high gain feedback design—Part 1: Almost controlled invariant subspaces*, IEEE Trans. Automat. Control, 26 (1981), pp. 235-252.

- [8] C. COMMAULT AND J. M. DION, *Structure at infinity of linear multivariable systems: A geometric approach*, 20th IEEE Conference on Decision and Control, San Diego, California, 1981, pp. 112-117.
- [9] A. S. MORSE, *Structural invariants of linear multivariable systems*, this Journal, 11 (1973), pp. 446-465.
- [10] J. M. DION, *Feedback block decoupling and infinite zero structure of linear systems*, Internat. J. Control, 37 (1983), pp. 521-533.
- [11] J. DESCUSSE, J. F. LAFAY AND M. MALABRE, *On the structure at infinity of linear block-decouplable systems: the general case*, IEEE Trans. Automat. Control, 28 (1981), pp. 1115-1118.
- [12] ———, *Further results on Morgan's problem*, Systems Control Lett., 4 (1984), pp. 203-208.
- [13] ———, *Solution of the static state feedback decoupling problem for linear systems with 2 outputs*, IEEE Trans. Automat. Control, 30 (1985), pp. 914-918.
- [14] M. HEYMANN, *Controllability subspaces and feedback simulation*, this Journal, 14 (1976), pp. 769-789.
- [15] W. M. WONHAM, *Linear multivariable control: a geometric approach*, 2nd edition, Springer-Verlag, Berlin, New York, 1979.
- [16] P. BRUNOVSKY, *A classification of linear controllable systems*, Kibernetika, 3 (1970), pp. 173-187.
- [17] M. MALABRE, *A complement about almost controllability subspaces*, Systems Control Lett., 3 (1983), pp. 119-122.
- [18] J. J. LOISEAU, *Some geometric considerations about the Kronecker normal form*, Internat. J. Control, 42 (1985), pp. 1411-1431.
- [19] M. MALABRE, *Structure à l'infini des triplets invariants. Application à la poursuite parfaite de modèle*, in Analysis and Optimization of Systems, Lectures Notes in Control and Information Sci., Vol. 44, Springer-Verlag, Berlin, New York, 1982, pp. 43-53.
- [20] J. J. LOISEAU, *Structural modifications of linear systems: a matrix pencil approach*, 7th International Symposium on Mathematical Theory of Network and Systems 85, Stockholm, June 10-14, 1985.
- [21] M. E. WARREN AND A. E. ECKBERG JR., *On the dimensions of controllability subspaces: a characterization via polynomial matrices and Kronecker invariants*, this Journal, 13 (1975), pp. 434-445.
- [22] J. S. THORP, *The singular pencil of a linear dynamical system*, Internat. J. Control, 18 (1973), pp. 577-596.

OPTIMAL CONTROL OF STRONGLY MONOTONE VARIATIONAL INEQUALITIES*

SHUZHONG SHI†

Abstract. By using the penalty method and Ekeland's variational principle, this paper proves the optimality condition for a solution to a nonconvex optimal control problem, in which the system is governed by a strongly monotone variational inequality.

Key words. optimal control, strongly monotone variational inequalities, optimality condition, penalty method, Ekeland's variational principle

AMS(MOS) subject classifications. 49B99, 49A29

1. Introduction. The optimal control problem for a system, governed by an elliptic variational inequality, is proposed by J. L. Lions [13], [14], and discussed in Mignot [15], Barbu [2], [3] and Mignot and Puel [16]. The formulation of this problem is as follows:

Let V and H be two Hilbert spaces ("state spaces") such that

$$V \subset H = H^* \subset V^*$$

where V^* is the dual of V , H^* is the dual of H , identified with H , and injections are dense and continuous; let U be another Hilbert space ("control space"). Suppose that $A \in L(V, V^*)$ is coercive, i.e.,

$$Av \in V, \quad \langle Av, v \rangle \geq c \|v\|_V^2, \quad c > 0$$

where $\langle \cdot, \cdot \rangle$ is the duality pairing on $V^* \times V$ and identified with the inner product on H , $B \in L(U, V^*)$ is compact, $K \subset V$ and $U_{ad} \subset U$ are two closed convex subsets respectively in V and in U , $g: K \rightarrow \mathbf{R}_+$ and $h: U_{ad} \rightarrow \mathbf{R}_+$ are two functions, and $f \in V^*$ is given. Then, the optimal control problem for an elliptic variational inequality is to find existence conditions and optimality (necessary) conditions of solutions to the following minimization problem:

$$(1.1) \quad \begin{aligned} & \min \{g(y) + h(u)\}, \\ & y \in K, \quad u \in U_{ad}, \\ & \langle Ay, z - y \rangle \geq \langle Bu + f, z - y \rangle \quad \forall z \in K. \end{aligned}$$

In Mignot [15] and Mignot and Puel [16],

$$\begin{aligned} g(y) &= \frac{1}{2} \|y - z_d\|_H^2, \quad z_d \in H, \\ h(u) &= \frac{N}{2} \|u\|_U^2, \quad N > 0, \end{aligned}$$

and in Barbu [2], [3], $U_{ad} = U$, g is Lipschitz on each bounded subset of K and h is coercive, lower semicontinuous and convex (then, it can implicitly consider the case $U_{ad} \neq U$).

* Received by the editors March 24, 1986; accepted for publication (in revised form) April 16, 1987. This research was supported in part by the FCAR.

† Nankai Mathematics Institute, Nankai University, Tianjin, China and Centre de recherches mathématiques, Université de Montréal, C.P. 6128, Succ. A, Montréal, Québec, Canada H3C 3J7.

Under the above assumptions, the existence of a solution to (1.1) is easily proved. In fact, assume that $\{u_n\} \subset U_{\text{ad}}$ is a minimizing sequence. Then, from the coercivity of h , $\{u_n\}$ is bounded. Without loss of generality, we can assume that $\{u_n\}$ converges weakly in U , and thus, by the compactness of B and using Lions and Stampacchia's theorem, the solution sequence to this inequality in (1.1), $\{y_n\}$, corresponding to $\{u_n\}$, converges strongly in V . Finally, the weak limit u of $\{u_n\}$ and the strong limit y of $\{y_n\}$ are an optimal control and an optimal state, respectively.

The difficulty is to find optimality conditions for a solution to (1.1). For this one, Mignot [15] and Mignot and Puel [16] utilize the conical derivative and Barbu [2], [3] uses smooth approximation.

The aim of this paper is, by penalty method, to prove the optimality condition for a solution to an optimal control problem, in which the system is governed by a strongly monotone variational inequality. The general form of this problem will be as follows:

$$(1.2) \quad \begin{aligned} & \min J(y, u), \\ & y \in K, \quad u \in U_{\text{ad}}, \\ & \langle F(y, u), z - y \rangle \geq 0 \quad \forall z \in K \end{aligned}$$

where

- (1.3) (1) K is a closed convex subset of a Banach space V , U_{ad} is a closed convex subset of another reflexive Banach space U ;
 (2) $J: K \times U_{\text{ad}} \rightarrow \mathbf{R}U\{+\infty\}$ is a lower semicontinuous function for the norm topology of K and the weak topology of U_{ad} , and when

$$\|u\|_U \rightarrow +\infty, \quad \inf_{y \in K} J(y, u) \rightarrow +\infty;$$

- (3) $F: K \times U_{\text{ad}} \rightarrow V^*$, dual of V , is such that for all $R > 0$, there exists a homeomorphism $\varphi_R: \mathbf{R}_+ \rightarrow \mathbf{R}_+$ such that

$$\begin{aligned} & \forall u \in B_R^{U_{\text{ad}}} := \{u \in U_{\text{ad}} \mid \|u\|_U < R\} \quad \forall y_1, y_2 \in K, \\ & \langle F(y_1, u) - F(y_2, u), y_1 - y_2 \rangle_{V^*, V} \geq \varphi_R(\|y_1 - y_2\|_V) \|y_1 - y_2\|_V; \end{aligned}$$

- (4) for any $\bar{u} \in U_{\text{ad}}$, if $y_{\bar{u}} \in K$ satisfies

$$(1.4) \quad \langle F(y_{\bar{u}}, \bar{u}), z - y_{\bar{u}} \rangle_{V^*, V} \geq 0 \quad \forall z \in K,$$

then $u \mapsto F(y_{\bar{u}}, u)$ is weakly continuous at $u = \bar{u}$, i.e., if $u_n \rightarrow \bar{u}$ in U weakly, then $F(y_{\bar{u}}, u_n) \rightarrow F(y_{\bar{u}}, \bar{u})$ in V^* strongly.

It is also not difficult to show the existence of a solution to (1.2). In fact, we can prove the following theorem.

THEOREM 1.1. *Assume that (1.3) (1)–(4) hold. Then the problem (1.2) has a solution $(\bar{y}, \bar{u}) \in K \times U_{\text{ad}}$.*

Proof. By Browder's [5] and Hartman and Stampacchia's [10] theorem and (1.3)(3), for any $u \in U_{\text{ad}}$, there exists a unique $y_u \in K$ such that (1.4) holds. Then, suppose that $\{u_n\} \subset U_{\text{ad}}$ satisfies

$$(1.5) \quad \lim_{n \rightarrow \infty} J(y_n, u_n) = \inf_{u \in U_{\text{ad}}} J(y_u, u)$$

where

$$y_n := y_{u_n}.$$

From (1.3)(2), $\{u_n\}$ is bounded, i.e., for a sufficiently large $R > 0$,

$$(1.6) \quad \|u_n\|_U < R.$$

Since U is reflexive, by extracting a subsequence, we can suppose that

$$(1.7) \quad u_n \rightarrow \bar{u} \quad \text{in } U_{\text{ad}} \text{ weakly.}$$

Set $\bar{y} := y_{\bar{u}}$. Then, from (1.4) we have

$$(1.8) \quad \langle F(y_n, u_n), z - y_n \rangle_{V^*, V} \geq 0 \quad \forall z \in K, \quad n = 1, 2, \dots$$

$$(1.9) \quad \langle F(\bar{y}, \bar{u}), z - \bar{y} \rangle_{V^*, V} \geq 0 \quad \forall z \in K.$$

Taking $z = \bar{y}$ in (1.8) and $z = y_n$ in (1.9), and adding these two inequalities, we obtain that

$$(1.10) \quad \langle F(y_n, u_n) - F(\bar{y}, \bar{u}), y_n - \bar{y} \rangle_{V^*, V} \leq 0.$$

But by (1.3)(3) and (1.6), we have

$$(1.11) \quad \langle F(y_n, u_n) - F(\bar{y}, u_n), y_n - \bar{y} \rangle_{V^*, V} \geq \varphi_R(\|\bar{y} - y_n\|_V) \|\bar{y} - y_n\|_V,$$

and on the other hand, by (1.3)(4) and (1.7), we have

$$(1.12) \quad F(\bar{y}, u_n) \rightarrow F(\bar{y}, \bar{u}) \quad \text{in } V^* \text{ strongly.}$$

From (1.10)–(1.12), it follows that $\varphi_R(\|\bar{y} - y_n\|_V) \rightarrow 0$, which is equivalent to

$$(1.13) \quad y_n \rightarrow \bar{y} \quad \text{in } K \text{ strongly.}$$

By (1.3)(2), (1.5), (1.7) and (1.13), it is easy to see that (\bar{y}, \bar{u}) is a solution to (1.2). \square

We can see in Theorem 1.1 that sufficient conditions for the existence of a solution to (1.2) are very mild. We do not even use the reflexivity of V . However, to obtain the optimality conditions of solution to (1.2), we will have to assume some regularity on J and on F . Hereafter, we show that under some regularity hypotheses, the following optimality conditions for a solution to (1.2) hold:

$$(1.14) \quad \begin{aligned} &\langle F'(\bar{y}, \bar{u}), \bar{p} \rangle_{V^*, V} = 0, \\ &F'_y(\bar{y}, \bar{u})^* \bar{p} \in N_K(\bar{y}) + \partial_y J(\bar{y}, \bar{u}), \\ &F'_u(\bar{y}, \bar{u})^* \bar{p} \in N_{U_{\text{ad}}}(\bar{u}) + \partial_u J(\bar{y}, \bar{u}) \end{aligned}$$

where $\bar{p} \in V$ (“costate”), $N_K(\bar{y})$ is the normal cone to K at \bar{y} , $N_{U_{\text{ad}}}(\bar{u})$ is the normal cone to U_{ad} at \bar{u} , $*$ means the adjoint and $\partial_y J$ and $\partial_u J$ refer to the generalized gradient or the subdifferential of J in respect to y and to u . Certainly, our main results will include the results on (1.1) as a special case.

In § 2, the preliminary material is provided. Section 3 is contributed to the “global Palais–Smale condition,” which will be applied to our main results. The main theorems concerning (1.13) are given in § 4 and some explanatory examples are cited at the end.

2. Preliminaries. We shall use the following three theorems as tools.

BERGE’S MAXIMUM THEOREM 2.1 [4], [1, p. 120]. *Let X and Y be two Hausdorff spaces, $f: X \times Y \rightarrow \mathbf{R}$ continuous and $W: Y \rightrightarrows X$ a continuous set-valued map with non-empty compact values. Then,*

- (1) $\varphi: Y \rightarrow \mathbf{R}$, defined by $\varphi(y) := \max_{x \in W(y)} f(x, y)$, is continuous.
- (2) $\Phi: Y \rightrightarrows X$, defined by $\Phi(y) := \{x \in X \mid f(x, y) = \varphi(y)\}$ is upper semicontinuous.

Remark. Recall that a set-valued map $W: Y \rightrightarrows X$ from a Hausdorff space Y to another Hausdorff space X is called upper (respectively, lower) semicontinuous, if for all closed (respectively, open) subset A of X , $W^{-1}(A) := \{y \in Y \mid W(y) \cap A \neq \emptyset\}$ is closed (respectively, open). If W is both upper and lower semicontinuous, then it is called continuous.

Ekeland's Variational Principle 2.2 [9], [1, p. 255]. Let (E, d) be a complete metric space and $G: E \rightarrow \mathbf{R}_+ \cup \{+\infty\}$ lower semicontinuous, $\neq +\infty$. If for a given $\varepsilon > 0$, $x_0 \in E$ satisfies

$$G(x_0) \leq \inf_E G + \varepsilon,$$

then there exists $x_\varepsilon \in E$ such that

$$G(x_\varepsilon) \leq G(x_0), \quad d(x_0, x_\varepsilon) \leq 1,$$

for all $x \neq x_\varepsilon$, $G(x) > G(x_\varepsilon) - \varepsilon d(x, x_\varepsilon)$.

Lopsided Minimax Theorem 2.3 [1, p. 319]. Let E and F be two closed convex subsets of Hausdorff topological vector spaces and $f: E \times F \rightarrow \mathbf{R}$. If

- (1) for all $y \in F$, $x \mapsto f(x, y)$ is convex;
- (2) for all $x \in E$, $y \mapsto f(x, y)$ is upper semicontinuous and concave;
- (3) F is compact;

then, there exists $\bar{y} \in F$ such that

$$\inf_{x \in E} f(x, \bar{y}) = \inf_{x \in E} \max_{y \in F} f(x, y).$$

Adopting a notation from [6], for a closed convex subset $K \subset V$ and $y^* \in V^*$, we define

$$(2.1) \quad \|y^*\|_y^K := \sup_{p \in (K - y) \cap B_V} \langle y^*, p \rangle_{V^*, V},$$

where B_V is the closed unit ball of V . Then, the variational inequality (1.4) can be rewritten as

$$(2.2) \quad \|-F(y_{\bar{u}}, \bar{u})\|_{y_{\bar{u}}}^K = 0,$$

and the minimization problem (1.2) with an "inequality constraint" can be represented as the following minimization problem with an "equality constraint":

$$(2.2)' \quad \begin{aligned} & \min J(y, u), \\ & y \in K, \quad u \in U_{\text{ad}}, \\ & \|-F(y, u)\|_y^K = 0. \end{aligned}$$

Therefore, it can be approximated by

$$(2.3) \quad \begin{aligned} & \min \{J(y, u) + N_n \|-F(y, u)\|_y^K\}, \\ & y \in K, \quad u \in U_{\text{ad}} \end{aligned}$$

where $N_n > 0$ is a "penalty factor" and $N_n \rightarrow +\infty$. By using Berge's Maximum Theorem 2.1, we have the following.

PROPOSITION 2.4. Let V be a reflexive Banach space with its dual V^* , K a closed convex subset of V , U_{ad} any Hausdorff topological space and $F: K \times U_{\text{ad}} \rightarrow V^*$ a continuous map. Then

- (1) $\varphi: K \times U_{\text{ad}} \rightarrow \mathbf{R}$, defined by

$$\varphi(y, u) := \|-F(y, u)\|_y^K$$

is continuous;

- (2) $M: K \times U_{\text{ad}} \rightrightarrows B_V$, defined by

$$M(y, u) := \{p \in (K - y) \cap B_V \mid \langle -F(y, u), p \rangle_{V^*, V} = \|-F(y, u)\|_y^K\}$$

is upper semicontinuous for the norm topology of K and the weak topology of B_V .

Proof. Set $X = V$, $Y = K \times U_{\text{ad}}$, $f: V \times (K \times U_{\text{ad}}) \rightarrow \mathbf{R}$, defined by

$$f(p; y, u) := \langle -F(y, u), p \rangle_{V^*, V}$$

and $W: K \times U_{\text{ad}} \rightrightarrows B_V$, defined by

$$W(y, u) := (K - y) \cap B_V.$$

Then, for the norm topology of K and the weak topology of B_V , f is continuous and W is continuous with nonempty (weakly) compact values. So, the proposition is a consequence of Berge's Maximum Theorem 2.1. \square

Our idea is as follows: by using Ekeland's Variational Principle 2.2 to the problems (2.3), we shall obtain an almost minimizer sequence $\{(y_n, u_n)\}$ to the problem (1.2). Then, by using Lopsided Minimax Theorem 2.3, we shall find an "almost costate (or Lagrange multiplier)" sequence $\{p_n\}$. Finally, we shall extract a subsequence of $\{(y_n, u_n; p_n)\}$, which will converge to $(\bar{y}, \bar{u}, \bar{p})$, satisfying (1.14). This idea has been applied in Chang and Shi [7] to obtain a local minimax theorem.

A word about the normal cone. Let K be a convex set of a reflexive Banach space V and $y \in K$. Then, the closed convex cone of V^*

$$(2.4) \quad N_K(y) := \{y^* \in V^* \mid \langle y^*, z - y \rangle_{V^*, V} \leq 0 \ \forall z \in K\}$$

is called the normal cone to K at y , and the closed convex cone of V

$$(2.5) \quad \begin{aligned} T_K(y) &:= \{p \in V \mid \langle y^*, p \rangle_{V^*, V} \leq 0 \ \forall y^* \in N_K(y)\} \\ &= \text{cl} \left(\bigcup_{\lambda > 0} \lambda(K - y) \right) \end{aligned}$$

is called the tangent cone to K at y . $N_K(y)$ may be considered as the subdifferential at $y \in K$ of the indicator function

$$(2.6) \quad \delta_K(y) := \begin{cases} 0 & \text{if } y \in K, \\ +\infty & \text{otherwise,} \end{cases}$$

which is lower semicontinuous and convex, provided K is closed and convex. Hence, the last two inclusions of (1.14) may be rewritten as

$$(2.7) \quad \begin{aligned} F'_y(\bar{y}, \bar{u})^* \bar{p} &\in \partial \delta_K(\bar{y}) + \partial_y J(\bar{y}, \bar{u}), \\ F'_u(\bar{y}, \bar{u})^* \bar{p} &\in \partial \delta_{U_{\text{ad}}}(\bar{u}) + \partial_u J(\bar{y}, \bar{u}). \end{aligned}$$

If we set

$$(2.8) \quad \tilde{J}(y, u) := J(y, u) + \delta_K(\bar{y}) + \delta_{U_{\text{ad}}}(\bar{u}),$$

we shall have a more symmetric form of (2.7) as

$$(2.9) \quad \begin{aligned} F'_y(\bar{y}, \bar{u})^* \bar{p} &\in \partial_y \tilde{J}(\bar{y}, \bar{u}), \\ F'_u(\bar{y}, \bar{u})^* \bar{p} &\in \partial_u \tilde{J}(\bar{y}, \bar{u}) \end{aligned}$$

under some hypotheses.

On the other hand, variational inequality (1.4) may be rewritten as either (2.2), or, by (2.4),

$$-F(y_u, u) \in N_K(y_u).$$

In the same way, the last two inclusions of (1.14) or (2.7) may be rewritten as two (quasi) variational inequalities. For instance, if $u \mapsto J(y, u)$ is convex, then

$$F'_u(\bar{y}, \bar{u})^* \bar{p} \in N_{U_{\text{ad}}}(\bar{u}) + \partial_u J(\bar{y}, \bar{u})$$

is equivalent to

$$J'_u(\bar{y}, \bar{u}; v - \bar{u}) - \langle F'_u(\bar{y}, \bar{u})^* \bar{p}, v - \bar{u} \rangle_{U^*, U} \geq 0 \quad \forall v \in U_{\text{ad}}$$

where

$$J'_u(\bar{y}, \bar{u}; v - \bar{u}) := \lim_{t \rightarrow 0^+} \frac{J(\bar{y}, \bar{u} + t(v - \bar{u})) - J(\bar{y}, \bar{u})}{t}$$

and even equivalent to

$$J(\bar{y}, v) - J(\bar{y}, \bar{u}) - \langle F'_u(\bar{y}, \bar{u})^* \bar{p}, v - \bar{u} \rangle_{U^*, U} \geq 0 \quad \forall v \in U_{\text{ad}}.$$

3. Global Palais–Smale condition. To obtain the optimality condition (1.14), we want some compactness condition of Palais–Smale type. The Palais–Smale condition for a C^1 -function on a Banach space has been generalized to a C^1 -function on a closed convex subset of Banach space by Chang and Eells [6] as follows:

Let Q be a closed convex subset of Banach space U and $f \in C^1(Q; \mathbf{R})$, i.e., a continuously differentiable function on an open neighbourhood of Q . We say that f satisfies the Palais–Smale (P.S.) condition if

- (3.1) Any sequence $\{u_n\} \subset Q$, along which $\{f(u_n)\}$ is bounded and $\| -f'(u_n) \|_{u_n}^Q \rightarrow 0$, possesses a convergent subsequence.

This definition is readily to be generalized to a lower semicontinuous function $f: Q \rightarrow \mathbf{R}U\{+\infty\}$ as follows: we say that f satisfies (P.S.) condition, if

- (3.2) Any sequence $\{u_n\} \subset Q$, along which $\{f(u_n)\}$ is bounded and

$$\| -\underline{f}'(u_n, \cdot) \|_{u_n}^Q := \sup_{w \in (Q - u_n) \cap B_U} \{ -\underline{f}'(u_n; w) \} \rightarrow 0,$$

possesses a convergent subsequence

where

$$(3.3) \quad \underline{f}'(u; w) := \liminf_{t \rightarrow 0^+} \frac{f(u + tw) - f(u)}{t}$$

is Dini lower derivative. A lower semicontinuous function $f: Q \rightarrow \mathbf{R}U\{+\infty\}$, which satisfies (P.S.) condition (3.2), has many interesting properties. For instance, by using Ekeland's Variational Principle 2.2, it is easy to show the following.

PROPOSITION 3.1. *Let $f: Q \rightarrow \mathbf{R}_+ U\{+\infty\}$ be a lower semicontinuous function satisfying the (P.S.) condition. Then f achieves its minimum on Q .*

The (P.S.) condition (3.1) or (3.2) is used to characterize the behaviour near a “critical point,” which means a point $u \in Q$ such that $\| -f'(u) \|_u^Q$ or $\| -\underline{f}'(u; \cdot) \|_u^Q = 0$. If we want to characterize a similar behaviour near a “noncritical point,” it is natural to propose the following definition.

DEFINITION 3.2. Let $f: Q \rightarrow \mathbf{R}U\{+\infty\}$ be a lower semicontinuous function and $u^* \in U^*$. We say that f satisfies u^* –(P.S.) if

- (3.4) Any $\{u_n\} \subset Q$, along which $\{f(u_n)\}$ is bounded and $\| -\underline{f}'(u_n, \cdot) + \langle u^*, \cdot \rangle_{U^*, U} \|_{u_n}^Q \rightarrow 0$, possesses a convergent subsequence.

If for all $u^* \in U^*$, f satisfies u^* –(P.S.) condition, then we say that f satisfies the *global (P.S.) condition*.

We shall need later that $u \mapsto J(y, u)$ satisfies the global (P.S.) condition. Such functions form a very large class. For instance, if U is reflexive and locally uniformly convex, a function, satisfying the global (P.S.) condition, may be of the following general form:

$$(3.5) \quad J(u) = J_1(u) + J_2(u) + J_3(u)$$

where

- (1) $J_1 \in C^1(Q, \mathbf{R})$ has a compact derivative $J'_1: Q \rightarrow U^*$;
- (2) $J_2: Q \rightarrow \mathbf{R}U\{+\infty\}$ is any lower semicontinuous convex function;
- (3) $J_3(u) := \Phi(\|u - u_0\|_U)$ with $u_0 \in U$ and $\Phi: \mathbf{R}_+ \rightarrow \mathbf{R}_+$, a convex homeomorphism.

We shall show some more general conclusions.

PROPOSITION 3.3. *Assume that $J: Q \rightarrow \mathbf{R}U\{+\infty\}$ satisfies that*

- $$(3.6) \quad \begin{aligned} (1) & \quad J \text{ is coercive, i.e., } J(u) \rightarrow +\infty \text{ as } \|u\|_U \rightarrow +\infty; \\ (2) & \quad J = J_0 + J_1, \text{ where } J_0 \text{ satisfies the global (P.S.) condition and } J_1 \in C^1(Q; \mathbf{R}) \\ & \quad \text{has a compact derivative } J'_1: Q \rightarrow U^*. \end{aligned}$$

Then J also satisfies the global (P.S.) condition.

Proof. Suppose that for $u^* \in U^*$, a sequence $\{u_n\} \subset Q$ satisfies that

- $$(3.7) \quad \begin{aligned} (1) & \quad \{J(u_n)\} \text{ is bounded;} \\ (2) & \quad \|\underline{J}'(u_n; \cdot) + \langle u^*, \cdot \rangle_{U^*, U}\|_{u_n}^Q = \|\underline{J}'_0(u_n; \cdot) + \langle -J'_1(u_n) + u^*, \cdot \rangle_{U^*, U}\|_{u_n}^Q \rightarrow 0. \end{aligned}$$

Then, from (3.6)(1) and (3.7)(1), $\{u_n\}$ is bounded, i.e.,

$$(3.8) \quad \|u_n\|_U \leq C_1$$

where $C_1 > 0$ is constant. Since J'_1 is compact, $u \mapsto \|J'_1(u)\|_{U^*}$ is bounded on $C_1 B_U$, and thus,

$$(3.9) \quad |J_1(u_n)| \leq |J_1(u_0)| + \sup_{v \in C_1 B_U} \|J'_1(v)\|_{U^*} \cdot \|u_n - u_0\|_U \leq C_2$$

where C_2 is another constant. Hence, from (3.7)(1) and (3.9), we have that

$$(3.10) \quad \{J_0(u_n)\} \text{ is bounded.}$$

On the other hand, also by the compactness of J'_1 , there exists a subsequence

$$\{u_{n_k}\} \subset \{u_n\}$$

such that

$$(3.11) \quad J'_1(u_{n_k}) - u^* \rightarrow \hat{u}^* \quad \text{in } U^* \text{ strongly,}$$

and joining up with (3.7)(2), it follows that

$$(3.12) \quad \|\underline{J}'_0(u_{n_k}; \cdot) + \langle \hat{u}^*, \cdot \rangle_{U^*, U}\|_{u_{n_k}}^Q \rightarrow 0.$$

Since J_0 satisfies the global (P.S.) condition, from (3.10) and (3.12), we deduce that $\{u_{n_k}\}$ has a convergent subsequence. \square

We now show a convex function class, in which each function satisfies the global (P.S.) condition.

DEFINITION 3.4. Let U be a reflexive Banach space, $Q \subset U$ closed convex and $J: Q \rightarrow \mathbf{R}U\{+\infty\}$. We say that J belongs to the (H)-class, if J is lower semicontinuous, convex and satisfies the following (H) condition:

(H) Any $\{u_n\} \subset D(J) := \{u \in Q | J(u) < +\infty\}$, along which $u_n \rightarrow u$ in Q weakly and $\underline{J}'_0(u_n; u - u_n) \rightarrow 0$, possesses a strongly convergent subsequence.

We prove that every coercive (H)-class function satisfies the global (P.S.) condition. In fact, we have a more general proposition as follows.

PROPOSITION 3.5. Let U be a reflexive Banach space, $Q \subset U$ closed convex and $J: Q \rightarrow \mathbf{R}U\{+\infty\}$. If

- (3.13) (1) J is coercive;
 (2) $J = J_0 + J_2$, where $J_0: Q \rightarrow \mathbf{R}U\{+\infty\}$ belongs to (H)-class and $J_2: Q \rightarrow \mathbf{R}U\{+\infty\}$ is any lower semicontinuous convex function,

then, J satisfies the global (P.S.) condition.

Proof. Suppose that for $u^* \in U^*$, a sequence $\{u_n\} \subset Q$ satisfies (3.7). Then, $\{u_n\} \subset D(J_0) \cap D(J_2)$ is bounded, and since U is reflexive, by extracting a subsequence, we can assume that

$$(3.14) \quad u_n \rightarrow u \quad \text{in } Q \text{ weakly.}$$

If $\{u_n\}$ does not possess any strongly convergent subsequence, then, without loss of generality, we can also assume that

$$(3.15) \quad \|u_n - u\| \geq 1/d > 0, \quad n = 1, 2, \dots$$

Therefore,

$$\begin{aligned} l_n &:= -\| -\underline{J}'(u_n; \cdot) + \langle u^*, \cdot \rangle_{U^*, U} \|_{u_n}^Q \\ &= \inf_{v \in Q, \|v - u_n\|_U \leq 1} \{ \underline{J}'(u_n; v - u_n) - \langle u^*, v - u_n \rangle_{U^*, U} \} \\ (3.16) \quad &\leq \frac{1}{\|u_n - u\|_U} \{ \underline{J}'(u_n; u - u_n) - \langle u^*, u - u_n \rangle_{U^*, U} \} \\ &\leq d \{ \underline{J}'(u_n; u - u_n) - \langle u^*, u - u_n \rangle_{U^*, U} \}. \end{aligned}$$

Since $J = J_0 + J_2$ and J_0 and J_2 are convex, we have

$$\begin{aligned} \underline{J}'(u_n; u - u_n) &= J'_0(u_n; u - u_n) + J'_2(u_n; u - u_n) \\ (3.17) \quad &= \inf_{t>0} \frac{J_0(u_n + t(u - u_n)) - J_0(u_n)}{t} + \inf_{t>0} \frac{J_2(u_n + t(u - u_n)) - J_2(u_n)}{t}. \end{aligned}$$

So, from (3.16) and (3.17), we obtain that

$$\begin{aligned} 1/dl_n + \langle u^*, u - u_n \rangle_{U^*, U} + J_2(u_n) - J_2(u) &\leq J'_0(u_n; u - u_n) \\ &\leq J_0(u) - J_0(u_n) \end{aligned}$$

and it follows that

$$(3.18) \quad J'_0(u_n; u - u_n) \rightarrow 0,$$

because from (3.7)(2) and (3.16), $1/dl_n \rightarrow 0$, from (3.14), $\langle u^*, u - u_n \rangle_{U^*, U} \rightarrow 0$ and J_0 and J_2 are lower semicontinuous convex, hence, weakly lower semicontinuous. By the (H) condition and (3.14), (3.18), we obtain that $\{u_n\}$ has a strongly convergent subsequence. This contradiction completes the proof. \square

The following proposition shows two sufficient conditions for (H), and then, with the coercivity, also for the global (P.S.) condition.

PROPOSITION 3.6. *If $J: Q \rightarrow \mathbf{R}U\{+\infty\}$ is lower semicontinuous convex and in one of the following two cases:*

- (3.19) (1) $J(u_1) - J(u_2) - J'(u_2; u_1 - u_2) \geq \Psi(\|u_1 - u_2\|_U)$ for all $u_1, u_2 \in D(J)$, where $\Psi \in C(\mathbf{R}_+; \mathbf{R}_+)$ is a homeomorphism;
 (2) $J(u) = \Phi(\|u - u_0\|)$ for all $u \in D(J)$, where $u_0 \in U$ and $\Phi \in C(\mathbf{R}_+; \mathbf{R}_+)$ is a convex homeomorphism, provided U is locally uniformly convex;

then J belongs to the (H)-class.

Proof. (1) If $\{u_n\} \subset D(J)$ weakly converges to u and $J'(u_n; u - u_n) \rightarrow 0$, then we have that

$$J(u) - J(u_n) - J'(u_n; u - u_n) \geq \Psi(\|u_n - u\|_U) \geq 0.$$

Since J is weakly lower semicontinuous, it follows that

$$\Psi(\|u_n - u\|_U) \rightarrow 0,$$

which is equivalent to $\|u_n - u\|_U \rightarrow 0$.

(2) If (3.19)(2) holds, then it is easy to see that for all $u \in D(J)$, the subdifferential of J at u

$$\partial J(u) := \{u^* \in U^* \mid J(u') - J(u) \geq \langle u^*, u' - u \rangle_{U^*, U} \quad \forall u' \in D(J)\} \neq \emptyset.$$

Hence, for any $u^* \in \partial J(u)$, we have

$$\langle u^*, u_n - u \rangle_{U^*, U} \leq J(u_n) - J(u) \leq J'(u_n; u - u_n),$$

and thus, when $\{u_n\} \subset D(J)$ weakly converges to u and $J'(u_n; u - u_n) \rightarrow 0$, it follows that

$$J(u_n) - J(u) := \Phi(\|u_n - u_0\|_U) - \Phi(\|u - u_0\|_U) \rightarrow 0,$$

which is equivalent to

$$(3.20) \quad \|u_n - u_0\|_U \rightarrow \|u - u_0\|_U.$$

Since U is locally uniformly convex, (3.20) implies that $\{u_n\}$ strongly converges to u . \square

COROLLARY. *If U is reflexive and $A \in L(U, U^*)$ is coercive, then*

$$J(u) := \langle A(u - u_0), u - u_0 \rangle_{U^*, U}$$

with any $u_0 \in U$ satisfies the global (P.S.) condition for any closed convex subset $Q \subset U$.

Proof. Without loss of generality, we assume $u_0 = 0$. For this J , we have that

$$\begin{aligned} J(u_1) - J(u_2) - J'(u_2; u_1 - u_2) &= \langle Au_1, u_1 \rangle_{U^*, U} - \langle Au_2, u_2 \rangle_{U^*, U} \\ &\quad - \langle (A + A^*)u_2, u_1 - u_2 \rangle_{U^*, U} \\ &= \langle A(u_1 - u_2), (u_1 - u_2) \rangle_{U^*, U} \geq c \|u_1 - u_2\|_U^2. \end{aligned} \quad \square$$

We end here the study of the global (P.S.) condition.

4. Main theorems. In order to simplify our discussion, we first assume that

$$(4.1) \quad J(y, u) = g(y) + h(u), \quad F(y, u) = F(y) + E(u)$$

and then, we shall generalize the result obtained from (4.1) to the general case.

THEOREM 4.1. Assume that

- (4.2) (1) V and U are two reflexive Banach spaces with their duals V^* and U^* , respectively; $K \subset V$ and $U_{\text{ad}} \subset U$ are two closed convex subsets, respectively, in V and in U ;
 (2) $g: K \rightarrow \mathbf{R}_+$ is locally Lipschitz;

$$h = h_1 + h_2: U_{\text{ad}} \rightarrow \mathbf{R}_+ U\{+\infty\} \quad \text{where } h_1 \in C^1(U_{\text{ad}}; \mathbf{R})$$

and $h_2: U_{\text{ad}} \rightarrow \mathbf{R}_+ U\{+\infty\}$ is lower semicontinuous convex; in addition, h is coercive and satisfying the global (P.S.) condition;

- (3) $F \in C^1(K; V^*)$, i.e., a continuously differentiable operator from an open neighbourhood of K to V^* , satisfies for all $y_1, y_2 \in K$, $\langle F(y_1) - F(y_2), y_1 - y_2 \rangle_{V^*, V} \geq c \|y_1 - y_2\|_V^2$, $c > 0$; $E \in C^1(U_{\text{ad}}; V^*)$ has a weakly continuous derivative: $E': U_{\text{ad}} \rightarrow L(U, V^*)$ (i.e., E' is continuous for the weak topology of U_{ad}) and for all $u \in U_{\text{ad}}$; $E'(u) \in L(U, V^*)$ is compact (completely continuous).

Then, there exist $(\bar{y}, \bar{u}) \in K \times U_{\text{ad}}$, which satisfies

$$(4.3) \quad \langle F(\bar{y}) + E(\bar{u}), z - \bar{y} \rangle_{V^*, V} \geq 0 \quad \forall z \in K$$

and is a solution to the following problem:

$$(4.4) \quad \begin{aligned} & \min \{g(y) + h(u)\}, \\ & y \in K, \quad u \in U_{\text{ad}}, \\ & \langle F(y) + E(u), z - y \rangle_{V^*, V} \geq 0 \quad \forall z \in K \end{aligned}$$

and $\bar{p} \in V$ such that

$$(4.5) \quad \begin{aligned} & \langle F(\bar{y}) + E(\bar{u}), \bar{p} \rangle_{V^*, V} = 0, \\ & F'(\bar{y})^* \bar{p} \in N_K(\bar{y}) + \partial g(\bar{y}), \\ & E'(\bar{u})^* \bar{p} \in N_{U_{\text{ad}}}(\bar{u}) + \partial h(\bar{u}) \end{aligned}$$

where ∂ refers to the generalized gradient in a broad sense, especially, $\partial h(u) := h'_1(u) + \partial h_2(u)$.

Proof. Suppose that $N_n > 0$ and $G_n: K \times U_{\text{ad}} \rightarrow \mathbf{R} U\{+\infty\}$ is defined by

$$(4.6) \quad G_n(y, u) := g(y) + h(u) + N_n \| -F(y) - E(u) \|_y^K.$$

From Proposition 2.4, G_n is lower semicontinuous. By using Ekeland's Variational Principle 2.2, for any $\varepsilon_n > 0$, there exists $(y_n, u_n) \in K \times U_{\text{ad}}$ such that

$$(4.7) \quad 0 \leq \inf_{K \times U_{\text{ad}}} G_n \leq G_n(y_n, u_n) \leq \inf_{K \times U_{\text{ad}}} G_n + \varepsilon_n,$$

$$(4.8) \quad \forall (y, u) \neq (y_n, u_n) \quad G_n(y, u) > G_n(y_n, u_n) - \varepsilon_n (\|y - y_n\|_V^2 + \|u - u_n\|_U^2)^{1/2}.$$

Set

$$(4.9) \quad \begin{aligned} M_n(y, u) &:= N_n M(y, u) \\ &= \{p_n \in N_n[(K - y) \cap B_V] \mid \langle -F(y) - E(u), p_n \rangle_{V^*, V} \\ &\quad = N_n \| -F(y) - E(u) \|_y^K \}, \end{aligned}$$

which is an upper semicontinuous set-valued map for the norm topology of $K \times U_{\text{ad}}$ and the weak topology of $N_n B_V$, provided by Proposition 2.4. Obviously, for any $(y, u) \in K \times U_{\text{ad}}$, $M_n(y, u)$ is nonempty, weakly compact and convex.

Taking any $s \in K - y_n$, any $w \in U_{\text{ad}} - u_n$ and any sequence $t_k \rightarrow 0^+$, from (4.8), we have

$$G_n(y_n + t_k s, u_n + t_k w) > G_n(y_n, u_n) - \varepsilon_n t_k (\|s\|_V^2 + \|w\|_U^2)^{1/2},$$

and then taking any

$$p_{nk}^{sw} \in M_n(y_n + t_k s, u_n + t_k w),$$

it follows that

$$(4.10) \quad \begin{aligned} & g(y_n + t_k s) + h(u_n + t_k w) - \langle F(y_n + t_k s) + E(u_n + t_k w), p_{nk}^{sw} \rangle_{V^*, V} \\ & > g(y_n) + h(u_n) - \langle F(y_n) + E(u_n), p_{nk}^{sw} \rangle_{V^*, V} - \varepsilon_n t_k (\|s\|_V^2 + \|w\|_U^2)^{1/2}. \end{aligned}$$

We apply the mean value theorem to the function

$$\varphi(t) := \langle F(y_n + ts) + E(u_n + tw), p_{nk}^{sw} \rangle_{V^*, V},$$

and then, there exists a $\theta_k \in]0, 1[$ such that

$$(4.11) \quad \begin{aligned} & \langle F(y_n + t_k s) + E(u_n + t_k w) - F(y_n) - E(u_n), p_{nk}^{sw} \rangle_{V^*, V} \\ & = \langle F'(y_n + \theta_k t_k s) + E'(u_n + \theta_k t_k w) w, p_{nk}^{sw} \rangle_{V^*, V} t_k. \end{aligned}$$

Hence, from (4.10) and (4.11), we obtain

$$(4.12) \quad \begin{aligned} & \frac{g(y_n + t_k s) - g(y_n)}{t_k} + \frac{h(u_n + t_k w) - h(u_n)}{t_k} - \langle F'(y_n + \theta_k t_k s) \\ & + E'(u_n + \theta_k t_k w) w, p_{nk}^{sw} \rangle_{V^*, V} > -\varepsilon_n (\|s\|_V^2 + \|w\|_U^2)^{1/2}. \end{aligned}$$

Since $N_n B_V$ is weakly compact, by extracting a subsequence, we can assume that

$$(4.13) \quad p_{nk}^{sw} \rightarrow p_n^{sw} \quad \text{in } N_n \bar{B}_V \text{ weakly, as } k \rightarrow \infty,$$

and since M_n is upper semicontinuous for the weak topology of $N_n \bar{B}_V$, it follows that

$$(4.14) \quad p_n^{sw} \in M_n(y_n, u_n).$$

Therefore, letting $k \rightarrow \infty$ in (4.12), we obtain

$$(4.15) \quad g^0(y_n; s) + h'(u_n; w) - \langle F'(y_n) s + E'(u_n) w, p_n^{sw} \rangle_{V^*, V} \geq -\varepsilon_n (\|s\|_V^2 + \|w\|_U^2)^{1/2}$$

where

$$(4.16) \quad g^0(y_n; s) := \limsup_{\substack{t \rightarrow 0^+ \\ v \rightarrow y_n}} \frac{g(v + ts) - g(v)}{t}$$

is Clarke directional derivative [8].

Now, consider the function $H : [(K - y_n) \times (U_{\text{ad}} - u_n)] \times M_n(y_n, u_n) \rightarrow \mathbf{R}$, defined by

$$(4.17) \quad \begin{aligned} H(s, w; p) &:= g^0(y_n; s) + h'(u_n; w) \\ &- \langle F'(y_n) s + E'(u_n) w, p \rangle_{V^*, V} + \varepsilon_n (\|s\|_V^2 + \|w\|_U^2)^{1/2}. \end{aligned}$$

Then, $(s, w) \mapsto H(s, w; p)$ is convex, $p \mapsto H(s, w; p)$ is weakly continuous affine and $M_n(y_n, u_n)$ is weakly compact. Therefore, we can use Lopsided Minimax Theorem 2.3 and conclude that there exists $p_n \in M_n(y_n, u_n)$ such that

$$(4.18) \quad \inf_{(s, w)} H(s, w; p_n) = \inf_{(s, w)} \max_{p \in M_n(y_n, u_n)} H(s, w; p).$$

But from (4.14), (4.15) and (4.17), the right side of (4.18) is ≥ 0 , and thus, we deduce that

$$\inf_{(s,w)} H(s, w; p_n) \geq 0,$$

which implies that for $(y_n, u_n) \in K \times U_{\text{ad}}$ and $p_n \in M_n(y_n, u_n)$, we have that

$$(4.19) \quad g^0(y_n; z - y_n) - \langle F'(y_n)^* p_n, z - y_n \rangle_{V^*, V} \geq -\varepsilon_n \|z - y_n\|_V \quad \forall z \in K,$$

$$(4.20) \quad h'(u_n; v - u_n) - \langle E'(u_n)^* p_n, v - u_n \rangle_{U^*, U} \geq -\varepsilon_n \|v - u_n\|_U \quad \forall v \in U_{\text{ad}}.$$

We shall show that when $N_n \rightarrow \infty$ and $\varepsilon_n \rightarrow 0$, $\{(y_n, u_n, p_n)\}$ has a subsequence, which converges to (y, u, p) , resolving to (4.4) and satisfying (4.5).

By (4.2)(3) and Browder, Hartman and Stampacchia's theorem, for any $u \in U_{\text{ad}}$, there exists $y_u \in K$ such that (4.3) holds, or

$$(4.21) \quad \| -F(y_u) - E(u) \|_{y_u}^K = 0.$$

Then, from (4.7), we have

$$0 \leq G_n(y_n, u_n) \leq G_n(y_u, u) + \varepsilon_n \quad \forall u \in U_{\text{ad}}$$

or

$$(4.22) \quad 0 \leq g(y_n) + h(u_n) + N_n \| -F(y_n) - E(u_n) \|_{y_n}^K \leq g(y_u) + h(u) + \varepsilon_n \quad \forall u \in U_{\text{ad}}.$$

Therefore, when $N_n \rightarrow \infty$ and $\varepsilon_n \rightarrow 0$, we have that

$$(4.23) \quad \delta_n := \| -F(y_n) - E(u_n) \|_{y_n}^K = -\langle F(y_n) + E(u_n), p_n \rangle_{V^*, V} \rightarrow 0,$$

and that $\{h(u_n)\}$ is bounded. The coercivity of h implies that $\{u_n\}$ is bounded and without loss of generality, we can assume that

$$(4.24) \quad u_n \rightarrow \bar{u} \quad \text{in } U_{\text{ad}} \text{ weakly.}$$

Set $\bar{y} := y_{\bar{u}}$. Then we have that (4.21) holds for $(y_{\bar{u}}, \bar{u})$ and (4.3) holds. On the other hand, from (4.23), we have that

$$(4.25) \quad \langle F(y_n) + E(u_n), z - y_n \rangle_{V^*, V} \geq -\delta_n \max(1, \|z - y_n\|_V) \quad \forall z \in K, \quad n = 1, 2, \dots$$

Taking $z = y_n$ in (4.3) and $z = \bar{y}$ in (4.25), and adding these two inequalities, we obtain

$$(4.26) \quad \langle F(y_n) - F(\bar{y}), y_n - \bar{y} \rangle_{V^*, V} + \langle E(u_n) - E(\bar{u}), y_n - \bar{y} \rangle_{V^*, V} \\ \geq \delta_n \max(1, \|\bar{y} - y_n\|_V).$$

By (4.2)(3), we have

$$(4.27) \quad \langle F(y_n) - F(\bar{y}), y_n - \bar{y} \rangle_{V^*, V} \geq C \|\bar{y} - y_n\|_V^2$$

and it is easy to show that E is weakly continuous; then, from (4.24)

$$(4.28) \quad E(u_n) \rightarrow E(\bar{u}) \quad \text{in } V^* \text{ strongly.}$$

From (4.26)-(4.28), we deduce that

$$(4.29) \quad y_n \rightarrow \bar{y} \quad \text{in } V \text{ strongly.}$$

We now show that $\{p_n\}$ is bounded and then, by extracting a subsequence, we can assume that

$$(4.30) \quad p_n \rightarrow \bar{p} \quad \text{in } V \text{ weakly.}$$

From (4.2)(3), we have that if for $y \in K$ and $t > 0$, $y + tp \in K$, then

$$\langle F(y + tp) - F(y), tp \rangle_{V^*, V} \geq Ct^2 \|p\|_V^2,$$

and it follows that

$$(4.31) \quad \langle F'(y)p, p \rangle_{V^*, V} \geq c \|p\|_V^2 \quad \forall y \in K, \quad \forall p \in T_K(y).$$

By joining up with (4.19) and (4.29), we deduce that

$$\begin{aligned} (C_{\bar{y}} + \varepsilon_n) \|p_n\|_V &\geq g^0(y_n; p_n) + \varepsilon_n \|p_n\|_V \\ &\geq \langle F'(y_n)p_n, p_n \rangle_{V^*, V} \geq C \|p_n\|_V^2 \end{aligned}$$

where $C_{\bar{y}}$ is the locally Lipschitz constant of g near \bar{y} . Hence, $\{p_n\}$ is bounded and we can assume that (4.30) holds.

Since E' is weakly continuous and $E'(\bar{u})$ is compact, (4.30) and (4.24) lead to

$$(4.32) \quad E'(u_n)^* p_n = (E'(u_n) - E'(\bar{u}))^* p_n + E'(\bar{u})^* p_n \rightarrow E'(\bar{u})^* \bar{p} \quad \text{in } U^* \text{ strongly,}$$

because $E'(\bar{u})^*$ is also compact. Thus, from (4.20), we deduce that

$$(4.33) \quad \|-h'(u_n; \cdot) + \langle E'(u)^* \bar{p}, \cdot \rangle_{U^*, U}\|_{U_{ad}}^U \rightarrow 0.$$

Since h satisfies the global (P.S.) condition, we finally obtain that

$$(4.34) \quad u_n \rightarrow \bar{u} \quad \text{in } U_{ad} \text{ strongly,}$$

provided by extracting a subsequence.

At last, we take the limit in (4.22), (4.23), (4.19) and (4.20) and thus, the proof will be complete. Here, we merely say one word about (4.20), for which the limit is not evident. By (4.2)(2) and (4.20), we have

$$\begin{aligned} h'_1(u_n; v - u_n) + h_2(v) - h_2(u_n) - \langle E'(u_n)^* p_n, v - u_n \rangle_{U^*, U} \\ \geq h'_1(u_n; v - u_n) + h'_2(u_n; v - u_n) - \langle E'(u_n)^* p_n, v - u_n \rangle_{U^*, U} \\ \geq -\varepsilon_n \|v - u_n\|_U \quad \forall v \in U_{ad}, \end{aligned}$$

and taking $n \rightarrow \infty$, we obtain

$$(4.35) \quad h'_1(\bar{u}; v - \bar{u}) - h_2(v) - h_2(\bar{u}) - \langle E'(\bar{u})^* \bar{p}, v - \bar{u} \rangle_{U^*, U} \geq 0 \quad \forall v \in U_{ad}$$

because

$$\liminf_{n \rightarrow \infty} h_2(u_n) \geq h_2(\bar{u}).$$

But (4.35) is equivalent to

$$\begin{aligned} h'_1(\bar{u}; v - \bar{u}) + h'_2(\bar{u}; v - \bar{u}) - \langle E'(\bar{u})^* \bar{p}, v - \bar{u} \rangle_{U^*, U} \\ = h'(\bar{u}; v - \bar{u}) - \langle E'(\bar{u})^* \bar{p}, v - \bar{u} \rangle_{U^*, U} \geq 0 \quad \forall v \in U_{ad}. \end{aligned} \quad \square$$

From Proposition 3.5, for any $\bar{u} \in U_{ad}$

$$(4.36) \quad l_{\bar{u}}(u) = \frac{1}{2} \|u - \bar{u}\|_U^2$$

satisfies the global (P.S.) condition. Then, we can use Theorem 4.1, Definition 3.4 and Proposition 3.5 to obtain the necessary conditions for a solution to problem (4.4), under a weakened assumption to h .

THEOREM 4.2. *Assume that*

- (4.37) (1) (4.2)(1) and (3) hold;
 (2) $g: K \rightarrow \mathbf{R}_+$ is locally Lipschitz; $h: U_{\text{ad}} \rightarrow U\{+\infty\}$ is the sum of a $h_1 \in C^1(U_{\text{ad}}; \mathbf{R}_+)$ with compact derivative and a lower semicontinuous convex function $h_2: U_{\text{ad}} \rightarrow \mathbf{R}_+ U\{+\infty\}$.

Then $(\bar{y}, \bar{u}) \in K \times U_{\text{ad}}$ is a solution to the problem (4.4) only if (4.3) holds and there exists $\bar{p} \in V$ such that (4.5) holds.

Proof. Consider the following problem:

$$(4.38) \quad \begin{aligned} & \min \{g(y) + h(u) + \tfrac{1}{2}\|u - \bar{u}\|_U^2\}, \\ & y \in K, \quad u \in U_{\text{ad}}, \\ & \langle F(y) + E(u), z - y \rangle_{V^*, V} \geq 0 \quad \forall z \in K. \end{aligned}$$

Then, obviously, (\bar{y}, \bar{u}) is the unique solution to (4.38). By Definition 3.4 and Proposition 3.5, $\tilde{h}(u) := h(u) + \tfrac{1}{2}\|u - \bar{u}\|_U^2$ satisfies (4.2)(2). Hence, by Theorem 4.1, there exists $\bar{p} \in V$ such that

$$(4.5) \quad \begin{aligned} & \langle F(\bar{y}) + E(\bar{u}), \bar{p} \rangle_{V^*, V} = 0, \\ & F'(\bar{y})^* \bar{p} \in N_K(\bar{y}) + \partial g(\bar{y}), \\ & E'(\bar{u})^* \bar{p} \in N_{U_{\text{ad}}}(\bar{u}) + \partial \tilde{h}(\bar{u}). \end{aligned}$$

But,

$$\partial \tilde{h}(\bar{u}) = \partial(h + l_{\bar{u}})(\bar{u}) = \partial h(\bar{u}) + \partial l_{\bar{u}}(\bar{u}) = \partial h(\bar{u}).$$

Therefore, (4.5)' is the same as (4.5). \square

By the same way, we can prove the following general theorem.

THEOREM 4.3. *Assume that*

- (4.39) (1) (4.2)(1);
 (2) $F \in C^1(K \times U_{\text{ad}}; V^*)$ satisfies
 (i) $\forall R > 0, \quad \forall u \in B_{R^{\text{ad}}}^U := \{u \in U_{\text{ad}} \mid \|u\|_U < R\},$
 $\exists C_R > 0 \quad \forall y_1, y_2 \in K,$
 $\langle F(y_1, u) - F(y_2, u), y_1 - y_2 \rangle_{V^*, V} \geq C_R \|y_1 - y_2\|_V^2,$
 (ii) $\forall y \in K, u \mapsto F'_u(y, u)$, partial derivative with respect to u , is weakly continuous,
 (iii) $\forall u \in U_{\text{ad}}, F'_u(y_u, u)$ is compact, provided $\| -F(y_u, u) \|_u^K = 0$;
 (3) $J: K \times U_{\text{ad}} \rightarrow \mathbf{R}_+ U\{+\infty\}$ satisfies
 (i) $y \mapsto J(y, u)$ is locally Lipschitz and for any $R > 0$ and any $u \in B_{R^{\text{ad}}}^U$, $J(\cdot, u)$ has same locally Lipschitz constants;
 (ii) $J(y, u) = J_1(y, u) + h(u)$, where for all $y \in K$, $J_1(y, \cdot) \in C^1(U_{\text{ad}}; \mathbf{R}_+)$, $(y, u) \mapsto J'_{1u}(y, u)$ is continuous and $h: U_{\text{ad}} \rightarrow \mathbf{R}_+ U\{+\infty\}$ is lower semicontinuous convex;
 (iii) $\inf_{y \in K} J(y, u) \rightarrow +\infty$ as $\|u\|_U \rightarrow +\infty$ and for any $y \in K$, $J(y, \cdot)$ satisfies the global (P.S.) condition on U_{ad} .

Then, there exist $(\bar{y}, \bar{u}) \in K \times U_{\text{ad}}$, which satisfies

$$(4.40) \quad \langle F(\bar{y}, \bar{u}), z - \bar{y} \rangle_{V^*, V} \geq 0 \quad \forall z \in K,$$

and is a solution to the following problem:

$$(4.41) \quad \begin{aligned} & \min J(y, u), \\ & y \in K, \quad u \in U_{\text{ad}}, \\ & \langle F(y, u), z - u \rangle_{V^*, V} \geq 0 \quad \forall z \in K \end{aligned}$$

and $\bar{p} \in V$ such that

$$(4.42) \quad \begin{aligned} & \langle F(\bar{y}, \bar{u}), \bar{p} \rangle_{V^*, V} = 0, \\ & F'_y(\bar{y}, \bar{u})^* \bar{p} \in N_K(\bar{y}) + \partial_y J(\bar{y}, \bar{u}), \\ & F'_u(\bar{y}, \bar{u})^* \bar{p} \in N_{U_{\text{ad}}}(\bar{u}) + \partial_u J(\bar{y}, \bar{u}). \end{aligned}$$

If (4.39) (iii) is not necessarily satisfied and $(y, u) \mapsto J'_{1u}(y, u)$ is continuous for the weak topology of U_{ad} , then (4.40) and (4.42) are necessary conditions for a solution (\bar{y}, \bar{u}) to the problem (4.41).

Finally, we devote the end of this paper to some explanatory examples.

Example 1. (1) $V := H_0^1(\Omega)$ and $U := L^2(\Omega)$ with a bounded regular open subset $\Omega \subset \mathbf{R}^n$;

(2) $g(y) := \int_{\Omega} G(x, y(x)) \, dx$, where

$$\begin{aligned} G(x, y) &:= \int_0^y \hat{g}(x, t) \, dt \quad \text{with } \hat{g}(\cdot, \cdot) \in C(\Omega \times \mathbf{R}; \mathbf{R}_+), \\ \hat{g}(\cdot, t) &:= -\hat{g}(\cdot, -t), \\ |\hat{g}(x, t)| &\leq C_g |t|^{s_g} + C_{g_0}, \quad s_g \in \begin{cases} [1, (n+2)/(n-2)] & \text{if } n \geq 3, \\ [1, \infty[& \text{otherwise,} \end{cases} \end{aligned}$$

$h(u) := \int_{\Omega} H(x, u(x)) \, dx + \frac{1}{2} \int_{\Omega} |u(x)|^2 \, dx$, where

$$\begin{aligned} H(x, u) &:= \int_0^u \hat{h}(x, t) \, dt \quad \text{with } \hat{h}(\cdot, \cdot) \in C(\Omega \times \mathbf{R}; \mathbf{R}_+), \\ \hat{h}(\cdot, t) &:= -\hat{h}(\cdot, -t), \\ |\hat{h}(x, t)| &\leq C_h |t|^{s_h} + C_{h_0}, \quad s_h \in \begin{cases} [1, (n+2)/n] & \text{if } n \geq 3, \\ [1, \infty[& \text{otherwise,} \end{cases} \\ H(X, u) &\geq \alpha u^2 - C \quad \text{with } \alpha > -1; \end{aligned}$$

(3) $F(y) := -\Delta y(\cdot) + \varphi(\cdot, y(\cdot))$, where

$$\begin{aligned} & \varphi(\cdot, \cdot) \in C^1(\Omega \times \mathbf{R}; \mathbf{R}), \, y \mapsto \varphi(x, y) \text{ is increasing,} \\ |\varphi(x, t)| &\leq C_{\varphi} |t|^{s_{\varphi}} + C_{\varphi_0}, \quad s_{\varphi} \in \begin{cases} [1, (n+2)/(n-2)] & \text{if } n \geq 3, \\ [1, \infty[& \text{otherwise,} \end{cases} \end{aligned}$$

$E(u) := \Psi(\cdot, u(\cdot))$, where

$$\begin{aligned} & \Psi(\cdot, \cdot) \in C^1(\Omega \times \mathbf{R}; \mathbf{R}), \\ |\Psi(x, t)| &\leq C_{\Psi} |t|^{s_{\Psi}} + C_{\Psi_0}, \quad s_{\Psi} \in \begin{cases} [1, (n+2)/n] & \text{if } n \geq 3, \\ [1, \infty[& \text{otherwise;} \end{cases} \end{aligned}$$

(4) Formula (4.3) becomes

$$\int_{\Omega} \{ \nabla \bar{y}(x) \cdot \nabla (z(x) - \bar{y}(x)) + (\varphi(x, \bar{y}(x)) + \Psi(x, \bar{u}(x)))(z(x) - \bar{y}(x)) \} \, dx \geq 0$$

$$\forall z \in K \subset H_0^1(\Omega),$$

Formula (4.5) becomes

$$\begin{aligned} \int_{\Omega} \{ \nabla \bar{y}(x) \cdot \nabla \bar{p}(x) + (\varphi(x, \bar{y}(x)) + \Psi(x, \bar{u}(x))) \bar{p}(x) \} dx &= 0, \\ \int_{\Omega} \{ \nabla \bar{p}(x) \cdot \nabla (z(x) - \bar{y}(x)) + [\varphi'_y(x, \bar{y}(x)) \bar{p}(x) - g(x, \bar{y}(x))] (z(x) - \bar{y}(x)) \} dx &\leq 0 \\ \forall z \in K \subset H_0^1(\Omega), \\ \int_{\Omega} [\varphi'_u(x, \bar{u}(x)) \bar{p}(x) - h(x, \bar{u}(x)) - \bar{u}(x)] (v(x) - \bar{u}(x)) dx &\leq 0 \\ \forall v \in U_{ad} \subset L^2(\Omega). \end{aligned}$$

Example 2. (1) $V := H_0^{1,p}(\Omega)$, $1 < p \leq 2$ and $U := L^q(\Omega)$, $q > 1$ with a same $\Omega \subset \mathbf{R}^n$,

$$K \subset \{ y \in H_0^{1,p}(\Omega) \mid \|y\|_{H_0^{1,p}(\Omega)} \geq 1 \};$$

(2) $g(y)$ is the same as in Example 1(1), but

$$\begin{aligned} s_g \in \begin{cases} \left[1, \frac{n(p-1)+p}{n-p} \right] & \text{if } n \geq 3, \\ [1, \infty[& \text{otherwise,} \end{cases} \\ h(u) := \int_{\Omega} H(x, u(x)) dx + 1/q \int_{\Omega} |u(x)|^q dx \end{aligned}$$

where $H(\cdot, \cdot)$ is the same as in Example 1(2), but

$$\begin{aligned} s_h \in \begin{cases} \left[1, q \frac{n(p-1)+p}{n-p} \right] & \text{if } n \geq 3, \\ [1, \infty[& \text{otherwise,} \end{cases} \\ H(x, u) \geq \alpha |u|^q - C \quad \text{with } \alpha > -1; \\ (3) \quad F(y) := - \sum_{i=1}^n \frac{\partial}{\partial x_i} \left[\left| \frac{\partial y}{\partial x_i} \right|^{p-2} \frac{\partial y}{\partial x_i} \right] + \varphi(\cdot, y(\cdot)) \end{aligned}$$

where $\varphi(\cdot, \cdot)$ is the same as in Example 1(3), but

$$s_{\varphi} \in \begin{cases} \left[1, \frac{n(p-1)+p}{n-p} \right] & \text{if } n \geq 3, \\ [1, \infty[& \text{otherwise.} \end{cases}$$

Notice that

$$\langle F(y_1) - F(y_2), y_1 - y_2 \rangle_{H^{-1,p'}(\Omega), H_0^{1,p}(\Omega)} \geq \left(\frac{\|y_1 - y_2\|_{H_0^{1,p}(\Omega)}^2}{\|y_1\|_{H_0^{1,p}(\Omega)}^{2-p} + \|y_2\|_{H_0^{1,p}(\Omega)}^{2-p}} \right).$$

$E(u)$ is the same as in Example 1(3), but

$$s_{\Psi} \in \begin{cases} \left[1, q \frac{n(p-1)+p}{np} \right] & \text{if } n \geq 3, \\ [1, \infty[& \text{otherwise;} \end{cases}$$

(4) Formula (4.3) becomes

$$\int_{\Omega} \left\{ \sum_{i=1}^n \left| \frac{\partial y}{\partial x_i} \right|^{p-2} \frac{\partial y}{\partial x_i} \frac{\partial}{\partial x_i} (z(x) - \bar{y}(x)) + [\varphi(x, \bar{y}(x)) + \Psi(x, \bar{u}(x))] (z(x) - \bar{y}(x)) \right\} dx$$

$$\forall z \in K \subset H_0^{1,p}(\Omega);$$

the first and second representations of (4.5) are similar, and the third representation becomes

$$\int_{\Omega} \{ [\Psi'_u(x, \bar{u}(x)) \bar{p}(x) - h(x, \bar{u}(x)) - |\bar{u}(x)|^{q-2} \bar{u}(x)] (v(x) - \bar{u}(x)) \} dx \leq 0$$

$$\forall v \in U_{ad} \subset L^q(\Omega)$$

REFERENCES

- [1] J.-P. AUBIN AND I. EKKLAND, *Applied Nonlinear Analysis*, Wiley-Interscience, New York, 1984.
- [2] V. BARBU, *Necessary conditions for nonconvex distributed control problems governed by elliptic variational inequalities*, J. Math. Anal. Appl., 80 (1981), pp. 566-597.
- [3] ———, *Optimal Control of Variational Inequalities*, Pitman, Boston, 1984.
- [4] C. BERGE, *Espaces topologiques et fonctions multivoques*, Dunod, Paris, 1959.
- [5] F. E. BROWDER, *Nonlinear monotone operators and convex sets in Banach spaces*, Bull. Amer. Math. Soc., 71 (1965), pp. 780-785.
- [6] KUNG-CHING CHANG AND J. EELLS, *Unstable minimal surface coboundaries*, to appear.
- [7] KUNG-CHING CHANG AND SHUZHONG SHI, *A local minimax theorem without compactness*, in Nonlinear and Convex Analysis, Lin and Simons, eds., Marcel Dekker, New York, 1987, pp. 211-233.
- [8] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
- [9] I. EKKLAND, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324-353.
- [10] P. HARTMAN AND G. STAMPACCHIA, *On some nonlinear elliptic differential functional equations*, Acta Math., 115 (1966), pp. 153-188.
- [11] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.
- [12] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non-linéaires*, Dunod-Gauthier-Villars, Paris, 1969.
- [13] ———, *Some Aspects of the Optimal Control of Distributed Parameter Systems* CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1972.
- [14] ———, *Various topics in the theory of optimal control of distributed systems*, in Lecture Notes in Econom. and Math. Systems, 105, Springer, Berlin-New York, 1976, pp. 166-303.
- [15] F. MIGNOT, *Contrôle dans les inéquations variationnelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130-185.
- [16] F. MIGNOT AND J. P. PUEL, *Optimal control in some variational inequalities*, this Journal, 22 (1984), pp. 466-476.

APPROXIMATING THE LINEAR QUADRATIC OPTIMAL CONTROL LAW FOR HEREDITARY SYSTEMS WITH DELAYS IN THE CONTROL*

MARK H. MILMAN†

Abstract. A factorization approach is presented for deriving approximations to the optimal feedback gains for the linear regulator-quadratic cost problem associated with time-varying functional differential equations with control delays. The approach is based on a discretization of the state penalty which leads to a simple structure for the feedback control law. General properties of the Volterra factors of Hilbert-Schmidt operators are then used to obtain convergence results for the controls, trajectories and feedback kernels. Two algorithms are derived from the basic approximation scheme, including a fast algorithm, in the time-invariant case. A numerical example is also considered.

Key words. optimal control, delays, factorization

AMS(MOS) subject classification. 93

1. Introduction. This paper is concerned with the application of factorization techniques to approximating the optimal feedback gain for the finite time linear regulator quadratic cost problem for systems governed by retarded functional differential equations (RFDE) with control delays. Feedback control laws for these systems have been previously derived for both the finite and infinite time problems in several articles under various hypotheses (see for example [6], [13], [14], [17], [24], [25]). The departure point for this paper is the representation derived in [24], in which the feedback kernel is realized in terms of solutions to a certain nonlinear integral equation corresponding to a Volterra factorization problem. Heavy use is made of the variation of constants formula for RFDE's (due to Banks) and general factorization results for Hilbert-Schmidt operators [11], [22]. Difficulties, due to the presence of an unbounded input operator because of the control delay, are circumvented in the factorization approach. Other applications of Volterra factorization include, for example, filtering and smoothing of nonstationary processes over a finite interval [15], [16], inverse problems in the spectral theory of differential operators [8], [18], solutions to two point boundary value problems and Fredholm equations of the first and second kinds [11], [21].

Although the specific approximation problem we consider in this paper has not to the author's knowledge been treated in the literature, several articles (e.g., [5], [10], [19]) have considered approximating the feedback kernel in systems without control delay terms. The approach in each of these articles involves expressing the RFDE as an evolution equation in the state space $R^N \times L_2$, and then approximating the resulting dynamical system. Delfour [5] discretizes in both the spatial and time variables, while Kunisch [19] and Gibson [10] discretize in only the spatial variable. Delfour considers the time-varying problem and obtains weak convergence of the solutions to the approximating Riccati equations. In [10] and [19] the open-loop semigroup is first approximated by discretizing the history space, and then the approximation theory of

* Received by the editors February 14, 1984; accepted for publication (in revised form) May 6, 1987. This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

† Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109.

[9] is used for subsequent convergence analysis. This analysis is based on exploiting the relationship developed in [9] between the open-loop semigroup and the Riccati equation defining the feedback law. (We note that the integral Riccati equations of [9] are also equivalent to certain factorization problems [23], so that the convergence analysis in [10] and [19] can also be performed within a factorization context.) Ultimately, Kunisch demonstrates weak L_2 -convergence of the kernels, while Gibson establishes a strong L_2 -convergence. Neither provide a priori rates of convergence. Although Gibson and Kunisch restrict their attention to the time-invariant case, the approach is extendable to time-varying problems by using the approximations of Banks and Rosen [3]. We also note that Gibson importantly considers the infinite time problem.

The convergence analysis presented in [10] and [19] depends heavily on the fact that the control map has finite rank. This condition does not hold in the control delay problem and straightforward extensions of these convergence results to the control delay case are not apparent.

Our approach to the problem does not rely on analysis of the associated Riccati equations. It will be shown that if the cost on the state in the regulator problem is a discrete sum with no integral term, then the associated factorization problem is solved by matrix inversion, and the exact feedback kernel can be defined in terms of the fundamental matrix solution, quadrature, and the solution to finite-dimensional linear equations. (This form of the solution generalizes a result of Manitius [20] for the problem with terminal state penalty and no control delays.) Thus an approximation scheme for the problem containing an integral state penalty term can be developed by approximating this term by quadrature and solving exactly for the feedback kernel of the resulting discretized state cost problem. Using this approach together with factorization arguments we will be able to establish $O(1/n)$ L_∞ -convergence for the approximate feedback kernels in the time-varying control delay case.

This result is (analytically) somewhat sharper than Gibson's in that the L_∞ convergence of the kernels applies on the square as well as the diagonal, and also that a priori rates can be provided. The principle reason this sharper result can be obtained is that we exploit the fact that only the fundamental matrix solution is required to define the feedback kernel, and thus it is never necessary to consider the less tractable problem of approximating the entire semigroup. We now briefly outline the organization of the paper.

In § 2 the necessary mathematical preliminaries are developed and discussed. Because the approach does not follow along a Riccati synthesis, we will recapitulate in this section some of the relevant discussion from [24] concerning the relationship between Volterra factorization and the RFDE control problem.

Section 3 contains the L_∞ -convergence results for the feedback kernels, controls and trajectories. Instead of considering specific quadrature schemes approximating the state cost, all the results are proved with respect to a sequence of Borel measures satisfying certain convergence hypotheses. The key tool of this section is a factorization lemma which asserts that the factorization problem is well posed (in an appropriate sense) in the space of integral operators with essentially bounded kernels.

In § 4, the explicit form of the optimal feedback kernel associated with discrete state cost is derived. The resulting approximation scheme developed from the cost discretizations is then used as an analytical tool to obtain further results regarding the feedback kernel. For example, using essentially matrix manipulations, a Wiener-Hopf integral equation for the optimal feedback kernel is derived which is shown to be the

control delay generalization of the Wiener-Hopf equation Manitius [20] had previously derived for the feedback kernel via a maximum principle.

In § 5 two algorithms representing implementation of the basic approximation scheme of § 4 are derived. In the time-invariant case a fast algorithm is derived by exploiting the near Toeplitz structure of the system of equations that defines the feedback kernel. A simple numerical example is also presented.

2. Preliminaries. Let $[-r, T]$ denote a closed and bounded interval in the real line with $r \geq 0$ and $T > 0$, and let Σ denote the class of Borel subsets of $[-r, T]$. For an arbitrary Banach space Y , $|y|$ will denote the norm of an element $y \in Y$, $B(Y, Z)$ will denote the space of bounded linear maps from Y into another Banach space Z , and for brevity we write $B(Y)$ for $B(Y, Y)$. Subscripts will sometimes be attached to the norm of an element to remove any ambiguities that might arise due to the fact that several different topologies will be used in the paper. The notation A^* (respectively A') will be used to denote the adjoint (transpose) of an operator (matrix).

In the sequel the Banach space of continuous functions $C([-r, T], R^N)$ will be denoted X , the Hilbert space $L_2([-r, T], R^M)$ will be denoted U , and H will denote the Hilbert space $L_2([-r, T], R^N)$. Now define the resolution of the identity $E: \Sigma \rightarrow B(U)$ by multiplication by the characteristic function, i.e., $[E(\omega)u](t) = \chi(\omega)(t)u(t)$ ($\chi(\omega)(t) = 0$ if $t \notin \omega$, $\chi(\omega)(t) = 1$ if $t \in \omega$), and let P' denote the family of projections $E([-r, t])$. The complementary family $I - P'$, will be denoted P_t . Note that P' is strongly continuous, i.e., $t \rightarrow P_t u$ is continuous for each $u \in U$.

In this section we shall review some of the results in [24] pertaining to the linear regulator problem with dynamics

$$\begin{aligned} \dot{x}(t) &= \int_{-r}^0 d_\theta \eta(t, \theta) x(t + \theta) + (BP_0 u)(t), & t \geq 0, \\ x(t) &= \phi(t), & t \in [-r, 0] \end{aligned} \quad (2.1)$$

and quadratic cost functional

$$J(u, x) = \int_{-r}^T \langle x(s), Q(s)x(s) \rangle d\mu(s) + \int_{-r}^T |u(s)|^2 ds. \quad (2.2)$$

In (2.1), we assume that $\phi(t)$ is continuous, $x(\cdot) \in X$, $u \in U$, $B \in B(U, H)$ and P_0 is the projection $E([0, T])$. (It would not affect subsequent convergence analysis of the feedback kernels to allow an arbitrary projection P_{t_0} , $t_0 \in [-r, T]$, in place of P_0 . The choice $t_0 = 0$ reflects the problem formulation in which control policies cannot be implemented until time $t = 0$.) The only constraint we impose on B at this time is that it be causal, i.e., for each $t \in [-r, T]$, if $u_1 = u_2$ almost everywhere on $[0, t]$ then $(Bu_1)(s) = (Bu_2)(s)$ for almost every $s \leq t$. The matrix valued function η is assumed measurable on $R \times R$ and is normalized so that $\eta(t, \theta) = 0$ for $0 \leq \theta$ and $\eta(t, \theta) = \eta(t, -r)$ for $\theta \leq -r$. It is further assumed that $\eta(t, \cdot)$ is left continuous for each t and there exists a function $m \in L_1(0, T)$ such that

$$|\text{Var } \eta(t, \cdot)| \leq m(t) \quad (2.3)$$

where $|\cdot|$ denotes any matrix norm. In the cost (2.2), μ denotes an arbitrary positive regular Borel measure on $[-r, T]$, and $Q(\cdot)$ is Borel measurable with $Q(s) \geq 0$ μ almost everywhere and is μ -essentially bounded.

At this time it is convenient to introduce some operators that will play a prominent role in subsequent analysis. We define:

$$(2.4) \quad L \in B(X); \quad Lx: t \rightarrow \begin{cases} 0, & t \in [-r, 0], \\ \int_0^t \int_{-r}^0 d_\theta(s, \theta)x(s + \theta) ds, & t \geq 0, \end{cases}$$

$$(2.5) \quad F \in B(U, X); \quad Fu: t \rightarrow \int_{-r}^t F(t, s)u(s) ds$$

and the adjoint-like

$$(2.6) \quad F^\# \in B(X, U); \quad F^\#x: t \rightarrow \int_t^T F'(s, t)Q(s)x(s) d\mu(s)$$

where $F(t, s) = [P_0 B^* Y'(t, \cdot)]'(s)$ and $Y(t, s)$ is the fundamental matrix solution of the homogeneous problem (see [12]). $Y(\cdot, \cdot)$ satisfies the Volterra equation

$$(2.7) \quad Y(t, s) = \begin{cases} I - \int_s^t Y(t, \sigma)\eta(\sigma, s - \sigma) d\sigma, & s \leq t, \\ 0, & s > t, \end{cases}$$

and the solution to (2.1) can be realized as

$$(2.8) \quad \begin{aligned} x(t) = Y(t, 0)\phi(0) &+ \int_{-r}^0 d_\beta \int_0^t Y(t, \sigma)\eta(\sigma, \beta - \sigma) d\sigma \phi(\beta) \\ &+ \int_0^t Y(t, \sigma)(BP_0 u)(\sigma) d\sigma. \end{aligned}$$

Also, $\sup |Y(t, s)| < \infty$, $Y(t, s)$ is absolutely continuous for $t \geq s$ for each s , and $Y(t, \cdot)$ is of bounded variation for each t . We note from the definition of $F(t, s)$ that $F(t, s) = 0$ for $s < 0$. Hence, $FP_0 = F$ and $P_0 F^\# = F^\#$.

Using a completing the squares argument, the open loop control law for (2.1)–(2.2) can be easily derived in terms of the operators defined above.

THEOREM 2.1. *The optimal control \hat{u} for the regulator problem (2.1)–(2.2) is $\hat{u} = M\tilde{\phi}$, where $M \in B(X, U)$, $\tilde{\phi} \in X$,*

$$M = -(I + F^\# F)^{-1} F^\# (I - L)^{-1},$$

$$\tilde{\phi}(t) = \begin{cases} \phi(0), & t \geq 0, \\ \phi(t), & t \in [-r, 0]. \end{cases}$$

Proof. The proof is shown in [24].

In [24] the feedback control law for (2.1)–(2.2) was derived from the open loop control law above using certain embedding and Volterra factorization arguments. This paper also requires factorization arguments, and now we digress a bit to give some pertinent background. (The interested reader is referred to [11], [22] and [24] for more details.)

Let $G: [-r, T] \rightarrow B(U)$, and assume that G is strongly continuous, i.e., $t \rightarrow G(t)u$ is continuous for each $u \in U$. Now let $K \in B(U)$ be a Hilbert–Schmidt operator and consider Riemann sums of the form

$$\sum_i E(\theta_i)KG(t_i), \quad -r = t_0 < \cdots < t_n = T,$$

$\theta_i = [t_i, t_{i+1}]$, and $t'_i \in \theta_i$. These sums can be shown to converge in the operator norm as the mesh of the partitions tend to zero [11]. This limit is expressed as the projection integral

$$(2.9) \quad \int dE KG(t),$$

and it can further be shown that

$$(2.10) \quad \left\| \int dE KG(t) \right\|_{HS} \leq \|K\|_{HS} \sup |G(t)|.$$

Here $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm.

Two important projection integrals on the space \mathcal{K} of Hilbert-Schmidt operators in $B(U)$ are obtained from the selections $G_+(t) = P'$ and $G_-(t) = P_t$. The resulting operators, p_+ and p_- , respectively,

$$p_+(K) = \int dE KG_+(t) \quad \text{and} \quad p_-(K) = \int dE KG_-(t)$$

are bounded projections on \mathcal{K} . If $K \in R(p_+)(R(p_-))$ we say that K is causal (anticausal). The elements of $R(p_+)$ and $R(p_-)$ are quasinilpotents. In the sequel we shall also write $K_+(K_-)$ for $p_+(K)$ ($p_-(K)$). Note also that $P'U$ and P_tU are invariant subspaces of $p_+(K)$ and $p_-(K)$, respectively.

We note that in the space U a Hilbert-Schmidt map K is necessarily an integral operator with kernel, say $K(t, s)$. In this case the projections $p_+(K)$ and $p_-(K)$ are simply the Volterra operators

$$p_+(K)u : t \rightarrow \int_{-r}^t K(t, s)u(s) ds, \quad p_-(K)u : t \rightarrow \int_t^T K(t, s)u(s) ds.$$

With this bit of background we can now state the basic factorization results that will be used in the sequel.

THEOREM 2.2 (Gohberg-Krein). *Let $K \in \mathcal{K}$. Then there exist unique operators $X_{\pm} \in R(p_{\pm})$ such that*

$$(2.11) \quad I + K = (I + X_-)(I + X_+)$$

if and only if $(I + P_t K P_t)$ is invertible for each $t \in [-r, T]$. Furthermore, $W_- = (I + X_-)^{-1} - I$ is given by the projection integral

$$W_- = - \int dE KG(t)$$

with

$$G(t) = P_t(I + P_t K P_t)^{-1}.$$

The decomposition of $(I + K)$ into the product in (2.11) is called the Volterra or special (right) factorization of $I + K$, and we will sometimes refer to X_{\pm} as the causal (anticausal) factor of K . Note that uniqueness of the factorization implies that $X_- = (X_+)^*$ when K is self-adjoint.

Two results that will be useful in subsequent convergence analysis are stated as corollaries below.

COROLLARY 2.3. *Suppose $K \in \mathcal{K}$ is self-adjoint with*

$$\sup_t |(I + P_t K P_t)^{-1}| \leq k.$$

Then $I + K$ has the factorization (2.14) and $W_{\pm} = (I + X_{\pm})^{-1} - I$ satisfy

$$|W_{\pm}|_{HS} \leq k|K|_{HS}.$$

COROLLARY 2.4. Let $K_i \in \mathcal{H}$, $i = 1, 2$ be self-adjoint with

$$\sup_i |(I + P_i K_i P_i)^{-1}| \leq k_i.$$

Let $W_i = (I + X_i)^{-1} - I$ where X_i denotes the causal factor of K_i . Then

$$|W_1 - W_2|_{HS} \leq \min_{i \neq j} k_i |K_1 - K_2|_{HS} (1 + k_j |K_j|_{HS}).$$

Proof. It suffices to prove the result for the adjoint operators. Thus let $G_i(t) = P_i(I + P_i K_i P_i)^{-1}$, and note that

$$\begin{aligned} |W_1^* - W_2^*|_{HS} &\leq \left| \int dE K_1 (G_1 - G_2) \right|_{HS} + \left| \int dE (K_1 - K_2) G_2 \right|_{HS} \\ &\leq |K_1|_{HS} |K_1 - K_2|_{HS} k_1 k_2 + |K_1 - K_2|_{HS} k_2 \\ &= k_2 |K_1 - K_2|_{HS} (1 + |K_1|_{HS} k_1). \end{aligned}$$

Symmetry of the argument with respect to K_1 and K_2 gives the result. \square

In [22] the Lebesgue analogue of the projection integral was introduced. This outgrowth of the projection integral provides the basis for the results of [24], which in turn is the departure point for the present paper. We will not, however, have any need for working directly with this variation of the projection integral. The current background is sufficient for defining the feedback solution to (2.1)–(2.2) and performing subsequent analysis. (Complete details can be found in [24].)

Define the Hilbert space $H_{\mu} = L_2([-r, T], R^N; \mu)$ as the space of μ -square integrable functions on $[-r, T]$ with values in R^N . It is evident that the map $Q: H_{\mu} \rightarrow H_{\mu}$ defined by $(Qx)(t) = Q(t)x(t)$ is bounded and that $F^{\#}$ has the representation $F^{\#} = F^*Q$ where F^* is the adjoint of F considered as an element of $B(U, H_{\mu})$. Hence, $F^{\#}F = F^*QF \geq 0$. In [24] it is verified that $F^{\#}F$ is Hilbert-Schmidt, so that Theorem 2.2 implies that $(I + F^{\#}F)$ has the factorization

$$(2.12) \quad I + F^{\#}F = (I + X^*)(I + X)$$

with X causal.

Since $I + X^*$ is invertible (recall that X^* is a quasinilpotent) let $W^* = (I + X^*)^{-1} - I$. And furthermore, since W^* is Hilbert-Schmidt, it has an $M \times M$ matrix kernel $W_-(t, s)$. Next define the $M \times N$ matrix valued function $P(t, \alpha)$ on $[-r, T] \times [-r, T]$ by

$$(2.13) \quad P(t, \alpha) = \int_{\alpha}^T K(t, s) Y(s, \alpha) d\mu(s)$$

where

$$(2.14) \quad K(t, s) = F'(s, t)Q(s) + \int_t^s W_-(t, \sigma)F'(s, \sigma)Q(s) d\sigma$$

and $Y(t, s)$ defined as in (2.7). The function $P(t, \alpha)$ provides the feedback solution to the regulator problem. This is made precise in the following.

THEOREM 2.5. *The optimal feedback control for (2.1)–(2.2) is given by*

$$\begin{aligned}\hat{u}(t) = & -P(t, t)\hat{x}(t) - \int_t^{\min(T, t+r)} P(t, \alpha) \int_{t-r}^t d_\beta(\alpha, \beta - \alpha) \hat{x}(\beta) d\alpha \\ & - \int_t^T P(t, \alpha)(BP'u)(\alpha) d\alpha\end{aligned}$$

where \hat{x} denotes the optimal trajectory, and for each t , P' denotes the projection on U , $(P'u)(\varepsilon) = \chi[-r, t](s)u(s)$. Furthermore, $P(t, \alpha)$ is square integrable (Lebesgue measure) on both the diagonal and the square $[-r, T] \times [-r, T]$.

Proof. The proof is shown in [24].

3. Convergence results. The specific optimization problem we shall be considering is the following:

$$(3.1) \quad \min J(u, x) = \langle x(T), Q_0 x(T) \rangle + \int_{-r}^T \langle x(s), Q(s)x(s) \rangle + |u(s)|^2 ds$$

subject to the constraint

$$(3.2) \quad x(t) = \int_{-r}^0 d_\theta(t, \theta)x(t+\theta) + (BP_0 u)(t), \quad t \geq 0,$$

$$(3.3) \quad x(t) = \phi(t), \quad t \in [-r, 0],$$

where $\phi(\cdot) \in C([-r, 0], \mathbb{R}^N)$ and

$$(3.4) \quad (Bu)(t) = \sum_{i=0}^k \chi[r, -r, T](t) B_i(t) u(t-r_i) + \int_{t-r}^t B(t, \theta) u(\theta) d\theta.$$

The assumptions on $\eta(\cdot, \cdot)$ are the same as in the preceding section. We shall assume that $0 = r_0 > -r_1 > \dots > -r_k = -r$ and $\sup |B_i(t)| = b_i < \infty$, $\sup |B(t, \theta)| = b < \infty$. In the cost (3.1) we impose continuity on $Q(s)$.

Interpreting these assumptions in the context of § 2, we have $\mu = \lambda + \delta$ where λ denotes Lebesgue measure, δ is the Dirac measure with support on $\{T\}$, and $Q(\cdot)$ is uniformly continuous on $[-r, T]$ with $Q(T) = Q_0$.

Now consider the following sequence $\{J_n\}$ of approximations to the cost $J(u, x)$:

$$(3.5) \quad J_n(u, x) = \int_{-r}^T \langle x(s), Q(s)x(s) \rangle d\mu_n(s) + \int_{-r}^T |u(s)|^2 ds$$

where μ_n is a sequence of positive regular Borel measures such that:

$$(H1) \quad \mu_n(T) = 1 \text{ for all } n.$$

$$(H2) \quad \text{Given } \varepsilon > 0 \text{ there exist } m \text{ such that } n \geq m \text{ implies } |\mu_n[a, b]| - |b - a| < \varepsilon \\ \text{for all } a < b; a, b \in [-r, T].$$

In this section we will discuss the convergence properties of the solution and feedback laws corresponding to the cost approximation above. Henceforth we refer to the optimization problem with cost (3.1) as problem \mathcal{P} , and the problem with cost (3.5) as problem \mathcal{P}_n . Unless otherwise noted, subscripts appearing on operators, functions, etc. (e.g., F_n) will indicate that these terms are associated with problem \mathcal{P}_n .

LEMMA 3.1. Let μ be defined as above and let μ_n satisfy (H1) and (H2). Then $\mu_n - \delta \rightarrow \lambda$ (Lebesgue measure) in the w^* topology of $C^*(-r, T)$.

Proof. The proof is shown in [34].

Since by definition $F(t, s) = [P_0 B^* Y'(t, \cdot)]'(s)$ (cf. (2.5)–(2.6)) from (3.4) it follows that

$$(3.6) \quad F(t, s) = \sum_i f_i(t, s)$$

where

$$f_0(t, s) = Y(t, s)B_0(s) + \int_s^{\min(s+r), t} Y(t, \theta)B(\theta, s) d\theta$$

and

$$f_i(t, s) = Y(t, s+r_i)B_i(s+r_i), \quad i = 1, \dots, k.$$

Now let $\gamma = \sup |Y(t, s)|$. Then using [12, p. 149] and the bounds in (3.4), we have

$$(3.7) \quad \sup |f_0(t, s)| \leq \gamma(b_0 + br), \quad \sup |f_i(t, s)| \leq b_i \gamma.$$

Also, for $t_2 > t_1 \geq s + r_i$,

$$(3.8) \quad |f_0(t_2, s) - f_0(t_1, s)| \leq \exp |m|_1 \int_{t_1}^{t_2} m(\sigma) d\sigma \{b_0 + br\} + b\gamma(t_2 - t_1)$$

and

$$(3.9) \quad |f_i(t_2, s) - f_i(t_1, s)| \leq b_i \exp |m|_1 \int_{t_1}^{t_2} m(\sigma) d\sigma, \quad i = 1, \dots, k.$$

It is not difficult to show (see [24]) that $F^* = F^*j$, $F_n^* = F^*j_n$ where F^* is the B -space adjoint of F , i.e., $F^*: X^* \rightarrow U$, and j, j_n are the mappings of X into X^* ,

$$j(x)y = \int \langle y(s), Q(s)x(s) \rangle d\mu(s),$$

$$j_n(x)y = \int \langle y(s), Q(s)x(s) \rangle d\mu_n(s).$$

Now it follows easily from definition and the estimates above that F is compact. Thus, using the w^* -convergence of $j_n(x) \rightarrow j(x)$ for each x (from Lemma 3.1), it can then be deduced that $F_n^* \rightarrow F^*$ strongly. Consequently from the compactness of F it also follows that $F_n^* F \rightarrow F^* F$ uniformly. Noting the form of the open loop control law (in Theorem 2.1), these general considerations are enough to demonstrate the L_2 -convergence of the approximate optimal controls and the uniform convergence of the corresponding optimal trajectories resulting from approximations based on \mathcal{P}_n . However, the major aim of this section is to produce the stronger L_∞ -convergence of the approximations for the feedback kernels as well as the controls, and this requires a somewhat more specific analysis.

Let Z denote the space $L_\infty([-r, T], R^M)$. From the definition of $F(t, s)$ it is evident that F^* and F_n^* are also in $B(X, Z)$. Our first result sharpens the convergence of $F_n^* \rightarrow F^*$ discussed above. This result (and the method of proof) will form the basis for the L_∞ -convergence arguments later.

LEMMA 3.2. $F_n^* \rightarrow F^*$ strongly in $B(X, Z)$.

Proof. Let $x \in X$. By definition

$$[F^* - F_n^*]x: t \rightarrow \sum_{i=0}^k \int f'_i(s, t)Q(s)x(s) d(\mu - \mu_n)(s).$$

For each i define

$$\tilde{f}_i(s, t) = \begin{cases} f_i(s, t), & s > t + r_i, \\ f_i(t + r_i, t), & s \leq t + r_i. \end{cases}$$

Similarly define

$$Q_i(s, t) = \begin{cases} Q(s), & s > t + r_i, \\ Q(t + r_i), & s \leq t + r_i \end{cases}$$

and

$$x_i(s, t) = \begin{cases} x(s), & s > t + r_i, \\ x(t + r_i), & s \leq t + r_i. \end{cases}$$

Considered as families of functions parameterized by t , $\{\tilde{f}_i(\cdot, t)\}$ is equicontinuous by virtue of (3.7)–(3.9), and $\{Q_i(\cdot, t)\}$ and $\{x_i(\cdot, t)\}$ are equicontinuous by virtue of the uniform continuity of $Q(\cdot)$ and $x(\cdot)$, respectively. Furthermore these families are clearly uniformly bounded. Hence we have the set

$S = \{x_t \in C([-r, T], R^M) : x_t(s) = f_i(s, t)Q_i(s, t)x_i(s, t), t \in [-r, T - r_i], i = 0, 1, \dots, k\}$ is relatively compact in $C([-r, T], R^N)$. Now note that

$$\begin{aligned} |[F^* - F_n^*](x)(t)| \leq \sum_{i=0} \left\{ \left| \int_{-r}^T f_i(s, t)Q(s, t)x_i(s, t)d(\mu - \mu_n)(s) \right| \right. \\ \left. + \left| \int_{-r}^{t+r} f_i(s, t)Q_i(s, t)x_i(s, t)d(\mu - \mu_n)(s) \right| \right\}. \end{aligned}$$

By the compactness of S it follows from Lemma 3.1 that the first integral above converges to zero uniformly with respect to t . And since the second integral has constant integrand for each t and i , uniform convergence is obtained here by using (H2). \square

Now let $H(t, s)$ and $H_n(t, s)$ denote the kernels of F^*F and F_n^*F , respectively. Fubini's theorem implies that

$$(3.10) \quad |H_n(t, s) - H(t, s)| \leq \left| \sum_{i,j} \int f'_i(\sigma, t)Q(\sigma)f_i(\sigma, s)d(\mu - \mu_n)(\sigma) \right|.$$

Using an argument similar to the one in the lemma above, we can prove the following result (see [34]).

LEMMA 3.3. *With the notation above, $H_n(t, s) \rightarrow H(t, s)$ uniformly on $[-r, T] \times [-r, T]$. In particular, $F_n^*F \rightarrow F^*F$ in the Hilbert-Schmidt topology (in $B(U)$) and in the $B(Z)$ topology.*

To obtain L_∞ -convergence of the feedback kernels we will use a result analogous to the one above regarding the convergence of the kernels of the Volterra factors of $I + F_n^*F$. Already we can use Corollary 2.4 and Lemma 3.3 to obtain Hilbert-Schmidt convergence of the factors (hence, L_∞ -convergence of their kernels). But our ultimate aim is to demonstrate the stronger L_∞ -convergence of the kernels. We shall need the following two results which are of some interest in their own right.

PROPOSITION 3.4. *Let K be an integral operator on U with essentially bounded kernel $K(t, s)$ where $\text{ess sup}_{t,s} |K(t, s)| = \beta < \infty$. Suppose that*

$$(3.11) \quad \sup_t |(I + P_t K P_t)^{-1}| = \alpha < \infty$$

so that $I + K$ has the factorization

$$(3.12) \quad I + K = (I + X^*)(I + X)$$

with X causal (Theorem 2.2). Then $W = (I + X)^{-1} - I$ and W^* are integral operators with kernels $W_{\pm}(t, s)$ satisfying the bound

$$(3.13) \quad \operatorname{ess\,sup}_{t,s} |W_{\pm}(t, s)| < \rho(\alpha, \beta)$$

where

$$\rho(\alpha, \beta) = \beta[1 + \beta(T+r)(1 + \alpha\beta(T+r))^2].$$

Proof. First note that Corollary 2.3 implies $W = (I + X)^{-1} - I$ is Hilbert-Schmidt with $|W|_{HS} \leq \alpha|K|_{HS}$. Now the factorization (3.12) implies that

$$(I + X^*) = (I + K)(I + W).$$

Subtracting the identity from the above and applying the projection p_- (i.e., taking anticausal parts) results in, for $\theta \geq t$,

$$X_-(t, \theta) = K(t, \theta) + \int_{\theta}^T K(t, \sigma) W_+(\sigma, \theta) d\sigma \quad \text{a.e. } t, \theta$$

where $X_-(t, \theta)$ and $W_+(t, \theta)$ are the kernels of X^* and W , respectively. Hence for almost every t, θ ,

$$|X_-(t, \theta)| \leq \beta \left\{ 1 + \int_{-r}^T |W_+(\sigma, \theta)| d\sigma \right\}.$$

Consequently,

$$\begin{aligned} \int_{-r}^T |X_-(t, \theta)|^2 d\theta &\leq \beta^2 \int_{-r}^T \left\{ 1 + \int_{-r}^T |W_+(\sigma, \theta)| d\sigma \right\}^2 d\theta \\ &= \beta^2 \left\{ (T+r) + 2 \int_{-r}^T \int_{-r}^T |W_+(\sigma, \theta)| d\sigma d\theta \right. \\ &\quad \left. + \int_{-r}^T \left[\int_{-r}^T |W_+(\sigma, \theta)| d\sigma \right]^2 d\theta \right\} \\ &\leq \beta^2 (T+r) (1 + |W_+|_{HS})^2 \\ &\leq \beta^2 (T+r) (1 + \alpha|K|_{HS})^2 \\ &\leq \beta^2 (T+r) [1 + \alpha\beta(T+r)]^2 \quad \text{a.e. } t. \end{aligned}$$

Let $\tilde{\alpha} = \beta^2(T+r)[1 + \alpha\beta(T+r)]^2$. Then since W^* satisfies the identity $W^* = -X^* - X^*W^*$, we have

$$|W_-(t, \theta)| \leq |X_-(t, \theta)| + \int_{-r}^T |X_-(t, \sigma) W_-(\sigma, \theta)| d\sigma \quad \text{a.e. } t, \theta.$$

Hence,

$$\begin{aligned} \int_{-r}^T |W_-(t, \theta)|^2 d\theta &\leq \int_{-r}^T |X_-(t, \theta)|^2 d\theta + 2 \int_{-r}^T |X_-(t, \theta)| \int_{-r}^T |X_-(t, s)| |W_-(s, \theta)| ds d\theta \\ &\quad + \int_{-r}^T \left[\int_{-r}^T |X_-(t, s)| |W_-(s, \theta)| ds \right]^2 d\theta \\ &\leq \tilde{\alpha} + 2\tilde{\alpha} |W_-|_{HS} + \tilde{\alpha} |W_-|_{HS}^2 \\ &\leq \tilde{\alpha} (1 + \alpha|K|_{HS})^2 \quad \text{a.e. } t. \end{aligned}$$

Thus,

$$(3.14) \quad \operatorname{ess\,sup}_t \int_{-r}^T |W_-(t, \theta)|^2 d\theta \leq \beta^2(T+r)[1 + \alpha\beta(T+r)]^4.$$

Now (3.12) also implies that

$$I + X = (I + W^*)(I + K).$$

Subtracting the identity and applying p_- yields

$$W^* + K_- + [W^*K]_- = 0.$$

Hence,

$$\begin{aligned} |W_-(t, \theta)| &\leq |K(t, \theta)| + \int_{-r}^T |W_-(t, s)| |K(s, \theta)| ds \\ &\leq \beta \left\{ 1 + (T+r)^{1/2} \operatorname{ess\,sup}_t \left[\int_{-r}^T |W_-(t, s)|^2 ds \right]^{1/2} \right\}. \end{aligned}$$

The result follows from (3.14). \square

As the result above may be regarded as the L_∞ analogue of Corollary 2.3, the next proposition is the L_∞ analogue of Corollary 2.4. The notation $X, W, X_\pm(t, s)$ and $W_\pm(t, s)$ will have the same meaning below as in the preceding proposition.

PROPOSITION 3.5. *Let K be as in the proposition above and let $\{K_n\}$ denote a sequence of integral operators on U with essentially bounded kernels $K_n(t, s)$ such that*

$$\operatorname{ess\,sup}_{t,s} |K_n(t, s)| \leq \delta$$

and

$$\lim_n \operatorname{ess\,sup}_{t,s} |K_n(t, s) - K(t, s)| = 0.$$

Assume further that $K_n \geq 0$ for each n so that $I + K_n$ has the factorization

$$I + K_n = (I + X_n^*)(I + X_n) \quad (\text{with } X_n \text{ causal}).$$

Let $W_n(t, s)$ denote the kernel of the integral operator $(I + X_n)^{-1} - I$. Then

$$\lim_n \operatorname{ess\,sup}_{t,s} |W_n(t, s) - W_+(t, s)| = 0.$$

Proof. First write $I + K_n = I + K + (K_n - K)$ so that

$$(3.15) \quad I + K_n = (I + X_-)(I + A_n)(I + X_+)$$

where $A_n = (I + X^*)^{-1}(K_n - K)(I + X)^{-1}$. By Proposition 3.4 $\operatorname{ess\,sup}_{t,s} |W_\pm(t, s)| \leq \alpha$ for some $\alpha < \infty$. Now since $I + K_n$ has the factorization, so does $I + A_n$. Specifically, $I + A_n = (I + Y_n^*)(I + Y_n)$ where $I + Y_n = (I + X_n)(I + W)$. It then follows from the identity

$$I + P_t A_n P_t = (I + P_t Y_n^* P_t)(I + P_t Y_n P_t)$$

that

$$\begin{aligned} \sup_t |(I + P_t A_n P_t)^{-1}| &\leq \sup_t |(I + P_t Y_n P_t)^{-1}|^2 \\ &\leq (1 + |W_n|)^2 \sup |(I + P_t W P_t)^{-1}|^2 \\ &\leq (1 + |K_n|_{HS})^2 [\exp \{ \tfrac{1}{2}(1 + |K|_{HS}^2) \}]^2. \end{aligned}$$

Here we have used Corollary 2.3 to obtain the first term in the product, and the fact that W is Hilbert-Schmidt and quasinilpotent together with [7, p. 1039] and

Corollary 2.3 to obtain the second term. Now define $Z_n = (I + Y_n)^{-1} - I$ and let $Z_n(t, s)$ denote its kernel. Then since

$$\operatorname{ess\,sup}_{t,s} |A_n(t, s)| \leq \beta_n [1 + \alpha(T+r) + \alpha^2(T+r)^2]$$

where $\beta_n = \operatorname{ess\,sup}_{t,s} |K(t, s) - K_n(t, s)|$ and $\alpha \geq \operatorname{ess\,sup}_{t,s} |W_{\pm}(t, s)|$, Proposition 3.4 implies that

$$\operatorname{ess\,sup}_{t,s} Z_n(t, s) \leq \rho(\tilde{\alpha}_n, \tilde{\beta}_n)$$

with

$$\tilde{\beta}_n = \beta_n [1 + \alpha(T+r) + \alpha^2(T+r)^2]$$

and

$$\tilde{\alpha}_n = [1 + |K_n|_{HS}]^2 [\exp \{ \frac{1}{2} (1 + |K|_{HS}^2) \}]^2.$$

Finally, note that

$$W_n = W + Z_n + WZ_n,$$

so that

$$\operatorname{ess\,sup}_{t,s} |W_n(t, s) - W_+(t, s)| \leq \rho(\tilde{\alpha}_n, \tilde{\beta}_n) [1 + (T+r)\alpha].$$

But $\rho(\tilde{\alpha}_n, \tilde{\beta}_n) = O(\beta_n)$. This completes the proof. \square

Before proceeding to the main result of the section, we state the following “open loop” result concerning convergence of the approximate control sequence and trajectories. Below, Z again denotes the space $L_{\infty}([-r, T], R^N)$. (See [34] for proof.)

THEOREM 3.6. *Let B denote the unit ball in $C([-r, 0], R^N)$, and for $\phi \in B$ let $\hat{u}_n(\phi)$ and $\hat{u}(\phi)$ denote the optimal controls for problems \mathcal{P}_n and \mathcal{P} , respectively. Also let $\hat{x}_n(\phi)$ and $\hat{x}(\phi)$ denote the corresponding trajectories. Then uniformly on B ,*

$$(i) \quad \lim_n |\hat{u}_n(\phi) - \hat{u}(\phi)|_z = 0,$$

$$(ii) \quad \lim_n |\hat{x}_n(\phi) - \hat{x}(\phi)|_x = 0.$$

Next we present the convergence properties of the sequence of approximating feedback kernels.

THEOREM 3.7. *Let $P_n(t, \alpha)$ and $P(t, \alpha)$ denote the feedback kernels of Theorem 2.5 associated with problems \mathcal{P}_n and \mathcal{P} . Then,*

$$(i) \quad \lim_n \operatorname{ess\,sup}_{t,\alpha} |P_n(t, \alpha) - P(t, \alpha)| = 0,$$

$$(ii) \quad \lim_n \operatorname{ess\,sup}_t |P_n(t, t) - P(t, t)| = 0.$$

Proof. From (2.13)–(2.14) we can write

$$(3.16) \quad \begin{aligned} |P_n(t, \alpha) - P(t, \alpha)| \leq & \sum_{i=0}^k \left\{ \left| \int_{\alpha}^T [K_{n,i}(t, s) - K_i(t, s)] Y(s, \alpha) d\mu_n(s) \right| \right. \\ & \left. + \left| \int_{\alpha}^T K_i(t, s) Y(s, \alpha) d(\mu - \mu_n)(s) \right| \right\} \end{aligned}$$

where

$$K_{n,i}(t, s) = f'_i(s, t)Q(s) + \int_t^{s-r_i} W_n(t, \theta) f'_i(s, \theta) Q(s) d\theta$$

and

$$K_i(t, s) = f'_i(s, t)Q(s) + \int_t^{s-r_i} W(t, \theta) f'_i(s, \theta) Q(s) d\theta.$$

Now note that

$$|K_{n,i}(t, s) - K_i(t, s)| \leq \int_t^{s-r_i} |W_n(t, \theta) - W(t, \theta)| |f'_i(s, \theta)| |Q(s)| d\theta.$$

Lemma 3.3 and Proposition 3.5 imply that

$$\lim_n \operatorname{ess\,sup}_{t, \theta} |W_n(t, \theta) - W(t, \theta)| = 0.$$

And since $|f'_i(s, \theta)|$ and $|Q(s)|$ are uniformly bounded, routine arguments yield a measurable set $\Omega x[-r, T]$ whose complement has zero Lebesgue measure such that $K_{n,i}(t, s) \rightarrow K_i(t, s)$ uniformly on $\Omega x[-r, T]$.

By the uniform boundedness of $Y(s, \alpha)$ and the sequence of measures μ_n , it follows that the first integral in (3.16) tends to zero uniformly on $\Omega x[-r, T]$. To prove convergence of the second integral in (3.16) we argue as in Lemma 3.2. Define the family of functions $\{\tilde{Y}\}$ parameterized by α and β ,

$$\tilde{Y}(s, \alpha, \beta) = \begin{cases} Y(s, \alpha), & s \geq \beta, \\ Y(\beta, \alpha), & s < \beta \end{cases}$$

and the family of functions $\{\tilde{K}_i\}$ parameterized by t and τ ,

$$\tilde{K}_i(t, \tau, s) = \begin{cases} K_i(t, s), & s \geq \tau, \\ K_i(t, \tau), & s < \tau. \end{cases}$$

It is straightforward to verify using the properties of K_i and Y that the set

$$\{\tilde{K}_i(t, \delta, \cdot) \tilde{Y}(\cdot, \alpha, \delta) : \delta = \max\{\alpha, t + r_i\}\}$$

is relatively compact in $C([-r, T], R^{M \times N})$. Thus the argument in Lemma 3.2 applies here to demonstrate in the second integral that

$$\lim_n \operatorname{ess\,sup}_{t, \alpha} \left| \int K_i(t, s) Y(s, \alpha) d(\mu_n - \mu)(s) \right| = 0.$$

The theorem is proved. \square

4. Applications. In this section we begin by deriving the optimal feedback kernel associated with an arbitrary discrete state cost penalty. It will be evident that given the fundamental matrix $Y(t, s)$, the feedback kernel in this case can be derived by quadrature and matrix inversion. This feedback structure when combined with the results of the preceding section leads to approximations to the optimal feedback kernel of problem \mathcal{P} (recall (3.1)–(3.3)), a Wiener–Hopf characterization of this kernel, and a priori bounds on its magnitude.

Let β denote a positive discrete measure on $[-r, T]$ of the form

$$\int f d\beta = f(s_n) + \sum_i a_i f(s_i), \quad s_i \in [-r, T], \quad s_n = T.$$

Inserting this measure into (2.2) results in the cost

$$(4.1) \quad J(u, x) = \langle x(T), Q(T)x(T) \rangle + \sum_i a_i \langle x(s_i), Q(s_i)x(s_i) \rangle + \int_{-r}^T |u(s)|^2 ds.$$

The optimal feedback kernel for this cost has the following semiseparable¹ structure.

THEOREM 4.1. Define the matrix functions $G(t)$ and $\bar{Y}(t)$,

$$G(t) = [\sqrt{a_0}F'(s_0, t); \cdots; \sqrt{a_{n-1}}F'(s_{n-1}, t); F'(s_n, t)]',$$

$$\bar{Y}(t) = [\sqrt{a_0}Y'(s_0, t); \cdots; \sqrt{a_{n-1}}Y'(s_{n-1}, t); Y'(s_n, t)]'.$$

Also define the block diagonal matrix \tilde{Q} by

$$\tilde{Q} = \text{diag}(Q(s_0), \cdots, Q(s_n)).$$

Then the optimal feedback kernel $P(t, \alpha)$ for the problem with dynamics (2.1) and cost (4.1) can be expressed as

$$(4.2) \quad P(t, \alpha) = G'(t)\tilde{Q}[I + U(t)]^{-1}\bar{Y}(\alpha)$$

where

$$(4.3) \quad U(t) = \int_t^T G(s)G'(s)\tilde{Q}ds.$$

Proof. It is straightforward to verify that the map F^*F resulting from the measure β in (4.1) is an integral operator with separable kernel $G'(t)\tilde{Q}G(s)$. Now Theorem 2.5 implies that

$$P(t, \alpha) = \sum_{s_i \cong \alpha} a_i K(t, s_i) Y(s_i, \alpha).$$

Here

$$K(t, s_i) = F'(s_i, t)Q(s_i) + \int_t^T W_-(t, r)F'(s_i, r)Q(s_i)dr$$

and $W_-(t, r)$ denotes the kernel of the anticausal operator W^* in the factorization

$$(I + F^*F)^{-1} = (I + W)(I + W^*).$$

Using the fact that F^*F has a separable kernel, it can be verified [11, p. 188] that

$$W_-(t, r) = -G'(t)\tilde{Q}[I + U(t)]^{-1}G(r)$$

with $U(t)$ defined as in (4.3). Thus we can write

$$P(t, \alpha) = \sum \left\{ \sqrt{a_i}F'(s_i, t)Q(s_i) - G'(t)\tilde{Q}[I + U(t)]^{-1} \cdot \int_t^T G(r)\sqrt{a_i}F'(s_i, r)Q(s_i)dr \right\} \sqrt{a_i}Y(s_i, \alpha).$$

Noting the definitions of G , \bar{Y} and U , it follows that

$$\begin{aligned} P(t, \alpha) &= [G'(t)\tilde{Q} - G'(t)\tilde{Q}[I + U(t)]^{-1}U(t)]\bar{Y}(\alpha) \\ &= G'(t)\tilde{Q}[I + U(t)]^{-1}\bar{Y}(\alpha). \end{aligned}$$

□

¹ The kernel $K(t, s)$ of an integral operator K is semiseparable if there exist matrix functions $H_i(t)$ and $G_i(s)$, $i = 1, 2$, such that $K(t, s) = H_1(t)G_1(s)$ for $s < t$, and $K(t, s) = H_2(t)G_2(s)$ for $s \geq t$. The kernel is separable if we can choose $H_1 = H_2$ and $G_1 = G_2$. In this case the associated operator K has finite rank.

Note that if in (4.1) we take for $i \neq n$, $a_i = 0$, and let the operator B denote a multiplication operator, $Bu: t \rightarrow B(t)u(t)$, then (4.2) reduces to Manitius' result for terminal state penalty [20],

$$P(t, \alpha) = B'(t)Y'(T, t)Q(T) \left[I + \int_t^T Y(T, r)B(r)B'(r)Y'(T, r)Q(T) dr \right]^{-1} Y(T, \alpha).$$

Thus Theorem 4.1 can be viewed as an extension of this result to problems with control delay and arbitrary discrete state penalty.

Now let μ_n be a sequence of discrete positive measures satisfying (H1) and (H2), and let $P_n(t, \alpha)$ denote the corresponding feedback kernels. The theorem implies each $P_n(t, \alpha)$ has the semiseparable form (4.2), while Theorem 3.7 implies the L_∞ -convergence of $P_n(t, \alpha)$ to $P(t, \alpha)$ (the optimal feedback kernel for problem \mathcal{P}) and also the L_∞ -convergence of $P_n(t, t)$ to $P(t, t)$. Introducing subscripts in the obvious way, define for each n ,

$$V_n(t) = [I + U_n(t)]^{-1} - I,$$

so that the identity

$$V_n(t) = -U_n(t)[I + U_n(t)]^{-1}$$

holds. Multiplying by $G'_n(t)\tilde{Q}_n$ we obtain

$$G'_n(t)\tilde{Q}_n V_n(t) = -G'_n(t)\tilde{Q}_n [I + U_n(t)]^{-1} U_n(t).$$

If we use the definitions of U_n and multiply by \bar{Y}_n it follows that

$$(4.4) \quad G'_n(t)\tilde{Q}_n V_n(t)\bar{Y}_n(\alpha) = \int_t^T G'_n(t)\tilde{Q}_n [I + U_n(t)]^{-1} G_n(r)G'_n(r)\tilde{Q}_n dr \bar{Y}_n(\alpha).$$

From (4.2) and the definition of V_n we have for $\alpha \geq t$,

$$P_n(t, \alpha) - G_n(t)Q_n\bar{Y}_n(\alpha) = G_n(t)Q_n\bar{Y}_n(\alpha).$$

And substituting (4.4) into the above,

$$(4.5) \quad P_n(t, \alpha) = G'_n(t)\tilde{Q}_n\bar{Y}_n(\alpha) - \int_t^T G'_n(t)\tilde{Q}_n [I + U_n(t)]^{-1} G_n(r)G'_n(r)\tilde{Q}_n dr \bar{Y}_n(\alpha).$$

Now let P_n denote the integral operator with kernel $P_n(t, \alpha)$. We then recognize $G'_n(t)\tilde{Q}_n [I + U_n(t)]^{-1} G_n(r)$, with $r \geq t$, as the kernel of the operator $[P_n B]_-$. Also $G'_n(t)\tilde{Q}_n\bar{Y}_n(\alpha)$ is recognized as the kernel of $F_n^* Y$, where Y is the operator in $B(H, X)$ defined

$$(4.6) \quad Yu: t \rightarrow \int_0^t Y(t, s)u(s) ds.$$

Thus (4.5) represents the Wiener-Hopf equation

$$P_n = [F_n^* Y]_- - [(P_n B) - F_n^* Y]_-.$$

Let P denote the operator with kernel $P(t, \alpha)$. Then since $P_n \rightarrow P$ and $F_n^* Y \rightarrow F^* Y$ in the Hilbert-Schmidt topology (the latter convergence is essentially Lemma 3.3), by continuity of the projection p_- on the space of Hilbert-Schmidt operators we obtain

$$(4.7) \quad P = [F^* Y]_- - [(PB) - F^* Y]_-.$$

This discussion is formalized in the following corollary.

COROLLARY 4.2. *The Wiener-Hopf equation (4.7) has a unique Hilbert-Schmidt solution P . A version of the kernel of P is the optimal feedback kernel for the optimization problem (3.1)–(3.2). Furthermore, this version of the kernel can be approximated in the L_∞ topology on the diagonal as well as the square by the semiseparable kernels $P_n(t, \alpha)$.*

Proof. We only need to establish the uniqueness assertion. Suppose there exist two solutions of (4.7) and let δ denote their difference. We then obtain

$$\delta + [(\delta B)_- FY]_- = 0.$$

Clearly it suffices to show that $(\delta B)_- = 0$. Define $\tilde{\delta} = -(\delta B)_- FY$ so that $\tilde{\delta}_- = \delta$. Then since B is causal it follows that $[\tilde{\delta} B]_- = [\delta B]_-$ and $\tilde{\delta} + [\tilde{\delta} B]_- F^* Y = 0$. Multiplying this latter equality on the right by B and noting that $YB = F$ (cf. (3.6) and (4.6)), we obtain the identity

$$(4.8) \quad [\tilde{\delta} B]_- + [(\tilde{\delta} B)_- F^* F]_- = 0.$$

We will show that zero is the only solution to the equation

$$X + [XF^* F]_- = 0,$$

thus proving $[\delta B]_- = 0$, and the result.

Now (4.8) is equivalent to

$$[X(I + F^* F)]_- = 0,$$

with X anticausal. Since $F^* F \geq 0$, there exists a causal Hilbert-Schmidt map V such that $I + F^* F = (I + V^*)(I + V)$. Thus X solves (4.8) if and only if Z solves

$$(4.9) \quad Z + [ZV]_- = 0.$$

(These solutions are related: $X = Z(I + V^*)^{-1}$.) Next consider the operator Ω on the space of Hilbert-Schmidt operators defined by $\Omega: Z \rightarrow [ZV]_-$ with V as above. Then (4.9) is equivalent to $(I + \Omega)Z = 0$. By induction we find

$$\Omega^n(Z) = [ZV^n]_-,$$

so that

$$|\Omega^n(Z)| \leq |Z|_{HS} |V^n|.$$

But V is quasinilpotent, and hence so is Ω . Thus the only solution to (4.9) is $Z = 0$, and the theorem is proved. \square

We note that the existence portion of the corollary was proved in a different manner in [24].

When B is a multiplication operator, $[PB]_- = PB$, so (4.7) reduces to

$$P = [F^* Y]_- - [PBF^* Y]_-.$$

The kernel of $F^* Y$ is of the form $B'(t)A(t, s)$ where

$$A(t, s) = \int_{\max(t, s)}^T Y'(r, t)Q(r)Y(r, s) dr + Y'(T, t)Q(T)Y(T, s).$$

Let A denote the operator with kernel $A(t, s)$ and note that $F^* Y = B^* A$. Next consider the following modification to the Wiener-Hopf equation (4.7),

$$(4.10) \quad \Pi = A_- - [\Pi B B^* A]_-.$$

Using the same techniques as in the proof of Corollary 4.2 it is possible to show that (4.10) has a unique Hilbert-Schmidt solution Π_0 . Also note that the corollary implies that $B^*\Pi_0 = P$. The Wiener-Hopf equation (4.10) is equivalent to the parameterized family of Fredholm equations Manitius [20] derived via a maximum principle for obtaining the feedback kernels:

$$(4.11) \quad \Pi(t, s) = A(t, s) - \int_t^T \Pi(t, \theta) B(\theta) B'(\theta) A(\theta, s) d\theta.$$

Corollary 4.2 extends this Wiener-Hopf characterization of the feedback kernel to problems with control delays, and simultaneously provides approximate solutions.

The special factorization has been previously exploited in solving Wiener-Hopf equations on finite intervals of the type (4.10), (4.11) that arise in inverse problems in the spectral theory of differential operators [8], [18], and in the filtering and smoothing problems for nonstationary processes [15], [16]. The corollary is in a sense a solution finding the right Wiener-Hopf problem.

Now we return to the original problem (3.1)–(3.3) and consider a specific sequence of measures for generating approximations to the optimal feedback kernel.

Let $\{\mu_n\}$ denote the sequence of measures

$$(4.12) \quad \int f d\mu_n = f(T) + (T+r)/n \sum_{i=0}^{n-1} f(i(T+r)/n - r).$$

This sequence is easily shown to satisfy (H1) and (H2). Now suppose the majorizing function $m(t)$ (cf. (2.3)) is bounded and the weighting function $Q(\cdot)$ has a bounded derivative. Letting $P_n(t, \alpha)$ denote the feedback kernel corresponding to the cost with measure μ_n , it is straightforward (although tedious) to derive a constant C from the estimates in § 3 such that

$$(4.13) \quad \begin{aligned} \text{ess sup}_{t, \alpha} |P(t, \alpha) - P_n(t, \alpha)| &\leq C/n, \\ \text{ess sup}_t |P(t, t) - P_n(t, t)| &\leq C/n. \end{aligned}$$

We can also use the approximations $P_n(t, \alpha)$ to obtain the following a priori bound on $|P(t, \alpha)|$ (see [34]):

$$|P(t, \alpha)| \leq (T+r+1) \sup_s |Q(s)| \sup_s |F(s, t)|_{HS} \sup_s |Y(s, \alpha)|_{HS}$$

where $|\cdot|_{HS}$ denotes the Hilbert-Schmidt matrix norm (the square root of the sum of the squares of the matrix elements). When the system (3.1)–(3.4) is time-invariant and stable, a bound independent of T can be established. If we assume for simplicity that

$$Bu: t \rightarrow B_0 u(t) + \int_{t-r}^t B(t-s) u(s) ds$$

and $|B_0|, \sup |B(t)| < b$, then it can be shown that [34]

$$|P(t, \alpha)| \leq |Q| b(1 + \sqrt{2r}) \int_0^\infty |Y(t)|^2 dt.$$

(Note that for the system to be time-invariant, $Q = \text{constant}$ and $Y(t, s) = Y(t-s)$.)

5. Algorithms. In this section we will examine in greater detail algorithms based on Theorem 4.1, with particular attention to time-invariant systems. We note that it has already been observed that combining (4.2) with the discretizations (4.12) produces

(in most cases) an $O(1/n)L_\infty$ -convergence of the feedback kernels,² regardless of the presence of control delays. However computing these kernels presupposes having the fundamental matrix $Y(\cdot, \cdot)$ in hand. In the general time-varying case with discrete state cost at the nodes $\{s_i\}$, $i = 1, \dots, n$, this amounts to solving the $n+1$ Volterra equations

$$(5.1) \quad Y(s_i, \sigma) = I - \int_{\sigma}^{s_i} Y(s_i, u) \eta(u, \sigma - u) du.$$

In the time-invariant case these computations reduce to the single equation

$$(5.2) \quad Y(t) = I - \int_0^t Y(u) \eta(u - t) du.$$

Once we have solutions (5.1) or (5.2), the feedback structure (4.2) is straightforward and can be computed from quadrature, matrix inversion and multiplication.

Of course we are not constrained to directly solving (5.1) or (5.2), and we can use other methods for obtaining the fundamental solution—e.g., state approximation methods [2], [3], the method of steps [26], [27] or other available methods for solving Volterra functional differential equations [28]–[30]. Having said this, we assume throughout this section that the functions $Y(t, s)$ and $F(t, s)$ have been computed. (A few issues associated with these computations will be addressed later.)

One straightforward implementation of (4.2) consists in defining the grid $\{s_i\}_{i=0}^n$ so that $s_i - s_{i-1} = \Delta = T/n$, taking $a_i = \Delta$, and replacing the integral in (4.3) by a first order Euler quadrature with nodes $\{s_i\}$. To see where this leads us, first write $P(t, \alpha)$ in the more symmetrical fashion

$$P(t, \alpha) = G'(t) \tilde{Q}^{1/2} [I + U(t)]^{-1} \tilde{Q}^{1/2} \bar{Y}(\alpha),$$

where

$$U(t) = \tilde{Q}^{1/2} \int_t^T G(s) G'(s) ds \tilde{Q}^{1/2}.$$

Let $\hat{U}(s_i)$ denote the approximation to $U(s_i)$,

$$\hat{U}(s_i) = \sum_{j \geq i} \Delta \tilde{Q}^{1/2} G(s_j) G'(s_j) \tilde{Q}^{1/2}$$

and form the approximations $\{\hat{P}(s_i, s_j)\}_{j \geq i}$ to $\{P(s_i, s_j)\}_{j \geq i}$,

$$(5.3) \quad \hat{P}(s_i, s_j) = G'(s_i) \tilde{Q}^{1/2} [I + \hat{U}(s_i)]^{-1} \tilde{Q}^{1/2} \bar{Y}(s_j).$$

Now note that

$$[I + \hat{U}(s_{i-1})]^{-1} = [I + \hat{U}(s_i) + \Delta \tilde{Q}^{1/2} G(s_i) G'(s_i) \tilde{Q}^{1/2}]^{-1}$$

and that

$$\text{Rank}(\Delta \tilde{Q}^{1/2} G(s_i) G'(s_i) \tilde{Q}^{1/2}) \leq M.$$

(Recall that M = dimension of the input space.) Thus, using the matrix identity,

$$(X + YY')^{-1} = X^{-1} - X^{-1} Y (Y' X^{-1} Y + I)^{-1} Y' X^{-1}$$

² It is unfortunate that although this convergence is obtained using a first-order quadrature scheme, it is not a priori evident that employing a higher order scheme would result in improved convergence. The stumbling block is that the convergence analysis we have used is based on properties of the fundamental matrix, which is generally only absolutely continuous, and thus precludes any straightforward extensions.

for compatible matrices X and Y , it follows that $(I + \hat{U}(s_{i-1}))^{-1}$ can be updated from $(I + \hat{U}(s_i))^{-1}$ in about $2MN^2n^2$ operations. Further, exploiting the semiseparable structure of $P(\cdot, \cdot)$, a rough operation count shows that $\{\hat{P}(s_i, s_j)\}_{j \geq i}$ can be computed in approximately $3MN^2n^3$ operations when $M \leq N \ll n$.

Considering that there are more than $MNn^2/2$ values in the matrices $\{\hat{P}(s_i, s_j)\}_{j \geq i}$, this algorithm is fairly efficient. However, we will subsequently show that it is possible to do substantially better in the time-invariant case. (We will also provide a more complete analysis in this case.)

For the remainder of this section we consider the problem defined by the dynamics

$$(5.4) \quad \begin{aligned} \dot{x}(t) &= \int_{-r}^0 d\eta(\theta) x(t+\theta) + \sum_{i=0}^k B_i u(t-r_i) + \int_{t-r}^t B(t-\theta) u(\theta) d\theta, & t \geq 0, \\ x(t) &= \phi(t), & t \in [-r, 0], \\ u(t) &= 0, & t \in [-r, 0] \end{aligned}$$

and cost

$$(5.5) \quad J(u, x) = \int_0^T \langle x(s), Qx(s) \rangle + |u(s)|^2 ds.$$

The assumptions here are the same as in (3.1)–(3.4), except that now everything is time-invariant. (Previously the lower limit in the integral defining the cost J was taken as $-r$. This served as a notational expediency in the preceding sections, which we dispense with in the present section.)

We will explicitly consider the discretizations of J ,

$$(5.6) \quad J_n = \sum_{i=0}^{n-1} \Delta \langle x(s_i), Qx(s_i) \rangle + \int_0^T |u(s)|^2 ds$$

where $\{s_i\}_{i=0}^n$ is a regular partition of the interval $[0, T]$ with mesh $\Delta = s_{i+1} - s_i = T/n$. (Although the mesh points of the partition change with n , we will not double subscript the s_i . This will not lead to any confusion in the sequel.) We will also assume that the point delays in the control, $r_i, i = 1, \dots, k$, correspond to some subset of the $\{s_i\}, i = 0, \dots, n$.

Again we let $P_n(\cdot, \cdot)$ denote the optimal feedback kernel for cost J_n and let $P(\cdot, \cdot)$ denote the optimal feedback kernel for cost J . In the time-invariant case $\text{Var} |\eta(\cdot)|$ and $Q(\cdot)$ are constant, so that (4.13) holds for the sequence $\{P_n(\cdot, \cdot)\}$. The particular algorithm which we will be developing is based on approximately $P_n(\cdot, \cdot)$ at the mesh points $\{s_i\}$; thus it is first necessary to prove that (4.13) actually holds everywhere. Before showing this we need some notation and a couple of simple observations.

For any matrix M , as before $|M|_{HS}$ denotes the Hilbert–Schmidt norm of M (the square root of the sum of the squares of its entries), and for specificity we write $|M|_2$ for the operator norm of M with respect to the corresponding Euclidean metrics. Note that $|M|_{HS} \geq |M|_2$.

The fundamental matrix solution $Y(t, \alpha)$ to (5.4) (cf. (2.7)) has the form $Y(t, \alpha) = Y(t - \alpha)$. We set

$$(5.7) \quad \gamma_Y = \sup_{t \in [0, T]} |Y(t)|_{HS},$$

and using [12, p. 149] we note that

$$(5.8) \quad |Y(t) - Y(s)|_{HS} = O(|t - s|), \quad t, s \in [0, T].$$

Similarly $F(t, s)$ (cf. (3.6)) is a difference kernel, $F(t, s) = F(t - s)$, and using (5.7) and (5.8) we have

$$(5.9) \quad \sup_{t \in [0, T]} |F(t)|_{HS} = \gamma_F < \infty$$

and

$$(5.10) \quad |F(t) - F(s)|_{HS} = O(|t - s|) \quad \text{when } t, s \in (r_i, r_{i+1}) \quad \text{for some } i, 0 \leq i \leq k-1.$$

LEMMA 5.1. *Let $P(t, \alpha)$ denote the optimal feedback kernel for the problem (5.4)–(5.5). Then*

$$|P(t, \alpha) - P(t', \alpha')| = O(|t - t'| + |\alpha - \alpha'|).$$

Proof. From Theorem 2.5 we have

$$P(t, \alpha) = \int_{\alpha}^T K(t, s) Y(s - \alpha) ds,$$

where

$$K(t, s) = F'(s - t)Q + \int_t^s W_{-}(t, \sigma) F'(\sigma - s) Q d\sigma.$$

Since $\sup_{t,s} |K(t, s)| < \infty$ (cf. Proposition 3.4 and (5.9)), it follows from (5.8) that

$$|P(t, \alpha) - P(t, \alpha')| < K_1 |\alpha - \alpha'|$$

for some constant K_1 independent of t, α, α' . Thus using (5.7) it remains to show the existence of a constant K_2 such that

$$\int |K(t_2, s) - K(t_1, s)| ds \leq K_2 |t_2 - t_1|.$$

Now $F(t)$ has only a finite number of jump discontinuities, so (5.9) and (5.10) imply

$$\int |F(s - t_2) - F(s - t_1)| ds = O(|t_2 - t_1|).$$

Hence, it is only necessary to verify that

$$(5.11) \quad \int |W_{-}(t_2, \sigma) - W_{-}(t_1, \sigma)| d\sigma = O(|t_2 - t_1|).$$

To this end let W^* denote the operator with kernel $W_{-}(t, \sigma)$ and let $X^* = (I + W^*)^{-1} - I$. Denote the kernel of X^* by $X_{-}(t, s)$. Also let W and X denote the adjoints of W^* and X^* respectively, with respective kernels $W_{+}(t, s)$ and $X_{+}(t, s)$. Now the factorization (recall Theorem 2.5)

$$I + F^* F = (I + X^*)(I + X)$$

implies

$$(5.12) \quad X^* = [F^* F]_{-} + [F^* F W]_{-},$$

where $F^* F$ has kernel $H(t, s)$,

$$H(t, s) = \int_{\max(t, s)}^T F'(\sigma - t) Q F(\sigma - s) d\sigma.$$

Note that

$$(5.13) \quad |H(t_2, s) - H(t_1, s)| = O(|t_2 - t_1| + |s_2 - s_1|).$$

Now (5.12) is equivalent to

$$X_-(t, s) = H(t, s) + \int_s^T H(t, \theta) W_+(\theta, s) d\theta \quad \text{a.e. } s, t.$$

Because $H(t, s)$ is continuous, $X_\pm(t, s)$, $W_\pm(t, s)$ are also continuous [11], and the equation above holds pointwise. Thus (5.13) and the triangle inequality imply

$$|X_-(t_2, s) - X_-(t_1, s)| = O(|t_2 - t_1|)$$

independently of s . Then (5.11) follows from the estimate above and the resolvent identity

$$W_-(t, s) + X_-(t, s) + \int_t^s X_-(t, \theta) W_-(\theta, s) d\theta = 0. \quad \square$$

PROPOSITION 5.2.

$$\sup_{t, \alpha \in (0, T]} |P_n(t, \alpha) - P(t, \alpha)| = O(1/n).$$

Proof. Using the identity

$$A[I + BA]^{-1} = A^{1/2}[I + A^{1/2}BA^{1/2}]^{-1}A^{1/2}$$

for $A, B \geq 0$, write the feedback kernel (from Theorem 4.1) as

$$(5.14) \quad P_n(t, \alpha) = \underline{F}'_n(t)[I + \underline{U}_n(t)]^{-1}\underline{Y}_n(\alpha)$$

where

$$(5.15) \quad \underline{F}(t) = \Delta^{1/2}[F'(s_0 - t)Q^{1/2} \cdots F'(T - t)Q^{1/2}]',$$

$$(5.16) \quad \underline{U}_n(t) = \int_t^T \underline{F}_n(\sigma) \underline{F}'_n(\sigma) d\sigma$$

and

$$(5.17) \quad \underline{Y}_n(\alpha) = \Delta^{1/2} \begin{bmatrix} Q^{1/2} Y(s_0 - \alpha) \\ \vdots \\ Q^{1/2} Y(T - \alpha) \end{bmatrix}.$$

Note the following bounds independent of n (recall (5.7) and (5.9)):

$$(5.18) \quad \sup_{\alpha} |\underline{Y}_n(\alpha)|_{HS}^2 \leq T |Q^{1/2}|_{HS}^2 \gamma_Y^2,$$

$$(5.19) \quad \sup_t |\underline{F}_n(t)|_{HS}^2 \leq T |Q^{1/2}|_{HS}^2 \gamma_F^2,$$

$$(5.20) \quad \sup_t |(I + \underline{U}_n(t))^{-1}|_2 \leq 1.$$

Now fix $(t_0, \alpha_0) \in (0, T] \times (0, T]$ with $t_0 \leq \alpha_0$. Then there exist indices i, j such that $(t_0, \alpha_0) \in (s_{i-1}, s_i] \times (s_{j-1}, s_j]$, $i \leq j$. From (5.8) we obtain

$$(5.21) \quad \begin{aligned} |\underline{Y}_n(s_j) - \underline{Y}_n(\alpha_0)|_{HS} &\leq \left[\sum_{k \geq j}^n \Delta |Y(s_k - s_j) - Y(s_k - \alpha_0)|_{HS}^2 \right]^{1/2} |Q^{1/2}| \\ &= \left[\sum_{k \geq j}^n \Delta |Y(s_k - \alpha_0 + \alpha_0 - s_j) - Y(s_k - \alpha_0)|_{HS}^2 \right]^{1/2} |Q^{1/2}| \\ &= O(1/n). \end{aligned}$$

Similarly, expressing $F(\cdot)$ as a sum (as in (3.6)) and using (5.10) we obtain

$$(5.22) \quad \|\underline{F}_n(s_i) - \underline{F}_n(t_0)\|_{HS} = O(1/n).$$

Next observe that

$$(5.23) \quad \|(I + \underline{U}_n(t_0))^{-1} - (I + \underline{U}_n(s_i))^{-1}\|_2 \leq \|(I + \underline{U}_n(t_0))^{-1}\|_2 \|\underline{U}_n(t_0) - \underline{U}_n(s_i)\|_{HS} \|(I + \underline{U}_n(s_i))^{-1}\|_2 \\ \leq \|\underline{U}_n(t_0) - \underline{U}_n(s_i)\|_{HS}.$$

But from (5.16) and (5.19) it follows that

$$(5.24) \quad \|\underline{U}_n(t_0) - \underline{U}_n(s_i)\|_{HS} = O(1/n).$$

Putting (5.21)–(5.24) together with the bounds (5.18)–(5.20), and using the triangle inequality, we get

$$(5.25) \quad \|P_n(t_0, \alpha_0) - P_n(s_i, s_j)\|_2 = O(1/n).$$

Then using the fact that (4.13) holds for a.e. t, α , and that the estimate (5.25) holds for any $(t_0, \alpha_0) \in (s_{i-1}, s_i] \times (s_{j-1}, s_j]$, Lemma 5.1 and the triangle inequality imply

$$\|P_n(t, \alpha) - P(t, \alpha)\|_2 = O(1/n)$$

for all $t, \alpha \in (0, T] \times (0, T]$. \square

Since the idea behind the algorithm defined in Theorems 5.3 and 5.4 is based on a relatively simple observation that gets somewhat obscured in the notation and proofs of the theorems, it is worthwhile here to briefly remark on this motivating idea.

If we return to the approximate gain defined in (5.3), it turns out that in the time-invariant case each $\hat{U}(s_i)$ is a principal minor of $\hat{U}(0)$. Thus what we would like to do is to invert $I + \hat{U}(0)$ via a recursion in which all of the principal minors of $I + \hat{U}(0)$ are also inverted. Now by inspection $\hat{U}(0)$ can be identified with the covariance matrix of the random process $\{\Delta^{1/2}\xi_i\}$,

$$\xi_i = \Delta \sum_{j=0}^{\infty} Q^{1/2} F((i-j)\Delta) \omega_j$$

where $E(\omega_i \omega_j') = \delta_{ij} I$. In the signal processing literature (see, for example, [31]) it is shown that processes of this type admit “fast” filter implementations due to their “near” Toeplitz covariance matrix structure. This is precisely the property we exploit.

THEOREM 5.3. Define the symmetric $N(n+1) \times N(n+1)$ matrix \hat{U}^n with $N \times N$ block entries \hat{U}_{ij}^n where (for $0 \leq j \leq i \leq n$)

$$\hat{U}_{ij}^n = \sum_{\sigma=0}^j \Delta^2 Q^{1/2} F((i-j)\Delta + \sigma\Delta) F'(\sigma\Delta) Q^{1/2}.$$

Let Z_n denote the matrix on $R^{N(n+1)}$,

$$Z_n = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 1 & \ddots & & 0 \\ \vdots & \ddots & \ddots & \\ 0 & & 1 & 0 \end{bmatrix},$$

and for each $p = 0, 1, \dots, n$, let $\Pi_{n,p}$ denote the projection

$$\Pi_{n,p}(x_1, \dots, x_{N(n+1)})' = (x_1 \cdots x_{N(p+1)}, 0, \dots, 0)'.$$

For $p, q = 0, 1, \dots, n$ define (recall $\Delta = T/n$)

$$\hat{P}_n(\Delta p, \Delta q) = \hat{X}_{n,n-p}' Z_n^{N(n+1-p)} Y_n(\Delta q)$$

where Y_n is defined in (5.17) and

$$\hat{X}_{n,p} = [I + \Pi_{n,p} \hat{U}^n \Pi_{n,p}]^{-1} \begin{bmatrix} \Delta^{1/2} Q^{1/2} F(0) \\ \vdots \\ \Delta^{1/2} Q^{1/2} F(\Delta p) \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Then

$$\max_{p \leq q} |\hat{P}_n(\Delta p, \Delta q) - P(\Delta p, \Delta q)|_2 = O(1/n), \quad p, q \in \{0, 1, \dots, n\}.$$

Proof. Recall the representation (5.14) for the optimal feedback kernel for cost J_n ,

$$P_n(t, \alpha) = F'_n(t) [I + U_n(t)]^{-1} Y_n(\alpha).$$

Thus we can write

$$(5.26) \quad P_n(\Delta p, \Delta q) = X'(\Delta p) Y_n(\Delta q)$$

where

$$(5.27) \quad X(\Delta p) = [I + U_n(\Delta p)]^{-1} \begin{bmatrix} 0 \\ \vdots \\ \Delta^{1/2} Q^{1/2} F(0) \\ \vdots \\ \Delta^{1/2} Q^{1/2} F((n-p)\Delta) \end{bmatrix}.$$

Note that $U_n(\Delta p)$ is an $N(n+1) \times N(n+1)$ symmetric matrix with $N \times N$ block entries $U_{i,j}^n(\Delta p)$, $i, j = 0, 1, \dots, n$ where (for $0 \leq j \leq i$)

$$(5.28) \quad U_{i,j}^n(\Delta p) = \begin{cases} \Delta Q^{1/2} \int_0^{(j-p)\Delta} F((i-j)\Delta + \sigma) F'(\sigma) d\sigma Q^{1/2}, & j \geq p, \\ 0 & \text{otherwise.} \end{cases}$$

For each $p = 0, 1, \dots, n$, let

$$(5.29) \quad X_{n,p} = [I + \Pi_{n,p} U_n(0) \Pi_{n,p}]^{-1} \begin{bmatrix} \Delta^{1/2} Q^{1/2} F(0) \\ \vdots \\ \Delta^{1/2} Q^{1/2} F(\Delta p) \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Observing that

$$U_{i+n-p, j+n-p}^n(\Delta(n-p)) = U_{i,j}^n(0), \quad i, j = 0, 1, \dots, p,$$

it follows from (5.27) and (5.29) that

$$X(\Delta(n-p)) = Z_n^{(n-p)N} X_{n,p}.$$

Therefore (using (5.26)),

$$P_n(\Delta p, \Delta q) = X'_{n,n-p} Z_n^{N(n+1-p)} Y_n(\Delta q).$$

Now,

$$|P_n(\Delta p, \Delta q) - P_n(\Delta p, \Delta q)|_2 \leq T^{1/2} |Q^{1/2}|_2 \gamma_Y |X_{n,n-p} - \hat{X}_{n,n-p}|_2.$$

(Here we have used (5.18).) But using (5.29) and the definitions of X and \hat{X} , we find that

$$|X_{n,n-p} - \hat{X}_{n,n-p}|_2 \leq T^{1/2} |Q^{1/2}|_2 \gamma_F |\hat{U}^n - U_n(0)|_{HS},$$

and

$$(5.30) \quad |\hat{U}^n - U_n(0)|_{HS} = \left[\sum_{i,j=0} |\hat{U}_{ij}^n - U_{ij}^n(0)|_{HS}^2 \right]^{1/2}.$$

It follows from (5.26) that

$$|\hat{U}_{ij}^n - U_{ij}^n(0)|_{HS} = O(\Delta^2)$$

independent of i, j . Thus the right side of (5.30) is $O(\Delta)$. Hence,

$$\max_{p,q} |P_n(\Delta p, \Delta q) - \hat{P}_n(\Delta p, \Delta q)|_2 = O(\Delta).$$

The result follows from the estimate above and Proposition 5.2. \square

Using the extended LWR algorithm [31] we will next show that $\{\hat{X}_{n,p}\}_{p=0}^n$ can be computed in a total of $O(n^2)$ operations. The theorem above represents the approximate gain in the form

$$\hat{P}_n(\Delta p, \Delta q) = X'_{n,n-p} Z_n^{N(n+1-p)} Y_n(\Delta q).$$

Now fixing p and letting q vary, we can view these products as a convolution. Since convolutions have fast implementations in $O(n \log n)$ operations, it will follow then that $\{\hat{P}(\Delta p, \Delta q)\}_{q \geq p}$ can be computed in $O(n^2 \log n)$ operations.

THEOREM 5.4. $\{\hat{X}_{n,p}\}_{p=0}^n$ can be computed in $O(n^2)$ operations.

Proof. By definition $\hat{X}_{n,p}$ satisfies

$$(5.31) \quad [I + \Pi_{n,p} \hat{U}^n \Pi_{n,p}] \hat{X}_{n,p} = \Pi_{n,p} F_n(0).$$

We will exhibit a fast recursion for solving (5.31). A simple calculation shows that

$$\hat{U}_{i+1,j+1}^n - \hat{U}_{i,j}^n = \Delta^2 Q^{1/2} F((i+1)\Delta) F'((j+1)\Delta) Q^{1/2}.$$

Therefore the matrix $\delta(\hat{U}^n)$ with block entries $\hat{U}_{i+1,j+1}^n - \hat{U}_{i,j}^n$, $i, j = 0, 1, \dots, n-1$, can be written:

$$\delta(\hat{U}^n) = \Delta^2 \begin{bmatrix} Q^{1/2} F(\Delta) \\ \vdots \\ Q^{1/2} F(n\Delta) \end{bmatrix} [F'(\Delta) Q^{1/2} \cdots F'(n\Delta) Q^{1/2}].$$

Consequently, $\text{rank}(\delta(\hat{U}^n)) \leq N \ll n$. Furthermore (and importantly [31]) $\delta(\hat{U}^n)$ has a factorization of the form

$$\delta(\hat{U}^n) = D \Sigma D'$$

with D an $nN \times N$ matrix and Σ an $N \times N$ signature matrix. In fact, this factorization

is easily done by inspection:

$$D = \Delta \begin{bmatrix} Q^{1/2}F(\Delta) & \overbrace{\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}}^{N-M} \\ \vdots & \\ Q^{1/2}F(n\Delta) & \end{bmatrix}, \quad \Sigma = I.$$

Therefore the LWR algorithm [31, p. 655] can be used to recursively compute the solutions M_p to

$$(5.32) \quad [I + \Pi_{n,p} \hat{U}^n \Pi_{n,p}] M_p = \begin{bmatrix} \hat{U}_{0,p+1}^n \\ \vdots \\ \hat{U}_{p,p+1}^n \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$p = 0, 1, \dots, n$ in a total of $O(N^2 n^2)$ operations. With these solutions in hand we can argue as in the scalar Toeplitz case (see, for example, [32]) to recursively solve the system (5.31). These details are supplied below.

So now consider solving the problem

$$(5.33) \quad [I + \Pi_{n,p} \hat{U}^n \Pi_{n,p}] \hat{X}_{n,p} = \Pi_{n,p} F_n(0),$$

given $\hat{X}_{n,p-1}$ and M_{p-1} (from (5.32)).

Let F_p denote the $N(p+1) \times M$ matrix composed of the first $N(p+1)$ rows of $F_n(0)$, and let f_p denote the $N \times M$ matrix composed of the $Np+1$ through $N(p+1)$ rows of $F_n(0)$. Thus we can write

$$F_{p+1} = \begin{bmatrix} F_p \\ f_p \end{bmatrix}.$$

Let R_p denote the $N(p+1) \times N(p+1)$ matrix composed of the northwest corner of $I + \hat{U}^n$, and let r_p denote the $N \times N$ southeast corner of R_p . Also let $\underline{M}_p \in R^{N(p+1) \times N}$ denote the matrix consisting of the first $N(p+1)$ rows of M_p , and similarly let N_p denote the $N(p+1) \times M$ matrix consisting of the first $N(p+1)$ rows of $\hat{X}_{n,p}$.

Now define

$$u_{p+1} = \begin{bmatrix} \hat{U}_{0,p+1}^n \\ \vdots \\ \hat{U}_{p,p+1}^n \end{bmatrix},$$

and consider the equation

$$(5.34) \quad \begin{bmatrix} R_p & u_{p+1} \\ u'_{p+1} & r_{p+1} \end{bmatrix} \begin{bmatrix} \mu \\ \nu \end{bmatrix} = \begin{bmatrix} F_p \\ f_p \end{bmatrix}.$$

Note that

$$(5.35) \quad \begin{bmatrix} \mu \\ \nu \end{bmatrix} = N_{p+1}.$$

From the top of the equations in (5.34) we obtain

$$(5.36) \quad \mu = R_p^{-1} F_p - R_p^{-1} u_{p+1} \nu = N_p - \underline{M}_p \nu.$$

Substituting this into the bottom equation, we get

$$[r_{p+1} - u'_{p+1} \underline{M}_p] \nu = f_p - u_{p+1} N_p.$$

Noting that

$$\begin{aligned} 0 &< \begin{bmatrix} I & 0 \\ -\underline{M}'_p & I_{N \times N} \end{bmatrix} \begin{bmatrix} R_p & u_{p+1} \\ u'_{p+1} & r_{p+1} \end{bmatrix} \begin{bmatrix} I & -\underline{M}_p \\ 0 & I_{N \times N} \end{bmatrix} \\ &= \begin{bmatrix} R_p & 0 \\ 0 & r_{p+1} - \underline{M}'_p u_{p+1} \end{bmatrix} \end{aligned}$$

($I_{N \times N} = N \times N$ identity matrix), we find that

$$(5.37) \quad \nu = [r_{p+1} - u'_{p+1} \underline{M}_p]^{-1} \{f_p - u'_{p+1} N_p\}.$$

Thus, given M_p and N_p , N_{p+1} can be computed from (5.36) and (5.37) in $O(M^2 p)$ operations ($p \gg N$). Hence, in particular $\{\hat{X}_{n,p}\}_{p=0}^n$ in (5.31) can be computed in a recursive manner in a total of $O(n^2)$ operations. \square

In [33] infinite-dimensional Chandrasekhar equations are derived for time-invariant hereditary systems without control delay. A fast algorithm based on approximating the Chandrasekhar equations is obtained and is shown to possess the same convergence properties as reported in [10]. Although there are no direct connections between the algorithm we developed here and the one in [33], there are some general connections between Chandrasekhar equations and the inversion of near Toeplitz systems [31].

We note that stability of the algorithms of this section with respect to the data $Y(t)$, $F(t)$ is easily demonstrated. For suppose $Y(\cdot)$ and $F(\cdot)$ were replaced by $Y_\varepsilon(\cdot)$ and $F_\varepsilon(\cdot)$ with

$$\sup_{t \in [0, T]} |Y(t) - Y_\varepsilon(t)|, |F(t) - F_\varepsilon(t)| < \varepsilon.$$

Using the obvious notation, we obtain the corresponding error estimates,

$$\sup_t |F_n - F_n^\varepsilon|, \sup_t |Y_n - Y_n^\varepsilon| = O(\varepsilon)$$

and

$$\max |\hat{U}_{ij}^n - \hat{U}_{ij}^{n,\varepsilon}|_{HS} = O(\varepsilon) \Delta,$$

all independent of n . Substituting these errors into the appropriate places in Theorem 5.3, it follows that an $O(\varepsilon)$ perturbation in the data $Y(t)$, $F(t)$ yields an $O(\varepsilon)$ perturbation in the estimate of the feedback kernel (all of these estimates in the sup norm).

Another feature to note is that although the implementation defined by Theorem 5.3 and Theorem 5.4 is recursive backward in time for the computation of the feedback kernel (just as one would suspect—e.g., the Riccati equation is solved backward in time), in terms of the algorithm's utilization of the fundamental matrix $Y(t)$, it is actually *forward* in time. Thus any approximating scheme for the computation of $Y(t)$ can be readily incorporated into the algorithm. This remark is also true for the implementation of Theorem 4.1 introduced in the beginning of this section.

One final remark concerning the algorithm is that in the event that the point delays $\{r_i\}_{i=1}^k$ do not correspond to a subset of the nodes $\{s_i\}_{i=0}^n$, the estimate in (5.22) is only $O(1/\sqrt{n})$. Thus the estimate in Theorem 5.3 would also be $O(1/\sqrt{n})$. Of course Theorem

4.1 has the flexibility to place nodes anywhere, and it may be possible to recover the stronger convergence by considering algorithms arising from different discretization strategies.

We conclude this section with the following simple scalar example:

$$\min J(u, x) = \int_0^2 |x(t)|^2 + |u(t)|^2 dt$$

subject to the constraint

$$\dot{x}(t) = x(t) + x(t-1) + u(t).$$

Since we are seeking the optimal feedback kernel $P(t, \alpha)$, it is not necessary to prescribe an initial condition above.

The algorithm described in the beginning of the section based on the approximation (5.3) and the recursive inversion of $I + \hat{U}(s_i)$ via the "matrix inversion lemma," was programmed using a few lines of Fortran code. Because the fundamental solution $Y(t)$ to the differential equation is easily derived for this problem, the exact solution was used in the algorithm. (Recall that an $O(\epsilon)$ error in the approximation of Y results in an $O(\epsilon)$ error in $P(t, \alpha)$.)

Discretizations with mesh width = .025, .01, .005, were considered. The results for this problem coincided fairly well with the theory. We observed essentially linear (uniform) convergence of the feedback kernels as predicted. Tables 5.1–5.3 contain these results. As an independent check for the correctness of these values, Table 5.4 contains values of the kernel obtained via the Riccati equation approach using the Kappel–Salamon [35] linear spline approximation of the history space with ten elements. The gain using these approximations appeared to have converged to about two significant figures.

This example is intended only to be illustrative of the general methodology. The fact that the algorithm of this example requires virtually no modification to handle the more complex case of control delays, underscores our comments in the introduction concerning the transparency of the factorization approach to the control delay problem. However, our numerical experience with possible algorithms that can be derived via

TABLE 5.1
 $\Delta = .025$.

	$P(t, \alpha)$	t			
		0.0	0.5	1.0	1.5
α	$t + 0.0$	2.7881	2.4237	1.7011	0.7462
	$t + 0.1$	2.3699	2.0632	1.4473	0.5799
	$t + 0.2$	2.0210	1.7566	1.2222	0.4252
	$t + 0.3$	1.7237	1.4972	1.0213	0.2789
	$t + 0.4$	1.4712	1.2790	0.8406	0.1381
	$t + 0.5$	1.2576	1.0972	0.6765	0.0000
	$t + 0.6$	1.0780	0.9411	0.5258	0.0000
	$t + 0.7$	0.9280	0.8037	0.3855	0.0000
	$t + 0.8$	0.8040	0.6822	0.2528	0.0000
	$t + 0.9$	0.7029	0.5472	0.1252	0.0000
	$t + 1.0$	0.6221	0.4775	0.0000	0.0000

TABLE 5.2
 $\Delta = .01$.

$P(t, \alpha)$		t			
		0.0	0.5	1.0	1.5
α	$t + 0.0$	2.7300	2.3951	1.6943	0.7521
	$t + 0.1$	2.3279	2.0378	1.4110	0.5845
	$t + 0.2$	1.9844	1.7341	1.2165	0.4285
	$t + 0.3$	1.6919	1.4770	1.0162	0.2810
	$t + 0.4$	1.4335	1.2609	0.8361	0.1391
	$t + 0.5$	1.2335	1.0808	0.6727	0.0000
	$t + 0.6$	1.0570	0.9270	9.5228	0.0000
	$t + 0.7$	0.9096	0.7916	0.3832	0.0000
	$t + 0.8$	0.7878	0.6721	0.2513	0.0000
	$t + 0.9$	0.6886	0.5659	0.1244	0.0000
	$t + 1.0$	0.6093	0.4710	0.0000	0.0000

TABLE 5.3
 $\Delta = .005$.

$P(t, \alpha)$		t			
		0.0	0.5	1.0	1.5
α	$t + 0.0$	2.7140	2.3855	1.6919	0.7541
	$t + 0.1$	2.3140	2.0293	1.4388	0.5859
	$t + 0.2$	1.9724	1.7265	1.2145	0.4295
	$t + 0.3$	1.6814	1.4703	1.0144	0.2817
	$t + 0.4$	1.4344	1.2549	0.8346	0.1394
	$t + 0.5$	1.2556	1.0753	0.6714	0.0000
	$t + 0.6$	1.0500	0.9223	0.5217	0.0000
	$t + 0.7$	0.9036	0.7876	0.3824	0.0000
	$t + 0.8$	0.7825	0.6887	0.2508	0.0000
	$t + 0.9$	0.6839	0.5631	0.1241	0.0000
	$t + 1.0$	0.6051	0.4688	0.0000	0.0000

TABLE 5.4
Spline approximation.

$P(t, \alpha)$		t			
		0.0	0.5	1.0	1.5
α	$t + 0.0$	2.6915	2.3742	1.6860	0.7491
	$t + 0.1$	2.2986	2.0177	1.4355	0.5894
	$t + 0.2$	1.9639	1.7215	1.2165	0.4366
	$t + 0.3$	1.6749	1.4685	1.0150	0.2776
	$t + 0.4$	1.4243	1.2489	0.8284	0.1257
	$t + 0.5$	1.2114	1.0605	0.6658	0.0217
	$t + 0.6$	1.0443	0.9234	0.5230	0.0005
	$t + 0.7$	0.8994	0.7862	0.3881	0.0032
	$t + 0.8$	0.7775	0.6620	0.2404	0.0005
	$t + 0.9$	0.6786	0.5567	0.1079	0.0004
	$t + 1.0$	0.6013	0.4690	0.0323	0.0000

these methods is limited, and beyond the operation counts and convergence characteristics provided in §§ 4 and 5, it would be premature to make any assessment regarding how well these algorithm work in general. (See additional comments in § 6.)

6. Concluding remarks. Our focus has been on the control problem for the RFDE on a finite interval. A natural question which arises is whether the approach can be adapted to treat the infinite time problem. Although several technical difficulties have to be dealt with, progress on this problem has been reported in [36]. We note that a factorization based approach to the regulator problem has been previously investigated by Davis [4].

Our treatment of the numerical aspects of the control problem has been from a fairly general perspective, and one might suspect that stronger results may be established in a more specific setting. One area, for example, that was not pursued either analytically or numerically was the use of higher order methods for time-invariant problems, where there appears to be room to improve upon the results. Along these lines, in the case of time-invariant systems without control delay we have obtained some new results utilizing (4.7) as a point of departure that substantially sharpens the analytical and numerical results of the present paper [37]. We essentially exhibit a fast Chandrasekhar-type set of equations that admits a second order implementation for construction of the kernel $P(t, \alpha)$ in the general case of incommensurate delays, and in the case of a single delay, arbitrarily high order methods are shown to be admissible. It is not clear at this time whether these results have extension to problems with control delay.

Acknowledgment. The author thanks Professors Alan Schumitzky and J. S. Gibson for some very fruitful discussions on the subject of this paper, and to Professor Gibson for providing the values shown in Table 5.4.

REFERENCES

- [1] H. T. BANKS, *Representations for solutions of linear functional differential equations*, J. Differential Equations, 5 (1969), pp. 399–409.
- [2] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: Numerical methods based on averaging approximations*, this Journal, 16 (1978), pp. 169–208.
- [3] H. T. BANKS AND I. G. ROSEN, *Spline approximations for linear nonautonomous delay systems*, J. Math. Anal. Appl., 96 (1983), pp. 226–268.
- [4] J. H. DAVIS, *Wiener-Hopf methods for open-loop unstable distributed systems*, this Journal, 17 (1979), pp. 713–728.
- [5] M. C. DELFOUR, *The linear quadratic optimal control problem for hereditary differential systems: Theory and numerical solution*, Appl. Math. Optim., 3 (1977), pp. 101–162.
- [6] ———, *The linear quadratic optimal control problem with delays in state and control variables: A state space approach*, this Journal, 24 (1986), pp. 835–883.
- [7] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part II*, Wiley-Interscience, New York, 1963.
- [8] I. M. GELFAND AND B. M. LEVITAN, *On the determination of a differential equation from its spectral function*, AMS Trans. Ser. 2 (1955), pp. 253–304.
- [9] J. S. GIBSON, *The Riccati integral equations for optimal control problems in Hilbert space*, this Journal, 17 (1979), pp. 537–565.
- [10] ———, *Linear-quadratic optimal control of hereditary differential systems: Infinite dimensional Riccati equations and numerical approximations*, this Journal, 21 (1983), pp. 95–139.
- [11] I. C. GOHBERG AND M. G. KREIN, *Theory and Applications of Volterra Operators in Hilbert Space*, American Mathematical Society, Providence, RI, 1970.
- [12] J. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.
- [13] A. ICHIKAWA, *Quadratic control of evolution equations with delays in control*, this Journal, 20 (1982), pp. 645–668.

- [14] K. ITO, *Regulator problem for hereditary differential systems with control delays*, ICASE Report 82-3, NASA Langley Research Center, Hampton, VA, 1982.
- [15] T. KAILATH AND P. FROST, *An innovations approach to least squares estimation, Part II: Linear smoothing in additive white noise*, IEEE Trans. Automat. Control, AC-13 (1968), pp. 655-660.
- [16] T. KAILATH, *A note on least squares estimation by the innovations method*, this Journal, 10 (1972), pp. 477-486.
- [17] H. N. KOIVO AND E. B. LEE, *Controller synthesis for linear systems with retarded state and control variables and quadratic cost*, Automatica, 8 (1972), pp. 203-208.
- [18] M. G. KREIN, *On inverse problems for a nonhomogeneous cord*, Dokl. Akad. Nauk SSSR, 82 (1952), pp. 669-672.
- [19] K. KUNISCH, *Approximation schemes for the linear-quadratic optimal control problem associated with delay equations*, this Journal, 20 (1982), pp. 506-450.
- [20] A. MANITIUS, *Optimal control of linear time-lag processes with quadratic performance indexes*, in Proc. Fourth IFAC Congress, Warsaw, Poland, 1969, pp. 16-28.
- [21] A. McNABB AND A. SCHUMITZKY, *Factorization of operators—III: Initial value methods for linear two point boundary value problems*, J. Math. Anal. Appl., 31 (1970), pp. 391-405.
- [22] M. MILMAN AND A. SCHUMITZKY, *On a class of operators on Hilbert space with applications to factorization and systems theory*, J. Math. Anal. Appl., 99 (1984), pp. 494-512.
- [23] M. MILMAN, *Special factorization and Riccati integral equations*, J. Math. Anal. Appl., 100 (1984), pp. 155-187.
- [24] M. MILMAN, J. FOSTER AND A. SCHUMITZKY, *Optimal feedback control of infinite dimensional linear systems*, J. Math. Anal. Appl., 119 (1986), pp. 259-281.
- [25] R. B. VINTER AND R. H. KWONG, *The infinite time quadratic control problem for linear systems with state and control delays: an evolution equation approach*, this Journal, 19 (1981), pp. 139-153.
- [26] R. BELLMAN, *On the computational solution of differential-difference equations*, J. Math. Anal. Appl., 2 (1961), pp. 108-110.
- [27] R. BELLMAN AND K. L. COOKE, *On the computational solution of a class of functional differential equations*, J. Math. Anal. Appl., 12 (1965), pp. 495-500.
- [28] J. J. OBERLE AND H. J. PESCH, *Numerical treatment of delay differential equations by Hermite interpolation*, Numer. Math., 37 (1981), pp. 235-255.
- [29] C. W. CRYER AND L. TAVERNINI, *The numerical solution of Volterra functional differential equations by Euler's method*, SIAM J. Numer. Anal., 9 (1972), pp. 105-129.
- [30] A. FELDSTEIN AND R. GOODMAN, *Numerical solution of ordinary and retarded differential equations with discontinuous derivatives*, Numer. Math., 21 (1973), pp. 1-13.
- [31] B. FRIEDLANDER, T. KAILATH, M. MORF AND L. LJUNG, *Extended Levinson and Chandrasekhar equations for general discrete-time linear estimation problems*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 653-659.
- [32] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.
- [33] J. A. BURNS, K. ITO AND R. K. POWERS, *Chandrasekhar equations and computational algorithms for distributed parameter systems*, Proc. 23rd Conference on Decision and Control, Las Vegas, NV, 1984.
- [34] M. MILMAN, *Approximating the Linear-Quadratic Optimal Control Law for Hereditary Systems with Delays in the Control*, JPL Pub. 87-6, Jet Propulsion Laboratory, Pasadena, CA, 1987.
- [35] F. KAPPEL AND D. SALAMON, *Spline approximation for retarded systems and the Riccati equation*, this Journal, 24 (1986), pp. 1082-1117.
- [36] M. MILMAN, *Representations for the optimal feedback control law for stable hereditary systems*, Proc. Internat. Conference on Control of Distributed Parameter Systems, Vorau, Austria, 1986, to appear.
- [37] M. MILMAN AND R. E. SCHEID, JR., *Analytical and numerical methods based on Volterra factorization for control of differential delay systems*, Proc. Internat. Symposium on Math. Theory of Networks and Systems, Phoenix, AZ, 1987, to appear.

THE SENSITIVITY OF THE STABLE LYAPUNOV EQUATION*

GARY HEWER† AND CHARLES KENNEY‡

Abstract. We present an analysis of the sensitivity of the solution of the Lyapunov equation $A^*X + XA = -W$, where A is stable. This analysis leads to a spectral norm bound on the relative perturbation of the solution which is optimal for a certain class of estimates and which is essentially equivalent to the Frobenius norm bound obtained from the associated Kronecker product system. The latter bound can be expressed in terms of $\text{sep}(A^*, -A)$ and is known to accurately reflect the sensitivity of the Lyapunov problem, but it is hard to interpret in terms of the original matrix A . In contrast, the spectral norm bound which we derive is directly related to the minimal L_2 damping of the dynamical system $\dot{z} = Az$. Moreover, this dynamical link with the sensitivity problem leads to a new method of systematically investigating the norm behavior of e^{At} as well as providing a wealth of information about control theoretic aspects of $\dot{z} = Az$, when A is the closed loop state matrix.

Key words. Lyapunov equation, matrix exponential, condition number

AMS(MOS) subject classification. 93

1. Introduction. In this paper we consider the sensitivity of the solution X to the Lyapunov equation

$$(1.1) \quad A^*X + XA = -W$$

where A, X and $W \in \mathbb{C}^{n \times n}$ and A^* denotes the conjugate transpose of A . We do not make any special assumptions about W such as symmetry or positive definiteness, but we do assume throughout that A is stable, that is the eigenvalues of A have negative real part:

$$(1.2) \quad \text{Re } \lambda_i(A) < 0, \quad i = 1, \dots, n.$$

This condition on A ensures that (1.1) has a unique solution (see [20]). Now let $\Delta A, \Delta X$ and ΔW be such that

$$(1.3) \quad (A + \Delta A)^*(X + \Delta X) + (X + \Delta X)(A + \Delta A) = -(W + \Delta W).$$

Our goal is to develop bounds on $\|\Delta X\|$ which are reasonably sharp, in a sense to be specified later, and which are easy to interpret. Here $\|\cdot\|$ is the spectral or 2-norm defined by

$$\|M\| = \max \frac{\|Mx\|}{\|x\|}$$

for $M \in \mathbb{C}^{n \times n}$ where the maximum is taken over all nonzero $x \in \mathbb{C}^n$ and $\|x\|^2 = \sum_i |x_i|^2$.

In order to place our work in the proper context, it is helpful to review some established results. For example, a central fact in dealing with (1.1) is that it is equivalent to a linear system of order n^2 :

$$(1.4) \quad Px = -w$$

* Received by the editors March 25, 1985; accepted for publication (in revised form) January 3, 1987. This research was supported by the Office of Naval Research under contract 1486AF00001, by Naval Systems Air Command and by Naval Weapons Center Independent Research Funds.

† RF Missile Systems Branch, Weapons Department, Naval Weapons Center, China Lake, California 93555.

‡ Electrical and Computer Engineering Department, University of California, Santa Barbara, California 93106.

where

$$P = I \otimes A^* + A^T \otimes I,$$

$$x = \text{Vec}(X),$$

$$w = \text{Vec}(W).$$

Here \otimes indicates the Kronecker product, $\text{Vec}(M)$ is the n^2 vector formed by stacking the columns of M (see [12]) and $A^T(A^*)$ is the (conjugate) transpose of A . The equivalence of (1.1) and (1.4) means that sensitivity information about (1.1) can be obtained by applying the well-developed theory of perturbations of linear systems to (1.4). This approach was taken in [10] for the more general Sylvester equation

$$AX + XB = C$$

and when applied to (1.1)–(1.3) we can state a Lyapunov version of their results as (see [10, § IV]) the following theorem.

THEOREM 1.1. *Assume that (1.1)–(1.3) hold with $W \neq 0$. Let $\varepsilon > 0$ be such that*

$$(1.5) \quad \begin{aligned} \|\Delta A\|_F &\leq \varepsilon \|A\|_F, \\ \|\Delta W\|_F &\leq \varepsilon \|W\|_F, \\ 4\varepsilon \|A\|_F \|P^{-1}\| &\leq 1, \end{aligned}$$

then

$$(1.6) \quad \frac{\|\Delta X\|_F}{\|X\|_F} \leq 8\varepsilon \|A\|_F \|P^{-1}\|$$

where $\|\cdot\|_F$ denotes the Frobenius norm defined by

$$\|M\|_F^2 = \sum |M_{ij}|^2.$$

(Note the work in [10] was restricted to real matrices but the extension to complex matrices is straightforward.)

This result is sharp in the sense that there are matrices and perturbations as above such that the bound in (1.6) is nearly an equality. Moreover numerical experiments show that the right side of (1.6) is an accurate predictor of sensitivity of the Lyapunov equation for small perturbations, and that $\|P^{-1}\|$ can be efficiently estimated by techniques similar to those used in [5] (see [4]).

For these reasons we will adopt the bound in Theorem 1.1 as a standard against which we measure our work and that of others ([21], [13], [14]); however, we must make allowances when comparing results based on the Frobenius norm and results based on the spectral norm. If we use (see [11])

$$(1.7) \quad \|M\| \leq \|M\|_F \leq \sqrt{n} \|M\|$$

where $M \in \mathbb{C}^{n \times n}$, then we have

$$(1.8) \quad \frac{\|\Delta X\|}{\sqrt{n} \|X\|} \leq \frac{\|\Delta X\|_F}{\|X\|_F} \leq \frac{\sqrt{n} \|\Delta X\|}{\|X\|}.$$

In view of (1.8) we will say that a bound on $\|\Delta X\|/\|X\|$ is reasonably sharp if it is no more than a small multiple (say less than 10 times) of the right-hand side of (1.6) when we disregard any factors of \sqrt{n} resulting from the conversions between norms. All the results which we survey in this paper, including our own, satisfy this condition.

The above applies to small perturbations but is easily extended to large perturbations. For example, let $\Delta W = 0$ and assume that (1.1)–(1.3) are satisfied with $X + \Delta X \neq 0$. Then we have by (1.4)

$$\begin{aligned} P\Delta x &= -\Delta P(x + \Delta x), \\ \|\Delta x\| &\leq \|P^{-1}\| \|\Delta P\| \|x + \Delta x\|, \end{aligned}$$

so

$$(1.9) \quad \frac{\|\Delta X\|_F}{\|X + \Delta X\|_F} \leq 2\|\Delta A\|_F \|P^{-1}\|.$$

In [14] Jonckheere has given a clever way to exploit inequalities such as (1.9): instead of thinking of (1.9) as a bound on sensitivity we consider the problem of finding a minimum perturbation ΔA , such that $A + \Delta A$ is unstable. That is, we define

$$\delta = \sup \{ \tilde{\delta} \mid \|\Delta A\|_F < \tilde{\delta} \Rightarrow A + \Delta A \text{ is stable} \}$$

and let $\|\Delta A_0\|_F = \delta$ such that $A + \Delta A_0$ is unstable. Now by choosing W appropriately and letting $\Delta A \rightarrow \Delta A_0$ with $\|\Delta A\|_F < \delta$ we have that $\|\Delta X\|_F \rightarrow \infty$. This means that (1.9) becomes

$$(1.10) \quad 1 \leq 2\|\Delta A_0\|_F \|P^{-1}\|,$$

and so

$$(1.11) \quad \frac{1}{2\|P^{-1}\|} \leq \|\Delta A_0\|_F = \delta.$$

We thus have the Jonckheere-type result (for more details see the proof of Theorem 2.2).

THEOREM 1.2. *Let $A \in \mathbb{C}^{n \times n}$ be stable and let*

$$(1.12) \quad \|\Delta A\|_F < \frac{1}{2\|P^{-1}\|}.$$

Then $A + \Delta A$ is stable.

Clearly $\|P^{-1}\|$ is important to the sensitivity question, but how are we to interpret it? One result in this direction is

$$(1.13) \quad \|P^{-1}\|^{-1} = \min \frac{\|A^*X + XA\|_F}{\|X\|_F} \equiv \text{sep}(A^*, -A),$$

where the minimum in (1.13) is taken over all nonzero $X \in \mathbb{C}^{n \times n}$ (see [27] and [28]). However the lack of an interpretation of $\|P^{-1}\|$ in terms of the original system has led to the search for new sensitivity bounds which are more readily understood.

We now consider three results along these lines for normal matrices:

$$(1.14) \quad A^*A = AA^*.$$

In describing these results we will assume that $W = W^*$ and $\Delta W = 0$. Lyapunov problems involving normal matrices occur naturally in some flexible body aerospace systems (see [15]). Laub [21] gives the result that for $P = I \otimes A^T + A^T \otimes I$ with A and W real

$$(1.15) \quad \|P\| \|P^{-1}\| = \frac{\max |\lambda_i^*(A) + \lambda_j(A)|}{\min |\lambda_i^*(A) + \lambda_j(A)|}.$$

When A is also stable we then have

$$(1.16) \quad \frac{\|dX\|_F}{\|X\|_F} \leq \frac{\|dA\|_F}{\min |\text{Re } \lambda_i(A)|},$$

where dA is any differential perturbation of A (that is (1.16) holds in the limit as $\|dA\|_F \rightarrow 0$). Inequality (1.16) has a clear interpretation: if any of the eigenvalues of A are near the imaginary axis we can expect the Lyapunov equation to be sensitive to small perturbations. In [13], Hammarling derived a similar bound for arbitrary ΔA :

$$(1.17) \quad \frac{\|\Delta X\|}{\|X + \Delta X\|} \leq \frac{\|\Delta A\|}{\|A\|} 2\sqrt{n} \frac{\max |\lambda_i(A)|}{\min |\lambda_i^*(A) + \lambda_j(A)|}$$

which for stable, real and normal matrices A reduces to

$$(1.18) \quad \frac{\|\Delta X\|}{\|X + \Delta X\|} \leq \frac{\|\Delta A\|\sqrt{n}}{\min |\operatorname{Re} \lambda_i(A)|}.$$

Inequality (1.18) has the same interpretation as (1.16).

In [14], Jonckheere derives a result for arbitrary ΔA which is sharper than Hammarling's:

$$(1.19) \quad \frac{\|\Delta X\|}{\|X + \Delta X\|} \leq 2\|\Delta A\| \|M\|$$

where

$$M = (M_{ij}) = \left(\frac{1}{\lambda_i(A) + \lambda_j(A)} \right)$$

but which is harder to interpret in terms of the eigenvalues of A .

That each of the above estimates is reasonably sharp is easily seen when we note that if A is normal then so is P and (see [20])

$$(1.20) \quad \lambda_{ij}(P) = \lambda_i(A^*) + \lambda_j(A) = \lambda_i^*(A) + \lambda_j(A).$$

Thus

$$(1.21) \quad \|P^{-1}\| = \frac{1}{\min |\lambda_{ij}(P)|} = \frac{1}{\min |\lambda_i^*(A) + \lambda_j(A)|}.$$

Our work on the sensitivity of the Lyapunov equation has centered on matrices A which are stable but not necessarily normal. We have also dropped the assumptions that $W = W^*$ and $\Delta W = 0$. Our main result is that if (1.1)–(1.3) are satisfied then

$$(1.22) \quad \frac{\|\Delta X\|}{\|X + \Delta X\|} \leq 2\|A + \Delta A\| \|H\| \left[\frac{\|\Delta A\|}{\|A + \Delta A\|} + \frac{\|\Delta W\|}{\|W + \Delta W\|} \right]$$

where H satisfies

$$(1.23) \quad A^*H + HA = -I$$

and we assume that $\|W + \Delta W\|$, $\|A + \Delta A\|$ and $\|X + \Delta X\|$ are all nonzero. Inequality (1.22) is proved in § 2 where we also show that it is reasonably sharp as defined above and that it is optimal for a certain class of estimates. We may compare (1.22) with the results of Laub, Hammarling and Jonckheere for the normal case by noting that when A is stable, real and normal there exists a unitary matrix Q such that

$$(1.24) \quad A = QDQ^*$$

where $D = \operatorname{diag}(\lambda_i(A))$. In this case the solution H to $A^TH + HA = -I$ is given by $H = QDQ^*$ where

$$(1.25) \quad D = \operatorname{diag} \left(\frac{-1}{\lambda_i^*(A) + \lambda_i(A)} \right).$$

This gives

$$(1.26) \quad \|H\| = \frac{1}{\min |\lambda_i^*(A) + \lambda_i(A)|} = \frac{1}{2 \min |\operatorname{Re} \lambda_i(A)|}.$$

Thus for $\Delta W = 0$ and arbitrary ΔA we have

$$(1.27) \quad \frac{\|\Delta X\|}{\|X + \Delta X\|} \leq \frac{\|\Delta A\|}{\min |\operatorname{Re} \lambda_i(A)|}$$

which is the large variation analogue of Laub's differential perturbation result (see (1.15) and (1.16)).

In § 3 we show that the norm of H in the sensitivity bound (1.22) can be interpreted in terms of the damping behavior of the solutions to the dynamical system

$$(1.28) \quad \dot{z} = Az, \quad z(0) = z_0 \in \mathbb{C}^n.$$

We say that a solution \tilde{z} to (1.28) with $\tilde{z}(0) = \tilde{z}_0$, $\|\tilde{z}_0\| = 1$ is *minimally damped* in the L_2 sense (or minimally L_2 damped) if

$$(1.29) \quad \|\tilde{z}\|_{L_2} = \max_{\|z_0\|=1} \|z\|_{L_2}$$

where z in (1.29) satisfies (1.28) and

$$(1.30) \quad \|z\|_{L_2} \equiv \left[\int_0^\infty \|z(t)\|^2 dt \right]^{1/2}$$

is the usual L_2 norm of z over $0 \leq t < \infty$. In § 3 we show that the term $\|H\|$ in (1.22) is the square of the minimal L_2 damping of the dynamical system (1.28) as measured by the right-hand side of (1.29):

$$(1.31) \quad \|H\| = \max_{\|z_0\|=1} \|z\|_{L_2}^2.$$

Thus the sensitivity of the Lyapunov equation and the minimal L_2 damping of the system (1.28) are related through $\|H\|$ and (1.22). This connection is of interest in light of the fact that the Lyapunov equation originally arose from the study of the stability of the solutions of the dynamical system (1.28).

The method used to establish (1.31) also provides a wealth of information about (1.28) which is useful from a control theory point of view. For example, any normalized eigenvector of H corresponding to the largest eigenvalue, $\lambda \max(H)$ represents the initial conditions of a minimally L_2 damped solution to (1.28). This result is easily extended to a finite time interval $t_0 \leq t \leq t_1$ by considering the solution $H_{t_0 t_1}$ to

$$(1.32) \quad A^* H_{t_0 t_1} + H_{t_0 t_1} A = e^{A^* t_1} e^{A t_1} - e^{A^* t_0} e^{A t_0},$$

in which case $\|H_{t_0 t_1}\|$ determines the minimal L_2 damping over the interval $t_0 \leq t \leq t_1$:

$$(1.33) \quad \|H_{t_0 t_1}\| = \max_{\|z_0\|=1} \|z\|_{L_2[t_0, t_1]}^2,$$

where z satisfies (1.28) and

$$(1.34) \quad \|z\|_{L_2[t_0, t_1]} \equiv \left[\int_{t_0}^{t_1} \|z(t)\|^2 dt \right]^{1/2}.$$

As in the case of the infinite time interval $0 \leq t < \infty$, the eigenvectors of $H_{t_0 t_1}$ corresponding to $\lambda \max(H_{t_0 t_1})$ represent the initial conditions of minimally L_2 damped solutions for $t_0 \leq t \leq t_1$.

In the remainder of § 3 we consider the problem of approximating the norm behavior of e^{At} . Our approach is based on the heuristic argument that since the curve $y = \|e^{At}\|$ is the upper envelope of the family of curves $y = \|z(t)\|$ where $\dot{z} = Az$ and $\|z(0)\| = 1$, the minimally L_2 damped solutions to (1.28) should provide a reasonable approximation to $y = \|e^{At}\|$. Various numerical experiments show that this approach works well, especially when we use the minimally L_2 damped solutions associated with the time interval $0 \leq t \leq t_1$ to approximate the short term behavior of $y = \|e^{At}\|$. The final section is devoted to bounding the sensitivity of the matrix exponential along the lines of Van Loan [22] and Kagström [16], but with an emphasis on bounds obtained by using the solutions to various Lyapunov equations.

2. Lyapunov sensitivity. In this section we develop bounds on the sensitivity of the solution to the Lyapunov equation and compare these bounds with established results. Our main result is the following theorem.

THEOREM 2.1. *Let (1.1)–(1.3) be satisfied and let H satisfy*

$$(2.1) \quad A^*H + HA = -I$$

then

$$(2.2) \quad \frac{\|\Delta X\|}{\|X + \Delta X\|} \leq 2\|A + \Delta A\| \|H\| \left[\frac{\|\Delta A\|}{\|A + \Delta A\|} + \frac{\|\Delta W\|}{\|W + \Delta W\|} \right]$$

where we assume that $\|X + \Delta X\|$, $\|A + \Delta A\|$ and $\|W + \Delta W\|$ are all nonzero.

Proof. Since A is assumed to be stable, H in (2.1) is positive definite ($H > 0$) and may be represented as

$$(2.3) \quad H = \int_0^\infty e^{A^*t} e^{At} dt.$$

Now expand (1.3) and use (1.1) to get

$$(2.4) \quad A^*\Delta X + \Delta XA = -(\Delta W + \Delta A^*(X + \Delta X) + (X + \Delta X)\Delta A).$$

This may be rewritten as

$$(2.5) \quad \Delta X = \int_0^\infty e^{A^*t} (\Delta W + \Delta A^*(X + \Delta X) + (X + \Delta X)\Delta A) e^{At} dt.$$

Let u and v denote the left and right singular vectors of unit length of ΔX such that

$$(2.6) \quad u^* \Delta X v = \|\Delta X\|.$$

Combining (2.5) and (2.6) we get

$$\begin{aligned} \|\Delta X\| &= \int_0^\infty u^* e^{A^*t} (\Delta W + \Delta A^*(X + \Delta X) + (X + \Delta X)\Delta A) e^{At} v dt \\ (2.7) \quad &\leq \|\Delta W + \Delta A^*(X + \Delta X) + (X + \Delta X)\Delta A\| \int_0^\infty \|e^{At} u\| \|e^{At} v\| dt \\ &\leq (\|\Delta W\| + 2\|\Delta A\| \|X + \Delta X\|) \int_0^\infty \|e^{At} u\| \|e^{At} v\| dt. \end{aligned}$$

By the Cauchy-Schwarz inequality we have

$$(2.8) \quad \int_0^\infty \|e^{At} u\| \|e^{At} v\| dt \leq \left[\int_0^\infty \|e^{At} u\|^2 dt \right]^{1/2} \left[\int_0^\infty \|e^{At} v\|^2 dt \right]^{1/2}.$$

Moreover

$$\int_0^\infty \|e^{A^t} u\|^2 dt = \int_0^\infty u^* e^{A^* t} e^{A^t} u dt = u^* \int_0^\infty e^{A^* t} e^{A^t} dt u = u^* H u.$$

Using $H = H^*$ and $H > 0$, together with the fact that u has unit length gives

$$u^* H u \leq \|H\|.$$

Thus

$$(2.9) \quad \int_0^\infty \|e^{A^t} u\|^2 dt \leq \|H\|.$$

Similarly

$$(2.10) \quad \int_0^\infty \|e^{A^t} v\|^2 dt \leq \|H\|.$$

Using (2.8)–(2.10) in (2.7) we get

$$(2.11) \quad \|\Delta X\| \leq (\|\Delta W\| + 2\|\Delta A\| \|X + \Delta X\|) \|H\|.$$

From (1.3) we have that

$$(2.12) \quad \|W + \Delta W\| \leq 2\|A + \Delta A\| \|X + \Delta X\|$$

so that (2.11) can be written as

$$\frac{\|\Delta X\|}{\|X + \Delta X\|} \leq 2\|A + \Delta A\| \|H\| \left[\frac{\|\Delta A\|}{\|A + \Delta A\|} + \frac{\|\Delta W\|}{\|W + \Delta W\|} \right]$$

which is the desired result. \square

Remark. Different bounds on $\|\Delta X\|$ can be obtained by using

$$\|e^{A^t} u\| \leq \|e^{A^t}\|, \quad \|e^{A^t} v\| \leq \|e^{A^t}\|$$

in (2.7). This gives

$$\|\Delta X\| \leq (\|\Delta W\| + 2\|\Delta A\| \|X + \Delta X\|) \int_0^\infty \|e^{A^t}\|^2 dt.$$

Now any integrable bound on $\|e^{A^t}\|^2$ can be used to estimate $\int_0^\infty \|e^{A^t}\|^2 dt$. For example, the bound

$$\|e^{A^t}\| \leq e^{\mu(A)t}$$

is integrable if

$$\mu(A) \equiv \frac{1}{2} \lambda_{\max}(A^* + A) < 0.$$

In this case we obtain

$$\|\Delta X\| \leq (\|\Delta W\| + 2\|\Delta A\| \|X + \Delta X\|) \frac{1}{2|\mu(A)|}.$$

For other bounds on $\|e^{A^*t}\|$, see [22] and [16]. However, the bound obtained in (2.11) is optimal for this class of estimates because

$$\|H\| = \left\| \int_0^\infty e^{A^*t} e^{At} dt \right\| \leq \int_0^\infty \|e^{A^*t}\|^2 dt.$$

We may compare (2.2) with (1.6) for small perturbations by assuming that for $\varepsilon > 0$

$$(2.13) \quad \|\Delta A\| \leq \varepsilon \|A\|,$$

$$(2.14) \quad \|\Delta W\| \leq \varepsilon \|W\|,$$

$$(2.15) \quad 8\varepsilon \|A\| \|H\| \leq \frac{1-\varepsilon}{1+\varepsilon}.$$

Then after some algebra, (2.2) gives

$$(2.16) \quad \frac{\|\Delta X\|}{\|X\|} \leq 8\varepsilon \|A\| \|H\| (1-\varepsilon) \approx 8\varepsilon \|A\| \|H\|,$$

for ε small.

If we disregard the difference in norms, (2.16) has the form of (1.6) except that $\|P^{-1}\|$ is replaced by $\|H\|$. These two terms may be compared by using (1.13) and (1.7)

$$\|P^{-1}\|^{-1} = \min \frac{\|A^*X + XA\|_F}{\|X\|_F} \leq \frac{\|A^*H + HA\|_F}{\|H\|_F} = \frac{\|I\|_F}{\|H\|_F} \leq \frac{\sqrt{n}}{\|H\|}.$$

Thus

$$(2.17) \quad \|H\| \leq \sqrt{n} \|P^{-1}\|.$$

In order to assess how accurately (1.6) and (2.16) measure the sensitivity of the Lyapunov equation we have selected four control problems of the form

$$(2.18) \quad x = A_0x + Bu, \quad y = Cx,$$

from the literature and formed the closed loop state matrix A by using standard quadratic regulator techniques (see [1]):

$$(2.19) \quad A = A_0 - BB^TP$$

where P is the positive extremal solution to the algebraic Riccati equation

$$(2.20) \quad 0 = C^TC + A_0^TP + PA_0 - PBB^TP$$

and (A, B, C) is observable and controllable. These examples were chosen because we had used them in some earlier studies of the algebraic Riccati equation.

This method ensures that A in (2.19) is stable:

$$(2.21) \quad \operatorname{Re} \lambda_i(A) < 0.$$

A variant of this procedure is to let $P = P(\rho)$ be the positive extremal solution to

$$(2.22) \quad 0 = C^TC + (A_0 + \rho I)^TP + P(A_0 + \rho I) - PBB^TP$$

for $\rho > 0$. This gives

$$\operatorname{Re} \lambda_i(A_0 + \rho I - BB^TP(\rho)) < 0$$

so that

$$(2.23) \quad \operatorname{Re} \lambda_i(A_0 - BB^TP(\rho)) < -\rho.$$

Because of (2.23), Anderson and Moore (see [1, Chap. 4]) refer to ρ as a prescribed degree of stability.

In Table 1 we compare the principal terms $\|A\| \|H\|$ and $\|A\|_F \|P^{-1}\|$ of (2.16) and (1.6) respectively, for $A = A_0 - BB^TP(\rho)$ for various values of ρ . These terms are

TABLE 1
Sensitivity estimates for $A^T X + XA = -W$.

Example	$\ H\ $	$\ P^{-1}\ $	$\ A\ \ H\ $	$\ A\ _F \ P^{-1}\ $	$\frac{\ \Delta X\ }{\ X\ }$	$\frac{\ \Delta X\ _F}{\ X\ _F}$	ρ
Moore-Laub	1.60	1.57	11.7	12.2	1.7×10^{-16}	1.8×10^{-16}	0
$n = 4$	7.79	7.85	3.8×10^2	4.1×10^2	1.4×10^{-15}	1.4×10^{-15}	2
$m = 2$	21.6	21.6	3.7×10^3	4.5×10^3	6.2×10^{-15}	6.2×10^{-15}	4
$l = 1$	38.2	38.1	1.6×10^4	2.0×10^4	2.0×10^{-14}	2.0×10^{-14}	6
	55.3	55.3	4.6×10^4	5.4×10^4	3.5×10^{-14}	3.5×10^{-14}	8
	72.7	72.7	9.9×10^4	1.1×10^5	2.9×10^{-13}	2.9×10^{-13}	10
Kailath	52.8	53.0	1.0×10^5	1.0×10^5	5.4×10^{-13}	5.4×10^{-13}	0
$n = 5$	3.2×10^2	2.9×10^2	1.9×10^6	1.8×10^6	3.1×10^{-13}	3.1×10^{-13}	2
$m = 2$	9.3×10^2	8.5×10^2	1.2×10^7	1.1×10^7	2.2×10^{-12}	2.1×10^{-12}	4
$l = 3$	1.6×10^3	1.5×10^3	3.3×10^7	3.0×10^7	1.9×10^{-12}	1.8×10^{-12}	6
	2.4×10^3	2.2×10^3	7.0×10^7	6.4×10^7	1.3×10^{-12}	1.3×10^{-12}	8
	3.1×10^3	2.9×10^3	1.2×10^8	1.1×10^8	7.4×10^{-12}	7.3×10^{-12}	10
Chung-Shapiro	73.1	68.0	4.2×10^3	4.4×10^3	4.5×10^{-15}	5.3×10^{-15}	0
$n = 6$	35.1	30.9	1.9×10^4	1.9×10^4	5.6×10^{-13}	5.8×10^{-13}	2
$m = 2$	140	136	2.7×10^5	3.1×10^5	1.9×10^{-12}	1.9×10^{-12}	4
$l = 3$	349	345	1.5×10^6	1.7×10^6	2.2×10^{-11}	2.3×10^{-11}	6
	652	648	4.9×10^6	5.8×10^6	8.0×10^{-11}	8.4×10^{-11}	8
	1.0×10^3	1.0×10^3	1.2×10^7	1.5×10^7	1.5×10^{-10}	1.5×10^{-10}	10
Davison-Maki	687	684	2.3×10^5	2.6×10^5	2.8×10^{-13}	2.8×10^{-13}	0
$n = 9$	8.0×10^4	5.4×10^4	8.2×10^8	5.6×10^8	3.6×10^{-10}	3.5×10^{-10}	1
$m = 1$	1.4×10^6	9.1×10^5	1.6×10^{11}	1.1×10^{11}	1.6×10^{-8}	1.6×10^{-8}	2
$l = 9$	8.8×10^6	6.2×10^6	5.9×10^{12}	4.3×10^{12}	1.5×10^{-6}	1.4×10^{-6}	3
	5.4×10^7	4.6×10^7	1.5×10^{14}	1.2×10^{14}	1.2×10^{-5}	1.2×10^{-5}	4

almost identical over the entire example set. The same is true for the terms $\|H\|$ and $\|P^{-1}\|$ which are included so as to test (2.17). The computation of $P(\rho)$ in (2.22) was performed by the numerically stable subroutine RICOND of Laub [21] and the spectral norms were evaluated by using the Linpack singular value decomposition subroutine DSVDC (see [9]).

In order to see how well (2.16) and (1.6) measure sensitivity, we used the numerically stable subroutine AXXBC of Golub, Nash and Van Loan [10] to solve

$$A^T X + XA = -W,$$

where X was chosen to be the n th order Hilbert matrix $X = (x_{ij}) = 1/(i+j-1)$ and W was obtained by forming the matrix $-(A^T X + XA)$. The computed solution X_c is then the exact solution to a perturbed problem $(A + \Delta A)^T X_c + X_c(A + \Delta A) = -(W + \Delta W)$ (see [10]) and we can estimate the sensitivity of the Lyapunov equation from the terms $\|X - X_c\|/\|X\|$ and $\|X - X_c\|_F/\|X\|_F$. All computations were performed on a VAX-780 in double precision for which the machine epsilon is about 10^{-17} . The results are given in Table 1 and we can say that for the problems tested the right-hand sides of (1.6) and (2.16) are accurate predictors of the sensitivity of the Lyapunov equation in that

$$\frac{\|X - X_c\|_F}{\|X\|_F} \approx 8\varepsilon \|A\|_F \|P^{-1}\|,$$

$$\frac{\|X - X_c\|}{\|X\|} \approx 8\varepsilon \|A\| \|H\|.$$

Since Theorem 2.1 applies to arbitrary perturbations we can obtain a Jonckheere-type bound on the size of the smallest destabilizing perturbation for A .

THEOREM 2.2. *Let A be stable and let H satisfy $A^*H + HA = -I$. Let ΔA satisfy $\|\Delta A\| < 1/2\|H\|$. Then $A + \Delta A$ is stable.*

Proof. We will use the fact that for any square matrix P the equation

$$Px = w$$

has no solution x if $P^*w = 0$ and $w \neq 0$. For the Lyapunov equation, this means that

$$(A + \Delta A)^*(X + \Delta X) + (X + \Delta X)(A + \Delta A) = -W$$

has no solution $X + \Delta X$ if $P^*w = 0$ where

$$P = I \otimes (A + \Delta A)^* + (A + \Delta A)^T \otimes I$$

and $w = \text{Vec}(W) \neq 0$ (see [12]).

Now define

$$(2.24) \quad \delta \equiv \sup \{ \tilde{\delta} \mid \|\Delta A\| < \tilde{\delta} \Rightarrow A + \Delta A \text{ is stable} \}.$$

The stability of A and the continuity of its eigenvalues ensure that $\delta > 0$. We may find ΔA_0 such that $A + \Delta A_0$ is unstable but $A + t\Delta A_0$ is stable for $0 \leq t < 1$ where $\|\Delta A_0\| = \delta$. Now let $X = X(t)$ be the solution to

$$(2.25) \quad (A + t\Delta A_0)^*X(t) + X(t)(A + t\Delta A_0) = -W, \quad 0 < t < 1,$$

where $P_1^*w = 0$ for $w = \text{Vec}(W) \neq 0$ and $P_1 = I \otimes (A + \Delta A_0)^* + (A + \Delta A_0)^T \otimes I$.

This ensures that $\lim_{t \rightarrow 1} \|X(t)\| = \infty$ because (2.25) has no solution at $t = 1$ by the choice of W . If we set $\Delta X = X(t) - X(0)$ and $\Delta W = 0$ we obtain from Theorem 2.1

$$\frac{\|X(t) - X(0)\|}{\|X(t)\|} \leq 2t\|\Delta A_0\|\|H\|,$$

which in the limit as $t \rightarrow 1$ gives

$$1 \leq 2\|\Delta A_0\|\|H\|.$$

Thus

$$\frac{1}{2\|H\|} < \|\Delta A_0\| = \delta$$

which is the desired result. \square

Note that the above argument suffices to prove Theorem 1.2 if we use (1.9) to bound $\|X(t) - X(0)\|_F / \|X(t)\|_F$ and then take the limit as $t \rightarrow 1$.

It is interesting that a more general version of Theorem 2.2 can be proved by a completely different argument which relies on the following well-known stability theorem (see [20]).

THEOREM 2.3. *Let \tilde{W} be Hermitian and positive definite and let \tilde{A}, \tilde{H} satisfy*

$$(2.26) \quad \tilde{A}^*\tilde{H} + \tilde{H}\tilde{A} = -\tilde{W}.$$

(2.27) *Then \tilde{A} is stable if and only if \tilde{H} is Hermitian positive definite.*

Using this theorem we can prove the next theorem.

THEOREM 2.4. *Let A be stable and let H, W satisfy*

$$(2.28) \quad A^*H + HA = -W$$

where W is Hermitian positive definite. Let ΔA satisfy

$$(2.29) \quad \|\Delta A\| < \frac{\lambda_{\min}(W)}{2\|H\|}.$$

Then $A + \Delta A$ is stable.

Proof. By Theorem 2.3 we have that H is Hermitian and positive definite. Now define

$$(2.30) \quad \tilde{W} = W + \Delta A^* H + H \Delta A,$$

$$(2.31) \quad \tilde{A} = A + \Delta A,$$

$$(2.32) \quad \tilde{H} = H.$$

Then (2.26) is satisfied and \tilde{W} is Hermitian positive definite because

$$\lambda_{\min}(\tilde{W}) \geq \lambda_{\min}(W) - 2\|\Delta A\| \|A\| \|H\| > 0$$

by (2.29). Therefore by Theorem 2.3, \tilde{A} is stable. \square

Theorem 2.1 and the numerical results in Table 1 show that $\|H\|$ plays a role in spectral norm sensitivity estimates which is analogous to the role of $\|P^{-1}\|$ in Frobenius norm sensitivity estimates. This analogy extends to the lower bounds $1/2\|P^{-1}\|$ of Theorem 1.2 and $1/2\|H\|$ of Theorem 2.2 for the smallest destabilizing perturbations of A for the Frobenius and spectral norms respectively. In the next section we seek to understand the roles of H and $\|H\|$ in the sensitivity problem by establishing a relationship between H and the dynamical system $\dot{z} = Az$.

3. Interpretation. In this section we explore the connection between H satisfying

$$(3.1) \quad A^* H + H A = -I$$

and the norm behavior of solutions $z = z(t)$ to the dynamical system

$$(3.2) \quad \dot{z} = Az, \quad z(0) = z_0 \in \mathbb{C}^n.$$

This connection gives us one way of understanding the sensitivity estimate (2.2) and provides a great deal of information about the norm behavior of e^{At} . As a by-product we also obtain a means of identifying solutions to (3.2) which are of interest in control theory.

Our main tool in establishing the connection between (3.1) and (3.2) is the following theorem which portrays H as the “area matrix” associated with (3.2).

THEOREM 3.1. *Let A be stable and let H satisfy (3.1). For $z_0 \in \mathbb{C}^n$ define $f(t) = \|z(t)\|^2$ where z satisfies (3.2). Then the area under f for $0 \leq t \leq \infty$ is equal to $z_0^* H z_0$.*

Proof. Since A is stable, H can be written as

$$H = \int_0^\infty e^{A^* t} e^{A t} dt.$$

The area under the curve f for $0 \leq t \leq \infty$ is given by

$$\begin{aligned} \int_0^\infty f(t) dt &= \int_0^\infty \|z(t)\|^2 dt \\ &= \int_0^\infty \|e^{A t} z_0\|^2 dt \\ &= \int_0^\infty z_0^* e^{A^* t} e^{A t} z_0 dt \\ &= z_0^* \int_0^\infty e^{A^* t} e^{A t} dt z_0 = z_0^* H z_0. \end{aligned}$$

\square

As in (1.29) we will say that \tilde{z} is minimally damped in the L_2 sense (or minimally L_2 damped) if \tilde{z} satisfies (3.2) with $\tilde{z}(0) = \tilde{z}_0$ of unit norm and

$$(3.3) \quad \|\tilde{z}\|_{L_2} = \max_{\|z_0\|=1} \|z\|_{L_2},$$

where z in (3.3) satisfies (3.2) and

$$\|z\|_{L_2} = \left[\int_0^\infty \|z(t)\|^2 dt \right]^{1/2}$$

is the usual L_2 norm of z over $0 < t < \infty$. As an immediate consequence of $z_0^* H z_0 \leq \|H\|$ for $\|z_0\| = 1$, we have by Theorem 3.1 that $\|H\|$ is the square of the minimal L_2 damping for the system (3.2):

$$(3.4) \quad \|H\| = \max_{\|z_0\|=1} \|z\|_{L_2}^2,$$

where equality is attained when z_0 is a normalized eigenvector of H corresponding to $\lambda_{\max}(H)$. Thus minimally L_2 damped solutions of (3.2) determine the norm of H and hence govern the sensitivity of the Lyapunov equation by (2.2). This generalizes the result for the real normal case: If A is real, stable and normal, the sensitivity of the Lyapunov equation (see 1.16) is governed by $1/|\alpha|$ where $\alpha = \max_i \operatorname{Re} \lambda_i(A)$. Moreover it is easily shown in this case that $1/|\alpha|$ determines the minimal L_2 damping for $z = Az$:

$$(3.5) \quad \frac{1}{2|\alpha|} = \max_{\|z_0\|=1} \|z\|_{L_2}^2.$$

This follows from the fact that for such matrices

$$(3.6) \quad \|z(t)\| \leq \|z_0\| e^{\alpha t}$$

and that there exists an initial vector z_0 such that

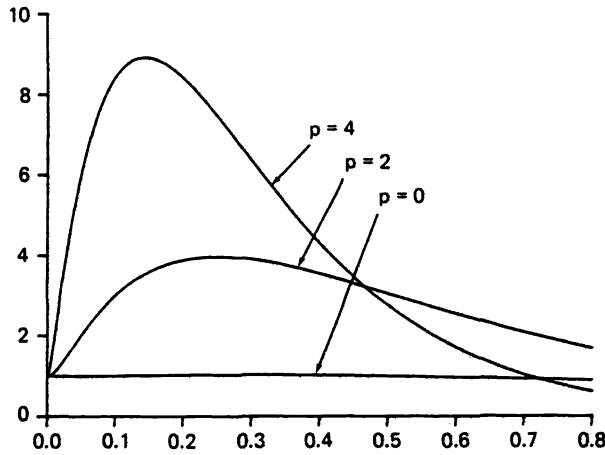
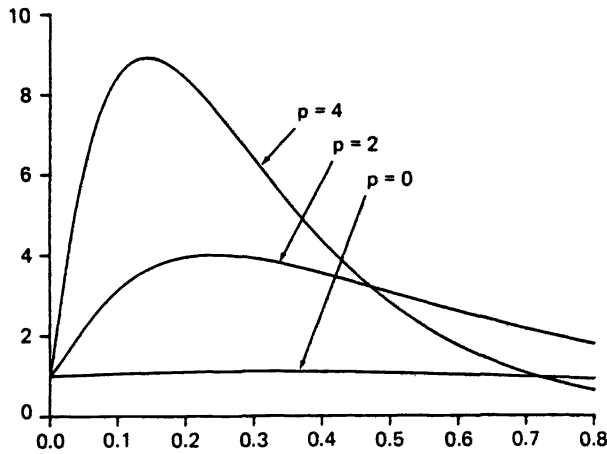
$$\|z_0\| = 1 \quad \text{with} \quad \|z(t)\| = e^{\alpha t} \quad \text{where} \quad \dot{z} = Az.$$

From the proof of Theorem 3.1 we see that the initial conditions z_0 , which give rise to minimally L_2 damped solutions of (3.2), are simply eigenvectors of H corresponding to the largest eigenvalue $\lambda_{\max}(H)$. The identification of such initial conditions is important in a control theory context because the corresponding solutions are either weakly damped or have large norm transient behavior relative to other solutions of (3.2).

The area interpretation of $\|H\|$ is also very useful in understanding the numerical results given in Table 1 for the "prescribed stability" procedure. In that procedure the closed loop state matrix $A = A(\rho)$ is forced to have eigenvalues to the left of $x = -\rho$ where ρ is the margin parameter. This would seem to be desirable in that solutions to $\dot{z} = Az$ are forced to decay more rapidly as ρ increases:

$$\|z(t)\| \leq C e^{-\rho t} \|z_0\|$$

where C is a constant depending on ρ but independent of the initial condition $z_0 = z(0)$. However an inspection of Table 1 reveals that $\|H\|$ increases as ρ gets larger, thus indicating that the constant $C = C(\rho)$ is increasing with ρ fast enough to mask the exponential decay for small t values. This is illustrated in Figs. 1(a), 2(a), 3(a) where we plot the maximal area curves $\|z(t)\|$ as a function of t for several values of ρ . Here $\dot{z} = A(\rho)z$ and $z(0)$ is a normalized eigenvector of $H = H(\rho)$ corresponding to $\lambda_{\max}(H)$ where $A^*H + HA = -I$. These figures clearly show that the price of fast asymptotic decay can be large norm transient behavior.

FIG. 1(a). Moore-Laub: norm $z(t)$ versus time where $\dot{z} = Az$.FIG. 1(b). Moore-Laub: spectral norm of $\exp(At)$ versus time.

In Theorem 3.1 we considered the time interval $0 \leq t < \infty$. In general however we are more often interested in a restricted time interval $t_0 \leq t \leq t_1$. For example, the interval $0 \leq t \leq t_1$ is of prime importance in analyzing the transient behavior of solutions to (3.2). The next theorem shows that we can obtain area results for such finite intervals.

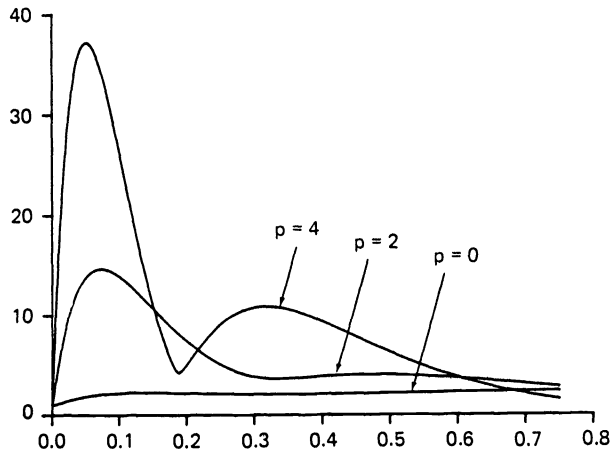
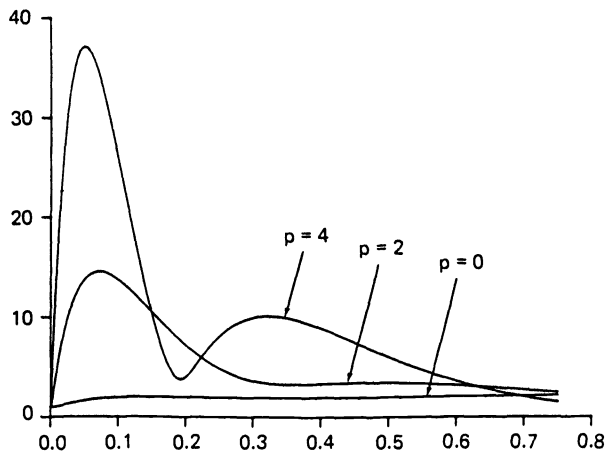
THEOREM 3.2. Let A be stable and let H satisfy

$$(3.7) \quad A^*H + HA = e^{A^*t_1} e^{At_1} - e^{A^*t_0} e^{At_0}.$$

For $z_0 \in \mathbb{C}^n$, define $f(t) = \|z(t)\|^2$ where z satisfies (3.2). Then the area under f for $t_0 \leq t \leq t_1$ is equal to $z_0^* H z_0$.

Proof. Since A is stable we may represent H as

$$(3.8) \quad \begin{aligned} H &= - \int_0^\infty e^{A^*t} (e^{A^*t_1} e^{At_1} - e^{A^*t_0} e^{At_0}) e^{At} dt \\ &= \int_0^\infty e^{A^*(t+t_0)} e^{A(t+t_0)} dt - \int_0^\infty e^{A^*(t+t_1)} e^{A(t+t_1)} dt \end{aligned}$$

FIG. 2(a). Chung-Shapiro: norm $z(t)$ versus time where $\dot{z} = Az$.FIG. 2(b). Chung-Shapiro: spectral norm of $\exp(At)$ versus time.

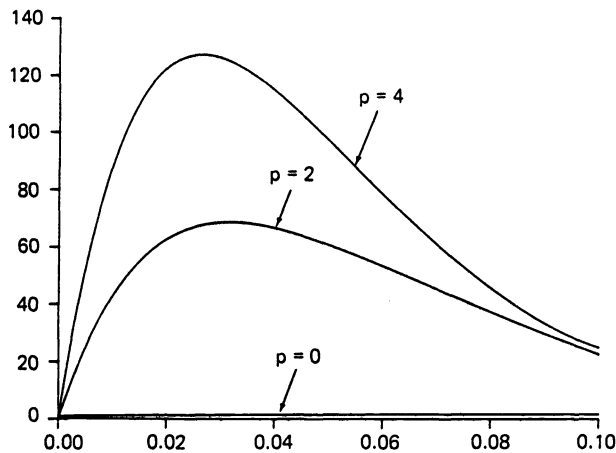
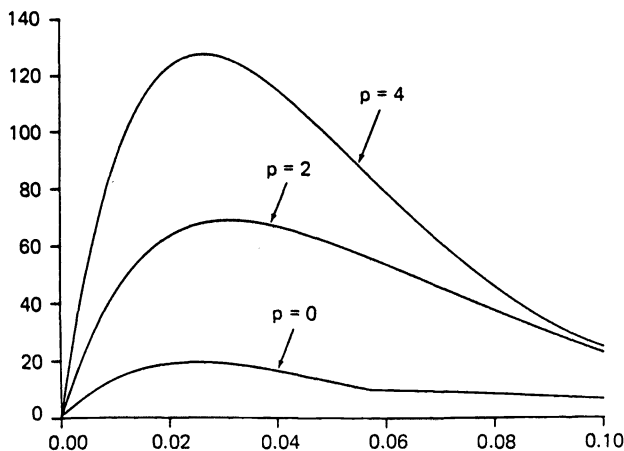
$$\begin{aligned}
 (3.9) \quad &= \int_{t_0}^{\infty} e^{A^*t} e^{At} dt - \int_{t_1}^{\infty} e^{A^*t} e^{At} dt \\
 &= \int_{t_0}^{t_1} e^{A^*t} e^{At} dt.
 \end{aligned}$$

Now as in the proof of Theorem 3.1 we have

$$\begin{aligned}
 \int_{t_0}^{t_1} f(t) dt &= \int_{t_0}^{t_1} \|z(t)\|^2 dt = \int_{t_0}^{t_1} \|e^{At} z_0\|^2 dt \\
 &= \int_{t_0}^{t_1} z_0^* e^{A^*t} e^{At} z_0 dt = z_0^* H z_0.
 \end{aligned}$$

□

As in the case for $0 \leq t \leq \infty$, the eigenvectors of H in (3.7) corresponding to the eigenvalue $\lambda_{\max}(H)$ represent the initial conditions of solutions to the dynamical system (3.2) which are least stable during the time interval $t_0 \leq t \leq t_1$ in the sense of having maximal L_2 norms.


 FIG. 3(a). Kailath: norm $z(t)$ versus time where $\dot{z} = Az$.

 FIG. 3(b). Kailath: spectral norm of $\exp(At)$ versus time.

If we turn to the problem of approximating the norm curve $N(t) = \|e^{At}\|$ of the matrix exponential, we find the area approach convenient because $N^2(t)$ is the upper envelope of the family of curves $f = f(t, z_0)$ where we let z_0 vary over all vectors of unit length:

$$(3.10) \quad \|e^{At}\|^2 = \max \|e^{At} z_0\|^2 = \max f(t, z_0).$$

Thus we have that for any z_0 of unit length, $f(t, z_0)$ is a lower bound on $\|e^{At}\|^2$ and in particular when H satisfies (3.1) and z_0 is an eigenvector of H corresponding to $\lambda_{\max}(H)$ then $f^{1/2} = f^{1/2}(t, z_0)$ generally provides a good approximation to $\|e^{At}\|$. This can be seen by comparing the maximal area curves $f^{1/2}(t, z_0) = \|z(t)\|$ in Figs. 1(a), 2(a), 3(a) with the curves $\|e^{At}\|$ in Figs. 1(b), 2(b), 3(b). In all cases except $\rho = 0$ for the Kailath problem (Figs. 3(a), 3(b)) the agreement is excellent.

In the Kailath problem with $\rho = 0$, the short term behavior of $\|e^{At}\|$ is determined by solutions of (3.2) which exhibit large transients before decaying rapidly. However the asymptotic behavior of $\|e^{At}\|$ is very closely matched by the norm of the minimally L_2 damped solution to (3.2). The short term behavior of $\|e^{At}\|$ for this problem can be well approximated by using minimally L_2 damped solutions for restricted time

intervals: in Fig. 4(b) we plot $\|e^{At}\|$ and in Fig. 4(a) we plot $\|z(t)\|$ for two solutions z to (3.2) whose initial conditions are eigenvectors of H , corresponding to $\lambda_{\max}(H)$, where H satisfies (3.7) for $0 \leq t \leq 0.04$ (curve 1 in Fig. 4(a)) and $0.06 \leq t \leq 0.1$ (curve 2 in Fig. 4(a)) respectively. The next theorem shows that if we restrict ourselves to an interval $t_0 \leq t \leq t_1$ then the approximation to $\|e^{At}\|$ by the maximal area curve $\|z(t)\|$ becomes exact as $t_1 \rightarrow t_0$.

THEOREM 3.3. Let H_ε denote the solution to

$$(3.11) \quad A^* H_\varepsilon + H_\varepsilon A = e^{A^* t_1} e^{A t_1} - e^{A^* t_0} e^{A t_0}$$

where $t_1 = t_0 + \varepsilon$ and A is stable. Then

$$(3.12) \quad \lim_{\varepsilon \rightarrow 0} \frac{H_\varepsilon}{\varepsilon} = e^{A^* t_0} e^{A t_0}.$$

Proof. As in the proof of Theorem 3.2, H_ε in (3.11) satisfies

$$(3.13) \quad H_\varepsilon = \int_{t_0}^{t_0 + \varepsilon} e^{A^* t} e^{A t} dt.$$

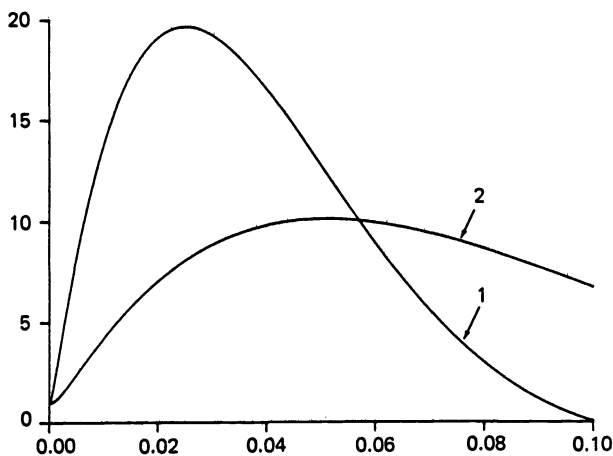


FIG. 4(a). Kailath: norm $z(t)$ versus time where $\dot{z} = Az$.

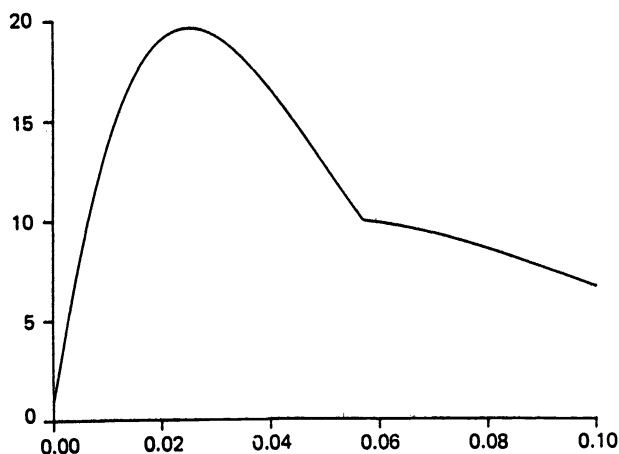


FIG. 4(b). Kailath: spectral norm of $\exp(At)$ versus time.

Thus, by the continuity of the matrix function $e^{A^*t} e^{At}$ we have

$$\lim_{\varepsilon \rightarrow 0} \frac{H_\varepsilon}{\varepsilon} = \lim_{\varepsilon \rightarrow 0} \frac{\int_{t_0}^{t_0+\varepsilon} e^{A^*t} e^{At} dt}{\varepsilon} = e^{A^*t_0} e^{At_0}. \quad \square$$

We may interpret Theorem 3.3 as

$$(3.14) \quad H_\varepsilon \approx \varepsilon e^{A^*t_0} e^{At_0}$$

for small ε . In particular, this means that the eigenspace of H_ε corresponding to $\lambda_{\max}(H_\varepsilon)$ tends to the singular vector space of e^{At_0} corresponding to $\sigma_{\max}(e^{At_0})$. Also notice that if we wish to find a solution $z = z(t)$ to $\dot{z} = Az$ which satisfies $\|z(t_0)\| = \|e^{At_0}\|$ for given value of t_0 , then we need only choose z_0 to be a normalized singular vector of e^{At_0} corresponding to $\sigma_{\max}(e^{At_0})$ because we then have $\|e^{At_0}z_0\| = \|e^{At_0}\|$ and $z(t) = e^{At}z_0$.

A nice feature of the area approach is that many established results can be proved quite directly. For example, the following theorem was implicit in the work of Pao [25] and explicitly proved by Bass in [2] (see also (4.9) in § 4 for a proof based on the logarithmic norm $\mu(A) = \frac{1}{2}\lambda_{\max}(A + A^*)$).

THEOREM 3.4. *Let A be stable and let H satisfy $A^*H + HA = -W$ where W is positive definite and Hermitian. Then*

$$(3.15) \quad \max_{0 \leq t < \infty} \|e^{At}\| \leq K^{1/2}(H)$$

where $K(H) = \|H\| \|H^{-1}\|$.

Proof. Let $t \geq 0$ and let $x_0 \in \mathbb{C}^N$ satisfy $\|e^{At}\| = \|e^{At}x_0\|$, $\|x_0\| = 1$. Then

$$\begin{aligned} \int_t^\infty x_0^* e^{A^*s} W e^{As} x_0 ds &\leq \int_0^\infty x_0^* e^{A^*s} W e^{As} x_0 ds = x_0^* H x_0 \\ &\leq \lambda_{\max}(H) = \|H\|, \end{aligned}$$

since $H = \int_0^\infty e^{A^*s} W e^{As} ds$, $H = H^*$ and $H > 0$.

On the other hand,

$$\begin{aligned} \int_t^\infty x_0^* e^{A^*s} W e^{As} x_0 ds &= \int_0^\infty x_0^* e^{A^*(t+s)} W e^{A(t+s)} x_0 ds \\ &= x_0^* e^{A^*t} \int_0^\infty e^{A^*s} W e^{As} ds e^{At} x_0 \\ &= x_0^* e^{A^*t} H e^{At} x_0 \geq \lambda_{\min}(H) \|e^{At}x_0\|^2 = \frac{1}{\|H^{-1}\|} \|e^{At}\|^2. \end{aligned}$$

Thus $\|e^{At}\|^2 \leq \|H\| \|H^{-1}\|$ for all $t \geq 0$. \square

The above theorem gives an upper bound on the “hump” height (see [23])

$$(3.16) \quad h \equiv \max_{0 \leq t < \infty} \|e^{At}\|.$$

The next theorem shows that the bound (3.15) also applies to all matrices $A = A + \Delta A$ in a neighborhood of A .

THEOREM 3.5. *Let A be stable and let $W = W^*$ be positive definite such that*

$$(3.17) \quad A^*H + HA = -W.$$

Let ΔA satisfy

$$(3.18) \quad \|\Delta A\| < \frac{\lambda_{\min}(W)}{2\|H\|}.$$

Then

$$(3.19) \quad \max_{0 \leq t \leq \infty} \|e^{(A+\Delta A)t}\| \leq K^{1/2}(H).$$

Proof. As in the proof of Theorem 2.4 let

$$\tilde{A} = A + \Delta A, \quad \tilde{W} = W + \Delta A^*H + H\Delta A.$$

The $\tilde{A}^*H + H\tilde{A} = -\tilde{W}$ and A is stable, W is Hermitian positive definite. The result now follows from Theorem (3.4).

The connection between the Lyapunov equation and the norm behavior of the matrix exponential is further brought out by the next lemma.

LEMMA 3.1. *The following are equivalent:*

$$(3.20) \quad (1) \quad \max_{t \geq 0} \|e^{At}\| \leq 1,$$

$$(3.21) \quad (2) \quad \mu(A) \leq 0,$$

(3) *There exists positive definite Hermitian H commuting with A such that*

$$(3.22) \quad A^*H + HA = -W$$

where $W = W^*$ and $W \geq 0$.

Proof. Since (1) and (2) are well known to be equivalent [22] we show that (2) \Rightarrow (3) \Rightarrow 1.

Assume (2): $\mu(A) = \frac{1}{2}\lambda_{\max}(A + A^*) \leq 0$. Then $W \equiv -(A + A^*)$ is positive semi-definite Hermitian and $H = I$ satisfies $A^*H + HA = -W$ and commutes with A . That is (2) \Rightarrow (3).

Now assume (3). Since H is positive definite Hermitian there is a unique positive definite Hermitian square root of H ; say $H^{1/2}$. Moreover since H commutes with A so does $H^{1/2}$ (see [8]). In particular

$$(3.23) \quad A = H^{1/2}H^{-1/2}A = H^{1/2}AH^{-1/2}$$

so that

$$(3.24) \quad \|e^{At}\| = \|e^{H^{1/2}AH^{-1/2}t}\| \leq e^{\mu(H^{1/2}AH^{-1/2}t)}.$$

The result now follows when we note that

$$\begin{aligned} \mu(H^{1/2}AH^{-1/2}) &= \frac{1}{2}\lambda_{\max}(H^{-1/2}A^*H^{1/2} + H^{1/2}AH^{-1/2}) \\ &= \frac{1}{2}\lambda_{\max}(H^{-1/2}(A^*H + HA)H^{-1/2}) \\ &= \frac{1}{2}\lambda_{\max}(-H^{-1/2}WH^{-1/2}) \leq 0. \end{aligned}$$

□

As a simple numerical test of the bound $\|e^{At}\| \leq K^{1/2}(H)$ we have taken

$$A = \begin{bmatrix} -1 & a \\ 0 & -1 \end{bmatrix}$$

for increasing values of a with $W = I$. The results are given in Table 2.

TABLE 2
Testing $\max \|e^{At}\| \leq K^{1/2}(H)$ for $W = I$.

a	$\max_{0 < t} \ e^{At}\ $	$K^{1/2}(H)$
0	1.0	1.0
10^1	3.7	10^1
10^2	3.7×10^1	10^2
10^3	3.7×10^2	10^3

TABLE 3
Further tests $\max \|e^{At}\| \leq K^{1/2}(H)$.

Example	$\max_{0 \leq t < \infty} \ e^{At}\ $	$K^{1/2}(H)$	$\mu(A)$	$t \text{ max}$	ρ
Moore-Laub	1.12	4.08	.630	.290	0
$n = 4$	4.01	12.1	18.9	.239	2
$m = 2$	8.94	26.1	78.9	.145	4
$l = 1$	14.8	41.9	2.0×10^2	.096	6
	20.8	58.3	4.0×10^2	.070	8
	27.0	74.8	6.6×10^2	.054	10
Kailath	19.6	81.4	9.2×10^2	.026	0
$n = 5$	68.8	2.1×10^2	2.9×10^3	.030	2
$m = 2$	1.3×10^2	3.7×10^2	6.2×10^3	.027	4
$l = 3$	1.8×10^2	5.1×10^2	1.0×10^4	.023	6
	2.3×10^2	6.6×10^2	1.5×10^4	.020	8
	2.9×10^2	7.6×10^2	2.0×10^4	.018	10
Chung-Shapiro	2.47	60.7	17.2	1.10	0
$n = 6$	14.4	46.8	2.7×10^2	.087	2
$m = 2$	37.2	1.1×10^2	9.4×10^2	.052	4
$l = 3$	66.5	1.8×10^2	2.1×10^3	.041	6
	99.4	2.7×10^2	3.8×10^3	.033	8
	1.3×10^2	3.6×10^2	5.9×10^3	.029	10
Davison-Maki	18.1	36.1	1.7×10^2	.087	0
$n = 9$	3.2×10^2	1.4×10^3	5.1×10^3	.069	1
$m = 1$	2.1×10^3	9.1×10^3	5.9×10^4	.041	2
$l = 9$	8.0×10^3	2.9×10^4	3.4×10^5	.028	3
	2.5×10^4	8.5×10^4	1.4×10^6	.022	4

As Table 2 indicates, the bound (3.15) is quite reasonable for this example. Moreover it can be shown that for large a

$$\max_{0 \leq t} \|e^{At}\| \simeq \frac{a}{e}, \quad K^{1/2}(H) \simeq a.$$

As a second numerical test we found $\max \|e^{At}\|$, $K^{1/2}(H)$ and $\mu(A)$ for the examples from § 2 for $W = I$. With the single exception of $\rho = 0$ for the Chung-Shapiro problem we see in Table 3 that $K^{1/2}(H)$ is always within a factor of 10 of $\max \|e^{At}\|$. It is of interest to note for these problems that $\max \|e^{At}\|$ and $\mu(A)$ are generally increasing with ρ and that the time $t \text{ max}$ at which the hump in $\|e^{At}\|$ occurs is decreasing with ρ .

4. Sensitivity of the matrix exponential. In [22], Van Loan developed several upper bounds for the exponential sensitivity measure

$$(4.1) \quad \Phi(t) \equiv \frac{\|e^{(A+E)t} - e^{At}\|}{\|e^{At}\|}$$

by using the variation of parameters identity

$$(4.2) \quad e^{(A+E)t} - e^{At} = \int_0^t e^{A(t-s)} E e^{(A+E)s} ds$$

(see [3]). This approach was also adopted by Kagström [16] who sharpened some of the bounds in [22] and developed new bounds based on similarity transformations.

In recognition of the work of these two investigators we shall use the phrase "Kagström-Van Loan sensitivity estimate" to denote any bound on Φ which is obtained from (4.2). In this section we present an estimate of this type as well as a second bound on Φ based on two results of Van Loan from [23] instead of (4.2). This latter bound is given in terms of the commutator $[A, E] = AE - EA$ and reduces to a standard result when $[A, E] = 0$.

Both Van Loan and Kagström used

$$(4.3) \quad \|e^{At}\| \leq K(Y)e^{\mu(B)t}$$

where $A = YBY^{-1}$, in order to obtain bounds based on the Jordan and Schur decompositions of A . In the next theorem we use (4.3) in connection with the Lyapunov equation

$$(4.4) \quad A^*H + HA = -W.$$

THEOREM 4.1. *Let A be stable and let W be positive definite Hermitian. Let H satisfy (4.4). Then for $t \geq 0$*

$$(4.5) \quad \Phi(t) \leq K^{1/2}(H)(e^{\|E\|t} - 1)e^{-\nu t} e^{-\alpha t},$$

where

$$\begin{aligned} K(H) &= \|H\| \|H^{-1}\|, \\ \nu &= \frac{1}{2}\lambda_{\min}(H^{-1/2}WH^{-1/2}), \\ \alpha &= \max_i \operatorname{Re} \lambda_i(A), \end{aligned}$$

and $H^{1/2}$ denotes the unique positive definite Hermitian square root of H .

Proof. Since $A = H^{-1/2}(H^{1/2}AH^{-1/2})H^{1/2}$ we have by (4.3) that

$$(4.6) \quad \|e^{At}\| \leq K(H^{1/2})e^{\mu(H^{1/2}AH^{-1/2})t}$$

for any $t \geq 0$. But $H^{1/2}$ is positive definite Hermitian so

$$(4.7) \quad K(H^{1/2}) = K^{1/2}(H)$$

and

$$\begin{aligned} \mu(H^{1/2}AH^{-1/2}) &= \frac{1}{2}\lambda_{\max}(H^{-1/2}A^*H^{1/2} + H^{1/2}AH^{-1/2}) \\ (4.8) \quad &= \frac{1}{2}\lambda_{\max}(H^{-1/2}(A^*H + HA)H^{-1/2}) \\ &= \frac{1}{2}\lambda_{\max}(-H^{-1/2}WH^{-1/2}) = -\frac{1}{2}\lambda_{\min}(H^{-1/2}WH^{-1/2}). \end{aligned}$$

Using (4.7) and (4.8) in (4.6) we get

$$(4.9) \quad \|e^{At}\| \leq K^{1/2}(H)e^{-\nu t}.$$

Now expand (4.2) as an infinite series of multiple integrals:

$$(4.10) \quad e^{(A+E)t} = e^{At} + \int_0^t e^{A(t-s)}Ee^{As}ds + \int_0^t \int_0^s e^{A(t-s)}Ee^{A(s-s_1)}Ee^{As_1}ds_1ds + \dots$$

Apply (4.9) to (4.10) and integrate to get

$$\begin{aligned} (4.11) \quad \|e^{(A+E)t} - e^{At}\| &\leq K^{1/2}(H)e^{-\nu t} \left(\|E\|t + \|E\|^2 \frac{t^2}{2} + \dots + \|E\|^j \frac{t^j}{j!} + \dots \right) \\ &\leq K^{1/2}(H)e^{-\nu t}(e^{\|E\|t} - 1). \end{aligned}$$

Inequality (4.11) with $\|e^{At}\| \geq e^{\alpha t}$ gives the desired result. \square

Note $H^{-1}W$ is similar to $H^{-1/2}WH^{-1/2}$ (see [29]) so $\lambda_{\min}(H^{-1/2}WH^{-1/2}) = \lambda_{\min}(H^{-1}W)$.

An important aspect of Theorem 4.1 is the freedom involved in choosing W in (4.4). For example, if we set $W = I$ then

$$(4.12) \quad \nu \equiv \frac{1}{2} \lambda_{\min}(H^{-1/2}WH^{-1/2}) = \frac{1}{2} \lambda_{\min}(H^{-1}) = \frac{1}{2\|H\|}$$

so that (4.5) becomes

$$\Phi(t) \leq K^{1/2}(H)(e^{\|E\|t} - 1)e^{-t/2\|H\|} e^{-at}.$$

Moreover from (4.9) we have in this case

$$(4.13) \quad \|e^{At}\| \leq K^{1/2}(H)e^{-t/2\|H\|}$$

which points up the connection between damping and maximal area ($\|H\|$) discussed in § 3.

In general, however, we are faced with the problem of selecting W and it is natural to want to make this choice so as to maximize the damping term ν in (4.9). From $\|e^{At}\| \geq e^{\alpha(A)t}$ and (4.9) we see that no matter how we choose W we cannot make ν greater than $|\alpha(A)|$. Following Pao [25] let $H = H(\rho)$ be the solution to

$$(4.14) \quad (A + \rho I)^*H + H(A + \rho I) = -I$$

for $\rho < |\alpha(A)|$. This may be rewritten in the form of (4.4) if we take $W = I + 2\rho H$, in which case

$$(4.15) \quad \nu = \frac{1}{2} \lambda_{\min}(H^{-1/2}(I + 2\rho H)H^{-1/2}) = \frac{1}{2} \lambda_{\min}(H^{-1} + 2\rho I).$$

Now for A stable and $\rho < |\alpha(A)|$ we have $A + \rho I$ stable so that H satisfying (4.14) is positive definite Hermitian. We may thus bound ν below by ρ from (4.15):

$$(4.16) \quad \rho < \nu = \nu(\rho) < |\alpha(A)|.$$

Inequality (4.16) shows that we may come arbitrarily close to the optimal damping rate $|\alpha(A)|$. In some cases, for example when A is normal, $K^{1/2}(H)$ remains bounded as $\rho \rightarrow |\alpha(A)|$; however if A has a defective eigenstructure associated with eigenvalue λ where $\operatorname{Re} \lambda = \alpha(A)$ then $K^{1/2}(H) \rightarrow \infty$ as $\rho \rightarrow |\alpha(A)|$ (see [26]).

As noted in [22], when A and E commute we may obtain a simple bound on Φ :

$$(4.17) \quad \Phi(t) \leq \|e^{Et} - I\| \leq e^{\|E\|t} - 1.$$

The desire to obtain a general bound on Φ which reduces to (4.17) when A and E commute motivated the following.

THEOREM 4.2. *Let $t \geq 0$ then*

$$(4.18) \quad \Phi(t) \leq \|e^{Et} - I\| + (\cosh(\|[A, E]\|^{1/2}t) - 1)e^{(\mu(A) - \alpha(A))t} e^{\mu(E)t}$$

where $[A, E]$ is the Lie product

$$(4.19) \quad [A, E] = AE - EA$$

and

$$\cosh(x) \equiv \frac{e^x + e^{-x}}{2}.$$

The proof of this theorem relies on two results of Van Loan [23]:

$$\text{Lemma 1 (Van Loan)} \quad e^B e^C = e^{B+C} + \int_0^1 e^{sB} [e^{(1-s)(B+C)}, C] e^{sC} ds;$$

$$\text{Lemma 2 (Van Loan)} \quad \|[e^B, C]\| \leq e^{\mu(B)} \|[B, C]\|.$$

Proof of Theorem 4.2. We show that

$$\|e^{A+E} - e^A\| \leq \|e^A\| \|e^E - 1\| + e^{\mu(A)} e^{\mu(E)} [\cosh(\|[A, E]^{1/2}\|) - 1]$$

which proves the result upon replacing A, E by At, Et and using $\mu(tM) = \mu(M)t$ for $t \geq 0$ with $\|e^{At}\| \leq e^{\alpha(A)t}$.

In Lemma 1 set $B = A + E$ and $C = -A$

then

$$\begin{aligned} e^{A+E} e^{-A} &= e^E + \int_0^1 e^{s(A+E)} [e^{(1-s)E}, -A] e^{-sA} ds \\ &= e^E + \int_0^1 e^{s(A+E)} [A, e^{(1-s)E}] e^{-sA} ds. \end{aligned}$$

Multiply both sides by e^A on the right to get

$$e^{A+E} = e^E e^A + \int_0^1 e^{s(A+E)} [A, e^{(1-s)E}] e^{(1-s)A} ds.$$

Repeated substitution gives

$$\begin{aligned} e^{A+E} &= e^E e^A + \int_0^1 e^{Es} e^{As} [A, e^{(1-s)E}] e^{(1-s)A} ds \\ &\quad + \int_0^1 \int_0^1 e^{s_1 s E} e^{s_1 s A} [sA, e^{(1-s_1) s E}] e^{(1-s_1) s A} \\ &\quad \cdot [A, e^{(1-s)E}] e^{(1-s)A} ds_1 ds + \dots \end{aligned}$$

so that

$$e^{A+E} - e^A = e^E e^A - e^A + \int_0^1 e^{Es} e^{As} [A, e^{(1-s)E}] e^{(1-s)A} ds + \dots$$

Taking norms we have

$$\|e^{A+E} - e^A\| \leq \|e^E - I\| \|e^A\| + \int_0^1 \|e^{Es}\| \|e^{As}\| \|[A, e^{(1-s)E}]\| \|e^{(1-s)A}\| ds + \dots$$

Now use $\|e^{Mr}\| \leq e^{\mu(M)r}$ for $r \geq 0$ and Lemma 2 of Van Loan to get

$$\begin{aligned} \|e^{A+E} - e^A\| &\leq \|e^E - I\| \|e^A\| + e^{\mu(E)} e^{\mu(A)} \|[A, E]\| \int_0^1 (1-s) ds \\ &\quad + e^{\mu(E)} e^{\mu(A)} \|[A, E]\|^2 \int_0^1 (1-s)s^2 ds \int_0^1 (1-s_1) ds_1 + \dots \\ &= \|e^E - I\| \|e^A\| + e^{\mu(E)} e^{\mu(A)} \|[A, E]\| \frac{1}{2} + \|[A, E]\|^2 \frac{1}{24} \\ &\quad + \dots + \|[A, E]\|^n \frac{1}{(2n)!} + \dots \\ &= \|e^E - I\| \|e^A\| + e^{\mu(E)} e^{\mu(A)} \{\cosh(\|[A, E]^{1/2}\|) - 1\}. \end{aligned}$$

□

When A and E commute (4.18) reduces to (4.17) because

$$[A, E] = AE - EA = 0.$$

Conclusion. An analysis of the sensitivity of the solution to the Lyapunov equation $A^*X + XA = -W$ has been presented. This analysis leads to a spectral norm bound on the relative perturbation of the solution which is essentially equivalent to the Frobenius norm bound obtained from the associated Kronecker product system. The latter bound can be expressed in terms of $\text{sep}(A^*, -A)$ and is known to accurately reflect the sensitivity of the Lyapunov problem, but it is hard to interpret in terms of the original matrix A . In contrast, the spectral norm bound which we have obtained is directly related to the minimal L_2 damping of the dynamical system $\dot{z} = Az$. Moreover, this dynamical link with the sensitivity problem leads to a new method of systematically investigating the norm behavior of e^{At} as well as providing a wealth of information about control theoretic aspects of $\dot{z} = Az$. In a future paper, we show that the dynamical approach also works well in analyzing the sensitivity of the Riccati equation.

REFERENCES

- [1] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
- [2] R. W. BASS, *Robustified LQG synthesis to specifications*, in Fifth Meeting Coordinating Group on Modern Control Theory, U.S. Army Research and Development Command, Picatinny Arsenal, Dover, NJ, H. Cohen, ed., 1983.
- [3] R. BELLMAN, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1969.
- [4] R. BYERS, *A LINPACK-style condition estimator for the equation $AX - XB^T = C$* , IEEE Trans. Automat. Control, AC-29 (1984), pp. 926-928.
- [5] A. CLINE, C. MOLER, G. STEWART AND J. WILKINSON, *An estimate for the condition number of a matrix*, SIAM J. Numer. Anal., 16 (1979), pp. 368-375.
- [6] J. C. CHUNG AND E. Y. SHAPIRO, *Constrained eigenvalue/eigenvector assignment—application to flight control systems*, Proc. 1982 Conference on Information and Systems, Dept. of Electrical Engrg. and Comput. Sci., Princeton Univ., Princeton, NJ.
- [7] E. J. DAVISON AND M. C. MAKI, *The numerical solution of the matrix Riccati differential equation*, IEEE Trans. Automat. Control AC-18 (1973), pp. 71-73.
- [8] R. C. DIPRIMA AND C. R. JOHNSON, *The range of $A^{-1}A^*$ in $GL(N, C)$* , Linear Algebra Appl., 9 (1974), pp. 209-222.
- [9] J. J. DONGARRA, J. R. BUNCH, C. B. MOLER AND B. W. STEWART, *LINPACK User's Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.
- [10] G. GOLUB, S. NASH AND C. VAN LOAN, *A Hessenberg-Schur method for the problem $AX + XB = C$* , IEEE Trans. Automat. Control, 24 (1979), pp. 909-913.
- [11] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.
- [12] A. GRAHAM, *Kronecker Products and Matrix Calculus with Applications*, John Wiley, New York, 1981.
- [13] S. J. HAMMARLING, *Numerical solution of the stable, non-negative definite Lyapunov equation*, IMA J. Numer. Anal., 2 (1982), pp. 303-323.
- [14] E. JONCKHEERE, *New bound on the sensitivity of the solution of the Lyapunov equation*, Linear Algebra Appl., 60 (1984), pp. 57-64.
- [15] ———, *Principal component analysis of flexible systems-open-loop case*, Presented at Symposium on the Mathematical Theory of Networks and Systems. Beer Sheva, Israel, 1983.
- [16] B. KAGSTRÖM, *Bounds and perturbation bounds for the matrix exponential*, BIT, 17 (1977), pp. 39-57.
- [17] T. KAILATH, *Some new algorithms for recursive estimation in constant linear systems*, IEEE Trans. Inform. Theory, IT-19 (1973), pp. 750-760.
- [18] R. E. KALMAN AND J. E. BERTRAM, *Control system design via the "second method" of Lyapunov*, Trans. ASME J. Basic Engrg., 2 (1960), pp. 371-393.
- [19] D. G. LAINIOTIS, *Partitioned Riccati solutions and integration-free doubling algorithms*, IEEE Trans. Automat. Control, AC-20 (1976), pp. 677-688.
- [20] P. LANCASTER, *Explicit solutions of linear matrix equations*, SIAM Rev., 12 (1970), pp. 544-566.
- [21] A. J. LAUB, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 913-921.
- [22] C. VAN LOAN, *The sensitivity of the matrix exponential*, SIAM J. Numer. Anal., 14 (1977), pp. 971-981.

- [23] C. MOLER AND C. VAN LOAN, *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Rev., 20 (1978), pp. 801–836.
- [24] B. MOORE AND A. LAUB, *Computation of supremal (A, B) -invariant and controllability subspaces*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 783–792.
- [25] C. V. PAO, *Logarithmic derivatives of a square matrix*, Linear Algebra Appl., 6 (1973), pp. 159–164.
- [26] ———, *A further remark on the logarithmic derivatives of a square matrix*, Linear Algebra Appl., 7 (1973), pp. 275–278.
- [27] G. STEWART, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.
- [28] J. VARAH, *On the separation of two matrices*, SIAM J. Numer. Anal., 16 (1979), pp. 216–222.
- [29] H. WIMMER, *Generalizations of theorems of Lyapunov and Stein*, Linear Algebra Appl., 10 (1975), pp. 139–146.

A PARALLEL ALGORITHM FOR A CLASS OF CONVEX PROGRAMS*

SHIH-PING HAN† AND GANG LOU†

Abstract. A parallel algorithm is proposed in this paper for solving the problem $\min \{q(x) | x \in C_1 \cap \cdots \cap C_m\}$ where q is a uniformly convex function and C_i are closed convex sets in R^n . In each iteration of the method, we solve in parallel m independent subproblems, each minimizing a definite quadratic function over an individual set C_i . The method has attractive convergence properties and can be implemented as parallel algorithms for tackling definite quadratic programs, linear programs, systems of linear equations and systems of generalized nonlinear inequalities.

Key words. convex program, parallel algorithm

AMS(MOS) subject classifications. 65K05, 90C25, 90C30

1. Introduction. In this paper we study a method for solving the optimization problem:

$$(P) \quad \begin{array}{ll} \text{minimize} & q(x) \\ \text{subject to} & x \in C = C_1 \cap \cdots \cap C_m \end{array}$$

where q is uniformly convex and differentiable on R^n and C_i are closed convex sets. In particular, the method can be used for tackling definite quadratic programs and, therefore, can be incorporated into a sequential quadratic programming algorithm for solving some more general problems (see, for example, [1]–[3], [6]).

The method is an iterative process. The main computation in each iteration is to solve m subproblems of the following form:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2} \|z - w_i\|^2 \\ \text{subject to} & z \in C_i \end{array}$$

with vector w_i varying iteratively for each set C_i . The key feature here is that the m subproblems are independent of each other and can be solved simultaneously and, therefore, the method is suitable for parallel computation. As its special applications, the method can be implemented as parallel algorithms for solving systems of linear equations, linear programming problems and systems of nonlinear inequalities.

The method is actually intended to solve some dual problems of (P) and is closely related to the successive projection method presented in [4], [5]. Not as in the successive projection method, the subproblems here can be solved independently rather than one after another. Moreover, the objective function q in (P) is more general than the one treated in [4], [5].

We present the method in § 2. In § 3 we study the dual problems that our method is actually dealing with. Convergence analysis is given in § 4. Some applications of the method to systems of linear equations, linear programming and systems of generalized nonlinear inequalities are given in § 5. In § 6 we also present a modified version of the method which can avoid some difficulties that may occur in choosing a parameter of the method.

Results in Rockafellar's book [10] will be frequently referred to and most of the symbols and notation used here are also the same as in that book. Recall that the

* Received by the editors October 27, 1986; accepted for publication May 14, 1987. This work was supported in part by the National Science Foundation under grant DMS-8602419.

† Department of Mathematics, University of Illinois, Urbana, Illinois 61801.

indicator function of a set C is denoted by $\delta(\cdot|C)$ and defined as: $\delta(x|C) = 0$ if $x \in C$; $\delta(x|C) = \infty$ otherwise. Its conjugate $\delta^*(\cdot|C)$ is the support function of the set C as given by $\delta^*(y|C) = \sup \{\langle x, y \rangle | x \in C\}$. We also use $\text{ri } C$ to denote the relative interior of the set C . Unless specified otherwise, the symbol \sum is to denote the summation from 1 to m , where m is the number of the sets C_i in problem (P).

2. The method. For the proposed method to work properly we require the objective function q in (P) to be smooth and uniformly convex; that is, the function q is finite and differentiable everywhere on R^n and also there exists a positive number ρ such that for any x, y in R^n and for any λ in $(0, 1)$

$$(2.1) \quad \lambda q(x) + (1 - \lambda)q(y) \geq q(\lambda x + (1 - \lambda)y) + \rho\lambda(1 - \lambda)\|x - y\|^2.$$

Under these assumptions, the function q is, of course, strictly convex. Moreover, it is also co-finite. This can be seen easily from the definition of cofiniteness (see [10, p. 116]) or by checking the value of the recession function q^{0+} at any nonzero point

$$\begin{aligned} q^{0+}(y) &= \sup \{(1/\lambda)(q(\lambda y) - q(0)) | \lambda > 0\} \\ &\geq \sup \{\rho(\lambda - 1)\|y\|^2 + q(y) - q(0) | \lambda > 0\} \\ &= \infty. \end{aligned}$$

Therefore, as a consequence of Theorem 26.6 of [10], the conjugate q^* of q is also smooth, strictly convex and co-finite. The method can now be stated as follows.

The method. Let α be a sufficiently large number and let $y^{(0)} = y_1^{(0)} = \cdots = y_m^{(0)} = 0$ and $x^{(0)} = \nabla q^*(y^{(0)})$. For $k = 1, 2, \dots$, we do the following computation:

(a) For $i = 1, \dots, m$, find $z_i^{(k)}$ which solves

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|z + \alpha y_i^{(k-1)} - x^{(k-1)}\|^2 \\ &\text{subject to} \quad z \in C_i; \end{aligned}$$

(b) Set $y_i^{(k)} = y_i^{(k-1)} + (1/\alpha)(z_i^{(k)} - x^{(k-1)})$;

(c) Set $y^{(k)} = y_1^{(k)} + \cdots + y_m^{(k)}$;

(d) $x^{(k)} = \nabla q^*(y^{(k)})$.

Remarks. (1) The bulk of the computation is usually in step (a) where the m subproblems are independent to each other and can be solved in parallel. The efficiency of the method certainly depends on how effectively we can solve the subproblems. When the sets C_i are subspaces or half-spaces, the subproblems are relatively easy to handle as will be seen in the later examples.

(2) The computation of $x^{(k)}$ in step (d) is an evaluation of a gradient of the conjugate function q^* . When an explicit form of the conjugate is known, the calculation is straightforward. For instance, if the function q is definite quadratic as $q(x) = \frac{1}{2}\langle x, Qx \rangle + \langle c, x \rangle$ then we have $x^{(k)} = Q^{-1}(y^{(k)} - c)$. For a more general function whose conjugate is not readily available, we need to do an unconstrained minimization to find $x^{(k)}$ as follows:

$$x^{(k)} = \arg \min q(x) - \langle x, y^{(k)} \rangle.$$

(3) The number α should be sufficiently large. Theoretically, the method will work properly if α is chosen to be larger than m/ρ where ρ is a uniform convexity constant satisfying (2.1). In particular, when q is definite quadratic then ρ can be chosen as one half of the smallest eigenvalue of the underlying matrix. The parameter α need not be a fixed constant. When an uniform convexity constant ρ is not known in advance,

the parameter α can be adjusted and updated iteratively. A more detailed discussion on how to update the parameter is given in § 6.

To illustrate the usefulness of the method we consider some special cases.

(I). *Linearly constrained problem*:

$$\begin{aligned} & \text{minimize} && q(x) \\ & \text{subject to} && \langle a_i, x \rangle \leq \beta_i, \quad i = 1, \dots, j, \\ & && \langle a_i, x \rangle = \beta_i, \quad i = j+1, \dots, m. \end{aligned}$$

For this case the method becomes the following. Let $y^{(0)} = y_1^{(0)} + \dots + y_m^{(0)} = 0$ and $x^{(0)} = \nabla q^*(y^{(0)})$. For $k = 1, 2, \dots$, we compute:

$$(a) \quad y_i^{(k)} = -(\gamma_i / \alpha \langle a_i, a_i \rangle) a_i, \quad (i = 1, \dots, m)$$

where $\gamma_i = \max\{\beta_i - \langle a_i, x^{(k-1)} - \alpha y_i^{(k-1)} \rangle, 0\}$, $(i = 1, \dots, j)$; $\gamma_i = \beta_i - \langle a_i, x^{(k-1)} - \alpha y_i^{(k-1)} \rangle$, $(i = j+1, \dots, m)$;

$$(b) \quad y^{(k)} = y_1^{(k)} + \dots + y_m^{(k)};$$

$$(c) \quad x^{(k)} = \nabla q^*(y^{(k)}).$$

(II). *Projection problem* (cf. [4], [5]):

$$\begin{aligned} & \text{minimize} && \frac{1}{2} \langle x - d, Q(x - d) \rangle \\ & \text{subject to} && x \in C_1 \cap \dots \cap C_m. \end{aligned}$$

For this case the method can be implemented in the following way. Let $y^{(0)} = y_1^{(0)} + \dots + y_m^{(0)} = 0$ and $x^{(0)} = d$. For $k = 1, 2, \dots$, we carry out the following computation:

(a) for $i = 1, \dots, m$, find $z_i^{(k)}$ which solves

$$\begin{aligned} & \text{minimize} && \frac{1}{2} \|z + \alpha y_i^{(k-1)} - x^{(k-1)}\|^2 \\ & \text{subject to} && z \in C_i \end{aligned}$$

and set $y_i^{(k)} = y_i^{(k-1)} + (1/\alpha)(z_i^{(k)} - x^{(k-1)})$;

$$(b) \quad y^{(k)} = y_1^{(k)} + \dots + y_m^{(k)},$$

$$(c) \quad x^{(k)} = Q^{-1}y^{(k)} + d.$$

(III). *Definite quadratic programming*:

$$\begin{aligned} & \text{minimize} && \frac{1}{2} \langle x - d, Q(x - d) \rangle \\ & \text{subject to} && \langle a_i, x \rangle \leq \beta_i, \quad i = 1, \dots, j, \\ & && \langle a_i, x \rangle = \beta_i, \quad i = j+1, \dots, m. \end{aligned}$$

This is, of course, a special case of both (I) and (II). As in (I), because each y_i is a multiple of the corresponding vector a_i , we can represent vectors y_i by scalars. In so doing, the calculation of vectors reduces to that of scalars and the method becomes extremely simple.

Initially, we compute $p_i = (1/\alpha \langle a_i, a_i \rangle) Q^{-1} a_i$, $(i = 1, \dots, m)$ and $x^{(0)} = d$. Let $\gamma_i^{(0)} = 0$, $(i = 1, \dots, m)$, and $x^{(0)} = -d$. For $k = 1, 2, \dots$, we compute:

$$(a) \quad \gamma_i^{(k)} = \max\{\beta_i + \gamma_i^{(k-1)} - \langle a_i, x^{(k-1)} \rangle, 0\}, \quad (i = 1, \dots, j); \quad \gamma_i^{(k)} = \beta_i + \gamma_i^{(k-1)} - \langle a_i, x^{(k-1)} \rangle, \quad (i = j+1, \dots, m);$$

$$(b) \quad x^{(k)} = \sum \gamma_i^{(k)} p_i + d.$$

3. Dual problems. For analyzing the convergence of the method we study in this section some duality problems which the method actually solves. Without too much extra effort we can present the duality results for the slightly more general problem:

$$(P) \quad \min q(x) + f(x)$$

where $f = f_1 + \cdots + f_m$ and q, f_1, \cdots, f_m are all convex functions. The two duality problems of (P) to be considered are:

$$(\mathbb{D}) \quad \min g(y) := q^*(y) + f^*(-y),$$

$$(\mathbb{D}') \quad \min h(y_1, \cdots, y_m) := q^*(y_1 + \cdots + y_m) + \sum f_i^*(-y_i).$$

Clearly, problem (P) is a special case of (P) with functions f_i being the indicator functions of the sets C_i . The two corresponding duality problems of (P) are, respectively, as follows:

$$(\mathbb{D}) \quad \min g(y) := q^*(y) + \delta^*(y|C),$$

$$(\mathbb{D}') \quad \min h(y_1, \cdots, y_m) := q^*(y_1 + \cdots + y_m) + \sum \delta^*(y_i|C_i).$$

It is noted that the duality problems (D) and (D) are essentially the same as those considered in Fenchel's duality theory [10]. Problems (D') and (D') simply result from applying infimal convolution operation to (D) and (D), respectively.

The basic assumption for the following duality results to hold is that the function q is strictly convex, co-finite and differentiable everywhere on R^n . In view of Theorem 26.6 of [10], under this assumption, the conjugate q^* of q is also a function of this type and $\nabla q^* = (\nabla q)^{-1}$.

We first give a result on the connection between Problems (P) and (D).

THEOREM 3.1. *Let q be strictly convex, co-finite and differentiable on R^n and let f be proper closed convex. A vector \bar{x} solves (P) if and only if $\nabla q(\bar{x})$ solves (D). Dually, a vector \bar{y} solves (D) if and only if $\nabla q^*(\bar{y})$ solves (P). Furthermore, such \bar{x} and \bar{y} exist and are unique.*

Proof. \bar{x} solves (P) $\Leftrightarrow 0 \in \nabla q(\bar{x}) + \partial f(\bar{x})$

$$\Leftrightarrow 0 \in -\bar{x} + \partial f^*(-\nabla q(\bar{x}))$$

$$\Leftrightarrow 0 \in \nabla q^*(\nabla q(\bar{x})) - \partial f^*(-\nabla q(\bar{x}))$$

$$\Leftrightarrow \nabla q(\bar{x}) \text{ solves (D).}$$

The second statement follows from a similar argument. In view of $\text{dom } q = \text{dom } q^* = R^n$, the existence of solutions (P) and (D) is a result of Fenchel's duality theorem [10, Thm. 31.1]. The uniqueness of solutions of (P) and (D) follows from the strict convexity. Q.E.D.

The following theorem explores the relationship between (P) and (D').

THEOREM 3.2. *Let q be strictly convex, co-finite and differentiable on R^n and let f be proper closed convex. A vector $(\bar{y}_1, \cdots, \bar{y}_m)$ solves (D') if and only if there exists a vector \bar{x} such that*

$$(a) \quad \bar{x} = \nabla q^*(\sum \bar{y}_i);$$

$$(b) \quad -\bar{y}_i \in \partial f_i(\bar{x}), (i = 1, \cdots, m).$$

Furthermore, any vector \bar{x} satisfying (a) and (b) solves (P).

$$\text{Proof. } (\bar{y}_1, \cdots, \bar{y}_m) \text{ solves (D')} \Leftrightarrow 0 \in \nabla q^*(\sum \bar{y}_i) - \partial f_i^*(-\bar{y}_i), (i = 1, \cdots, m)$$

$$\Leftrightarrow \bar{x} \in \partial f_i^*(-\bar{y}_i), (i = 1, \cdots, m), \text{ and } \bar{x} = \nabla q^*(\sum \bar{y}_i)$$

$$\Leftrightarrow (a) \text{ and } (b).$$

If a vector \bar{x} satisfies (a) and (b) then $\sum -\bar{y}_i \in \sum \partial f_i(\bar{x}) \subset \partial f(\bar{x})$, which implies $0 \in \nabla q(\bar{x}) + \partial f(\bar{x})$. Consequently, \bar{x} solves (P). Q.E.D.

The theorem below concerns how problem (D) is related to problem (D').

THEOREM 3.3. *Let q be strictly convex, co-finite and differentiable on R^n and let f be proper closed convex, then*

$$\bar{y} \text{ solves } (\mathbb{D}) \Leftrightarrow \exists (\bar{y}_1, \dots, \bar{y}_m), (\bar{y}_1, \dots, \bar{y}_m) \text{ solves } (\mathbb{D}') \text{ and } \bar{y} = \sum \bar{y}_i.$$

Moreover, if f_1, \dots, f_p are polyhedral and f_{p+1}, \dots, f_m are closed convex such that $\text{dom } f_1 \cap \dots \cap \text{dom } f_p \cap \text{ri}(\text{dom } f_{p+1}) \cap \dots \cap \text{ri}(\text{dom } f_m) \neq \emptyset$ then

$$\inf g(y) = \inf h(y_1, \dots, y_m).$$

Proof. The necessity of the first part of the theorem is a direct consequence of Theorems 3.1 and 3.2. To prove the sufficiency we note that if \bar{y} solves (\mathbb{D}) then, by Theorem 3.1, the vector $\bar{x} := \nabla q^*(\bar{y})$ solves (\mathbb{P}) . Hence,

$$-\bar{y} = -\nabla q(\bar{x}) \in \partial f(\bar{x}) = \sum \partial f_i(\bar{x}).$$

Therefore, it follows that there exists $(\bar{y}_1, \dots, \bar{y}_m)$ such that $\bar{y} = \sum \bar{y}_i$ and $-\bar{y}_i \in \partial f_i(\bar{x})$, $(i = 1, \dots, m)$. Then, the result follows immediately from Theorem 3.2.

To prove the second part of the theorem, we note that $\inf h \geq \inf g$. When $\text{dom } f_1 \cap \dots \cap \text{dom } f_p \cap \text{ri}(\text{dom } f_{p+1}) \cap \dots \cap \text{ri}(\text{dom } f_m) \neq \emptyset$, then for any y there exist y_1, \dots, y_m such that $y_1 + \dots + y_m = y$ and $f^*(-y) = (f_1^* \square \dots \square f_m^*)(-y) = f_1^*(-y_1) + \dots + f_m^*(-y_m)$. Therefore, it can never occur that $\inf h > \inf g$ and, consequently, we have $\inf g(y) = \inf h(y_1, \dots, y_m)$. Q.E.D.

4. Convergence analysis. In this section we study the convergence of the method for solving problem (P). For the results of § 3 to be applicable here and for the method to work properly we assume throughout this section:

- (a) q is uniformly convex and differentiable on R^n ;
- (b) C_i are closed convex and $C = C_1 \cap \dots \cap C_m \neq \emptyset$.

Assumption (a) above implies that q is strictly convex and co-finite and, in conjunction with assumption (b), makes Theorem 3.1 applicable. Therefore, problems (P) and (D) have unique solutions, which will be denoted by \bar{x} and \bar{y} , respectively.

In the subsequent discussion we will also assume that the parameter α is chosen to be larger than m/ρ where ρ is an uniform convexity constant of q that satisfies condition (2.1).

For establishing our convergence theorems, we first give some useful lemmas below.

LEMMA 4.1. *The vector $y_i^{(k)}$ solves the problem*

$$(4.1) \quad \min (\alpha/2) \|y_i^{(k-1)} - y\|^2 + \langle x^{(k-1)}, y \rangle + \delta^*(y|C_i).$$

Proof. In the method, the vector $z_i^{(k)}$ solves

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} \|z + \alpha y_i^{(k-1)} - x^{(k-1)}\|^2 \\ & \text{subject to} \quad z \in C_i. \end{aligned}$$

Let $\varphi(z) = (1/2\alpha) \|z + \alpha y_i^{(k-1)} - x^{(k-1)}\|^2$. Therefore, the vector $z_i^{(k)}$ solves the problem:

$$\min \varphi(z) + \delta(z|C_i).$$

By applying Theorem 3.1 to the above problem, we have that the vector $y_i^{(k)} = y_i^{(k-1)} + (1/\alpha)(z_i^{(k)} - x^{(k-1)}) = \nabla \varphi(z_i^{(k)})$ is the solution of

$$\min \varphi^*(y) + \delta^*(y|C_i)$$

which is equivalent to (4.1). Q.E.D.

LEMMA 4.2. $\langle x^{(k)} - x^{(k-1)}, y^{(k)} - y^{(k-1)} \rangle \leq (m/2\rho) \sum \|y_i^{(k-1)} - y_i^{(k)}\|^2$.

Proof. By the uniform convexity of q , we have that for any two vectors u and v ,

$$\langle u - v, \nabla q(u) - \nabla q(v) \rangle \geq 2\rho \|u - v\|^2 \quad (\text{see, for example, [8, p. 86]})$$

which yields

$$\|y^{(k-1)} - y^{(k)}\| = \|\nabla q(x^{(k-1)}) - \nabla q(x^{(k)})\| \geq 2\rho \|x^{(k-1)} - x^{(k)}\|.$$

Therefore, it follows that

$$\begin{aligned} \langle x^{(k)} - x^{(k-1)}, y^{(k)} - y^{(k-1)} \rangle &\leq \|x^{(k-1)} - x^{(k)}\| \|y^{(k-1)} - y^{(k)}\| \\ &\leq (1/2\rho) \|y^{(k-1)} - y^{(k)}\|^2 \\ &\leq (1/2\rho) (\sum \|y_i^{(k-1)} - y_i^{(k)}\|)^2 \\ &\leq (1/2\rho) m (\sum \|y_i^{(k-1)} - y_i^{(k)}\|^2), \end{aligned}$$

where the last inequality is a result of the Cauchy-Schwarz inequality. Q.E.D.

LEMMA 4.3. Let h be defined as in (\mathbb{D}') , then

$$h(y_1^{(k-1)}, \dots, y_m^{(k-1)}) \geq h(y_1^{(k)}, \dots, y_m^{(k)}) + (\alpha\rho - m)/(2\rho) \sum \|y_i^{(k-1)} - y_i^{(k)}\|^2.$$

Proof. By convexity we have that

$$q^*(y^{(k-1)}) \geq q^*(y^{(k)}) + \langle \nabla q^*(y^{(k)}), y^{(k-1)} - y^{(k)} \rangle,$$

which, in conjunction with $x^{(k)} = \nabla q^*(y^{(k)})$, implies that

$$\begin{aligned} h(y_1^{(k-1)}, \dots, y_m^{(k-1)}) &\geq h(y_1^{(k)}, \dots, y_m^{(k)}) + \langle x^{(k)}, y^{(k-1)} - y^{(k)} \rangle \\ &\quad + \sum (\delta^*(y_i^{(k-1)} | -C_i) - \delta^*(y_i^{(k)} | -C_i)). \end{aligned}$$

Because $y_i^{(k)}$ solves (4.1), we get

$$\delta^*(y_i^{(k-1)} | -C_i) - \delta^*(y_i^{(k)} | -C_i) \geq \langle x^{(k-1)}, y_i^{(k)} - y_i^{(k-1)} \rangle + (\alpha/2) \|y_i^{(k-1)} - y_i^{(k)}\|^2.$$

Thus, combining the two inequalities above, we have

$$\begin{aligned} h(y_1^{(k-1)}, \dots, y_m^{(k-1)}) &\geq h(y_1^{(k)}, \dots, y_m^{(k)}) + \langle x^{(k)} - x^{(k-1)}, y^{(k-1)} - y^{(k)} \rangle \\ &\quad + (\alpha/2) \sum \|y_i^{(k-1)} - y_i^{(k)}\|^2. \end{aligned}$$

The desired result follows from the above inequality and Lemma 4.2. Q.E.D.

COROLLARY 4.4. (a) For $i = 1, \dots, m$, $\|y_i^{(k-1)} - y_i^{(k)}\| \rightarrow 0$, as $k \rightarrow \infty$;

(b) $\|y^{(k-1)} - y^{(k)}\| \rightarrow 0$, as $k \rightarrow \infty$;

(c) $\|x^{(k-1)} - x^{(k)}\| \rightarrow 0$, as $k \rightarrow \infty$.

Proof. By adding the inequalities in Lemma 4.3 from $k = 0$ to p , we have that

$$h(y_1^{(0)}, \dots, y_m^{(0)}) - h(y_1^{(p)}, \dots, y_m^{(p)}) \geq \omega \sum_{j=1}^p (\sum \|y_i^{(j-1)} - y_i^{(j)}\|^2)$$

where $\omega = (\alpha\rho - m)/(2\rho) > 0$. Because $h(y_1^{(p)}, \dots, y_m^{(p)}) \geq \inf h \geq g(\bar{y}) > -\infty$, the left-hand side of the inequality is bounded above by a finite number which is independent of p . This implies (a) and (b). Statement (c) follows from (b) and $\|y^{(k-1)} - y^{(k)}\| = \|\nabla q(x^{(k-1)}) - \nabla q(x^{(k)})\| \geq 2\rho \|x^{(k-1)} - x^{(k)}\|$. Q.E.D.

LEMMA 4.5. The sequences $\{x^{(k)}\}$ and $\{y^{(k)}\}$ are bounded.

Proof. Since $g(y^{(k)}) \leq h(y_1^{(k)}, \dots, y_m^{(k)}) \leq h(y_1^{(0)}, \dots, y_m^{(0)})$, we have that the sequence $\{y^{(k)}\}$ is in a lower level set of g . The level set is bounded because the infimum of g is attained at one point, namely \bar{y} .

With $\{y^{(k)}\} = \{\nabla q(x^{(k)})\}$ bounded, the boundedness of the sequence $\{x^{(k)}\}$ is a direct consequence of the co-finiteness of q (see [10, Lemma 26.7]). It is also evident from the following observation:

$$\begin{aligned}\|x^{(k)}\| &\leq (1/2\rho)\|\nabla q(x^{(k)}) - \nabla q(0)\| \\ &\leq (1/2\rho)(\|y^{(k)}\| + \|\nabla q(0)\|).\end{aligned}\quad \text{Q.E.D.}$$

We now give a convergence theorem of the method for the case that the sets C_i are polyhedral convex.

THEOREM 4.6. *If C_i are polyhedral convex and $\cap C_i \neq \emptyset$, then $\{x^{(k)}\}$ and $\{y^{(k)}\}$ converge to the solution \bar{x} of (P) and the solution \bar{y} of (D), respectively.*

Proof. Since $y_i^{(k)}$ solves (4.1), we have that

$$0 \in x^{(k-1)} + \alpha(y_i^{(k)} - y_i^{(k-1)}) + \partial\delta^*(y_i^{(k)}|C_i)$$

which implies that

$$(4.2) \quad -y_i^{(k)} \in \partial\delta(x^{(k)} + w_i^{(k)}|C_i)$$

where $w_i^{(k)} = x^{(k-1)} - x^{(k)} + \alpha(y_i^{(k)} - y_i^{(k-1)})$. Let x^* be an accumulation point of $\{x^{(k)}\}$ and assume $x^{(k_j)} \rightarrow x^*$. In view of Corollary 4.4, we have $w_i^{(k)} \rightarrow 0$ and, thus, $x^{(k_j)} + w_i^{(k_j)} \rightarrow x^*$. By the polyhedrality of C_i , it follows that for all sufficiently large k_j ,

$$-y_i^{(k_j)} \in \partial\delta(x^{(k_j)} + w_i^{(k_j)}|C_i) \subset \partial\delta(x^*|C_i).$$

Therefore, we have that for all sufficiently large k_j

$$-y^{(k_j)} \in \sum \partial\delta(x^*|C_i) \subset \partial\delta(x^*|C).$$

Hence, from $y^{(k_j)} = \nabla q(x^{(k_j)}) \rightarrow \nabla q(x^*)$ it follows that $-\nabla q(x^*) \in \partial\delta(x^*|C)$, which implies that x^* is the unique solution \bar{x} of (P). Therefore, no point other than \bar{x} can be an accumulation point of the sequence $\{x^{(k)}\}$. Then it follows from the boundedness of the sequence that $\{x^{(k)}\}$ converges to the solution \bar{x} of (P). Meanwhile, we have that $y^{(k)} = \nabla q(x^{(k)}) \rightarrow \nabla q(\bar{x}) = \bar{y}$. Q.E.D.

We now consider the case that C_i are general convex sets. For this situation the proofs of our convergence theorems are based on the existence of an accumulation point of the sequence $(y_1^{(k)}, \dots, y_m^{(k)})$.

PROPOSITION 4.7. *Let $\cap \text{ri } C_i \neq \emptyset$. Then any accumulation point of the sequence $\{(y_1^{(k)}, \dots, y_m^{(k)})\}$ is a solution of (D'). Furthermore, if such an accumulation point exists then the sequences $\{x^{(k)}\}$ and $\{y^{(k)}\}$ converge to the solutions of (P) and (D), respectively.*

Proof. Let $(y_1^{(k_j)}, \dots, y_m^{(k_j)}) \rightarrow (y_1^*, \dots, y_m^*)$. Since $\{x^{(k)}\}$ is bounded, by passing to a further subsequence, if necessary, we can find a point x^* such that $x^{(k_j)} \rightarrow x^*$. In view of (4.2) and $w_i^{(k)} \rightarrow 0$, we have that for $i = 1, \dots, m$,

$$-y_i^* \in \partial\delta(x^*|C_i).$$

On the other hand, it follows from $x^{(k_j)} = \nabla q^*(\sum y_i^{(k_j)})$ that $x^* = \nabla q^*(\sum y_i^*)$. Therefore, by Theorem 3.2, (y_1^*, \dots, y_m^*) solves (D').

To prove the second statement, we note that $h(y_1^{(k)}, \dots, y_m^{(k)}) \geq g(y^{(k)})$ and, by the monotone decrease of $\{h(y_1^{(k)}, \dots, y_m^{(k)})\}$ and by the first part of the proof,

$$\lim h(y_1^{(k)}, \dots, y_m^{(k)}) = \inf h = \inf g.$$

Thus, we have $\lim g(y^{(k)}) = \inf g$. Therefore, $\{y^{(k)}\}$ converges to \bar{y} because the infimum of g is attained at the unique point \bar{y} . At the same time, we have $x^{(k)} = \nabla q^*(y^{(k)}) \rightarrow \bar{x} = \nabla q^*(\bar{y})$. Q.E.D.

In view of Proposition 4.7, to show the convergence of $\{x^{(k)}\}$ to the solution of (P) we need to study when the sequence $\{(y_1^{(k)}, \dots, y_m^{(k)})\}$ has an accumulation point or, more conveniently, to consider when the sequence is bounded.

LEMMA 4.8. *If $\text{int}(C) \neq \emptyset$ then the sequence $\{(y_1^{(k)}, \dots, y_m^{(k)})\}$ is bounded.*

Proof. Let $z \in \text{int}(C)$. Then

$$\begin{aligned} q(z) + q^*(0) &= q(z) + h(y_1^{(0)}, \dots, y_m^{(0)}) \\ &\cong q(z) + h(y_1^{(k)}, \dots, y_m^{(k)}) \\ &\cong q(z) + q^*(y_1^{(k)} + \dots + y_m^{(k)}) + \sum \delta^*(y_i^{(k)} | -C_i) \\ &\cong \langle z, y_1^{(k)} + \dots + y_m^{(k)} \rangle + \sum \delta^*(y_i^{(k)} | -C_i) \\ &\cong \sum \delta^*(y_i^{(k)} | z - C_i). \end{aligned}$$

Therefore, it follows from $z \in C$ that $\delta^*(y_i^{(k)} | z - C_i) \geq 0$ and, hence, $y_i^{(k)}$ is in the level set $\{y | \delta^*(y | z - C_i) \leq q(z) + q^*(0)\}$, which is bounded because

$$0 \in z - \text{int } C_i = \text{int}(\text{dom } \delta(\cdot | z - C_i)). \quad \text{Q.E.D.}$$

THEOREM 4.9. *If $\text{int}(C) \neq \emptyset$, then $\{x^{(k)}\}$ and $\{y^{(k)}\}$ converge to the unique solutions of (P) and (D), respectively.*

From a practical viewpoint it is important to study when the sequence $\{(y_1^{(k)}, \dots, y_m^{(k)})\}$ converges because the vectors $y_1^{(k)}, \dots, y_m^{(k)}$ are really computed in practice. For this reason we conclude this section with a result on this matter.

THEOREM 4.10. *Let C_1, \dots, C_p be polyhedral convex such that $C_1 \cap \dots \cap C_p \cap \text{ri } C_{p+1} \cap \dots \cap \text{ri } C_m \neq \emptyset$. If $\text{aff}(\partial\delta(\bar{x} | C_i))$ are linearly independent in the sense:*

$$\sum z_i = 0 \quad \text{and} \quad z_i \in \text{aff}(\partial\delta(\bar{x} | C_i)) \quad (i = 1, \dots, m) \Rightarrow z_i = 0 \quad (i = 1, \dots, m),$$

then the sequence $\{(y_1^{(k)}, \dots, y_m^{(k)})\}$ converges to the unique solution of (D'). Consequently, the sequences $\{x^{(k)}\}$ and $\{y^{(k)}\}$ converge to the solutions of (P) and (D), respectively.

Proof. By the nonemptiness assumption, the infimum of (D') is attainable. Actually, the infimum is attained at a unique point. Suppose $(\bar{y}_1, \dots, \bar{y}_m)$ and (y_1^*, \dots, y_m^*) are two solutions of (D'), then it follows from Theorem 3.3 that $\bar{y} = \bar{y}_1 + \dots + \bar{y}_m = y_1^* + \dots + y_m^*$. Then, if we let $z_i = \bar{y} - y_i^*$, then

$$\sum z_i = 0 \quad \text{and} \quad z_i \in \text{aff}(\partial\delta(\bar{x} | C_i)) \quad (i = 1, \dots, m).$$

Therefore, by the linear independence assumption, we have $\bar{y}_i = y_i^*$ for $i = 1, \dots, m$ and, consequently, the solution of (D') is unique. The uniqueness of the solution, in turn, implies that the level sets of h are bounded and, hence, the sequence $\{(y_1^{(k)}, \dots, y_m^{(k)})\}$ is bounded. Therefore, convergence of the sequence $\{(y_1^{(k)}, \dots, y_m^{(k)})\}$ to the solution of (D) follows immediately from Proposition 4.7. The last statement of the theorem is also a direct result of Proposition 4.7. Q.E.D.

5. Applications. The proposed method can be used to solve some problems which are not explicitly of the form (P). We consider some of such applications in this section.

(1) *System of linear equations.* We consider the fundamental problem

$$Ax = b$$

where A is an $m \times n$ matrix with full row rank. To find a solution to the system, we can apply our method to the following problem:

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} \|x\|^2 \\ & \text{subject to} \quad Ax = b. \end{aligned}$$

By partitioning the matrix A and the vector b into row blocks as

$$A = \begin{pmatrix} A_1 \\ \vdots \\ A_m \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

we can set $C_i = \{x | A_i x = b_i\}$. For this case the objective function $q(x) = \frac{1}{2} \|x\|^2 = q^*(x)$; hence, we have

$$\begin{aligned} (5.1) \quad x^{(k)} &= y^{(k)} \\ &= \sum y_i^{(k)} \\ &= x^{(k-1)} + \sum (y_i^{(k)} - y_i^{(k-1)}). \end{aligned}$$

Meanwhile, both $y_i^{(k)}$ and $y_i^{(k-1)}$ are normal to the set C_i ; therefore, there exists a vector $v_i^{(k)}$ such that

$$y_i^{(k)} - y_i^{(k-1)} = A_i^T v_i^{(k)}.$$

By $z_i^{(k)} = x^{(k-1)} + \alpha(y_i^{(k)} - y_i^{(k-1)})$ and $z_i^{(k)} \in C_i$, we have that

$$(5.2) \quad A_i A_i^T v_i^{(k)} = (1/\alpha)(b_i - A_i x^{(k-1)}).$$

Thus, when the vectors $v_i^{(k)}$ are known, we can compute $x^{(k)}$ from $x^{(k-1)}$ by (5.1). To do this computationally, we first find a QR-decomposition $A_i = Q_i R_i$ of A_i and the system (5.2) becomes

$$R_i^T R_i v_i^{(k)} = (1/\alpha)(b_i - A_i x^{(k-1)}).$$

With vectors $u_i^{(k)}$ to represent $R_i v_i^{(k)}$, the method can be further simplified. We state the resulting method: Find the QR decompositions $A_i^T = Q_i R_i$, ($i = 1, \dots, m$). Let $x^{(0)}$ be an estimate to a solution of the system. For $k = 1, 2, \dots$, we do the following calculation:

- (a) for $i = 1, \dots, m$, solve $R_i^T u = (1/\alpha)(b_i - A_i x^{(k-1)})$ for $u_i^{(k)}$;
- (b) $x^{(k)} = x^{(k-1)} + \sum Q_i u_i^{(k)}$.

We note that m triangular linear systems are solved in parallel in each iteration. Of course, the matrices Q_i should be in product forms so that the products $Q_i u_i^{(k)}$ can be done very efficiently. It is also noted that, according to the theory given in the previous sections, the method should work properly with the parameter α larger than m .

(2) *Linear programming.* According to Mangasarian and Meyer [7], a linear program

$$\begin{aligned} & \text{minimize} \quad \langle c, x \rangle \\ & \text{subject to} \quad Ax \leq b \end{aligned}$$

can be tackled by solving the quadratic program

$$\begin{aligned} & \text{minimize} \quad \langle c, x \rangle + (\sigma/2) \|x\|^2 \\ & \text{subject to} \quad Ax \leq b \end{aligned}$$

where σ is any sufficiently small but positive number. Clearly, the quadratic program is a problem to which our method is applicable. Through this approach our method can be implemented as a parallel algorithm for linear programming. Unlike many other parallel algorithms for solving linear programs, the method does not require the matrix A to have any sparse structure.

(3) *Systems of generalized nonlinear inequalities.* Consider the following system of generalized nonlinear inequalities:

$$(5.3) \quad F(x) \in K$$

where F is a continuously differentiable function from R^n into R^m and K is a closed convex cone. For solving the problem Robinson [9] presented a method in which a sequence of points $\{u^{(k)}\}$ calculated iteratively by solving subproblems of the form:

$$(5.4) \quad \begin{aligned} &\text{minimize} \quad \|x - u^{(k)}\|^2 \\ &\text{subject to} \quad F(u^{(k)}) + F'(u^{(k)})(x - u^{(k)}) \in K. \end{aligned}$$

The method has a quadratic rate of convergence and is an extension of Newton's method to systems of nonlinear inequalities. If we can partition the function F and the cone K so that Problem (5.3) becomes

$$F_i(x) \in K_i, \quad i = 1, \dots, m,$$

then subproblem (5.4) reduces to

$$\begin{aligned} &\text{minimize} \quad \|x - u^{(k)}\|^2 \\ &\text{subject to} \quad x \in C_1 \cap \dots \cap C_m, \end{aligned}$$

where $C_i := \{x \mid F_i(u^{(k)}) + F'_i(u^{(k)})(x - u^{(k)}) \in K_i\}$. For this situation, our method again becomes applicable and the resulting method provides a way for solving Problem (5.3) in parallel.

6. Discussion. In many situations, as in all the examples given in § 5, the parameter α can be easily chosen. But, even when the objective function q is definite quadratic, it is sometimes very costly to determine this parameter in advance. However, as mentioned before, the parameter need not be fixed and can be adjusted iteratively. We present in this section such a procedure. It is noted that the method is more efficient with a smaller α ; therefore, even when a large workable α is available, it may still be more advantageous in some cases to start with a smaller parameter and let it be adjusted gradually.

The modified method. Let α_0 and β be positive numbers and let $\lambda > 1$. Let $y^{(0)} = y_1^{(0)} = \dots = y_m^{(0)} = 0$ and $x^{(0)} = \nabla q^*(y^{(0)})$. For $k = 1, 2, \dots$, we do the following computation:

- (a) $\tilde{\alpha} = \alpha_{k-1}$;
- (b) For $i = 1, \dots, m$, find \tilde{z}_i which solves

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} \|z + \tilde{\alpha} y_i^{(k-1)} - x^{(k-1)}\|^2 \\ &\text{subject to} \quad z \in C_i; \end{aligned}$$

- (c) Set $\tilde{y}_i = y_i^{(k-1)} + (1/\tilde{\alpha})(\tilde{z}_i - x^{(k-1)})$;
- (d) Set $\tilde{y} = \tilde{y}_1 + \dots + \tilde{y}_m$ and set $\tilde{x} = \nabla q^*(\tilde{y})$;
- (e) Check the condition

$$(6.1) \quad ((\tilde{\alpha}/2) - \beta) \sum \|y_i^{(k-1)} - \tilde{y}_i\|^2 \geq \langle \tilde{x} - x^{(k-1)}, \tilde{y} - y^{(k-1)} \rangle.$$

If (6.1) is not true then set $\tilde{\alpha} = \lambda \tilde{\alpha}$ and go to (b); otherwise, set $\alpha_k = \tilde{\alpha}$, $x^{(k)} = \tilde{x}$, $y^{(k)} = \tilde{y}$, and $y_i^{(k)} = \tilde{y}_i$, ($i = 1, \dots, m$).

In view of Lemma 4.2, when q is uniformly convex, condition (6.1) will hold after increasing α_k at most a finite number of times. Therefore, as in the proof of Lemma 4.3, we still have

$$h(y_1^{(k-1)}, \dots, y_m^{(k-1)}) \geq h(y_1^{(k)}, \dots, y_m^{(k)}) + \beta \sum \|y_i^{(k-1)} - y_i^{(k)}\|^2.$$

Consequently, all the convergence results hold equally well for this modified version of the method.

REFERENCES

- [1] R. FLETCHER, *Practical Methods of Optimization*, Vol. 2: *Constrained Optimization*, John Wiley, Chichester, 1981.
- [2] S.-P. HAN, *Superlinear convergent variable metric methods for general nonlinear programming*, Math. Programming, 11 (1976), pp. 263–282.
- [3] ———, *Variable metric methods for minimizing a class of nondifferentiable functions*, Math. Programming, 20 (1981) pp. 1–13.
- [4] ———, *A successive projection method*, Math. Programming, to appear.
- [5] ———, *A decomposition method and its application to convex programming*, Math. Oper. Res. to appear.
- [6] M. J. D. POWELL, *Algorithms for nonlinear constraints that use Lagrangian functions*, Survey of Math. Programming, 1 (1976), pp. 513–537.
- [7] O. L. MANGASARIAN AND R. R. MEYER, *Nonlinear perturbation of linear programs*, this Journal, 17 (1979), pp. 745–752.
- [8] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
- [9] S. M. ROBINSON, *Extension of Newton's method to nonlinear functions with values in a cone*, Numer. Math., 19 (1972), pp. 341–347.
- [10] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

FIXED POLES IN TRANSFER FUNCTION EQUATIONS*

G. CONTE†, A. M. PERDON† AND B. F. WYMAN‡

Abstract. In this paper we study the pole structure of the solutions $H(z)$, if any, of equations of the form $T(z) = H(z)G(z)$ or, dually, $T(z) = F(z)H(z)$, where $T(z)$ and $G(z)$, or $F(z)$, are given matrices of rational functions (multivariable transfer functions) over a field K . This study is motivated by various design problems in linear dynamical system theory (such as model matching or factorization problems), whose solutions are systems with an internal dynamics determined by the pole structure of $H(z)$. The methods we use are algebraic and module theoretic methods and the main tools are represented by the modules of the poles and of the zeros associated with a transfer function. The main result is a complete description of the "essential" pole structure which is common to all the solutions $H(z)$. This is given by means of a module whose invariant factors are explicitly computed in terms of fractional representations of the data $T(z)$ and $G(z)$ or $F(z)$. The essential pole structure is shown to consist, in a suitable algebraic sense, exactly of the poles of $T(z)$ which do not appear as poles of $G(z)$ (resp. $F(z)$) together with the zeros of $G(z)$ (resp. $F(z)$) which do not appear as poles of $[T'(z)G'(z)]'$ (resp. $[T(z)G(z)]$). The possibility of stating this fact, which agrees with the basic intuition, in a precise meaningful way, is due to the chosen algebraic framework. Furthermore, it is shown that no design limitations, except a technical one, apply to the remaining "inessential" part of the pole structure of any solutions.

Key words. fixed poles, zero module, algebraic system design

AMS(MOS) subject classifications. 93B25, 93B50

1. Introduction.

1.1. From the input/output point of view, a finite-dimensional, stationary, linear dynamical system Σ with m inputs and p outputs is given by a transfer function T_Σ which can be represented by a $p \times m$ matrix of rational functions. More generally, if $K(z)$ is the field of rational functions over a scalar field K , a transfer function is a $K(z)$ -linear transformation $T(z): U(z) \rightarrow Y(z)$, where $U(z) = K^m(z)$ and $Y(z) = K^p(z)$ are finite-dimensional $K(z)$ -vector spaces of input and output signals.

Given two transfer functions $T(z): U(z) \rightarrow Y(z)$ and $G(z): U(z) \rightarrow W(z)$ with the same space of inputs, we study the connecting transfer function (if any) $H(z): W(z) \rightarrow Y(z)$ such that $T(z) = H(z)G(z)$. This equation and the dual one, namely $T(z) = F(z)H(z)$ where $T(z): U(z) \rightarrow Y(z)$ and $F(z): W(z) \rightarrow Y(z)$ are given, arise naturally in various control problems, such as, for instance, system inversion, model matching and system factorization (see [9], [12], [13], [16]–[18], [20] for a general treatment and, more specifically, [3]–[5], [7], [8], [10], [11], [19], [22]).

The fundamental criterion for the existence of a solution $H(z)$, when $T(z)$ and $G(z)$ are given, is quite simple and well known: the nullspace condition $\text{Ker } G \subset \text{Ker } T$ is necessary and sufficient for the existence of $H(z)$. In many problems, however, the existence of a generic connecting transfer function is scarcely interesting; what is required is a solution $H(z)$ which has, in addition, some special property. For instance, we may seek a solution which is proper and/or stable or which, more generally, has no poles in a fixed subset S of $K \cup \{\infty\}$. Various methods based on the analysis of the invariant factors of the matrices involved ([14], [16]) or on other matrix theoretic

* Received by the editors December 8, 1986; accepted for publication (in revised form) May 19, 1987.

† Department of Mathematics, University of Genoa, Genoa, Italy. The work of these authors was partially supported by the Ministero Pubblica Istruzione.

‡ Department of Mathematics, Ohio State University, Columbus, Ohio 43210. The work of this author was supported by grants from the National Science Foundation and the Centre National de la Recherche Scientifique, which made it possible for him to work in Europe in the summer of 1985.

techniques ([7]–[9], [11], [12], [15]) can be used to check the existence of solutions of this kind. However, in the multivariable case, other interesting properties of the connecting transfer function $H(z)$ do not depend specifically on the location of its poles in $K \cup \{\infty\}$, but on the pole structure, in the sense of [12, 6.5.3], at the point of $K \cup \{\infty\}$. For instance, the existence of a cyclic structure, which implies single input controllability, on the modes associated to a certain frequency α is equivalent to a pole structure at α of the form $(0, \dots, 0, \sigma_\alpha)$. It is therefore natural to look not only for the existence of solutions having no poles in S , but also for more complete information on their pole structure. In this case, the analysis of the invariant factors or the use of other direct matrix theoretic methods do not appear capable of providing satisfactory results. One explicit result in this direction obtained by means of such techniques is Theorem 2.4 of [20]. It permits the computation of a lower bound for the total order of pole, at a given point α , of any solution $H(z)$. Clearly, in the multivariable case, this information on the pole structure is quite poor.

1.2. In this paper we study the pole structure of the solutions in a module theoretic framework, employing the notions of pole module and zero module of a linear transfer function (see [21], [22]). The advantage in this approach is the fact that it supplies a complete description of the pole structure we are considering. The basic result in this setting is that there is an “essential” pole structure which appears in every solution of $T(z) = H(z)G(z)$ and which is representable by means of a suitable module determined by $T(z)$ and $G(z)$. Using this module theoretic characterization one can not only check the presence in the solutions of poles in a certain region S , but one can also compute explicitly the list of multiplicities which form the structure. The results previously mentioned concerning the existence of solutions with specific polar properties are therefore effectively improved in this way.

Moreover, due to the richness of the module theoretic framework, a clean interpretation in terms of poles and zeros of $T(z)$ and $G(z)$ of the essential pole structure of the solutions is possible. Let us remark that, intuitively and roughly speaking, one can expect that the essential part of the poles of any solution $H(z)$ must supply the poles of $T(z)$ which do not appear already in $G(z)$ as well as the poles needed to cancel the zeros of $G(z)$ which do not appear in $[T(z)'G(z)']'$. (In general, it is necessary to consider $[T(z)'G(z)']'$ since some zero of $G(z)$ may fail to appear as a zero of $T(z)$ for nondynamical reasons, i.e., without being canceled by a pole of $H(z)$, if $T(z)$ is not injective). Indeed, these phenomena appear clearly in the scalar case. If $T(z) = p(z)/q(z)$ and $G(z) = p'(z)/q'(z)$ are given as fractions in lowest terms, then the unique solution has a representation $H(z) = p(z)q'(z)/p'(z)q(z)$ which is not necessarily reduced. Restricting the present discussion to the finite poles of $H(z)$, we have that they consist of roots of $q(z)$, which do not already appear as roots of $q'(z)$, that is poles of $T(z)$ which are not poles of $G(z)$, together with roots of $p'(z)$, which do not appear as roots of $p(z)$ (zeros of $G(z)$ which are not zeros of $T(z)$). A straightforward generalization of these observations to the multivariable case is not possible because numerator matrices may not be invertible and numerically coincident multivariable poles and zeros may not cancel. An attempt to extend the above interpretation of the essential poles to the multivariable case is found in [20]. The matrix theoretic techniques used in that paper do not give complete information on the pole structure. On the other hand, we can show here that the module which represents the essential pole structure of the solutions consists, in a precise algebraic sense, of two modules which have a natural direct interpretation as the module of poles of $T(z)$ which do not appear in $G(z)$ and as the module of zeros of $G(z)$ which do not appear in $[T(z)'G(z)']'$.

1.3. Let us describe the framework of the paper in greater detail. Essentially, we represent the zero structure and the pole structure of a transfer function $T(z)$ by means of appropriate modules. The finite pole structure is given by the state space $X(T)$ of the minimal realization of (the strictly proper part of) $T(z)$ provided with the module structure over the ring of polynomials $K[z]$ induced by the internal dynamics matrix. The finite zero structure is given by the zero module $Z(T)$ first introduced in [22] (see [21] for the general theory). In this setting, our aim is to describe, as completely as possible, the pole module $X(H)$, $H(z)$ being a solution of $T(z) = H(z)G(z)$.

After some preliminaries and notation we prove, in § 3, that for every $H(z)$ there exists an exact sequence

$$0 \rightarrow P \rightarrow X(H) \rightarrow C(H) \rightarrow 0,$$

where $P = P(T, G)$ is a module easily described in terms of $T(z)$ and $G(z)$. Hence, P represents, for any solution $H(z)$, a fixed or essential pole structure. The explicit computation of P in terms of fractional representations of $T(z)$ and $G(z)$ is made possible by the procedure presented in § 3.6. Moreover, we show that P is the middle term of an exact sequence

$$0 \rightarrow X \rightarrow P \rightarrow Z \rightarrow 0$$

where the modules X and Z can be naturally viewed as representing, respectively, the poles of $T(z)$ not in $G(z)$ and the zeros of $G(z)$ not in $[T(z)'G(z)']'$. Thus P consists, in a precise algebraic sense, exactly of the poles and zeros suggested by the basic intuition. Note that this interpretation, which generalizes the scalar case, makes sense because of the module theoretic framework in which we are working. In § 4, we show that the Cokernel $C(H)$ of inessential poles is subject only to technical restrictions on the number of invariant factors, while the location of the modes in $C(H)$ is arbitrary when $C(H) \neq 0$. Therefore, the fixed finite poles of every solution are entirely represented by P and solutions whose finite poles are all essential can be constructed. More generally, we describe how to construct, by means of matrix fraction techniques, solutions whose poles are all essential except, possibly, those at an arbitrarily fixed point of $K \cup \{\infty\}$. Easy counterexamples show, however, that globally essential solutions, i.e., solutions all of whose poles, finite and at infinity, are essential, may not exist. Two examples are described at the end of § 5.

Although for simplicity we limited our discussion in this introduction to the case of finite poles and zeros, the main text of the paper studies the structure of $H(z)$ with respect to an arbitrary subset of $K \cup \{\infty\}$ (see for example [21] for background to the general theory). In addition, the problem $T(z) = H(z)G(z)$ described above and the dual one $T(z) = F(z)G(z)$, where $T(z)$ and $F(z)$ are given and $H(z)$ is sought, are treated simultaneously. The above results generalize those obtained in the special case of system inversion, namely when $Y(z) = U(z)$, $T(z)$ is the identity and $G(z)$ is monic, by Wyman and Sain in [22], [23].

Some of the results in this paper appeared in [3]. In that paper the equation $T(z) = H(z)G(z)$ is first reduced to a simpler form, multiplying both members by the left least common multiple of the right coprime matrix denominators of $T(z)$ and $G(z)$. The use of this technical artifice is made unnecessary in the present paper, with the advantage of providing clearer proofs, different from those of [3], and a better understanding of the role played by poles and zeros of $T(z)$ and $G(z)$ in determining P . The procedure to compute explicitly the essential pole structure was announced in [6].

2. Preliminaries and notation. Let K be a field and let $K[z]$ and $K(z)$ denote, respectively, the ring of polynomials and the field of rational functions in the variable z over K . Given a point $\alpha \in K$, any $g(z) \in K(z)$ can be represented as $g(z) = (z - \alpha)^r (p(z)/q(z))$ with $r \in \mathbb{Z}$, $p(z)$ and $q(z)$ polynomials not divisible by $(z - \alpha)$. The *valuation of $g(z)$ at α* is then defined by $v_\alpha(g) = r$. If $g(z) \in K(z)$ is written as the quotient of polynomials $g(z) = p(z)/q(z)$, the *valuation of $g(z)$ at ∞* is defined by $v_\infty(g) = \deg q - \deg p$.

Given a proper subset $S \subset K \cup \{\infty\}$, we denote by O_S the ring

$$O_S = \{f(z) \in K(z) : v_\alpha(f) \geq 0 \text{ for all } \alpha \text{ in } S\}.$$

We say loosely that O_S is the ring of rational functions which are regular at every point of S . The ring O_S is a principal ideal domain. When S consists of one point, say $S = \{\alpha\}$, we use the notation O_α . We assume that the reader is familiar with the theory of modules over a principal ideal domain as described in [1], for example.

Given a K -vector space V , we denote by $V(z)$ the $K(z)$ -vector space $V \otimes_K K(z)$ and by $\Omega_S V$ the O_S -module $V \otimes_K O_S$. $\Omega_S V$ is, in an obvious way, an O_S -submodule of $V(z)$.

Given two K -vector spaces $U = K^m$ and $Y = K^p$, by a *transfer function* we mean a $K(z)$ -linear map $T(z) : U(z) \rightarrow Y(z)$ or, equivalently, the $p \times m$ matrix of rational functions which represents $T(z)$ with respect to the canonical $K(z)$ -basis of $U(z)$ and $Y(z)$.

For a given subset $S \subset K \cup \{\infty\}$, we associate (see [21]) to any transfer function $T(z)$:

(i) a right (left) S -coprime fractional representation $T(z) = N_S(z) D_S^{-1}(z)$ (respectively $T(z) = \tilde{D}_S^{-1}(z) \tilde{N}_S(z)$), where $N_S(z)$, $D_S(z)$ (respectively $\tilde{N}_S(z)$, $\tilde{D}_S(z)$) are matrices with elements in O_S whose common right (left) factors are S -unimodular, i.e., they have an inverse with elements in O_S , and $D_S(z)$ ($\tilde{D}_S(z)$) is nonsingular;

(ii) a finitely generated torsion O_S -module $Z_S(T)$, called the *module of zeros in S* , defined by

$$Z_S(T) = \frac{T^{-1}(\Omega_S Y) + \Omega_S U}{\text{Ker } T + \Omega_S U};$$

(iii) a finitely generated torsion O_S -module $X_S(T)$, called the *module of poles in S* , defined by

$$X_S(T) = \frac{\Omega_S U}{T^{-1}(\Omega_S Y) \cap \Omega_S U}.$$

By abuse of notation, using the natural isomorphism induced by $T(z)$, we will also use the representation

$$X_S(T) = \frac{T(\Omega_S U) + \Omega_S Y}{\Omega_S Y}.$$

Note that, when $S = K$, then O_S is the ring of polynomials $K[z]$, $Z_S(T)$ coincides with the (finite) zero module introduced by Wyman and Sain in [22] and $X_S(T)$ coincides with the state space of the minimal realization of (the strictly proper part of) $T(z)$. On the other hand, when $S = \{\infty\}$, then O_∞ is the ring of rational formal power series in z^{-1} and $Z_\infty(T)$ and $X_\infty(T)$ (for which the notation $P_\infty(T)$ should be preferred in order to avoid confusion with the generalized state space introduced previously in the literature) are, respectively, the infinite zero module and the infinite pole module introduced in [2].

In general, the following isomorphisms hold [21]:

$$Z_S(T) \simeq \text{Tor}(\Omega_S Y / N_S(\Omega_S U)) \simeq \text{Tor}(\Omega_S Y / \tilde{N}_S(\Omega_S U))$$

(where Tor denotes the torsion submodule over O_S) and

$$X_S(T) \simeq \Omega_S Y / D_S(\Omega_S Y) \simeq \Omega_S U / \tilde{D}_S(\Omega_S U).$$

In other words, the invariant factors of $Z_S(T)$ and of $X_S(T)$ coincide with the nontrivial elements in the Smith form, over O_S , of $N_S(z)$ and $\tilde{N}_S(z)$ or of $D_S(z)$ and $\tilde{D}_S(z)$, respectively. Since the zero and the pole structure of $T(z)$ at a point α of S are given by the ordered list of the valuations at α of the elements in the Smith form of $N(z)$ and, respectively, of $D(z)$ (see, for instance, [12, § 6.5], where the “classical” case $S = K$ and the “local” case $S = \{\infty\}$ are considered), all the invariant factor information about the zeros and the poles of $T(z)$ in S can be derived from $Z_S(T)$ and $X_S(T)$.

3. Fixed poles. Let the transfer functions $T(z): U(z) \rightarrow Y(z)$ and $G(z): U(z) \rightarrow W(z)$ be given. The design problem which consists in factoring $T(z)$ through $G(z)$, i.e., in finding $H(z): W(z) \rightarrow Y(z)$ such that

$$(3.1) \quad T(z) = H(z)G(z),$$

is solvable if and only if

$$(A) \quad \text{Ker } G \subset \text{Ker } T.$$

Dually, given $T(z)$ as above and $F(z): W(z) \rightarrow Y(z)$, the design problem which consists in factoring $T(z)$ through $F(z)$, i.e., in finding $H(z): U(z) \rightarrow W(z)$ such that

$$(3.2) \quad T(z) = F(z)H(z),$$

is solvable if and only if

$$(B) \quad \text{Im } T \subset \text{Im } F.$$

From now on, we deal with pairs of transfer functions $T(z)$, $G(z)$ (respectively, $T(z)$, $F(z)$) as above, which are assumed to verify condition (A) (respectively, (B)). The set of solutions $H(z)$ of (3.1), (3.2) is therefore not empty and, unless $G(z)(F(z))$ is nonsingular, it contains more than one element. Our aim in this section is to introduce a suitable algebraic notion of “fixed poles” for (3.1) and (3.2), which allows us to describe the pole structure of the solutions $H(z)$ with respect to an arbitrary fixed subset $S \subset K \cup \{\infty\}$. Therefore, let us start by considering the following two O_S -modules:

$$P_S = \frac{G^{-1}(\Omega_S W)}{G^{-1}(\Omega_S W) \cap T^{-1}(\Omega_S Y)} \quad \text{and} \quad \tilde{P}_S = \frac{T(\Omega_S U) + F(\Omega_S W)}{F(\Omega_S W)}$$

associated, respectively, with (3.1) and (3.2).

Now, we can state the following basic results.

PROPOSITION 3.3. *For any solution $G(z)$ of (3.1) there exists a natural inclusion $j: P_S \rightarrow X_S(H)$.*

PROPOSITION 3.4. *For any solution $H(z)$ of (3.2) there exists a natural projection $\pi: X_S(H) \rightarrow \tilde{P}_S$.*

Proof of Proposition 3.3. The map $j: P_S \rightarrow X_S(H)$ is induced in the following commutative diagram by (1):

$$\begin{array}{ccccc} G^{-1}(\Omega_S W) \cap T^{-1}(\Omega_S Y) & \longrightarrow & G^{-1}(\Omega_S W) & \longrightarrow & P_S \\ \downarrow G & (1) & \downarrow G & & \downarrow j \\ \Omega_S W \cap H^{-1}(\Omega_S Y) & \longrightarrow & \Omega_S W & \longrightarrow & X_S(H) \end{array}$$

Furthermore, since $G^{-1}(\Omega_S W) \cap T^{-1}(\Omega_S Y) = G^{-1}(\Omega_S W \cap H^{-1}(\Omega_S Y))$, the map j is easily seen to be injective.

Proof of Proposition 3.4. The map $\pi: X_S(H) \rightarrow \tilde{P}_S$ is induced in the following commutative diagram by (2).

$$\begin{array}{ccccc} \Omega_S W & \longrightarrow & H(\Omega_S U) + \Omega_S W & \longrightarrow & X_S(H) \\ \downarrow F & (2) & \downarrow F & & \downarrow \pi \\ F(\Omega_S W) & \longrightarrow & T(\Omega_S U) + F(\Omega_S W) & \longrightarrow & \tilde{P}_S \end{array}$$

Furthermore, since $T(\Omega_S U) + F(\Omega_S W) = F(H(\Omega_S U) + \Omega_S W)$, π is easily seen to be surjective. \square

When we compare with [4] and [20], the above propositions justify the following definitions.

DEFINITION 3.5. The modules P_S and \tilde{P}_S are called *modules of fixed poles* in S of (3.1) and (3.2), respectively.

A first consequence of the above results is that the invariant factors of P_S or \tilde{P}_S appear, in the same order, in the Smith form, over O_S , of the denominator matrices of any solution $H(z)$ and, clearly, the conditions $P_S = 0$ or $\tilde{P}_S = 0$ is necessary for the existence of solutions having no poles in S . However, much more can be said, since, as shown below, the fixed pole structure can be computed explicitly in terms of fractional representations of $T(z)$ and $G(z)$ or $F(z)$.

3.6. Computation of the fixed pole structure. Assume that the data $T(z)$ and $G(z)$ of (3.1) have right S -coprime fractional representations $T(z) = N(z)D^{-1}(z)$, $G(z) = N_G(z)D_G^{-1}(z)$ and let $M(z) = D(z)A(z) = D_G(z)B(z)$ be the least common left multiple, in the ring of matrices with elements in O_S , of $D(z)$ and $D_G(z)$. By the same technique used in the proofs of Propositions 3.3 and 3.4, it is not difficult to see that the map $[u(z)] \rightarrow [M(uz)]$ induces an isomorphism between the modules $(N_G B)^{-1}(\Omega_S W) / ((N_G B)^{-1}(\Omega_S W) \cap (NA)^{-1}(\Omega_S Y))$ and P_S . Now, let r be the rank of

$$\begin{bmatrix} N_G(z)B(z) \\ N(z)A(z) \end{bmatrix} \quad \text{and let} \quad \begin{bmatrix} N_G(z)B(z) \\ N(z)A(z) \end{bmatrix} = \begin{bmatrix} \overbrace{V_1(z)}^r \mid V_3(z) \\ V_2(z) \mid V_4(z) \end{bmatrix} \begin{bmatrix} \Delta(z) \mid 0 \\ -0 \mid 0 \end{bmatrix} \begin{bmatrix} U_1(z) \\ U_2(z) \end{bmatrix} \Big\}_r$$

be the Smith decomposition. Denoting by $S(z)$ the full row rank $r \times m$ matrix $S(z) = \Delta(z)U_1(z)$, which can be viewed as the maximum (nonsingular) common right divisor of $N_G(z)B(z)$ and $N(z)A(z)$, the equalities $N_G(z)B(z) = V_1(z)S(z)$ and $N(z)A(z) = V_2(z)S(z)$ hold. Then, writing $U' = K'$, we have $(N_G B)^{-1}(\Omega_S W) \cap (NA)^{-1}(\Omega_S Y) = S^{-1}(\Omega_S U')$ and hence $P_S \cong (V_1 S)^{-1}(\Omega_S W) / S^{-1}(\Omega_S U')$. Remarking that $V_1(u) = 0$ implies, for a suitable u' such that $u = S(u')$, $V_1 S(u') = N_G B(u') = 0$ and that, by (A), $\text{Ker } N_G B \subset \text{Ker } NA$, one can show that $V_1(z)$ has full column rank. Then, as before, the multiplication by $V_1(z)S(z)$ induces an isomorphism between P_S and $(\Omega_S W \cap \text{Im } V_1) / V_1(\Omega_S U')$. The last module is isomorphic to the torsion submodule

$\text{Tor}(\Omega_S W / V_1(\Omega_S U'))$ of $\Omega_S W / V_1(\Omega_S U')$ and, as a consequence, the invariant factors of P_S which form the fixed pole structure are the nontrivial element in the Smith form of $V_1(z)$.

The analogous result for the fixed pole structure in (3.2) can be obtained by duality.

Remark 3.7. Note that the equality between the nontrivial elements in the Smith form of $V_1(z)$ and the fixed pole structure does not follow directly from the above computation. In order to state its validity, we need the abstract result of Proposition 3.3, which is primarily due to the module theoretic framework we are employing.

Now, we can give a description of P_S and \tilde{P}_S in terms of zeros and poles of the data $T(z)$ and $G(z)$, or $F(z)$, which, in connection with the scalar case recalled in the Introduction, will clarify the nature of the fixed poles. Let us consider (3.1) first. Since (A) holds, the zero module of the transfer function $[T(z)'G(z)']': U(z) \rightarrow (Y \oplus W)(z)$ can be represented as

$$\begin{aligned} Z_S \begin{pmatrix} T \\ G \end{pmatrix} &= \frac{\begin{pmatrix} T \\ G \end{pmatrix}^{-1} \Omega_S(Y \oplus W) + \Omega_S U}{\text{Ker} \begin{pmatrix} T \\ G \end{pmatrix} + \Omega_S U} \\ &= \frac{T^{-1}(\Omega_S Y) \cap G^{-1}(\Omega_S W) + \Omega_S U}{\text{Ker } T \cap \text{Ker } G + \Omega_S U} \\ &= \frac{T^{-1}(\Omega_S Y) \cap G^{-1}(\Omega_S W) + \Omega_S U}{\text{Ker } G + \Omega_S U}. \end{aligned}$$

Therefore, as $Z_S(G) = (G^{-1}(\Omega_S W) + \Omega_S U) / (\text{Ker } G + \Omega_S U)$, we have a natural inclusion $i: Z_S \begin{pmatrix} T \\ G \end{pmatrix} \rightarrow Z_S(G)$ induced by the obvious inclusion of the numerator modules. On the other hand, we have a natural projection $p: X_S \begin{pmatrix} T \\ G \end{pmatrix} \rightarrow X_S(G)$ between the pole modules

$$X_S \begin{pmatrix} T \\ G \end{pmatrix} = \frac{\Omega_S U}{T^{-1}(\Omega_S Y) \cap G^{-1}(\Omega_S W) \cap \Omega_S U} \quad \text{and} \quad X_S(G) = \frac{\Omega_S U}{G^{-1}(\Omega_S W) \cap \Omega_S U}$$

induced by the obvious inclusion of the denominator modules.

Notation 3.8. We denote by Z_S and X_S the O_S -modules defined, up to isomorphism, respectively, by the short exact sequences

$$\begin{aligned} 0 \rightarrow Z_S \begin{pmatrix} T \\ G \end{pmatrix} \xrightarrow{i} Z_S(G) \rightarrow Z_S \rightarrow 0, \\ 0 \rightarrow X_S \rightarrow X_S \begin{pmatrix} T \\ G \end{pmatrix} \xrightarrow{p} X_S(G) \rightarrow 0. \end{aligned}$$

It is clear that the module Z_S represents the zeros of $G(z)$ which are not zeros of $[T(z)'G(z)']'$. In particular, the above sequence splits over K and, as a K -vector space, $Z_S(G) \simeq Z_S((T'G')') \oplus_K Z_S$. In other words Z_S represents exactly the zeros that one expects to be cancelled by the poles of any solution $H(z)$ of (3.1). Analogously, since $X_S((T'G')')$ describes the union of the poles of $T(z)$ and of $G(z)$, X_S represents the poles of $T(z)$ which are not poles of $G(z)$ (see also [3, Remark 3.5] for this and for a description of the modules in matrix terms). They are exactly the poles that one expects to appear as poles of any solution $H(z)$ of (3.1). This is made precise by the following proposition.

PROPOSITION 3.9. *There exists a natural inclusion $\varphi: X_S \rightarrow P_S$ and a natural projection $\psi: P_S \rightarrow Z_S$ such that the following sequence is exact:*

$$0 \rightarrow X_S \xrightarrow{\varphi} P_S \xrightarrow{\psi} Z_S \rightarrow 0.$$

Proof. By a natural isomorphism, X_S can be represented as

$$X_S = \frac{G^{-1}(\Omega_S W) \cap \Omega_S U}{T^{-1}(\Omega_S Y) \cap G^{-1}(\Omega_S W) \cap \Omega_S U}.$$

Then $\varphi: X_S \rightarrow P_S$ is the map induced in the following commutative diagram, where π denotes the canonical projections, by (1):

$$\begin{array}{ccccc} G^{-1}(\Omega_S W) \cap T^{-1}(\Omega_S Y) \cap \Omega_S U & \longrightarrow & G^{-1}(\Omega_S W) \cap \Omega_S U & \xrightarrow{\pi_X} & X_S \\ \downarrow \text{incl.} & & \downarrow \text{incl.} & & \downarrow \varphi \\ G^{-1}(\Omega_S W) \cap T^{-1}(\Omega_S Y) & \longrightarrow & G^{-1}(\Omega_S W) & \xrightarrow{\pi} & P_S \end{array} \quad (1)$$

and φ is injective by diagram chasing. Analogously, representing Z_S as

$$Z_S = \frac{G^{-1}(\Omega_S W) + \Omega_S U}{T^{-1}(\Omega_S Y) \cap G^{-1}(\Omega_S W) \cap \Omega_S U},$$

we have that ψ is the surjective map induced in the following commutative diagram by (2):

$$\begin{array}{ccccc} G^{-1}(\Omega_S W) \cap T^{-1}(\Omega_S Y) & \longrightarrow & G^{-1}(\Omega_S W) & \xrightarrow{\pi} & P_S \\ \downarrow \text{incl.} & & \downarrow \text{incl.} & & \downarrow \psi \\ G^{-1}(\Omega_S W) \cap T^{-1}(\Omega_S Y) + \Omega_S U & \longrightarrow & G^{-1}(\Omega_S W) + \Omega_S U & \xrightarrow{\pi_Z} & Z_S \end{array} \quad (2)$$

Now, let $\pi_X(u) \in X_S$. This implies, in particular, that $u \in \Omega_S U$ and that $\psi\varphi\pi_X(u) = \psi\pi(u) = 0$. Hence, $\varphi(X_S) \subset \text{Ker } \psi$. On the other hand, let $\pi(u) \in \text{Ker } \psi \subset P_S$: this implies, in particular, that $u \in G^{-1}(\Omega_S W)$ and that $u \in G^{-1}(\Omega_S W) \cap T^{-1}(\Omega_S Y) + \Omega_S U$. Therefore, $u = u' + u''$, with $u' \in G^{-1}(\Omega_S W) \cap T^{-1}(\Omega_S Y)$ and $u'' \in \Omega_S U \cap G^{-1}(\Omega_S W)$, and $\pi(u) = \pi(u'')$. As a consequence, $\pi(u) = \varphi\pi_X(u'') \in \varphi(X_S)$ and $\text{Ker } \psi \subset \varphi(X_S)$. Thus we have shown the exactness of the sequence $0 \rightarrow X_S \rightarrow P_S \rightarrow Z_S \rightarrow 0$. \square

The proposition we have just proved implies, in particular, that $P_S = X_S \oplus_K Z_S$ as a K -vector space. Then, summarizing Proposition 3.3 and Proposition 3.9, we can say that the fixed poles of (3.1) consist of the zeros of $G(z)$ which are not zeros of $T(z)$ and of the poles of $T(z)$ which are not poles of $G(z)$. This statement makes sense only if we refer to the module theoretic definitions of zeros and poles.

The situation concerning (3.2) can be treated dually. Without giving the details, let us simply remark that the inclusion $i_2: W \rightarrow U \oplus W$ induces a natural inclusion $\tilde{i}: X_S(F) \rightarrow X_S(T \ F)$ between the pole modules and, since (B) holds, a natural projection $\tilde{p}: Z_S(F) \rightarrow Z_S(T \ F)$ between the zero modules with which we are dealing.

Notation 3.10. (i) We denote by \tilde{X}_S the module defined, up to isomorphism, by the short exact sequence

$$0 \rightarrow X_S(F) \xrightarrow{\tilde{i}} X_S(T \ F) \rightarrow \tilde{X}_S \rightarrow 0.$$

(ii) We denote by \tilde{Z}_S the module defined, up to isomorphism, by the short exact sequence

$$0 \rightarrow \tilde{Z}_S \rightarrow Z_S(F) \xrightarrow{\tilde{P}} Z_S(T - F) \rightarrow 0.$$

Just as in the discussion following § 3.6, we can give an interpretation of \tilde{X} and \tilde{Z} in terms of poles of $T(z)$ which are not poles of $F(z)$ and of zeros of $F(z)$ which are not zeros of $(T(z) - F(z))$. They are exactly the poles that one expects to find as poles of any solution $H(z)$ of (3.2) and, in fact, the following can be proved.

PROPOSITION 3.11. *There exists a natural inclusion $\tilde{\varphi}: \tilde{Z}_S \rightarrow \tilde{P}_S$ and a natural projection $\tilde{\psi}: \tilde{P}_S \rightarrow \tilde{X}_S$ such that the following sequence:*

$$0 \rightarrow \tilde{Z}_S \xrightarrow{\tilde{\varphi}} \tilde{P}_S \xrightarrow{\tilde{\psi}} \tilde{X}_S \rightarrow 0$$

is exact.

As in the previous case, we have that $\tilde{P}_S = \tilde{X}_S \oplus_K \tilde{Z}_S$ as a K -vector space. Summarizing Propositions 3.4 and 3.11 we can say, in a precise algebraic sense, that the fixed poles of (3.2) consists of the zeros of $F(z)$ which are not zeros of $(T(z) - F(z))$ and of the poles of $T(z)$ which are not poles of $F(z)$.

4. Essential solutions. The results of the previous section imply, in particular, that for any solution $H(z)$ of (3.1) or (3.2) we have an exact sequence of the form

$$0 \rightarrow P_S \rightarrow X_S(H) \rightarrow C_S(H) \rightarrow 0$$

or one of the form

$$0 \rightarrow \tilde{C}_S(H) \rightarrow X_S(H) \rightarrow \tilde{P}_S \rightarrow 0.$$

We say that the modules P_S and \tilde{P}_S describe the *essential pole structure* in S of the solutions of (3.1) and (3.2) and that, for any $H(z)$, the cokernel module $C_S(H)$ or the kernel module $\tilde{C}_S(H)$ is the module of *inessential poles* in S (compare with [22], [23]). Since we have a complete knowledge of P_S or \tilde{P}_S , the aim of this section is to study the module of inessential poles, describing in particular what restrictions, if any, apply to it. We restrict our investigation to (3.1), since the analogous results concerning (3.2) can be obtained by transposition.

To begin with, let us state some technical results and introduce some notation.

Remark 4.1. Let $Q(z)$ be an S -unimodular matrix such that $Q(z)G(z)$ has the form $Q(z)G(z) = [G_1(z)' \ 0]'$, with $G_1(z)$ of full row rank. Writing $W(z) = W_1(z) \oplus W_2(z)$ with $W_1(z) = Q(z)G(U(z))$, we can represent any solution $H(z)$ of (3.1) as $H(z) = [H_1(z) \ H_2(z)]Q(z)$, where $H_1(z): W_1(z) \rightarrow Y(z)$ is such that $T(z) = H_1(z)G_1(z)$ and $H_2(z): W_2(z) \rightarrow Y(z)$ is arbitrary. Since $G_1(z)$ is full row rank, $H_1(z)$ is uniquely determined and given by $H_1(z) = T(z)G_1'(z)$, where $G'(z)$ is any right inverse of $G_1(z)$. Moreover, since $Q(z)$ is S -unimodular, the pole (and zero) structure of $H(z)$ in S is that of $[H_1(z) \ H_2(z)]$.

LEMMA 4.2. *Given the S -coprime fractional representations $H_1(z) = D_1^{-1}(z)N_1(z)$ and $H_2(z) = D_2^{-1}(z)N_2(z)$, let $M(z) = A(z)D_1(z) = B(z)D_2(z)$ be the minimal common left multiple of $D_1(z)$ and $D_2(z)$. Then $[H_1(z) \ H_2(z)] = M^{-1}(z)[A(z)N_1(z) \ B(z)N_2(z)]$ is an S -coprime fractional representation.*

Proof. See [3, Remark 3.5].

Now, let us consider again the diagram in the proof of Proposition 3.3. With the notation introduced above, and applying the isomorphism induced by $Q(z)$ to the

second row and adding the Cokernels of the vertical maps, we obtain the following commutative diagram:

$$\begin{array}{ccccc}
 G^{-1}(\Omega_S W) \cap T^{-1}(\Omega_S Y) & \longrightarrow & G^{-1}(\Omega_S W) & \longrightarrow & P_S \\
 \downarrow QG & & \downarrow QG & & \downarrow j \\
 (\Omega_S W_1 \oplus \Omega_S W_2) \cap [H_1 \ H_2]^{-1}(\Omega_S Y) & \longrightarrow & \Omega_S W_1 \oplus \Omega_S W_2 & \longrightarrow & X_S(H) \\
 \downarrow & (1) & \downarrow & & \downarrow \\
 K_S & \longrightarrow & \Omega_S W_2 & \longrightarrow & C_S(H)
 \end{array}$$

in which j , $X_S(H)$ and $C_S(H)$ are employed with an obvious abuse of notation. The last row is exact and, by the commutativity of (1), it is possible to evaluate the image of the module K_S in $\Omega_S W_2$. In fact, it consists of the elements $w \in \Omega_S W_2$ for which there exists $w' \in \Omega_S W_1$ such that $H_1(w') + H_2(w)$ belongs to $\Omega_S Y$. This is to say that $AN_1(w') + BN_2(w) = M(y)$ or, equivalently, that $BN_2(w) = A(N_1(-w') + D_1(y))$ for some $w' \in \Omega_S W_1$ and $y \in \Omega_S Y$. Since $N_1(z)$ and $D_1(z)$ are S -coprime, the last equality is equivalent to $BN_2(w) = A(y')$ for some $y' \in \Omega_S Y$; hence the image of K_S is given by $\Omega_S W_2 \cap (BN_2)^{-1}(A(\Omega_S Y))$. In conclusion, we have that $C_S(H) \simeq \Omega_S W_2 / (\Omega_S W_2 \cap (A^{-1}BN_2)^{-1}(\Omega_S Y)) \simeq X_S(A^{-1}BN_2)$ and, since $A(z)$ and $B(z)N_2(z)$ are easily seen to be S -coprime, $C_S(H) \simeq \Omega_S W / A(\Omega_S Y)$. We state explicitly the above results as follows.

PROPOSITION 4.3. *Let $Q(z)$ and $G_1(z)$ be as defined in Remark 4.1. Then, any solution $H(z)$ of (3.1) has the form $[H_1(z) \ H_2(z)]Q(z)$ where $H_1(z)$ is uniquely determined by $H_1(z) = T(z)G_1'(z)$, $G_1'(z)$ being any right inverse of $G_1(z)$, and $H_2(z)$ is an arbitrary $(\dim Y) \times (\dim W - \text{rank } G)$ transfer function. Moreover, if $H_1(z) = D_1^{-1}(z)N_1(z)$ and $H_2(z) = D_2^{-1}(z)N_2(z)$ are S -coprime fractional representations and if $M(z) = A(z)D_1(z) = B(z)D_2(z)$ is the minimum left common multiple of $D_1(z)$ and $D_2(z)$, we have $C_S(H) \simeq \Omega_S Y / A(\Omega_S Y)$ and, since $X_S(H) \simeq \Omega_S Y / M(\Omega_S Y)$, $P_S \simeq \Omega_S Y / D_1(\Omega_S Y)$.*

COROLLARY 4.4. *Given any finitely generated torsion O_S -module E , whose number of invariant factors is less than or equal to $\min \{\dim Y, \dim W - \text{rank } G\}$, there exists a solution $H(z)$ of (3.1) such that $C_S(H) \simeq E$. In particular, for any S there exists a solution $H(z)$ of (3.1) whose poles in S coincide with the fixed poles.*

Proof. By Proposition 4.3, if $A(z)$ is a $p \times p$ matrix such that $E \simeq \Omega_S Y / A(\Omega_S Y)$, it is sufficient to take $H_2(z) = D_1^{-1}(z)A^{-1}(z)$. Solutions whose poles coincide with the fixed ones are obtained whenever the arbitrary transfer function $H_2(z) = D_2^{-1}(z)N_2(z)$ is chosen in such a way that $D_1(z)$ is a left multiple of $D_2(z)$. \square

By Corollary 4.4 we can now say that no restrictions, except for the number of invariant factors, apply to the module of inessential poles $C_S(H)$ or $\tilde{C}_S(H)$. This implies, in particular, that in each solution $H(z)$ the fixed part of the pole structure in S is entirely described by P_S or \tilde{P}_S . The location of other poles in S , if any, is arbitrary. When the module of inessential poles is zero, the corresponding solution is said to be *essential* in S . Solutions of this kind have the "smallest" possible pole structure in S , but they are not necessarily essential also in $K \cup \{\infty\} \setminus S$. In general, there do not exist solutions that are essential at α for any $\alpha \in K \cup \{\infty\}$. An example of this is given by the equation $1 = [z^2 \ z + 1]H(z)$. The transfer function $[z^2 \ z + 1]$ has no zeros, finite or at infinity, and, since the left-hand member is 1, this implies that $\tilde{P}_S = 0$ for any S . However, no solution without poles can exist, since no linear combination with coefficients in K of z^2 and of $z + 1$ is equal to 1. Nevertheless, if

$G(z)$ is surjective or $F(z)$ is injective, the unique solution $H(z)$ of (3.1) or (3.2) is essential in any S , since $j: P_S \rightarrow X_S(H)$ or $\pi: X_S(H) \rightarrow \tilde{P}_S$ turn out to be respectively surjective or injective.

5. Conclusion. It has been shown that for each solution $H(z)$ of the equation $T(z) = H(z)G(z)$ and for each subset S of $K \cup \{\infty\}$ there exists an exact sequence of O_S -modules $0 \rightarrow P_S \rightarrow X_S(H) \rightarrow C_S \rightarrow 0$ where $X_S(H)$ is the module of pole in S of $H(z)$.

As suggested by the notation, P_S does not depend on the particular $H(z)$, but only on the data $T(z)$ and $G(z)$. It represents the fixed poles which appear in every solution and, as an abstract module, it is described by $P_S = G^{-1}(\Omega_S W) / (G^{-1}(\Omega_S W) \cap T^{-1}(\Omega_S Y))$.

A dynamical interpretation in module theoretic terms of the fixed poles is given by the exact sequence $0 \rightarrow X_S \rightarrow P_S \rightarrow Z_S \rightarrow 0$, where X_S and Z_S represent, respectively, the poles of $T(z)$ which are not poles of $G(z)$ and the zeros of $G(z)$ which are not zeros of $[T'(z)G'(z)]'$.

The invariant factors of P_S , which form the fixed part of the pole structure in S of any solution, are explicitly computed by means of the procedure of § 3.6, using fractional representations of $T(z)$ and $G(z)$.

The cokernel $C_S(H)$ represents the inessential poles in the solution. It is subject only to a technical restriction on the number of the invariant factors, while the location of the poles in $C_S(H)$ is arbitrary. These results give a complete description of the pole structure in S of $H(z)$.

Example 5.1. Consider the transfer functions:

$$T(z) = \begin{pmatrix} 1/(z+2) & -2/z(z+2) \\ 1/2(z+1) & -1/2z(z+1) \end{pmatrix},$$

$$G(z) = \begin{pmatrix} (z+2)/z(z+1) & (z+2)/z(z+1) \\ 1/z & (z^2+z+1)/z(z+1)^2 \\ (z+2)/z(z+1) & (z+2)/z(z+1) \end{pmatrix}.$$

Condition (A) is trivially satisfied since both $T(z)$ and $G(z)$ are full rank. Taking $S = K$, we compute the invariant factors of P_S , the module of finite fixed poles of the equation $T(z) = H(z)G(z)$, using the procedure viewed in § 3.6. Coprime fractional representations for $T(z)$ and $G(z)$ are the following: $T(z) = N(z)D^{-1}(z)$ and $G(z) = N_G(z)D_G^{-1}(z)$, where

$$N(z) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad D(z) = \begin{pmatrix} 2z & 3z+2 \\ -2z & z^2 \end{pmatrix},$$

$$N_G(z) = \begin{pmatrix} z+2 & 0 \\ z+1 & 1 \\ z+2 & 0 \end{pmatrix}, \quad D_G(z) = \begin{pmatrix} z^2+z & (z+1)^2 \\ 0 & -(z+1)^2 \end{pmatrix}.$$

The minimum common left multiple of $D(z)$, $D_G(z)$ is given by $M(z) = D(z)A(z) = D_G(z)B(z)$

$$M(z) = \begin{pmatrix} z(z+1)(z+2) & 2z(z+1)^2 \\ 0 & -2z(z+1)^2 \end{pmatrix}$$

with

$$A(z) = \begin{pmatrix} z^2/2 & (z+1)^2 \\ z & 0 \end{pmatrix}, \quad B(z) = \begin{pmatrix} z+2 & 0 \\ 0 & 2z \end{pmatrix}.$$

Then, since

$$N(z)A(z) = \begin{pmatrix} z(z+1) & 2(z+1)^2 \\ z(z+2)/2 & (z+1)^2 \end{pmatrix} \quad \text{and} \quad N_G(z)B(z) = \begin{pmatrix} (z+2)^2 & 0 \\ (z+1)(z+2) & 2z \\ (z+2)^2 & 0 \end{pmatrix}$$

turn out to be right coprime, their maximum common right divisor $S(z)$ is unimodular; hence $P_S \simeq (N_GB)^{-1}(\Omega_S W)/\Omega_S U \simeq \text{Tor}(\Omega_S W/N_GB(\Omega_S U))$ and the invariant factors of P are the nontrivial elements in the Smith form over O_S of $N_G(z)B(z)$, which is given by

$$\begin{pmatrix} 1 & 0 \\ 0 & z(z+2)^2 \\ 0 & 0 \end{pmatrix}.$$

This means that any solution $H(z)$ has a pole of total order 1 at 0, whose structure is $(0, 1)$, and a pole of total order 2 at -2 , whose structure is $(0, 2)$. This last one is due to the presence of a pole of total order 1 at -2 in $T(z)$, which does not appear in $G(z)$, and of a zero of total order 1 at -2 in $G(z)$, which does not appear in $[T'(z) \ G'(z)]^t$. Note that in this example the pole at 0 which appears in $T(z)$ and the one which appears in $G(z)$, both of order 1, are "distinct", in the sense that they give rise to a pole of order 2 in $[T'(z) \ G'(z)]^t$. This causes the presence of the pole at 0 of order 1 in every $H(z)$. Modifying the numerator matrix of $G(z)$ into

$$\bar{N}_G = \begin{pmatrix} 1 & 0 \\ z+1 & z+2 \\ 1 & 0 \end{pmatrix},$$

we obtain a second example in which the pole and zero structure of the data are the same as before. However, the computation shows that the invariant factors of P_S , in this case, are given by the Smith matrix

$$\begin{pmatrix} z+2 & 0 \\ 0 & z(z+2) \\ 0 & 0 \end{pmatrix}.$$

Thus, we have, as before, a pole of order 1 at 0 and a pole of order 2 at -2 in every solution $H(z)$, but now the structure at -2 , which is $(1, 1)$, is not cyclic. The methods mentioned in § 1.1 are not capable of pointing out this difference between the two cases.

REFERENCES

- [1] M. F. ATIYAH AND I. G. MACDONALD, *Introduction to Commutative Algebra*, Addison-Wesley, Reading, MA, 1969.
- [2] G. CONTE AND A. M. PERDON, *Infinite zero module and infinite pole module*, Proc. VII Internat. Conference on Analysis and Optimization of Systems, Nice, Lecture Notes in Control and Information Science 62, Springer-Verlag, 1984, pp. 302-315.
- [3] ———, *Zero module and factorization problems*, Cont. Math., 47 (1985), pp. 81-94.
- [4] ———, *On the causal factorization problem*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 811-813.
- [5] ———, *On the minimum delay problem*, Systems Control Lett., 5 (1985), pp. 213-215.
- [6] ———, *Zero and pole module in linear system design*, Proc. Colloque Σ_∞ , Paris, June, 1986.
- [7] E. EMRE, *Generalized model matching and (F, G) -invariant submodules for linear systems over rings*, Linear Algebra Appl., 50 (1983), pp. 133-166.
- [8] E. EMRE AND M. L. J. HAUTUS, *A polynomial characterization of (A, B) -invariant and reachability subspaces*, this Journal, 18 (1980), pp. 420-436.

- [9] G. D. FORNEY, *Minimal basis of rational vector spaces with applications to multivariable linear systems*, this Journal, 13 (1975), pp. 493-520.
- [10] J. HAMMER AND M. HEYMANN, *Strictly observable linear systems*, this Journal, 21 (1983), pp. 1-16.
- [11] ———, *Causal factorization and linear feedback*, this Journal, 19 (1981), pp. 445-468.
- [12] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
- [13] S. KUNG AND T. KAILATH, *Some notes on valuation theory in linear systems*, Proc. IEEE Conference on Decision and Control, San Diego, CA, 1978, pp. 515-517.
- [14] C. C. MACDUFFEE, *The Theory of Matrices*, Chelsea, New York, 1968.
- [15] M. MALABRE, *Sur le rôle de la structure l'infini et des sous-espaces presque invariants dans la résolution de problèmes de commande*, Thèse d'état, Université de Nantes, France, 1985.
- [16] A. S. MORSE, *System invariants under feedback and cascade control*, in *Mathematical System Theory*, Udine, 1975, Lecture Notes in Economic and Mathematical Systems 131, Springer-Verlag, New York, 1976.
- [17] ———, *Structure and design of linear model following systems*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 346-354.
- [18] ———, *Minimal solutions to transfer matrix equations*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 131-133.
- [19] S. H. WANG AND E. J. DAVISON, *A minimization algorithm for the design of linear multivariable systems*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 220-225.
- [20] W. A. WOLOVICH, P. ANTSAKLIS AND H. ELLIOT, *On the stability of solutions to minimal and nonminimal design problems*, IEEE Trans. Automat. Control, AC-27 (1977), pp. 88-94.
- [21] B. F. WYMAN, G. CONTE AND A. M. PERDON, *Local and global linear system theory*, in *Frequency Domain and State Space Methods for Linear Systems*, C. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1985, pp. 165-184.
- [22] B. WYMAN AND M. SAIN, *The zero module and essential inverse systems*, IEEE Trans. Circuits and Systems, CAS-28 (1981), pp. 112-126.
- [23] ———, *On the design of pole modules for inverse systems*, IEEE Trans. Circuits and Systems, CAS-32 (1985), pp. 977-988.

CHARACTERIZATION OF ALL CONTROLLED INVARIANT SUBSPACES FOR SPECTRAL SYSTEMS*

HANS ZWART†

Abstract. For a class of spectral systems a complete characterization of all controlled invariant subspaces contained in a closed subspace will be given. This characterization is given in terms of the invariant zeros of the transfer function. As a consequence of this we derive necessary and sufficient conditions for the existence of $V^*(\text{Ker } C)$, the largest controlled invariant subspace in the kernel of C .

Key words. infinite-dimensional linear systems, controlled invariance, invariant zeros, discrete spectral operators

AMS(MOS) subject classifications. primary 93C25; secondary 47A15

1. Introduction. In this paper we shall consider the following linear controlled system described by the set of equations:

$$(1.1a) \quad \dot{x} = Ax + Bu,$$

$$(1.1b) \quad y = Cx$$

where $x \in H$, $u \in U$ and $y \in Y$. H and Y are separable Hilbert spaces and U is \mathbb{C} or \mathbb{R} ; we take it to be the same as the field of H . Furthermore A is assumed to be a generator of a C_0 semigroup, $Bu = b \cdot u$ with b in H and C is a bounded linear operator from H to Y .

In 1969 Basile and Marro [1] introduced the concept of controlled invariance for the case that H and Y are finite-dimensional; starting in a subspace it is possible to find a control such that the system stays in that subspace. This concept catalyzed the beginning of the geometric theory, which led to the solution of various control problems (see Wonham [19]).

The concept of controlled invariance for infinite-dimensional Hilbert spaces was introduced by Schmidt and Stern in 1980 [16], and later this theory was extended by Pandolfi [12] and Curtain [4].

One concept that plays an important role in the geometric theory is that of the largest controlled invariant subspace in the kernel of C , usually denoted by $V^*(\text{Ker } C)$. From Basile and Marro [1] it follows that $V^*(\text{Ker } C)$ always exists if H is finite-dimensional. If H is infinite dimensional, however, this does not necessarily hold, as was shown by Pandolfi [12], at least for retarded systems. In Curtain [4] sufficient conditions were given such that $V^*(\text{Ker } C)$ exists.

The aim of this paper is to give sufficient and necessary conditions for the existence of $V^*(\text{Ker } C)$ for the class of spectral systems. We shall express this in terms of the invariant zeros of the transfer function $C(sI - A)^{-1}B$.

That there exists a close relationship between invariant subspaces in $\text{Ker } C$ and invariant zeros is well known in finite dimensions and can be illustrated by the following problem. We can ask a very simple question. What is the form of all one-dimensional controlled invariant subspaces in the kernel of C ?

* Received by the editors May 12, 1986; accepted for publication (in revised form) May 6, 1987. This research was supported by the Netherlands Organization for the Advancement of Pure Scientific Research (Z.W.O.).

† Institute of Mathematics, University of Groningen, P.O. Box 800, 9700 AV, Groningen, the Netherlands.

Let $\text{span}\{e\}$ be such a subspace; then since it is controlled invariant it is also $A+BF$ invariant, for some feedback law F (see Basile and Marro [1]). Thus e is an eigenvector of $A+BF$. We shall distinguish between two cases:

- (i) $Fe=0$. In this case e is an eigenvector of A and in the kernel of C .
- (ii) $Fe \neq 0$. Without loss of generality we may assume that $Fe=1$, so that $(A+BF)e = \alpha \cdot e$ implies $(\alpha I - A)e = b$. If we assume that α is in $\rho(A)$, then $e = (\alpha I - A)^{-1}b$. Since e is in the kernel of C , we obtain $C(\alpha I - A)^{-1}b = 0$. Thus α is a zero of the transfer function. So we see that a one-dimensional controlled invariant subspace is either an eigenvector in the kernel of C or it is $\text{span}\{e\}$ where e satisfies the equations $Ce=0$ and $(\alpha I - A)e = b$ for some α in \mathbb{C} . And if this α is an element of the resolvent set of A , then it is a zero of the transfer function.

For finite-dimensional systems the concept of zeros is very well understood. In infinite dimensions, however, only a few articles have been published; see Pohjolainen [14].

The proof of the existence of $V^*(\text{Ker } C)$ is based on the result of pole placement from Sun [17]. To get an idea that there exists a relationship between the existence of $V^*(\text{Ker } C)$ and the problem of pole placement we refer to the finite-dimensional case. It is well known (Wonham [19]) that in this case $\sigma(A+BF|_{V^*(\text{Ker } C)})$ is fixed for all F subject to $(A+BF)V^*(\text{Ker } C) \subset V^*(\text{Ker } C)$. In this paper we shall prove a similar result for spectral systems, see Theorem 6.1. Otherwise we have from the results of Sun [17] a restriction on $\sigma(A+BF)$, and hence especially on the fixed part of $\sigma(A+BF|_{V^*(\text{Ker } C)})$ and therefore on $V^*(\text{Ker } C)$.

The organization of the paper will be as follows. In § 2 we will recall some facts and properties of discrete spectral operators. In § 3 various concepts of invariance and of system invariance will be discussed. We will derive some properties of these concepts and for the class of discrete spectral operators we will give a full description of all invariant subspaces. In § 4 we will give the definition of invariant zeros. Properties of these invariant zeros for the class of spectral systems will be given in § 5. In § 6 the main theorems of this paper will be presented. In this section we will give a full description of all closed loop invariant subspaces in the kernel of C . In particular we will give necessary and sufficient conditions for the existence of $V^*(\text{Ker } C)$. Application of these theorems will be given in § 7. The examples in this section were calculated by L. Nooitgedagt [11].

2. Discrete spectral operators.

DEFINITION 2.1. Discrete operator. A linear operator A from \underline{H} to \underline{H} is discrete if there exists a number λ in its resolvent set for which the resolvent $R(\lambda; A) := (\lambda I - A)^{-1}$ is compact.

LEMMA 2.2. If A is discrete, then:

- (a) Its spectrum, $\sigma(A)$, is a denumerable set of points with no finite limit point.
- (b) The resolvent $R(\lambda, A)$ is compact for every λ not in $\sigma(A)$.
- (c) Every λ_0 in $\sigma(A)$ is a pole of finite order $\theta(\lambda_0)$ of the resolvent and if, for some positive integer k , x satisfies the equation

$$(A - \lambda_0 I)^k x = 0$$

then x satisfies the equation

$$(A - \lambda_0 I)^{\theta(\lambda_0)} x = 0.$$

The set of all vectors x satisfying the equation $(A - \lambda_0 I)^{\theta(\lambda_0)} x = 0$ is a finite-dimensional linear space, called the space of generalized eigenvectors of A corresponding to the eigenvalue λ_0 .

(d) If

$$(2.1) \quad E(\lambda_0) = \frac{1}{2\pi i} \cdot \int_{\Gamma} (\lambda I - A)^{-1} d\lambda,$$

where Γ is a small closed curve surrounding only the eigenvalue λ_0 and Γ is traversed once in the positive sense, then $E(\lambda_0)$ projects H onto the space of generalized eigenvectors corresponding to λ_0 .

Proof. See Lemma XIX 2.2 of Dunford and Schwartz [8].

Remark. The spectrum of A shall be denoted by $\{\lambda_n\} n \geq 1$.

DEFINITION 2.3. Discrete spectral operator. A discrete operator is spectral if the spectral projections $E(\lambda_j)$ defined by (2.1) satisfy:

(a) The family of sums of finite collections of projections $E(\lambda_j)$ is uniformly bounded; and

(b) No nonzero x in H satisfies all of the equations $E(\lambda_j)x = 0$, λ_j in $\sigma(A)$.

Remark. The spectral projections $E(\lambda_j)$ are not necessarily selfadjoint.

LEMMA 2.4. If A is a discrete spectral operator then the spectral projections $\{E(\lambda_j), \lambda_j$ in $\sigma(A)\}$ generate a uniformly bounded Boolean algebra with the completeness property:

$$(2.2) \quad \sum_{j=1}^{\infty} E(\lambda_j) = I$$

where the convergence is in the strong topology.

Proof. See § XVIII.1 of Dunford and Schwartz [8].

3. Invariant subspaces. In this section we shall discuss some concepts of (system) invariance. Let A be a generator of a C_0 semigroup, $T(t)$, on H . For a generator A we can define two concepts of invariance: semigroup invariance and generator invariance. In finite dimensions these two concepts are equivalent but as we shall see they are in general not equivalent for unbounded generators of C_0 semigroups. We shall first recall the definition of semigroup and generator invariance.

DEFINITION 3.1. Semigroup invariance or $T(t)$ -invariance. A closed linear subspace Y of H will be called semigroup invariant or $T(t)$ -invariant if $T(t)Y \subset Y$, for all $t \geq 0$.

DEFINITION 3.2. Generator invariance or A -invariance. A closed linear subspace Y of H will be called generator or A -invariant if $A(Y \cap D(A)) \subset Y$ where $D(A)$ is the domain of A .

We will recall some basic facts about these definitions. Semigroup invariance always implies generator invariance but the converse is in general not true for unbounded generators; see, for example, Schmidt and Stern [16] for a counterexample. But there is a class of subspaces where the two concepts of invariance is equivalent.

LEMMA 3.3. If Y is a closed linear subspace in the domain of A then generator invariance is equivalent to semigroup invariance.

Proof. See Lemma 2.3 of Curtain [4].

Since the generator A is completely determined by $T(t)$ and vice versa it may seem strange that the two concepts of invariance are not equivalent. But, as in the Hille-Yosida theorem, the relation between A and $T(t)$ -invariance is determined by the resolvent of A . Define \underline{C}_{∞} to be the largest connected subset of $\rho(A)$ that contains an interval of the form $[r, +\infty)$; since A generates a C_0 semigroup this set is nonempty.

LEMMA 3.4. Let Y be a closed linear subspace of H ; then the following concepts are equivalent:

- (a) Y is semigroup invariant;
- (b) $(\lambda I - A)^{-1}Y \subset Y$ for a λ in \underline{C}_{∞} ;

- (c) $(\lambda I - A)^{-1}Y \subset Y$ for all λ in C_∞ ;
 (d) The range of $(\lambda I - A)$ restricted to V is V for all $\lambda \in C_\infty$.

Proof.

(a) \rightarrow (b): see Pazy [13, pp. 121].

(b) \rightarrow (c): see Kurtz [10].

(c) \rightarrow (a): see Pazy [13, pp. 121].

(a) \leftrightarrow (d): see Kurtz [10].

LEMMA 3.5. Let Y be a closed linear subspace which is $T(t)$ -invariant; then $\overline{Y \cap D(A)} = Y$.

Proof. From p. 37 of Davies [6] we have that every element x in H is the limit of $\lambda \cdot (\lambda I - A)^{-1}x$, $\lambda \rightarrow +\infty$.

Assume that $x \in Y$; then by the previous lemma we have, for λ sufficiently large, that $(\lambda I - A)^{-1}x$ is in $Y \cap D(A)$.

Combining these results we have that x is the limit of a sequence in $Y \cap D(A)$. So $\overline{Y \cap D(A)} = Y$.

If a subspace Y of R^n is invariant with respect to a diagonal matrix then Y must be of the form $\text{span} \{e_i, i \in I \subset \{1 \cdots n\}\}$ where e_i is the i th basis vector of R^n . For the class of discrete spectral operators a similar theorem holds. First we shall recall a lemma of Dunford and Schwartz [8] that gives some invariance properties of the spectral projections, $E(\lambda_i)$.

LEMMA 3.6. Let A be a discrete spectral operator and λ_j an element of $\sigma(A)$. Then $D(A) \supset E(\lambda_j)H$, the subspace $E(\lambda_j)H$ is A -invariant, $AE(\lambda_j)x = E(\lambda_j)Ax$ for all x in $D(A)$ and $\sigma(A|E(\lambda_j)H) = \{\lambda_j\}$.

Proof. See Dunford and Schwartz [8, pp. 2294].

THEOREM 3.7. Let the discrete spectral operator A generate the C_0 semigroup $T(t)$; then a closed linear subspace Y of H is $T(t)$ -invariant if and only if

$$(3.1) \quad Y = \sum_{j=1}^{\infty} W_j$$

where W_i is a subspace of H which is contained in $E(\lambda_i)H$ and is A -invariant.

The summation (3.1) is in the strong topology, i.e., for all $x \in Y$ there exist $\{w_i; i \in \mathbb{N}\}$, with $w_i \in W_i$ such that $\sum_{i=1}^n w_i$ converges to x for $n \rightarrow \infty$; otherwise if $\{w_i, i \in \mathbb{N}\}$; $w_i \in W_i$ is such that $\sum_{i=1}^n w_i$ converges for $n \rightarrow \infty$, then the limit is in Y .

Furthermore the spectrum of A restricted to V is equal to the set of all $\lambda_i \in \sigma(A)$ subject to the corresponding W_i is not the zero subspace.

Proof. (If.) Since the dimension of $E(\lambda_i)H$ is finite, the dimension of W_i must also be finite. So W_i is a closed linear subspace of H . Furthermore $E(\lambda_i)H$ is contained in $D(A)$, and with Lemma 3.6 and Lemma 3.3 we may conclude that W_i is $T(t)$ -invariant.

Every x in Y is the limit of a sequence x_n , with $x_n \in \sum_{i=1}^n W_i$. So with the above we have $T(t)x_n$ is in $\sum_{i=1}^n W_i$. $T(t)$ is a bounded linear operator; thus $T(t)x_n$ converges to an element in H , but also to an element of Y since $T(t)x_n$ is in Y . So Y is $T(t)$ -invariant.

(Only if.) Let Y be a $T(t)$ -invariant subspace. Since $\sigma(A) = \{\lambda_i\}$, $i \in \mathbb{N}$ we have that the resolvent set, $\rho(A)$, is connected. With Lemma 3.4 this implies that Y is also $(\lambda I - A)^{-1}$ invariant for all λ in the resolvent set of A . So

$$E(\lambda_i)Y = \frac{1}{2\pi i} \cdot \int_{\Gamma} (\lambda \cdot I - A)^{-1} Y d\lambda \subset Y \quad (\text{see (2.1)}).$$

So $E(\lambda_i)\underline{Y} \subset (E(\lambda_i)\underline{H}) \cap \underline{Y}$. Using the fact that $E(\lambda_i)$ is a projection we get $(E(\lambda_i)\underline{H}) \cap \underline{Y} = (E(\lambda_i)\underline{H}) \cap (E(\lambda_i)\underline{Y}) \subset E(\lambda_i)\underline{Y}$. Thus $E(\lambda_i)\underline{Y} = (E(\lambda_i)\underline{H}) \cap \underline{Y}$.

If we set \underline{W}_i equal to $E(\lambda_i)\underline{Y}$ and $x_i := E(\lambda_i)x$; $x \in \underline{Y}$, then $A(\underline{W}_i) = AE(\lambda_i)\underline{Y} = AE(\lambda_i)E(\lambda_i)\underline{Y} = AE(\lambda_i)(E(\lambda_i)\underline{H} \cap \underline{Y}) \subset AE(\lambda_i)(\underline{Y} \cap D(A)) = E(\lambda_i)A(\underline{Y} \cap D(A)) \subset E(\lambda_i)\underline{Y} = \underline{W}_i$ (see Lemma 3.6). So $A\underline{W}_i \subset \underline{W}_i$.

By (2.2) and Definition 2.3 every x in \underline{H} can be uniquely written as $\sum_{i=1}^{\infty} x_i$. Hence $\underline{Y} = \sum_{i=1}^{\infty} \underline{W}_i$.

We shall now prove the last assertion.

Let \underline{J} denote the index set of all $\lambda_i \in \sigma(A)$ subject to $\dim(\underline{W}_i)$ is larger than zero. Since $\underline{W}_i \subset E(\lambda_i)\underline{H}$ is A invariant it must contain an eigenvector corresponding to λ_i ; thus $\{\lambda_i; i \in \underline{J}\} \subset \sigma(A|_{\underline{Y}})$. From Lemma 3.4 we have for that all $\lambda \in C_{\infty}$, $(\lambda I - A)^{-1}|_{\underline{Y}}$ is a bounded linear operator from \underline{Y} to \underline{Y} and since A is discrete it is also a compact operator. Furthermore, it is the inverse of $(\lambda I - A|_{\underline{Y}})$. So $A|_{\underline{Y}}$ is a discrete operator and hence the spectrum of $A|_{\underline{Y}}$ is a pure point spectrum. Now it is easy to show that $\{\lambda_i; i \in \underline{J}\} = \sigma(A|_{\underline{Y}})$.

With this theorem we can prove an interesting corollary.

Let A be of the following form: $A = \sum_{i=1}^{\infty} \lambda_i \langle \cdot, \phi_i \rangle_{\underline{H}} \cdot \phi_i$, with $\{\phi_i\}$ an orthonormal basis of \underline{H} , $\lambda_i \in \mathbb{R}$ and $\sum_{i=1}^{\infty} \lambda_i^{-2} < \infty$. If we set the domain of A equal to all x in \underline{H} such that $\sum_{i=1}^{\infty} |\lambda_i \langle x, \phi_i \rangle_{\underline{H}}|^2$ exists, then A is a discrete spectral operator with spectral projections $E(\lambda_i) = \langle \cdot, \phi_i \rangle \cdot \phi_i$. Furthermore if $\sup \{\lambda_i | i \in \mathbb{N}\} < \infty$, then A generates a C_0 semigroup $T(t)$ (see Curtain [3]) and

$$T(t) = e^{At} = \sum_{i=1}^{\infty} e^{\lambda_i t} \langle \cdot, \phi_i \rangle_{\underline{H}} \cdot \phi_i.$$

Let C be a bounded linear operator from \underline{H} to a Hilbert space \underline{Y} . This operator can be seen as an observation. An important concept in system theory is the nonobservable subspace \underline{Y}_0 , i.e., all trajectories that are nonobservable.

COROLLARY 3.8. *If A and C satisfy the properties as stated above, then \underline{Y}_0 is $\overline{\text{span}}\{\phi_i | C(\phi_i) = 0\}$. So $\underline{Y}_0 = \{0\}$ iff $C(\phi_i) \neq 0$ for all i .*

Proof. From Curtain [4] we obtain that \underline{Y}_0 is the largest subspace of \underline{H} that is semigroup invariant and in the kernel of C . From Theorem 3.7 and the special structure of A we have that $\underline{Y}_0 = \overline{\text{span}}\{\phi_i | i \in \underline{J} \subset \mathbb{N}; C(\phi_i) = 0\}$. Combining these results for \underline{Y}_0 we conclude this corollary.

We can also define a controlled version of generator invariance. Let b be an element in \underline{H} . With this element we can define an input operator B , i.e., $Bu = b \cdot u$ where u is an element of \underline{U} , the input space. Here we shall assume that \underline{U} is \mathbb{R} or \mathbb{C} .

DEFINITION 3.9. Bounded closed loop invariance. A closed linear subspace \underline{Y} of \underline{H} is called (bounded) closed loop invariant if there exists a bounded linear feedback law F from \underline{H} to \underline{U} such that $T_F(t)\underline{Y} \subset \underline{Y}$ for all $t \geq 0$. $T_F(t)$ is the semigroup generated by $A + BF$.

Remark. If there is no doubt about the feedback law then we will simply use $T_F(t)$ -invariance.

DEFINITION 3.10. Generator feedback invariance. A closed linear subspace \underline{Y} will be called generator feedback invariant if there exists a bounded linear feedback law F from \underline{H} to \underline{U} such that $(A + BF)(\underline{Y} \cap D(A)) \subset \underline{Y}$.

DEFINITION 3.11. (A, B) invariance. A closed linear subspace \underline{Y} will be called (A, B) -invariant if $A(\underline{Y} \cap D(A)) \subset \underline{Y} + \text{Im } B$.

If \underline{H} is finite-dimensional, then these three concepts of invariance are equivalent and a subspace which satisfies one of these concepts is called controlled invariant. However, if \underline{H} is infinite dimensional and A is an unbounded operator, then these

concepts no longer are equivalent (see Schmidt and Stern [16]) and here we shall mean by controlled invariance closed loop invariance.

In Curtain [4] the relation among Definitions 3.9–3.11 is studied. From this we have the following lemma.

LEMMA 3.12. *If V is a closed linear subspace contained in $D(A)$, then closed loop, generator feedback- and (A, B) -invariance are equivalent.*

Proof. See Lemma 4.6 of Curtain [4].

DEFINITION 3.13. $V^*(K)$. Let K be a closed linear subspace in H . By $V^*(K)$ we shall denote the largest (bounded) closed loop invariant subspace in K .

Remark. In general $V^*(K)$ need not exist; see, for example, Pandolfi [12].

Since we do not have equivalence between (A, B) - and closed loop invariance, we must derive new proofs for results that, in the finite dimensional case, were proved by this equivalence. One of these results is given in the next lemma.

LEMMA 3.14. *Let V_1 and V_2 be closed linear subspaces of H that are (bounded) closed loop invariant and assume further that V_2 is finite dimensional; then $V_1 + V_2$ is also (bounded) closed loop invariant.*

Proof. By the definition of closed loop invariance we have that there exist F_1 and F_2 both bounded such that $T_{F_i}(t)V_i \subset V_i$, $i = 1, 2$. Define F_0 to be such that

$$\begin{aligned} F_0|_{V_1} &= F_1|_{V_1}, \\ F_0|_{V'_2} &= F_2|_{V'_2} \quad \text{where } V'_2 = V_2 \cap (V_1 \cap V_2)^\perp, \\ F_0|_{(V_1 + V_2)^\perp} &= 0. \end{aligned}$$

This F_0 is a bounded linear operator. So there is a real λ_0 such that $[\lambda_0, \infty) \subset \rho(A + BF_i)$, $i = 0, 1, 2$. For a $\lambda \in [\lambda_0, \infty)$ we have

$$(3.2) \quad \begin{aligned} (\lambda I - A - BF_0)^{-1} &= (\lambda I - A - BF_i)^{-1} + (\lambda I - A - BF_0)^{-1} \\ &\quad \cdot B(F_0 - F_i)(\lambda I - A - BF_i)^{-1}, \quad i = 1, 2. \end{aligned}$$

If x is an element of V_1 then, by Lemma 3.4, we have that $(\lambda I - A - BF_1)^{-1}x$ is in V_1 . So by (3.2) and the definition of F_0 we have that for $x \in V_1$

$$(3.3) \quad (\lambda I - A - BF_0)^{-1}x = (\lambda I - A - BF_1)^{-1}x,$$

so $(\lambda I - A - BF_0)^{-1}V_1 \subset V_1$.

Since V_2 is closed loop invariant we have from Lemma 3.5 that $D(A)$ is dense in V_2 . And since V_2 is finite-dimensional, it must be a closed subspace contained in $D(A)$. If x is an element of V_2 , then it can be written as $x = x_1 + x_2$, where $x_1 \in V_1 \cap V_2 \subset D(A)$ and $x_2 \in V'_2 \subset D(A)$. And so we have that $B(F_0 - F_2)x = B(F_0 - F_2)(x_1 + x_2) = B(F_1 - F_2)x_1 + 0 = (A + BF_1)x_1 - (A + BF_2)x_1$. From this equation we see that $B(F_0 - F_2)V_2 \subset (V_1 + V_2) \cap \text{Im } B$. Since $B = b$ we must consider two cases.

(i): $b \notin V_1 + V_2$. In this case we have that $B(F_0 - F_2)V_2 = 0$. So with (3.2) we have

$$(3.4) \quad (\lambda I - A - BF_0)^{-1}x = (\lambda I - A - BF_2)^{-1}x \quad \text{for all } x \text{ in } V_2.$$

So $(\lambda I - A - BF_0)^{-1}(V_1 + V_2) \subset (V_1 + V_2)$ and with Lemma 3.4 we conclude that $(V_1 + V_2)$ is closed loop invariant.

(ii): $b \in V_1 + V_2$. In this case we can decompose b in $b_1 + b_2$ with $b_1 \in V_1$ and $b_2 \in V'_2$. Choose λ such that $|\tau| < 1$; where $\tau = (F_0 - F_2)(\lambda I - A - BF_2)^{-1}b_2$. This is possible since $A + BF_2$ generates a C_0 semigroup. From (3.3) we have that

$$(3.5) \quad (\lambda I - A - BF_0)^{-1}b_1 = (\lambda I - A - BF_1)^{-1}b_1$$

and from (3.2) we have that $(\lambda I - A - BF_0)^{-1}b_2 = (\lambda I - A - BF_2)^{-1}b_2 + (\lambda I - A - BF_0)^{-1} \cdot B(F_0 - F_2)(\lambda I - A - BF_2)^{-1}b_2 = (\lambda I - A - BF_2)^{-1}b_2 + \tau \cdot (\lambda I - A - BF_0)^{-1} \cdot b$. Adding this last equation to (3.5), we obtain

$$(3.6) \quad (\lambda I - A - BF_0)^{-1}b = (\lambda I - A - BF_1)^{-1}b_1 + (\lambda I - A - BF_2)^{-1}b_2 + \tau \cdot (\lambda I - A - BF_0)^{-1} \cdot b.$$

Now since $|\tau| < 1$

$$(3.7) \quad (\lambda I - A - BF_0)^{-1}b = (1 - \tau)^{-1} \cdot \{(\lambda I - A - BF_1)^{-1}b_1 + (\lambda I - A - BF_2)^{-1}b_2\}.$$

So $(\lambda I - A - BF_0)^{-1}b$ is an element of $Y_1 + Y_2$. Let x be an element of Y_2 ; then from (3.2), $(\lambda I - A - BF_0)^{-1}x = (\lambda I - A - BF_2)^{-1}x + (\lambda I - A - BF_0)^{-1}b \cdot (F_0 - F_2)(\lambda I - A - BF_2)^{-1}x$. Since $(F_0 - F_2)(\lambda I - A - BF_2)^{-1}x$ is in \underline{U} and $(\lambda I - A - BF_0)^{-1}b$ is in $Y_1 + Y_2$ we may conclude that $(\lambda I - A - BF_0)^{-1}x$ is in $Y_1 + Y_2$. Combining this with (3.3) we have proved this lemma.

Remark. This lemma can easily be generalized to the case that $\text{Im } B$ is finite-dimensional.

As in Theorem 3.7 we pose the question of what the $T_F(t)$ -invariant subspaces look like. This question is in general not solvable even if A is a discrete spectral operator since $A + BF$ need not be a discrete spectral operator. Furthermore we are not interested in all $T_F(t)$ -invariant subspaces but only in those which are in the kernel of C . Before we can give a complete description of all $T_F(t)$ -invariant subspace we must investigate the notion of the zeros of a transfer function.

4. Invariant zeros. In this section we will discuss the concept of invariant zeros which will play an important role in the investigation of $T_F(t)$ -invariant subspaces in $\text{Ker } C$, the kernel of C .

We will consider the following controlled system:

$$(4.1) \quad \begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx \end{aligned}$$

where A is a generator of a C_0 semigroup on \underline{H} , $Bu = bu$; $b \in \underline{H}$ and C is a bounded operator from \underline{H} to \underline{Y} , where \underline{Y} is a Hilbert space.

For the definition of invariant zeros it is convenient to introduce the following sequence of subspaces:

$$\begin{aligned} \mathcal{Z}_\mu^0 &:= \{0\}; \quad \mathcal{Z}_\mu^k := \{x \in D(A): \text{there exist } \gamma \text{ in } \mathbb{C} \text{ and } z \text{ in } \mathcal{Z}_\mu^{k-1} \\ &\quad \text{such that } (\mu I - A)x = \gamma \cdot b + z\} \cap \text{Ker } C, \quad k > 0. \end{aligned}$$

DEFINITION 4.1. Invariant zero. μ in \mathbb{C} is called an invariant zero if \mathcal{Z}_μ^1 is not the zero subspace.

Remark. If $\mu \in \rho(A)$ then μ is an invariant zero if and only if $C(\mu I - A)^{-1}b = 0$.

Remark. μ is an invariant zero if and only if there exist nonzero $x \in \underline{H}$ and $u \in \underline{U}$ such that

$$\begin{bmatrix} \mu - A & B \\ C \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = 0 \quad (\text{see Davison and Wang [7]}).$$

The next lemma will show that \mathcal{Z}_μ^k is a nested sequence of subspaces.

LEMMA 4.2. \mathcal{Z}_μ^k is a linear subspace in $\text{Ker } C$. $\mathcal{Z}_\mu^{k-1} \subset \mathcal{Z}_\mu^k$ and if $\mathcal{Z}_\mu^{k-1} = \mathcal{Z}_\mu^k$, then $\mathcal{Z}_\mu^{k+1} = \mathcal{Z}_\mu^k$; $k \geq 1$.

Proof. The first assertion can be proved by induction. We shall prove the second and third assertions. Since $\mathcal{Z}_\mu^0 = \{0\}$ we have proved the second assertion for $k = 1$. Suppose it holds for all $n \leq k$, then we will prove it for $n = k + 1$. Let x be an element of \mathcal{Z}_μ^k , then there exist $\gamma \in \mathcal{C}$ and $z \in \mathcal{Z}_\mu^{k-1}$ with $(\mu I - A)x = \gamma \cdot b + z$ and x is in $\text{Ker } C$; $z \in \mathcal{Z}_\mu^{k-1} \subset \mathcal{Z}_\mu^k$. So $x \in \mathcal{Z}_\mu^{k+1}$.

Suppose $\mathcal{Z}_\mu^{k-1} = \mathcal{Z}_\mu^k$. Let x be an element of \mathcal{Z}_μ^{k+1} ; thus $x \in \text{Ker } C$ and there exist $\gamma \in \mathcal{C}$ and $z \in \mathcal{Z}_\mu^k$ with $(\mu I - A)x = \gamma \cdot b + z$, $z \in \mathcal{Z}_\mu^k = \mathcal{Z}_\mu^{k-1}$. This implies that $x \in \mathcal{Z}_\mu^k$. So $\mathcal{Z}_\mu^{k+1} \subset \mathcal{Z}_\mu^k$ and $\mathcal{Z}_\mu^{k+1} \supset \mathcal{Z}_\mu^k$; thus $\mathcal{Z}_\mu^{k+1} = \mathcal{Z}_\mu^k$.

DEFINITION 4.3. Order (μ) . Let μ be in \mathcal{C} . By order (μ) we mean the smallest number $k \geq 0$ such that $\mathcal{Z}_\mu^{k+1} = \mathcal{Z}_\mu^k$. If such a number does not exist then order (μ) is ∞ .

Remark. μ is an invariant zero if and only if order (μ) is larger than zero.

DEFINITION 4.4. \mathcal{N}_μ^k . By \mathcal{N}_μ^k we mean the kernel of $(\mu I - A)^k$ intersected with the kernels of $C(\mu I - A)^1$ for $0 \leq 1 < k$. Thus $\mathcal{N}_\mu^k = \{x \in D(A^k) \text{ subject to } (\mu I - A)^k x = 0 \text{ and } C(\mu I - A)^1 x = 0 \text{ for } 0 \leq 1 < k\}$.

Remark. If \mathcal{N}_μ^1 is nonzero, then μ is called an output decoupling zero; see Davison and Wang [7].

The next lemma will give some properties of \mathcal{N}_μ^k similar to those of \mathcal{Z}_μ^k .

LEMMA 4.5. \mathcal{N}_μ^k has the following properties for $k \in \mathbb{N}$:

- (a) $\mathcal{N}_\mu^k \subset \mathcal{N}_\mu^{k+1}$.
- (b) If $\mathcal{N}_\mu^k = \mathcal{N}_\mu^{k+1}$, then $\mathcal{N}_\mu^{k+1} = \mathcal{N}_\mu^{k+2}$.
- (c) \mathcal{N}_μ^k is not the zero subspace for a $k > 0$ iff $\mu \in \sigma_p(A)$ and there exists a nonzero eigenvector, corresponding to μ , in the kernel of C .

Proof. The proof is the same as that for Lemma 4.2.

LEMMA 4.6. Suppose that $\mathcal{N}_\mu^k = \{0\}$, then $\mathcal{Z}_\mu^k = \text{span}\{x_1 \cdots x_n\}$ where n is the minimum of k and order (μ) and x_i satisfies:

$$(4.2) \quad \begin{aligned} x_i &\in \text{Ker } C, \\ (\mu I - A)x_1 &= b \quad \text{and} \\ (\mu I - A)x_i &= x_{i-1}. \end{aligned}$$

Proof. Let us first remark that since $\mathcal{N}_\mu^1 = \{0\}$, we have that the equation $(\mu I - A)x = y$ has at most one solution in $\text{Ker } C$.

If order $(\mu) = 0$, then $\mathcal{Z}_\mu^k = \{0\}$. So in this case the assertion is proved.

Assume now that order $(\mu) \neq 0$ and consider the assertion for $k = 1$. Then $n = 1$ and $(\mu I - A)x = b$ has exactly one solution $x_1 \in \text{Ker } C$ and the assertion holds for $k = 1$.

Now suppose that the assertion holds for $i \leq k - 1$. We will prove it for $i = k$.

If $k - 1 \geq \text{order}(\mu)$, then $\mathcal{Z}_\mu^{k-1} = \mathcal{Z}_\mu^k$ and $\min(k, \text{order}(\mu)) = \min(k - 1, \text{order}(\mu))$. So the assertion holds.

If $k - 1 < \text{order}(\mu)$, then $n := \min(k, \text{order}(\mu)) = k$ and $\mathcal{Z}_\mu^{k-1} \neq \mathcal{Z}_\mu^k$ so there exists $x \in \mathcal{Z}_\mu^k$ and $x \notin \mathcal{Z}_\mu^{k-1}$. Since $x \in \mathcal{Z}_\mu^k$ there exist $\gamma_0 \in \mathcal{C}$ and z in \mathcal{Z}_μ^{k-1} such that $(\mu I - A)x = \gamma_0 \cdot b + z$. By the induction hypothesis we have that there are constants γ_i such that $z = \sum_{i=1}^{k-1} \gamma_i \cdot x_i$. Since x is not in \mathcal{Z}_μ^{k-1} we must have that $\gamma_{k-1} \neq 0$. Defining x_k as $x_k := (x - \sum_{i=0}^{k-2} \gamma_i \cdot x_{i+1}) / \gamma_{k-1}$ we have

$$\gamma_{k-1} \cdot (\mu I - A)x_k = (\mu I - A)x - \sum_{i=0}^{k-2} \gamma_i \cdot (\mu I - A)x_{i+1}$$

(by the definition of x_i , $i < k$),

$$= \gamma_0 \cdot b + \sum_{i=1}^{k-1} \gamma_i \cdot x_i - \gamma_0 \cdot b - \sum_{i=1}^{k-2} \gamma_i \cdot x_i = \gamma_{k-1} \cdot x_{k-1}.$$

So x_k satisfies the conditions (4.2) and by the definition of x_k we have that $\text{span}\{x_1 \cdots x_k\} = \mathcal{Z}_\mu^{k-1} + \text{span}\{x_k\} \subset \mathcal{Z}_\mu^k$.

Now we shall show that all elements of \mathcal{Z}_μ^k are in $\text{span}\{x_1 \cdots x_k\}$. Suppose that x' is an arbitrary element in \mathcal{Z}_μ^k . Then by the definition of \mathcal{Z}_μ^k there exist $\gamma'_0 \in \mathbb{C}$ and $z' \in \mathcal{Z}_\mu^{k-1}$ such that $(\mu I - A)x' = \gamma'_0 \cdot b + z'$. By the induction hypothesis we have that

$$z' = \sum_{i=1}^{k-1} \gamma'_i \cdot x_i.$$

Defining y as $\gamma_{k-1}x' - \gamma'_{k-1}x$ we have that

$$\begin{aligned} (\mu I - A)y &= (\mu I - A)(\gamma_{k-1}x' - \gamma'_{k-1}x) \\ &= \gamma_{k-1} \left(\gamma'_0 \cdot b + \sum_{i=1}^{k-1} \gamma'_i \cdot x_i \right) - \gamma'_{k-1} \left(\gamma_0 \cdot b + \sum_{i=1}^{k-1} \gamma_i \cdot x_i \right) \\ &= (\gamma_{k-1} \cdot \gamma'_0 - \gamma'_{k-1} \cdot \gamma_0) \cdot b + \sum_{i=1}^{k-2} (\gamma_{k-1} \cdot \gamma'_i - \gamma'_{k-1} \cdot \gamma_i) x_i. \end{aligned}$$

By the definition of \mathcal{Z}_μ^{k-1} and the induction hypothesis for $i = k-2$, the above equation implies that $y \in \mathcal{Z}_\mu^{k-1}$. Thus

$$x' = (y + \gamma'_{k-1}x) / \gamma_{k-1} \in \mathcal{Z}_\mu^{k-1} + \text{span}\{x\} = \mathcal{Z}_\mu^{k-1} + \text{span}\{x_k\} = \text{span}\{x_1 \cdots x_k\}.$$

Since x' is an arbitrary element in \mathcal{Z}_μ^k we have proved that $\mathcal{Z}_\mu^k \subset \text{span}\{x_1 \cdots x_k\}$.

Thus $\mathcal{Z}_\mu^k = \text{span}\{x_1 \cdots x_k\}$.

COROLLARY 4.7. *If $\mu \in \rho(A)$, then $\mathcal{Z}_\mu^k = \text{span}\{(\mu I - A)^{-1}b, \dots, (\mu I - A)^{-n}b\}$, where n is the minimum of k and order (μ) .*

Proof. This is a simple corollary of Lemma 4.6.

The last lemma of this section and the first remark after Definition 4.1 will give an explication for the name invariant zero.

LEMMA 4.8. *If \mathcal{Z}_μ^k is finite-dimensional, then it is $T_F(t)$ -invariant.*

Proof. By the definition of \mathcal{Z}_μ^k this subspace is contained in $D(A)$. And by assumption we have that \mathcal{Z}_μ^k is finite-dimensional. So from Lemma 3.12 we only have to prove that \mathcal{Z}_μ^k is (A, B) -invariant.

Let x be an element of \mathcal{Z}_μ^k , then $(\mu - A)x = \gamma \cdot b + z$ where $z \in \mathcal{Z}_\mu^{k-1} \subset \mathcal{Z}_\mu^k$ and $\gamma \in \mathbb{C}$. So $Ax = \mu x - z - \gamma \cdot b$, and $A(\mathcal{Z}_\mu^k) \subset \mathcal{Z}_\mu^k + \text{Im } B$.

5. Invariant zeros for the class of spectral systems. In this section we shall consider system (4.1) with some additional assumptions and discuss the concept of invariant zeros for this special kind of system. The additional assumptions we make on the system (4.1) are:

(\nabla 1) The generator A is a discrete spectral operator with spectral decomposition

$$A = \sum_{i=1}^{\infty} \lambda_i \cdot E(\lambda_i)$$

where $\lambda_i \neq \lambda_j$ (for all $i \neq j$) and $\dim E(\lambda_i) = 1 (i \geq 1)$. Without loss of generality we may assume that the $E(\lambda_i) (i \geq 1)$ are selfadjoint operators in H (see Wermer [18]). The normalized eigenvector of A corresponding to $E(\lambda_i)$ will be denoted by $\phi_i (i \geq 1)$,

$$(\nabla 2) \quad b_i := \langle \phi_i, b \rangle_H \neq 0.$$

Let us remark that (\nabla 2) is the controllability assumption; see, e.g., Curtain and Pritchard [5].

If $\mu \in \sigma(A)$, then $(\lambda I - A)^{-1} = 1/(\lambda - \mu) \cdot E(\mu) + R_\mu(\lambda)$ where $R_\mu(\lambda)$ is analytic in μ and commutes with $E(\mu)$ (see Kato [9, III.6.5]). $R_\mu(\lambda)$ can be seen as the inverse of $\{\lambda - \sum_{i=1, \lambda_i \neq \mu}^{\infty} \lambda_i E(\lambda_i)\}$. Define P_μ as $I - E(\mu)$.

With assumptions $(\nabla 1)$ and $(\nabla 2)$ we shall give a full description of \mathcal{Z}_μ^k .

LEMMA 5.1. *Let $\mu \in \sigma(A)$ be an invariant zero; then $\mathcal{Z}_\mu^k = \text{span} \{x_1, \dots, x_n\}$ where n is the minimum of k and order (μ) and x_i satisfies:*

- (5.1) (a) $x_i \in \text{Ker } C$,
 (b) $x_1 = \phi$; ϕ is the eigenvector corresponding to μ , and
 (c) $x_i = (R_\mu(\mu))^{i-1} P_\mu b$; $i > 1$.

Proof. Let $x \in \mathcal{Z}_\mu^1$; then $(\mu I - A)x = \gamma \cdot b$ for some $\gamma \in \mathbb{C}$. Since $\mu \in \sigma(A)$, there exists a nonzero vector ϕ such that $(\mu I - A)\phi = 0$. So $0 = \langle (\mu I - A)\phi, x \rangle = \langle \phi, (\mu I - A)x \rangle = \langle \phi, b \rangle \cdot \gamma$. With $(\nabla 2)$ we conclude that $\gamma = 0$. Thus $(\mu I - A)x = 0$. This implies that $x \in N_\mu^1$ and that x is an eigenvector corresponding to μ . The multiplicity of all eigenvalues is one so $x \in \text{span} \{\phi\} = N_\mu^1$. Thus we have $\mathcal{Z}_\mu^1 = N_\mu^1 = \text{span} \{\phi\}$. And so the assertion holds for $k = 1$.

Suppose that k is larger than 0. Then since $\mathcal{Z}_\mu^1 \subset \mathcal{Z}_\mu^k$ and $P_\mu \mathcal{Z}_\mu^1 = 0$, \mathcal{Z}_μ^k can be decomposed in $\mathcal{Z}_\mu^k = \mathcal{Z}_\mu^1 \oplus P_\mu \mathcal{Z}_\mu^k = \text{span} \{\phi\} \oplus P_\mu \mathcal{Z}_\mu^k$. We shall prove that $P_\mu \mathcal{Z}_\mu^k$ can be interpreted as “ \mathcal{Z}_μ^{k-1} ” for the system $(P_\mu A, P_\mu b, CP_\mu)$ on the Hilbert space $P_\mu H$.

Let $x \in P_\mu \mathcal{Z}_\mu^k$; then $x = P_\mu x'$; $x' \in \mathcal{Z}_\mu^k$. Since $x' \in \mathcal{Z}_\mu^k$ there exist $z \in \mathcal{Z}_\mu^{k-1}$ and $\gamma \in \mathbb{C}$ such that

$$(5.2) \quad (\mu I - A)x' = \gamma \cdot b + z = \gamma \cdot P_\mu b + \gamma \cdot \langle \phi, b \rangle \cdot \phi + z.$$

Defining $z' = \gamma \cdot \langle \phi, b \rangle \cdot \phi + z$ we have

$$(5.3) \quad \langle \phi, z' \rangle = \langle \phi, (\mu I - A)x' \rangle - \langle \phi, \gamma \cdot P_\mu b \rangle = \langle (\mu I - A)\phi, x' \rangle - \langle P_\mu \phi, \gamma \cdot b \rangle = 0.$$

So $z' \in P_\mu \mathcal{Z}_\mu^{k-1}$. Since $(\mu I - A)P_\mu = (\mu I - A)$ we have that

$$(5.4) \quad (\mu I - AP_\mu)x = (\mu I - A)P_\mu x' = (\mu I - A)x' = \gamma \cdot P_\mu b + z', \quad z' \in P_\mu \mathcal{Z}_\mu^{k-1}.$$

Recall that $P_\mu \mathcal{Z}_\mu^1 = \{0\} = \mathcal{Z}_\mu^0$, so by induction and (5.4) $P_\mu \mathcal{Z}_\mu^k = “\mathcal{Z}_\mu^{k-1}”$ for the system $(P_\mu A, P_\mu b, CP_\mu)$.

Since $\mu \in \rho(P_\mu A)$ we can apply Corollary 4.7. So we have $P_\mu \mathcal{Z}_\mu^k = \text{span} \{\chi_1, \dots, \chi_n\}$ where $\chi_i = (R_\mu(\mu))^{i-1} P_\mu b$ and $n' = \min(k, \text{order}(\mu) - 1)$. If we define $x_{i+1} = \chi_i$, then these vectors satisfy (5.1).

COROLLARY 5.2. *If (A, B) satisfies $(\nabla 1)$ and $(\nabla 2)$, then $\dim(\mathcal{Z}_\mu^k) = n$ where n is the minimum of k and order (μ) .*

Proof. See Lemma 5.1 and Corollary 4.7.

LEMMA 5.3. *Let $C(\lambda I - A)^{-1}B \neq 0$. If (A, B) satisfies $(\nabla 1)$ and $(\nabla 2)$, then the order of every invariant zero is finite.*

Proof. Since C and B are bounded operators we have that $C(\lambda I - A)^{-1}B$ is an analytic function on $\rho(A)$. Let $\mu \in \sigma(A)$ be an invariant zero, then from the previous lemma $C(\lambda I - A)^{-1}B = C\{1/(\lambda - \mu) \cdot E(\mu)B\} + CR_\mu(\lambda)B = CR_\mu(\lambda)B$. This function is analytic in μ and has a zero with multiplicity order $(\mu) - 1$ at μ , Lemma 5.1.

So at all invariant zeros $C(\lambda I - A)^{-1}B$ is analytic or has an analytic extension. And from the well-known zero-set theorem for analytic functions (see Rudin [15, p. 209]) we have that the multiplicity of the zeros is finite, thus order (μ) is finite.

LEMMA 5.4. *If (A, B) satisfies $(\nabla 1)$, $(\nabla 2)$ and the transfer function $C(\lambda I - A)^{-1}B$ is not identically zero, then $Z_{\mu_i}^0 \cap \overline{\text{span}_{j \in \underline{I}/i} \{Z_{\mu_j}^0\}} = \{0\}$, where \underline{I} denotes the index set of all invariant zeros and $\theta_j = \text{order}(\mu_j)$.*

Proof. The complete proof is very long and rather technical; here we shall give the proof in the special case that the order of all invariant zeros is one and none of these zeros is in $\sigma(A)$. So we have to prove that $(\mu I - A)^{-1}b$ is not in $\overline{\text{span}_{\mu_i \neq \mu} \{(\mu_i I - A)^{-1}b\}}$. Suppose that this is true; then there exists a sequence γ_i^n such that $\|(\mu I - A)^{-1}b - \sum_{i=1}^n \gamma_i^n \cdot (\mu_i I - A)^{-1}b\| < 1/n$. This implies that

$$(5.5) \quad \left\| C(\mu I - A)^{-1} \left\{ (\mu I - A)^{-1}b - \sum_{i=1}^n \gamma_i^n \cdot (\mu_i I - A)^{-1}b \right\} \right\| < \|C(\mu I - A)^{-1}\| \cdot 1/n.$$

By the resolvent identity we have that

$$(5.6) \quad (\mu I - A)^{-1}(\mu_i I - A)^{-1}b = (\mu - \mu_i)^{-1} \{(\mu_i I - A)^{-1}b - (\mu I - A)^{-1}b\}.$$

Using this identity we have

$$(5.7) \quad \begin{aligned} & (\mu I - A)^{-1} \left\{ (\mu I - A)^{-1}b - \sum_{i=1}^n \gamma_i^n \cdot (\mu_i I - A)^{-1}b \right\} \\ &= (\mu I - A)^{-2}b - \sum_{i=1}^n \gamma_i^n \cdot (\mu - \mu_i)^{-1} \{(\mu_i I - A)^{-1}b - (\mu I - A)^{-1}b\}. \end{aligned}$$

Notice that since μ and μ_i are invariant zeros we have that $C(\mu I - A)^{-1}b$ and $C(\mu_i I - A)^{-1}b$ are both zero. With this property and (5.7) we can simplify the expression inside the norm signs of (5.5):

$$(5.8) \quad \begin{aligned} & C(\mu I - A)^{-1} \left\{ (\mu I - A)^{-1}b - \sum_{i=1}^n \gamma_i^n \cdot (\mu_i I - A)^{-1}b \right\} \\ &= C(\mu I - A)^{-2}b. \end{aligned}$$

Combining (5.8) with (5.5) implies that $C(\mu I - A)^{-2}b$ is zero, so μ is a zero of order two. This is in contradiction with the assumptions. So $(\mu I - A)^{-1}b$ is not in $\overline{\text{span}_{\mu_i \neq \mu} \{(\mu_i I - A)^{-1}b\}}$.

6. Characterization of all invariant subspaces. In this section we shall give a complete characterization of all $T_F(t)$ -invariant subspaces in $\text{Ker } C$. This is yet not possible for an arbitrary generator, but there is a large class, introduced by Sun [17], where we can give a complete answer. We shall start by defining this class.

We say that (A, B) satisfies condition ∇ if it satisfies $(\nabla 1)$ and $(\nabla 2)$, see § 5, and we have the following extra conditions on the eigenvalues of A :

$$(\nabla 3) \quad \inf_{i \neq j} |\lambda_i - \lambda_j| = \delta > 0,$$

$$(\nabla 4) \quad \sup_{1 \leq i < \infty} \sum_{\substack{j=1 \\ j \neq i}}^{\infty} \left| \frac{1}{\lambda_j - \lambda_i} \right|^2 = \tau < \infty.$$

Our main results are as follows.

THEOREM 6.1. *Suppose that $C(\lambda I - A)^{-1}B \neq 0$ and (A, B) satisfies condition ∇ . Then a closed linear subspace \underline{Y} in $\text{Ker } C$ is $T_F(t)$ -invariant for some bounded feedback*

law F iff there exists a subset of \mathbb{N} , denoted by \underline{J} , such that:

$$(6.1) \quad (a) \quad \underline{Y} = \sum_{j \in \underline{J}} \mathbb{Z}_{\mu_j}^k,$$

where μ_j is an invariant zero, $\mu_i \neq \mu_j$ if $i \neq j$;

(b) $\dim(\mathbb{Z}_{\mu_j}^k) > 0$ for all j in \underline{J} and $\dim(\mathbb{Z}_{\mu_j}^k) = 1$ for all but finitely many j in \underline{J} ; and

(c) There exists a subsequence $n_j (j \in \underline{J})$ in \mathbb{N} such that

$$(6.2) \quad \sum_{j \in \underline{J}} \left[\frac{\lambda_{n_j} - \mu_j}{b_{n_j}} \right]^2 < \infty$$

where λ_{n_j} is the n_j th eigenvalue of A and $b_{n_j} = \langle \phi_{n_j}, b \rangle_H$; ϕ_{n_j} is the eigenvector corresponding to λ_{n_j} .

Remark. From Theorem 3.7 and the proof of this theorem we have that $\sigma(A + BF|_{\underline{Y}}) = \{\mu_j; j \in \underline{J}\}$. So if \underline{Y} is closed loop invariant with $A + BF$ stable, then $\{\mu_j; j \in \underline{J}\} \subset \{z \in \mathbb{C}; \operatorname{re}(z) < 0\}$.

With this theorem we can solve the existence of $V^*(\operatorname{Ker} C)$.

THEOREM 6.2. Suppose that $C(\lambda I - A)^{-1}B \neq 0$ and (A, B) satisfies condition ∇ . Then $V^*(\operatorname{Ker} C)$ exists for a bounded feedback law iff the following conditions hold:

(i) For all but finitely many j 's the order of the invariant zero, μ_j , is one; and

(ii) There exists a subsequence $n_j (j \in \underline{J})$ in \mathbb{N} such that

$$(6.3) \quad \sum_{j \in \underline{J}} \left[\frac{\lambda_{n_j} - \mu_j}{b_{n_j}} \right]^2 < \infty$$

where \underline{J} is the index set of all invariant zeros and λ_{n_j}, b_{n_j} are the same as in Theorem 6.1.

Remark. If $C(\lambda I - A)^{-1}B \equiv 0$, then $V^*(\operatorname{Ker} C)$ exists and is equal to the controllability subspace (see Curtain [4]). Since (A, B) is controllable (condition $(\nabla 2)$) we have that $V^*(\operatorname{Ker} C) = H$.

Remark. For the proof of the sufficient part in both theorems, condition $(\nabla 4)$ can be omitted (see Clarke and Holland [2]).

Before we can prove these theorems we must first prove some lemmas.

LEMMA 6.3. If A satisfies conditions $(\nabla 1)$, $(\nabla 3)$ and $(\nabla 4)$, then $A + BF$ is a discrete spectral operator for all bounded feedback laws F . The spectral projections of $A + BF$ will be denoted by $E_F(\nu_i)$. Furthermore if i is sufficiently large, then $\dim(E_F(\nu_i)H) = 1$.

Proof. See Sun [17, Thm. 2.1 and p. 734]. See Kato [9] for the fact that $(\lambda I - A - BF)^{-1}$ is a compact operator.

LEMMA 6.4. Suppose that (A, B) satisfies assumption ∇ . If \underline{W} is an $(A + BF)$ -invariant subspace of $E_F(\mu)H$ in the kernel of C with $\dim(\underline{W}) = k, k > 0$, then μ is an invariant zero with $k \leq \operatorname{order}(\mu)$ and $\underline{W} = \mathbb{Z}_{\mu}^k$.

Proof. Since \underline{W} is a finite-dimensional, $(A + BF)$ -invariant subspace in $E_F(\mu)H$, there exists a nonzero vector e_1 in \underline{W} such that $(A + BF - \mu I)e_1 = 0$. Thus

$$(6.4) \quad (\mu - A)e_1 = bFe_1,$$

$e_1 \in \underline{W} \subset \operatorname{Ker} C$. So μ is an invariant zero and e_1 is in \mathbb{Z}_{μ}^1 . Since $\dim(\mathbb{Z}_{\mu}^1) = 1$ (see Corollary 5.2) $A + BF|_{\underline{W}}$ has only one eigenvector, and $\operatorname{span}\{e_1\} = \mathbb{Z}_{\mu}^1$.

Since \underline{W} is finite-dimensional, $(A + BF)\underline{W} \subset \underline{W}$, $\sigma(A + BF|_{\underline{W}}) = \mu$ and $A + BF|_{\underline{W}}$ has only one eigenvector we have that $\underline{W} = \operatorname{span}\{e_1, \dots, e_k\}$ where e_i satisfies: $(A + BF - \mu)e_1 = 0$ and $(A + BF - \mu)e_i = e_{i-1}, 2 \leq i \leq k$. This last equation implies that $(\mu I - A)e_i = bFe_i - e_{i-1}, 2 \leq i \leq k$.

By induction it is now easy to prove that $i \leq \operatorname{order}(\mu)$ and $\operatorname{span}\{e_1, \dots, e_i\} = \mathbb{Z}_{\mu}^i, 2 \leq i \leq k$.

Proof of Theorem 6.1. (Only if.) Let \underline{V} be $T_F(t)$ -invariant, then by Lemma 6.3 and Theorem 3.7 \underline{V} is of the following form:

$$(6.5) \quad \underline{V} = \sum_{i=1}^{\infty} \underline{W}_i,$$

where \underline{W}_i is a $(A + BF)$ -invariant subspace of $E_F(\mu_i)\underline{H}$.

Define \underline{J} to be the set of indices with \underline{W}_i not the zero subspace. By Lemma 6.3, we have for all but finitely i in \underline{J} , $\dim(\underline{W}_i) = \dim(E_F(\mu_i)\underline{H}) = 1$. Since \underline{V} is in $\underline{\text{Ker}} C$ we have that \underline{W}_i is in $\underline{\text{Ker}} C$, so by Lemma 6.4 we have that $\underline{W}_i = \underline{Z}_{\mu_i}^{k_i}$, $i \in \underline{J}$. Since $\{\mu_i; i \in \underline{J}\}$ is contained in the spectrum of $A + BF$ we have from Sun [17] that (6.2) must hold.

(If.) Let \underline{J}_1 be the subset of \underline{J} that contains all indices j such that $\dim(\underline{Z}_{\mu_j}^{k_j})$ is one, then by Corollary 5.2 $\underline{Z}_{\mu_j}^{k_j} = \underline{Z}_{\mu_j}^1$. By condition ∇ and (6.2) we may apply Theorem 1.1 of Sun [17]. So there exists a bounded feedback law F_1 such that $\sigma(A + BF_1) \supset \{\mu_j | j \in \underline{J}_1\}$. We will prove that $\underline{V}_1 := \overline{\text{span}_{j \in \underline{J}_1} \{\underline{Z}_{\mu_j}^1\}}$ is closed loop invariant with respect to the operator $T_{F_1}(t)$. Let e_j be an eigenvector of $A + BF_1$ with eigenvalue μ_j , then $(A + BF_1)e_j = \mu_j \cdot e_j$ or equivalently

$$(6.6) \quad (\mu_j I - A)e_j = b \cdot F_1 e_j.$$

Again we have to consider two cases:

(i) $\mu_j \notin \sigma(A)$. In this case we can premultiply (6.6) with $(\mu_j I - A)^{-1}$. So

$$(6.7) \quad e_j = (\mu_j I - A)^{-1} b \cdot F_1 e_j.$$

This equation implies that $Ce_j = C(\mu_j I - A)^{-1} b \cdot F_1 e_j$. And this last expression is zero since μ_j is an invariant zero. So $e_j \in \underline{\text{Ker}} C$. This combined with (6.6) implies that $\text{span}\{e_j\} \subset \underline{Z}_{\mu_j}^1$. From the fact that $\underline{Z}_{\mu_j}^1$ is one-dimensional we have that $\underline{Z}_{\mu_j}^1 = \text{span}\{e_j\}$.

(ii) $\mu_j \in \sigma(A)$. Let ϕ be the eigenvector of A corresponding to μ_j , then

$$(6.8) \quad 0 = \langle 0, e_j \rangle_H = \langle (\mu_j - A)\phi, e_j \rangle_H = \langle \phi, (\mu_j - A)e_j \rangle_H = \langle \phi, b \rangle_H \cdot F_1 e_j.$$

By $(\nabla 2)$, (6.8) implies that $F_1 e_j = 0$. This together with (6.6) and the fact that there exists only one eigenvector, of A , corresponding to the eigenvalue μ_j , implies that $\phi = \tau \cdot e_j$; $\tau \neq 0$. Since $\underline{Z}_{\mu_j}^1 \neq \{0\}$ and $\mu_j \in \sigma(A)$ we have from Lemma 5.1 that $\underline{Z}_{\mu_j}^1 = \underline{N}_{\mu_j}^1 = \text{span}\{\phi\} = \text{span}\{e_j\}$.

So for every j in \underline{J}_1 we have that $\underline{Z}_{\mu_j}^1$ is $T_{F_1}(t)$ -invariant, and so is \underline{V}_1 .

Let \underline{J}_2 be the subset of \underline{J} that contains all indices j such that $\dim(\underline{Z}_{\mu_j}^{k_j})$ is larger than one. Then by the condition in Theorem 6.1 we have that \underline{J}_2 is finite. And with Lemmas 4.8 and 3.14 we conclude that $\underline{V}_2 := \text{span}_{j \in \underline{J}_2} \{\underline{Z}_{\mu_j}^{k_j}\}$ is $T_{F_2}(t)$ -invariant, for some bounded operator F_2 . Since $\underline{V} = \underline{V}_1 + \underline{V}_2$ we can apply Lemma 3.14 to conclude the proof.

Proof of Theorem 6.2. (If.) Define \underline{V} to be the closure of the span over all $\underline{Z}_{\mu_j}^{\theta_j}$, $j \in \underline{J}$, where μ_j are the invariant zeros and θ_j is their order. From Theorem 6.1 we have that this subspace is bounded closed loop invariant and by definition it is the largest closed subspace with this property.

(Only if.) By Theorem 6.1 we have that $\underline{V}^*(\underline{\text{Ker}} C)$ is of the form $\sum_{j \in \underline{J}} \underline{Z}_{\mu_j}^{k_j}$. If there were a μ_{j_0} such that j_0 is not in \underline{J} or k_{j_0} is smaller than θ_{j_0} ; the order of μ_{j_0} , then by Lemma 5.4 $(\sum_{j \in \underline{J}} \underline{Z}_{\mu_j}^{k_j} + \underline{Z}_{\mu_{j_0}}^{\theta_{j_0}})$ is larger than $\sum_{j \in \underline{J}} \underline{Z}_{\mu_j}^{k_j}$ and is closed loop invariant by Lemma 3.14. This is in contradiction to the fact that $\underline{V}^*(\underline{\text{Ker}} C) = \sum_{j \in \underline{J}} \underline{Z}_{\mu_j}^{k_j}$ is the largest closed subspace with this property. So $\underline{V}^*(\underline{\text{Ker}} C) = \sum_{j \in \underline{J}} \underline{Z}_{\mu_j}^{\theta_j}$, where the summation is over all invariant zeros and $\theta_j = \text{order}(\mu_j)$. Conditions (i) and (ii) are direct consequences of Theorem 6.1.

7. Examples. In this section we will discuss some examples. The example that will be discussed in this section is the heated rod with various controls and observations. This can be schematized as below:



and for the mathematical model we take

$$(7.1) \quad \begin{aligned} \frac{\partial x}{\partial t} &= \frac{\partial^2 x}{\partial \xi^2} + b(\xi) \cdot u(t), \\ x(0, t) &= 0 = x(1, t), \\ y(t) &= \int_0^1 c(\xi) \cdot x(t, \xi) d\xi. \end{aligned}$$

This can be formulated as a system of the form (4.1), where we take $\underline{H} = L_2(0, 1)$ and the system operator A is given by

$$A = \frac{\partial^2}{\partial \xi^2}, \quad D(A) = \{h \in \underline{H} : h'' \in \underline{H} \text{ and } h(0) = 0 = h(1)\}.$$

A is selfadjoint and has eigenvalues $\{-n^2\pi^2; n = 1, \dots, \infty\}$ and eigenvectors $\{\phi_i(\xi) = \sqrt{2} \cdot \sin i\pi\xi; i = 1, \dots, \infty\}$.

In this section we take $b(\xi) = I_{[0, 2/\pi]}(\xi)$, the characteristic function of the interval $[0, 2/\pi]$, and we shall investigate the existence of $\underline{V}^*(\text{Ker } C)$ for the two measurement functions $c_1(\xi)$ and $c_2(\xi)$:

$$(7.2) \quad c_1(\xi) = \begin{cases} -20^2\xi^2 + 40\xi, & \xi \in \left[0, \frac{1}{20}\right], \\ 1, & \xi \in \left[\frac{1}{20}, 1 - \frac{1}{20}\right], \\ -20^2\xi^2 + (2 \cdot 20^2 - 40)\xi - (20^2 - 40), & \xi \in \left[1 - \frac{1}{20}, 1\right], \end{cases}$$

$$(7.3) \quad c_2(\xi) = I_{[0, 1]}(\xi).$$

It is easy to see that A satisfies $(\nabla 1)$, $(\nabla 3)$ and $(\nabla 4)$. Furthermore since

$$(7.4) \quad b_i = \langle \phi_i, b \rangle = \int_0^1 I_{[0, 2/\pi]}(x) \sin i\pi x dx = \frac{-\sqrt{2}}{i\pi} \{\cos 2 \cdot i - 1\} \neq 0,$$

(A, B) satisfies ∇ .

Since $\langle c_k, \phi_{2j} \rangle = 0$, $k = 1, 2; j \in \underline{N}$, we have that $-4j^2\pi^2; j \in \underline{N}$ are invariant zeros, for both measurements; these are the only invariant zeros in $\sigma(A)$. In Nooitgedagt [11] the position of the other invariant zeros, that are proper zeros of the system, is calculated. The values of these zeros are listed in Table 1.

TABLE 1
The “odd” zeros are located at eigenvalues of A .

zero-number	G_1	G_2	pole
2	-88.57748	-88.57290	-88.82644
4	-235.48318	-234.91203	-246.74011
6	-478.77778	-478.28330	-483.61062
8	-797.61298	-797.27853	-799.43796
10	-1184.41496	-1181.63195	-1194.22213
12	-1666.41632	-1665.76207	-1667.96314
14	-2217.35020	-2215.29725	-2220.66099
16	-2846.10865	-2840.74241	-2852.31567
18	-3562.80012	-3562.64551	-3562.92719
20	-4349.23549	-4343.64385	-4352.49554
38	-15011.66046	-15000.01489	-15011.66829
58	-34355.85119	-34350.98638	-34356.09292
78	-61596.20066	-61593.72592	-61596.20107
98	-96731.78108	-96719.44637	-96731.99273
198	-390846.20363	-390836.43821	-390846.2039

The first column contains the zeros of G_1 and G_2 and the last column the eigenvalue of A that is closest to that zero, where $G_k(s) = \langle c_k, (s - A)^{-1}b \rangle_H$, $k = 1, 2$.

Calculating the partial sum,

$$S(n) := \sum_{i=1}^n \left| \frac{\mu_i - \lambda_{i+1}}{b_{i+1}} \right|^2,$$

where $\lambda_{i+1} := -(i+1)^2 \pi^2$ is the eigenvalue closest to μ_i (see the Remark at the end of this section), for both transfer functions yield the values in Table 2. And, if we plot this as function of n , we obtain Figs. 1 and 2.

TABLE 2

n	G_1	G_2
2	1734.50	1799.62
4	6356.74	6902.80
6	13935.26	16111.46
8	25473.00	32265.61
10	39831.54	55928.94
12	72852.55	88338.93
14	88931.69	132997.56
16	103183.24	188896.95
18	114984.03	258928.29
20	137389.40	345927.28
38	145466.20	2404081.04
58	148906.37	7737351.82
78	151658.92	17903511.5
98	152989.63	34452776.5
118	154363.54	58960201.8
138	155064.37	92969495.8
158	155885.73	138039382
178	156317.35	195748288
198	156863.13	267624120

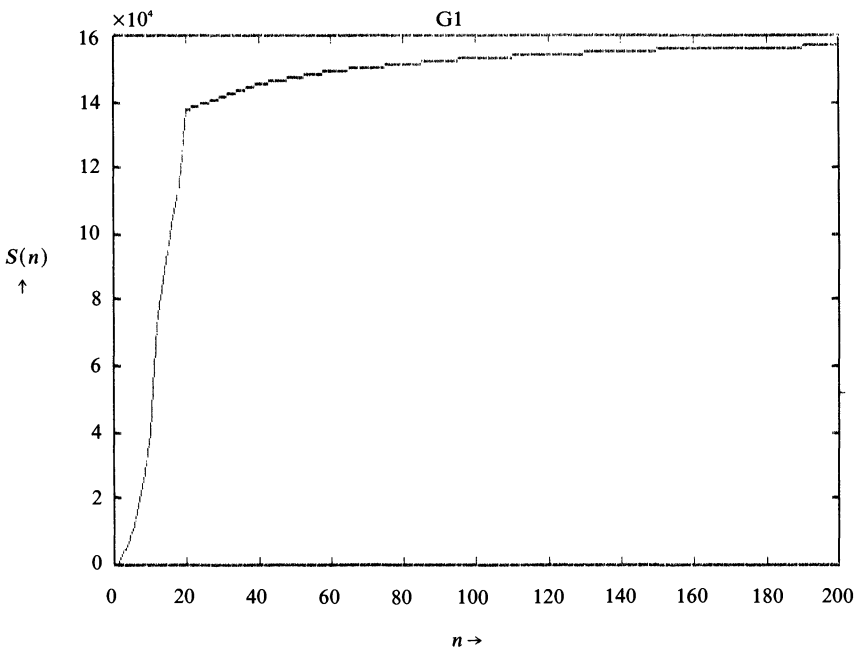


FIG. 1

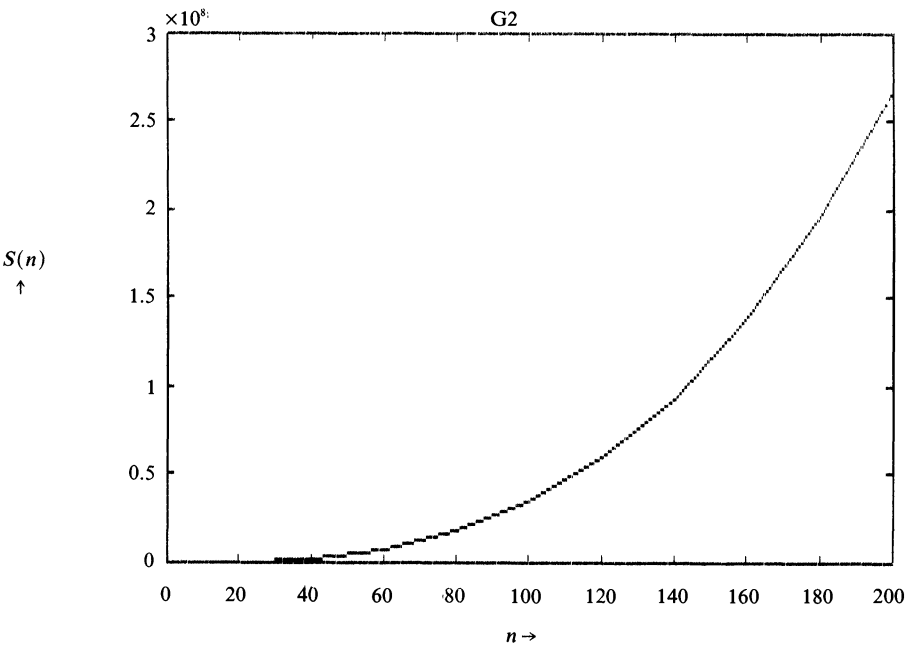


FIG. 2

Remark. For the calculated zeros, μ_i ; $1 \leq i \leq 198$, it can be shown that

$$\min_{j \in N} \left\{ \left| \frac{\mu_i - \lambda_j}{b_j} \right| \right\} = \left| \frac{\mu_i - \lambda_{i+1}}{b_{i+1}} \right|.$$

So

$$S(198) := \sum_{i=1}^{198} \left| \frac{\mu_i - \lambda_{i+1}}{b_{i+1}} \right|^2$$

is smaller than

$$\sum_{i=1}^{198} \left| \frac{\mu_i - \lambda_{j_i}}{b_{j_i}} \right|^2$$

for every subsequence $\{j_i; 1 \leq i \leq 198\}$ in N .

Thus concerning Theorem 6.2 we have (numerical) evidence that for $c_1(\xi) V^*(\text{Ker } C)$ exists but not for $c_2(\xi)$.

From Curtain [4] we have that the existence of $V^*(\text{Ker } C)$, $C = \langle c(\xi), \cdot \rangle_H$, for bounded F , is closely related to $c(\xi) \in D(A)$. In this example we see that, in spite of the fact that $c_1(\xi)$ and $c_2(\xi)$ are close in L_2 -norm, $V^*(\text{Ker } C)$ exists for c_1 but not for c_2 . Notice that $c_1 \in D(A)$ and $c_2 \notin D(A)$. Research is continuing for the case that $c(\xi) \notin D(A)$ and there we see that we must not restrict our attention to bounded feedback laws. In the near future I hope to publish some results on this subject.

8. Conclusions. In this paper we have shown that for the class of discrete spectral systems there exists a complete characterization of all (bounded) closed loop invariant subspaces. With this characterization we obtained necessary and sufficient conditions for the existence of $V^*(\text{Ker } C)$. We calculated these numerically for a simple example. On the other hand if we know a priori (for example, see Curtain [4]) that $V^*(\text{Ker } C)$ exists, then the conditions give information concerning the asymptotic distribution of the zeros of the transfer function.

Furthermore we have shown that (as for a finite-dimensional state space) if F is such that $V \subset \text{Ker } C$ is $T_F(t)$ invariant, then the spectrum of $(A + BF)|_V$ is a subset of the set of all (invariant) zeros of the transfer function $C(\lambda I - A)^{-1}B$ and it is fixed. So if a part of this lies in the unstable region, then it is impossible to have invariance and stability at the same time.

Acknowledgment. I would like to thank Ruth Curtain for her careful reading of this manuscript and for her valuable suggestions.

REFERENCES

- [1] G. BASILE AND G. MARRO, *Controlled and conditioned invariant subspaces in linear system theory*, J. Optim. Theory Appl., 3 (1969), pp. 306-315.
- [2] B. M. N. CLARKE AND W. F. HOLLAND, *Eigenstructure specification for linear systems in Hilbert space* I. *Spectral scalar operators*, Macquarie Mathematics Reports, no. 85-0060, School of Mathematics and Physics, Macquarie University, North Ryde, Australia, 1985.
- [3] R. F. CURTAIN, *Spectral systems*, Internat J. Control, 39 (1984), pp. 657-666.
- [4] ———, *Invariance concepts in infinite dimensions*, this Journal, 24 (1986), pp. 1009-1031.
- [5] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite dimensional linear systems theory*, Lecture Notes in Control and Inform. Sci. 8, Springer-Verlag, Berlin, 1978.
- [6] E. B. DAVIES, *One Parameter Semigroups*, Academic Press, New York, 1980.
- [7] E. J. DAVISON AND S. J. WANG, *Properties and calculation of transmission zeros of linear multivariable systems*, Automatica, 10 (1974), pp. 634-658.

- [8] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part III, Spectral Operators*, Wiley-Interscience, New York, 1971.
- [9] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.
- [10] T. G. KURTZ, *A general theorem on the convergence of operator semigroups*, Trans. Amer. Math. Soc., 148 (1970), pp. 23–32.
- [11] L. NOOITGEDAGT, *Computation of transmission zeros for distributed parameter systems and an application to spectral systems*, M.Sc. thesis, Dept. of Mathematics, University of Groningen, Groningen, the Netherlands, March, 1986.
- [12] I. PANDOLFI, *Disturbance decoupling and invariant subspaces for delay systems*, Appl. Math. Optim., 14 (1986), pp. 55–73.
- [13] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
- [14] S. POHJOLAINEN, *Computation of transmission zeros for distributed parameter systems*, Report 34, Dept. of Electrical Engineering, Mathematics, Tampere University of Technology, Tampere, Finland, 1979.
- [15] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
- [16] E. J. P. G. SCHMIDT AND R. F. STERN, *Invariance theory for infinite dimensional linear control systems*, Appl. Math. Optim., 6 (1980), pp. 113–122.
- [17] S. H. SUN, *On spectrum distribution of completely controllable linear systems*, Acta Math. Sinica, 21 (1978), pp. 193–205 (translation by L. F. Ho); this Journal, 19 (1981), pp. 730–743.
- [18] P. WERMER, *Commuting spectral operations on Hilbert spaces*, Pacific J. Math., 4 (1954), pp. 355–361.
- [19] W. M. WONHAM, *Linear Multivariable Control, a Geometric Approach*, Springer-Verlag, New York, 1978.

HOMOGENEOUS INDICES, FEEDBACK INVARIANTS AND CONTROL STRUCTURE THEOREM FOR GENERALIZED LINEAR SYSTEMS*

MARK A. SHAYMAN†

Abstract. We define a new set of indices for a generalized linear system. These indices, referred to as the *homogeneous indices*, are a natural generalization of the minimal column indices (Kronecker indices) of an ordinary state-space system. We prove that the homogeneous indices are a complete set of invariants for the action of a natural group of feedback transformations on generalized linear systems. We also show that the homogeneous indices determine exactly which closed loop invariant polynomials can be assigned by feedback, thereby generalizing the Control Structure Theorem of Rosenbrock.

Key words. generalized linear system, proportional and derivative feedback, feedback invariants, homogeneous indices, singular pencils

AMS(MOS) subject classifications. 93B10, 93B17, 93B25, 93B55, 93C35

1. Introduction. In the past several years, there has been considerable interest in generalized linear systems (also called “descriptor systems”), i.e., generalized state-space models of the form

$$(1.1) \quad E\dot{x}(t) = Ax(t) + Bu(t)$$

with the matrix E possibly singular (see, e.g., [1], [16], [17]). We represent this system by the matrix triple (E, A, B) and refer to it as a *regular system* if E is nonsingular and as a *singular system* if E is singular.

Recently, Shayman and Zhou [2] have presented a unified theory of control synthesis for generalized linear systems using constant-ratio proportional and derivative (CRPD) feedback. The framework includes the theory of static state feedback and output feedback for regular systems as a special case. The main elements of this theory include (1) a covering of the space of all systems, both regular and singular, by a family of open and dense subsets indexed by the unit circle; (2) a group of transformations which may be viewed as symmetries of the cover; (3) an admissible class of feedback transformations on each subset which is specifically adapted to that subset. A general procedure of control synthesis of CRPD feedback for generalized linear systems is obtained which uses the symmetry transformations to systematically reduce each synthesis problem to an ordinary static state feedback (or output feedback) synthesis problem for a corresponding regular system. This procedure was used to obtain natural generalizations of the Disturbance Decoupling Theorem, the Pole Assignment Theorem and Brunovsky’s canonical form.

In order to give a precise statement of the problems to be addressed in the present paper, we review the three main elements of the theory presented in [2]. We begin with the covering of the space of generalized systems. Let $\hat{\Sigma}(n, m)$ denote the space of all matrix triples $(E, A, B) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times m}$. Let $\Sigma(n, m)$ denote the open and dense subset of $\hat{\Sigma}(n, m)$ characterized by the requirement that $\det(sE - A)$ does not vanish identically. This condition guarantees uniqueness for the solutions of (1.1). In the literature, the systems belonging to $\Sigma(n, m)$ are generally referred to as “regular

* Received by the editors September 22, 1986; accepted for publication (in revised form) August 12, 1987. This research was partially supported by the National Science Foundation under grants ECS-8696108 and CDR-8500108, and by a grant from the Monsanto Company.

† Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, Maryland 20742.

systems.” However, we will reserve the word “regular” to refer to a generalized linear system (E, A, B) for which E is nonsingular. We refer to the systems in $\Sigma(n, m)$ as the *admissible* systems, and to the condition $\det(sE - A) \neq 0$ as the *admissibility assumption*.

We now define a covering of the space $\Sigma(n, m)$ of admissible systems. For each $\theta \in \mathbb{R}$, let $\Sigma_\theta(n, m)$ denote the subset of $\Sigma(n, m)$ given by

$$(1.2) \quad \Sigma_\theta(n, m) = \{(E, A, B) \in \Sigma(n, m) : \det(\cos \theta E - \sin \theta A) \neq 0\}.$$

It is easy to show that $\Sigma_{\theta+\pi}(n, m) = \Sigma_\theta(n, m)$, and that $\{\Sigma_\theta(n, m) : \theta \in [0, \pi)\}$ is a covering of $\Sigma(n, m)$ by open and dense subsets. By virtue of the periodicity, it is natural to regard the parameter θ as a point on the unit circle. Note that in the special case where $\theta = 0$, $\Sigma_0(n, m)$ consists of those triples (E, A, B) for which E is nonsingular, i.e., the regular systems.

Next, we define a group of symmetries of the cover $\{\Sigma_\theta(n, m) : \theta \in [0, \pi)\}$ —transformations which map these subsets into each other. For each $\phi \in \mathbb{R}$, define a mapping $R_\phi : \hat{\Sigma}(n, m) \rightarrow \hat{\Sigma}(n, m)$ by

$$(1.3) \quad R_\phi(E, A, B) = (\cos \phi E + \sin \phi A, -\sin \phi E + \cos \phi A, B).$$

It is straightforward to show that R_ϕ maps $\Sigma_\theta(n, m)$ isomorphically onto $\Sigma_{\theta+\phi}(n, m)$. In particular, this implies that each subset $\Sigma_\theta(n, m)$ in the covering is isomorphic to the set $\Sigma_0(n, m)$ of regular systems.

We now define a class of admissible feedback transformations for each subset $\Sigma_\theta(n, m)$. Specifically, we allow feedback of the form

$$(1.4) \quad u = F(\cos \theta x - \sin \theta \dot{x}) + v$$

to be applied to the systems belonging to the subset $\Sigma_\theta(n, m)$. In (1.4), θ is fixed while the $m \times n$ gain matrix F is arbitrary, and v represents a new external input. The fixed parameter θ specifies the ratio of state to derivative in the feedback law. Consequently, we refer to (1.4) as *constant-ratio proportional and derivative* (CRPD) state feedback. This specialized form of proportional and derivative feedback was suggested as a design tool for singular systems in the doctoral thesis of Zhou [3] (also Zhou, Shayman and Tarn [4]), and independently by Christodoulou [5].

Remark 1.1. As mentioned previously, in the special case where $\theta = 0$, $\Sigma_0(n, m)$ is the set of all regular systems. In this case, (1.4) is ordinary state feedback. Thus, the theory outlined above includes the theory of state feedback for regular systems as a special case.

There are three main contributions in the present paper. The first is the introduction of a new set of indices for a generalized linear system, which we refer to as the *homogeneous indices* of (E, A, B) . These indices are a natural generalization of the minimal column indices (“Kronecker indices”) of a regular system. In fact, we will show that if (E, A, B) is a controllable regular system, its homogeneous indices and its minimal column indices coincide.

The second contribution is a solution to the CRPD feedback equivalence problem for generalized linear systems. We determine necessary and sufficient conditions for two controllable systems in $\Sigma_\theta(n, m)$ to be transformable to each other via the CRPD feedback (1.4) together with change of basis in the state-space, change of basis in the input space and left-multiplication of (1.1) by a nonsingular matrix. We show that two such systems are feedback equivalent if and only if they have the same homogeneous indices. This generalizes the well-known result that two controllable regular systems are equivalent under the state feedback group if and only if they have identical Kronecker indices.

The third contribution in this paper is a generalization of Rosenbrock's Control Structure Theorem [6]. Rosenbrock's Theorem describes precisely which closed-loop invariant polynomials are attainable by applying state feedback to a given controllable regular system. Using the concept of homogeneous indices, we are able to describe exactly which closed-loop invariant polynomials are attainable by applying CRPD feedback (1.4) to a controllable system in $\Sigma_\theta(n, m)$. Rosenbrock's Theorem is recovered as a special case of our result by setting $\theta = 0$.

2. Homogeneous indices. We begin by reviewing Kronecker's definition of the minimal column indices of a singular pencil of matrices [7]. (See also [8, p. 37] and [18, p. 55].) Let M and N be real $m \times n$ matrices. The matrix pencil $\lambda M + N$ is called a *regular pencil* if $m = n$ and $\det(\lambda M + N)$ does not vanish identically. Otherwise, it is called a *singular pencil*.

The minimal column indices of a singular pencil are defined as follows: Let v_1 be a minimal degree nonzero polynomial solution to the equation

$$(2.1) \quad (\lambda M + N)v = 0.$$

Let v_2 be a minimal degree solution which is linearly independent (over the polynomial ring $\mathbb{R}[\lambda]$) of v_1 . Let v_3 be a minimal degree solution which is linearly independent of $\{v_1, v_2\}$. Proceeding in this way, we obtain a sequence v_1, \dots, v_p of solutions. Such a sequence is called a *fundamental series of solutions* of (2.1). Let $\varepsilon_1 \leq \varepsilon_2 \leq \dots \leq \varepsilon_p$ denote the degrees of v_1, \dots, v_p , respectively. Using the fact that column vectors over a polynomial ring are linearly independent if and only if they are linearly independent over the corresponding field of fractions, we can easily show [8, p. 38] that these nonnegative integers are independent of the choice of fundamental series. $(\varepsilon_1, \dots, \varepsilon_p)$ are called the *minimal column indices* of the singular pencil $\lambda M + N$.

Recall that two $m \times n$ pencils, $\lambda M + N$ and $\lambda \bar{M} + \bar{N}$, are said to be strictly equivalent [8, p. 24] if there exist nonsingular constant matrices P and Q of dimensions $m \times m$ and $n \times n$ such that

$$(2.2) \quad P(\lambda M + N)Q = \lambda \bar{M} + \bar{N}.$$

It is well known that strictly equivalent singular pencils have identical minimal column indices.

Let (E, A, B) be a regular system, i.e., $(E, A, B) \in \Sigma_0(n, m)$, and assume (E, A, B) is controllable. Let $(\varepsilon_1, \dots, \varepsilon_p)$ denote the minimal column indices of the singular pencil $[\lambda E - A, B]$. Let r denote the rank of B , and let M_i denote the matrix $[E^{-1}B, (E^{-1}A)(E^{-1}B), \dots, (E^{-1}A)^{i-1}(E^{-1}B)]$. Let $l_1 = \text{rank } M_1$ and let $l_i = \text{rank } M_i - \text{rank } M_{i-1}$, $i = 2, \dots, n$. Then $l_1 \geq \dots \geq l_n \geq 0$. The following facts are well known (see, e.g., [9]).

PROPOSITION 2.1. *Let (E, A, B) be a controllable regular system. Then,*

(a) $(\varepsilon_1, \dots, \varepsilon_p)$ is a partition of n into m parts, with $\varepsilon_1, \dots, \varepsilon_{m-r}$ zero and $\varepsilon_{m-r+1}, \dots, \varepsilon_m$ strictly positive. (Thus, $p = m$.)

(b) $\varepsilon_j = \text{Card} \{i: l_i \geq m - j + 1\}$ ($j = 1, \dots, m$).

Remark 2.1. Proposition 2.1 is no longer true if the assumption that (E, A, B) be a regular system is dropped. Rosenbrock has shown [10] that if the system $(E, A, B) \in \Sigma(n, m)$ has no finite or infinite input decoupling zero (i.e., is controllable), then the pencil $[\lambda E - A, B]$ has no finite elementary divisor and no minimal index for the rows. It has infinite elementary divisors, each of degree 1, equal in number to the rank defect of E . It has m minimal indices for the columns. Thus, $p = m$, and $(\varepsilon_1, \dots, \varepsilon_p)$ is a partition of rank E , rather than a partition of n as it is in the case of a regular system.

We now define the homogeneous indices of a generalized linear system. Let $(E, A, B) \in \hat{\Sigma}(n, m)$. We associate to (E, A, B) the degree one matrix polynomial in two variables given by $[\lambda E - \mu A, B]$. Abusing terminology slightly, we refer to $[\lambda E - \mu A, B]$ as a matrix pencil. Let z_1 be a column vector with entries in the ring $\mathbb{R}[\lambda, \mu]$ of polynomials in two variables which is a minimal degree nonzero solution to the equation

$$(2.3) \quad [\lambda E - \mu A, B]z = 0.$$

(For a polynomial in two variables, “degree” refers to the total degree, and the degree of a solution z is the highest degree of its components.) Let z_2 be a minimal degree solution which is linearly independent over $\mathbb{R}[\lambda, \mu]$ of z_1 . Let z_3 be a minimal degree solution which is linearly independent of $\{z_1, z_2\}$. Proceeding in this way, we obtain a sequence z_1, \dots, z_q of solutions, which we refer to as a *fundamental series of solutions* of (2.3). Since linear independence over $\mathbb{R}[\lambda, \mu]$ is equivalent to linear independence over the fraction field $\mathbb{R}(\lambda, \mu)$ of rational functions in two variables, it follows that q is at most equal to $n + m$. Let $\delta_1 \leq \dots \leq \delta_q$ denote the degrees of z_1, \dots, z_q , respectively. Using the fraction field $\mathbb{R}(\lambda, \mu)$, it follows by an argument which is analogous to the one given in [8, p. 38] for the minimal column indices that $\delta_1, \dots, \delta_q$ are well defined, i.e., independent of the choice of fundamental series. We will refer to $(\delta_1, \dots, \delta_q)$ as the *homogeneous indices* of the system (E, A, B) .

Remark 2.2. It should be noted that the pencil $[\lambda E - \mu A, B]$ is *not* the homogenization of the pencil $[\lambda E - A, B]$. Since $[\lambda E - A, B] = \lambda[E, 0] + [-A, B]$, the homogenization of $[\lambda E - A, B]$ would be $\lambda[E, 0] + \mu[-A, B] = [\lambda E - \mu A, \mu B]$. However, the term “homogeneous indices” seems appropriate since the submatrix $\lambda E - \mu A$ of $[\lambda E - \mu A, B]$ is the homogenization of the submatrix $\lambda E - A$ of $[\lambda E - A, B]$. The homogenization of $\lambda E - A$ to $\lambda E - \mu A$ plays a crucial role in the Weierstrass theory of regular matrix pencils [8, p. 26].

We now establish important properties of the homogeneous indices which will be needed later. Given a triple (E, A, B) , let $\text{HI}(E, A, B)$ denote its set of homogeneous indices, and let $\text{CI}(E, A, B)$ denote its set of minimal column indices, i.e., the minimal column indices of the singular pencil $[\lambda E - A, B]$.

The following result shows that the homogeneous indices of (E, A, B) are invariant under system rotation.

PROPOSITION 2.2. *If $(E, A, B), (\hat{E}, \hat{A}, B) \in \hat{\Sigma}(n, m)$ with $(\hat{E}, \hat{A}, B) = R_\phi(E, A, B)$, then $\text{HI}(E, A, B) = \text{HI}(\hat{E}, \hat{A}, B)$.*

Proof. Let z_1, \dots, z_p be a fundamental series of solutions of the equation

$$[\lambda E - \mu A, B]z = 0,$$

and let $\hat{z}_i(\hat{\lambda}, \hat{\mu}) = z_i((\cos \phi)\hat{\lambda} + (\sin \phi)\hat{\mu}, (-\sin \phi)\hat{\lambda} + (\cos \phi)\hat{\mu})$. It is easy to verify that $\hat{z}_1, \dots, \hat{z}_p$ is a fundamental series of solutions of the equation

$$[\hat{\lambda}\hat{E} - \hat{\mu}\hat{A}, B]\hat{z} = 0.$$

Since $\deg \hat{z}_i = \deg z_i$, the result follows immediately. \square

Remark 2.3. Proposition 2.2 describes a crucial difference between the homogeneous indices and the minimal column indices. In contrast to the homogeneous indices, the minimal column indices are *not* invariant under system rotation. For example, consider the system (E, A, B) with

$$E = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Let $\phi = \pi/2$, and let $(\hat{E}, \hat{A}, B) = R_\phi(E, A, B)$. Then $\hat{E} = A$ and $\hat{A} = -E$. It is easy to check that $\text{CI}(E, A, B) = (1)$ whereas $\text{CI}(\hat{E}, \hat{A}, B) = (2)$. On the other hand, we have $\text{HI}(E, A, B) = \text{HI}(\hat{E}, \hat{A}, B) = (2)$.

The next result shows that for a controllable regular system, the homogeneous indices coincide with the minimal column indices. The proof is deferred to the next section.

PROPOSITION 2.3. *Let (E, A, B) be a controllable regular system. Then,*

$$\text{HI}(E, A, B) = \text{CI}(E, A, B).$$

Remark 2.4. Using Propositions 2.2, 2.3 and 2.1, we can obtain a simple procedure for computing the homogeneous indices of a controllable generalized system. Let $C_\theta(n, m)$ denote the subset of $\Sigma_\theta(n, m)$ consisting of those systems which are controllable according to the definition of Yip and Sincovec [11]. (That is, there are no finite or infinite input decoupling zeros.) It is proved in [2] that controllability is invariant under system rotation. Thus, $R_\phi(C_\theta(n, m)) = C_{\theta+\phi}(n, m)$. Let $(E, A, B) \in C_\theta(n, m)$, and let $(\hat{E}, \hat{A}, B) = R_{-\theta}(E, A, B) \in C_0(n, m)$. By Propositions 2.2 and 2.3, we have $\text{HI}(E, A, B) = \text{HI}(\hat{E}, \hat{A}, B) = \text{CI}(\hat{E}, \hat{A}, B)$. Thus, the homogeneous indices of (E, A, B) can be determined by computing the l_i 's for the controllable regular system (\hat{E}, \hat{A}, B) , and then using Proposition 2.1(b) to obtain $\text{CI}(\hat{E}, \hat{A}, B)$. For example, let (E, A, B) be as in Remark 2.3, and let $\theta = -\pi/2$. Then, $(E, A, B) \in C_\theta(n, m)$ and $(\hat{E}, \hat{A}, B) = R_{-\theta}(E, A, B) = (A, -E, B)$. Since $\text{rank } B = 1$ and $\text{rank } [B, -EB] = 2$, we get $l_1 = 1, l_2 = 1$. Applying Proposition 2.1(b), we obtain $\text{HI}(E, A, B) = (2)$.

PROPOSITION 2.4. *If (E, A, B) is a controllable admissible system, then $\text{HI}(E, A, B)$ is a partition of n into m parts, of which $\text{rank } B$ parts are strictly positive.*

Proof. By assumption, $(E, A, B) \in C_\theta(n, m)$ for some θ . Using the notation of Remark 2.4, we have $\text{HI}(E, A, B) = \text{CI}(\hat{E}, \hat{A}, B)$. The result follows from this together with Proposition 2.1(a). \square

Remark 2.5. Proposition 2.4 describes an important difference between the homogeneous indices and the minimal column indices of a controllable system. The homogeneous indices sum to n regardless of whether the system is regular or singular. In contrast, the minimal column indices sum to $\text{rank } E$ (Remark 2.1), which is equal to n only if the system is regular.

Remark 2.3 shows that in contrast to the homogeneous indices, the minimal column indices are not invariant under system rotation. However, what is true is that if two controllable regular systems are related by a system rotation, then they have identical minimal column indices.

PROPOSITION 2.5. *If (E, A, B) and (\hat{E}, \hat{A}, B) are controllable regular systems with $(\hat{E}, \hat{A}, B) = R_\phi(E, A, B)$, then $\text{CI}(E, A, B) = \text{CI}(\hat{E}, \hat{A}, B)$.*

Proof. Using Propositions 2.2 and 2.3, we have $\text{CI}(E, A, B) = \text{HI}(E, A, B) = \text{HI}(\hat{E}, \hat{A}, B) = \text{CI}(\hat{E}, \hat{A}, B)$. \square

3. Feedback invariants. We begin by reviewing the definition of the state feedback group. (See, e.g., [9], [12]–[14].) Consider the ordinary state-space model

$$(3.1) \quad \dot{x}(t) = Ax(t) + Bu(t)$$

where $(A, B) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times m}$. We consider three types of elementary transformation on the system (3.1). They are (1) change of basis in the state-space, $x = Pz$ with P a nonsingular $n \times n$ matrix; (2) change of basis in the input space, $u = Qv$ with Q a nonsingular $m \times m$ matrix; (3) state feedback $u = Fx + v$. These operations transform the matrix pair (A, B) as follows:

$$(3.2) \quad (A, B) \rightarrow (P^{-1}AP, P^{-1}B),$$

$$(3.3) \quad (A, B) \rightarrow (A, BQ),$$

$$(3.4) \quad (A, B) \rightarrow (A + BF, B).$$

The transformation group generated by (3.2)–(3.4) can be conveniently represented in the following way. Recall that a right group action of a group G on a set X is a mapping $\eta: X \times G \rightarrow X$ satisfying the conditions $\eta(x, e) = x$ and $\eta(x, g_1 g_2) = \eta(\eta(x, g_1), g_2)$ where e denotes the identity element of G . If $x \in X$, the orbit of x , denoted xG , consists of the subset $\{\eta(x, g): g \in G\}$ of X .

Let $C(n, m)$ denote the space of all matrix pairs $(A, B) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times m}$ which are controllable. Let $H(n, m)$ denote the group consisting of all nonsingular $(n + m) \times (n + m)$ matrices of the form

$$\begin{bmatrix} P & 0 \\ F & Q \end{bmatrix}$$

with $Pn \times n$, $Fm \times n$, $Qm \times m$. We refer to $C(n, m)$ as the *space of controllable pairs* and to $H(n, m)$ as the *state feedback group*. Define a right group action of $H(n, m)$ on $C(n, m)$ by

$$(3.5) \quad \eta \left((A, B), \begin{bmatrix} P & 0 \\ F & Q \end{bmatrix} \right) = (P^{-1}AP + P^{-1}BF, P^{-1}BQ).$$

The transformations (3.2)–(3.4) correspond to the special cases of (3.5), where $F = 0$ and $Q = I$, $P = I$ and $F = 0$, $P = I$ and $Q = I$, respectively.

It is of interest to know when two systems (A_1, B_1) and (A_2, B_2) are related by a transformation in the state feedback group, i.e., belong to the same $H(n, m)$ -orbit. It is also useful to have a canonical form for this group action—to identify the “simplest” element on each orbit. This is provided by the following result of Brunovsky [12]. (See also [9], [6], [15].)

THEOREM 3.1 [12]. (a) $(A_1, B_1), (A_2, B_2) \in C(n, m)$ belong to the same $H(n, m)$ -orbit if and only if $\text{CI}(I, A_1, B_1) = \text{CI}(I, A_2, B_2)$.

(b) Let r be a positive integer with $r \leq \min(n, m)$, and let $n_1 \geq n_2 \geq \dots \geq n_r$ be a partition of n into r positive parts. The $H(n, m)$ -orbit consisting of those pairs $(A, B) \in C(n, m)$ for which $\text{CI}(I, A, B) = (0, \dots, 0, n_r, n_{r-1}, \dots, n_1)$ contains the canonical pair (A_c, B_c) given by

$$A_c = \begin{bmatrix} J_{n_1} & 0 & 0 & \cdots & 0 \\ 0 & J_{n_2} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & & \cdots & J_{n_r} \end{bmatrix},$$

$$B_c = \left[\begin{array}{ccccc|ccc} e_{n_1} & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & e_{n_2} & 0 & \cdots & 0 & & & \\ 0 & 0 & \ddots & & \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots & & & \vdots \\ 0 & 0 & & \cdots & e_{n_r} & 0 & \cdots & 0 \end{array} \right]$$

where J_k is a $k \times k$ matrix of the form

$$\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 1 \\ 0 & 0 & \cdots & & 0 \end{bmatrix}$$

and e_k is a k -dimensional column vector in which the only nonzero component is the last, which is 1.

We will refer to the pair (A_c, B_c) in Theorem 3.1 as the *Brunovsky canonical form* associated with the set of minimal column indices $(0, \dots, 0, n_r, \dots, n_1)$.

The following result is needed for the proof of Proposition 2.3.

LEMMA 3.1. *If $(A, B), (\hat{A}, \hat{B}) \in C(n, m)$ belong to the same $H(n, m)$ -orbit, then*

$$\text{HI}(I, A, B) = \text{HI}(I, \hat{A}, \hat{B}).$$

Proof. It suffices to consider the following two special cases:

$$(a) (\hat{A}, \hat{B}) = (P^{-1}AP, P^{-1}BQ);$$

$$(b) (\hat{A}, \hat{B}) = (A + BF, B).$$

First consider (a). We have

$$(3.6) \quad [\lambda I - \mu \hat{A}, \hat{B}] = P^{-1}[\lambda I - \mu A, B] \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix}.$$

Let z_1, \dots, z_p be a fundamental series of solutions of the equation $[\lambda I - \mu A, B]z = 0$. Let

$$\hat{z}_i = \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix}^{-1} z_i.$$

Then it is clear that $\hat{z}_1, \dots, \hat{z}_p$ is a fundamental series of solutions of the equation $[\lambda I - \mu \hat{A}, \hat{B}]\hat{z} = 0$. Since $\deg \hat{z}_i = \deg z_i$, it follows that $\text{HI}(I, A, B) = \text{HI}(I, \hat{A}, \hat{B})$.

Now consider (b). We have

$$(3.7) \quad [\lambda I - \mu \hat{A}, \hat{B}] = [\lambda I - \mu A, B] \begin{bmatrix} I & 0 \\ -\mu F & I \end{bmatrix}.$$

Let z_1, \dots, z_p be a fundamental series of solutions of $[\lambda I - \mu A, B]z = 0$, and let

$$\hat{z}_i = \begin{bmatrix} I & 0 \\ \mu F & I \end{bmatrix} z_i.$$

Then $\hat{z}_1, \dots, \hat{z}_p$ are solutions of $[\lambda I - \mu \hat{A}, \hat{B}]\hat{z} = 0$ which are linearly independent (over $\mathbb{R}[\lambda, \mu]$).

We claim that $\deg \hat{z}_i = \deg z_i$. Let $z_i = \begin{bmatrix} x_i \\ u_i \end{bmatrix}$ and let $\hat{z}_i = \begin{bmatrix} \hat{x}_i \\ \hat{u}_i \end{bmatrix}$. Then $\hat{x}_i = x_i$ and $\hat{u}_i = \mu Fx_i + u_i$. Thus, it suffices to show that $\deg \hat{u}_i = \deg u_i$. Let $d = \deg x_i$, and write

$$x_i = x_{i0}\lambda^d + x_{i1}\lambda^{d-1}\mu + \dots + x_{id}\mu^d + \bar{x}_i$$

with $\deg \bar{x}_i < d$. Suppose that $x_{i0}, \dots, x_{i,k-1}$ are zero, but x_{ik} is nonzero. Since $Bu_i = \mu Ax_i - \lambda x_i$, it follows that the coefficient vector of $\lambda^{d-k+1}\mu^k$ in u_i is nonzero. In particular, $\deg u_i \geq d+1$. If $\deg u_i > d+1$, then since $\deg \mu Fx_i \leq d+1$, it follows that $\deg \hat{u}_i = \deg u_i$. Therefore, we may assume $\deg u_i = d+1$. Since u_i contains $\lambda^{d-k+1}\mu^k$ but μFx_i does not contain $\lambda^{d-k+1}\mu^k$, it follows that $\deg \hat{u}_i = d+1 = \deg u_i$. Thus, $\deg \hat{z}_i = \deg z_i$ as claimed.

We claim that $\hat{z}_1, \dots, \hat{z}_p$ is a fundamental series of solutions of $[\lambda I - \mu \hat{A}, \hat{B}]\hat{z} = 0$. Suppose not. Then there is a smallest positive integer k for which $\deg \hat{z}_k$ is not minimal. Thus, there exists a solution \hat{y}_k with $\deg \hat{y}_k < \deg \hat{z}_k$ such that $\hat{z}_1, \dots, \hat{z}_{k-1}, \hat{y}_k$ are linearly independent. Let

$$y_k = \begin{bmatrix} I & 0 \\ -\mu F & I \end{bmatrix} \hat{y}_k.$$

Then, z_1, \dots, z_{k-1}, y_k is a linearly independent set of solutions of $[\lambda I - \mu A, B]z = 0$.

Essentially the same argument as given in the preceding paragraph shows that $\deg y_k = \deg \hat{y}_k$, so $\deg y_k < \deg z_k$. This contradicts the assumption that z_1, \dots, z_p is a fundamental series. Thus, $\hat{z}_1, \dots, \hat{z}_p$ is a fundamental series. Since $\deg \hat{z}_i = \deg z_i$, this proves that $\text{HI}(I, A, B) = \text{HI}(I, \hat{A}, \hat{B})$. \square

Remark 3.1. We can use (3.5) to define the action of $H(n, m)$ of all matrix pairs $(A, B) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times m}$, rather than on the set of controllable pairs $C(n, m)$. Since the proof of Lemma 3.1 does not use controllability, it follows that the statement of Lemma 3.1 remains true with $C(n, m)$ replaced by $\mathbb{R}^{n \times n} \times \mathbb{R}^{n \times m}$.

Proof of Proposition 2.3. Let (E, A, B) be a controllable regular system. Since left-multiplication is a strict equivalence transformation, it does not change either the minimal column indices or the homogeneous indices. Consequently, it suffices to show that $\text{HI}(I, E^{-1}A, E^{-1}B) = \text{CI}(I, E^{-1}A, E^{-1}B)$. Let (A_c, B_c) denote the Brunovsky canonical form of $(E^{-1}A, E^{-1}B)$. From Theorem 3.1(a) and Lemma 3.1, we have $\text{CI}(I, A_c, B_c) = \text{CI}(I, E^{-1}A, E^{-1}B)$ and $\text{HI}(I, A_c, B_c) = \text{HI}(I, E^{-1}A, E^{-1}B)$. Thus, it suffices to show that $\text{HI}(I, A_c, B_c) = \text{CI}(I, A_c, B_c)$.

Let $(0, \dots, 0, n_r, \dots, n_1)$ denote the minimal column indices of (I, A_c, B_c) . We can choose an $(n+m) \times (n+m)$ permutation matrix N such that

$$[\lambda I - \mu A_c, B_c]N = [\text{diag}\{K(n_1), \dots, K(n_r)\}, O_{n \times (m-r)}],$$

where $K(n_i)$ denotes the $n_i \times (n_i + 1)$ pencil $[\lambda I_{n_i} - \mu J_{n_i}, e_{n_i}]$. Although $[\lambda I - \mu A_c, B_c]N$ is not of the form $[\lambda E - \mu A, B]$, we can define a fundamental series of solutions for the equation $[\lambda I - \mu A_c, B_c]N\hat{z} = 0$ in the obvious way. Hence, homogeneous indices are well defined for the pencil $[\lambda I - \mu A_c, B_c]N$. If $\hat{z}_1, \dots, \hat{z}_p$ denotes such a fundamental series, then it is clear that $N\hat{z}_1, \dots, N\hat{z}_p$ is a fundamental series for the equation $[\lambda I - \mu A_c, B_c]z = 0$. Since $\deg N\hat{z}_i = \deg \hat{z}_i$, it follows that the homogeneous indices of $[\lambda I - \mu A_c, B_c]$ —i.e., $\text{HI}(I, A_c, B_c)$ —are equal to the homogeneous indices of $[\lambda I - \mu A_c, B_c]N$. Thus, it suffices to show that the homogeneous indices of $[\lambda I - \mu A_c, B_c]N$ are $(0, \dots, 0, n_r, \dots, n_1)$.

It is easy to see that for a block-diagonal pencil, the complete set of homogeneous indices is obtained as the union of the corresponding systems of homogeneous indices of the individual diagonal blocks. (This is analogous to the situation for the minimal column indices noted in [8, p. 39].) Consequently, the homogeneous indices of $[\lambda I - \mu A_c, B_c]N$ consist of $0, \dots, 0$ (multiplicity $m-r$) together with the homogeneous indices of the r pencils $[\lambda I_{n_i} - \mu J_{n_i}, e_{n_i}]$. Let $z_{n_i}^*$ denote the $(n_i + 1)$ -component vector with j th component $\lambda^{j-1} \mu^{n_i-j}$ ($j = 1, \dots, n_i$) and $(n_i + 1)$ th component $-\lambda^{n_i}$. It is easy to verify that every nonzero solution of $[\lambda I_{n_i} - \mu J_{n_i}, e_{n_i}]z_i = 0$ is an $\mathbb{R}[\lambda, \mu]$ -multiple of $z_{n_i}^*$. Consequently, $[\lambda I_{n_i} - \mu J_{n_i}, e_{n_i}]$ has the single homogeneous index (n_i) . Thus, the homogeneous indices of $[\lambda I - \mu A_c, B_c]N$ are $(0, \dots, 0, n_r, \dots, n_1)$, completing the proof. \square

We now review the definition of the CRPD state feedback groups given in [2]. We consider four types of elementary transformations on the system (1.1). They are (1) change of basis in the state-space, $x = Pz$ with P a nonsingular $n \times n$ matrix; (2) change of basis in the input space, $u = Qv$ with Q a nonsingular $m \times m$ matrix; (3) CRPD feedback $u = F(\cos \theta x - \sin \theta \dot{x}) + v$ with θ a fixed number in $[0, \pi)$ and F an arbitrary $m \times n$ matrix; (4) left-multiplication by a nonsingular $n \times n$ matrix R^{-1} . These operations transform the matrix triple (E, A, B) as follows:

$$(3.8) \quad (E, A, B) \rightarrow (P^{-1}EP, P^{-1}AP, P^{-1}B),$$

$$(3.9) \quad (E, A, B) \rightarrow (E, A, BQ),$$

$$(3.10) \quad (E, A, B) \rightarrow (E + \sin \theta BF, A + \cos \theta BF, B),$$

$$(3.11) \quad (E, A, B) \rightarrow (R^{-1}E, R^{-1}A, R^{-1}B).$$

The transformation group generated by (3.8)–(3.11) can be conveniently represented in the following way. For each θ , let $G_\theta(n, m)$ denote the group consisting of all nonsingular $(3n + m) \times (3n + m)$ matrices of the form

$$(3.12) \quad \begin{bmatrix} R & 0 & 0 & 0 \\ 0 & P & 0 & 0 \\ 0 & 0 & P & 0 \\ 0 & \sin \theta F & \cos \theta F & Q \end{bmatrix}.$$

We refer to the family of groups $\{G_\theta(n, m): \theta \in [0, \pi)\}$ as the CRPD *state feedback groups*.

Remark 3.2. If $\cos \theta F$ and $\sin \theta F$ in (3.12) are replaced with arbitrary $m \times n$ matrices F_1, F_2 , then we obtain a larger group which we denote by $G(n, m)$ and refer to as the *proportional and derivative state feedback group*. This corresponds to replacing the CRPD feedback (1.4) with the more general proportional and derivative feedback $u = F_1x - F_2\dot{x} + v$. Each of the CRPD feedback groups, $G_\theta(n, m)$, is a subgroup of $G(n, m)$.

Define a right group action of $G_\theta(n, m)$ on $\hat{\Sigma}(n, m)$ by

$$(3.13) \quad \left((E, A, B), \begin{bmatrix} R & 0 & 0 & 0 \\ 0 & P & 0 & 0 \\ 0 & 0 & P & 0 \\ 0 & \sin \theta F & \cos \theta F & Q \end{bmatrix} \right) \rightarrow (R^{-1}EP + \sin \theta R^{-1}BF, R^{-1}AP + \cos \theta R^{-1}BF, R^{-1}BQ).$$

Each of the transformations (3.8)–(3.11) is a special case of (3.13). Let $g_\theta(R, P, Q, F)$ denote the transformation on $\hat{\Sigma}(n, m)$ induced by the matrix (3.12) in $G_\theta(n, m)$. In other words, $g_\theta(R, P, Q, F)(E, A, B)$ denotes the right-hand side of (3.13).

Recall from Remark 2.4 that $C_\theta(n, m)$ denotes the controllable systems in the open and dense subset $\Sigma_\theta(n, m)$, and that $R_\phi(C_\theta(n, m)) = C_{\theta+\phi}(n, m)$. The following three results are proved in [2].

PROPOSITION 3.1 [2]. $\Sigma_\theta(n, m)$ is invariant under the action of $G_\theta(n, m)$.

PROPOSITION 3.2 [2]. The following is a commutative diagram:

$$\begin{array}{ccc} \Sigma_\theta(n, m) & \xrightarrow{g_\theta(R, P, Q, F)} & \Sigma_\theta(n, m) \\ R_\phi \downarrow & & \downarrow R_\phi \\ \Sigma_{\theta+\phi}(n, m) & \xrightarrow{g_{\theta+\phi}(R, P, Q, F)} & \Sigma_{\theta+\phi}(n, m) \end{array}$$

That is, $R_\phi \circ g_\theta(R, P, Q, F) = g_{\theta+\phi}(R, P, Q, F) \circ R_\phi$.

PROPOSITION 3.3 [2]. $C_\theta(n, m)$ is invariant under the action of $G_\theta(n, m)$.

By virtue of Proposition 3.3, we can restrict the action of $G_\theta(n, m)$ on $\hat{\Sigma}(n, m)$ to the invariant subset $C_\theta(n, m)$. The problem which we consider is the one of determining a complete set of invariants for the action of $G_\theta(n, m)$ on $C_\theta(n, m)$. Roughly speaking, this means finding a set of functions of (E, A, B) with the property that these functions have the same values on (E_1, A_1, B_1) as on (E_2, A_2, B_2) if and only if (E_1, A_1, B_1) and (E_2, A_2, B_2) are related by a transformation in the CRPD feedback group $G_\theta(n, m)$.

Remark 3.3. In the special case where $\theta = 0$, $C_0(n, m)$ consists of the controllable regular systems, and $G_0(n, m)$ can be regarded as the state feedback group $H(n, m)$

augmented by left-multiplication. In this case, it follows from Theorem 3.1(a) that the minimal column indices are a complete set of invariants. In other words, $(E_1, A_1, B_1), (E_2, A_2, B_2) \in C_0(n, m)$ are equivalent under the action of $G_0(n, m)$ if and only if $\text{CI}(E_1, A_1, B_1) = \text{CI}(E_2, A_2, B_2)$.

The following result is the main result of this section. It presents the solution to the problem posed above, and represents a natural generalization of Theorem 3.1.

THEOREM 3.2. (a) $(E_1, A_1, B_1), (E_2, A_2, B_2) \in C_\theta(n, m)$ belong to the same $G_\theta(n, m)$ -orbit if and only if

$$\text{HI}(E_1, A_1, B_1) = \text{HI}(E_2, A_2, B_2).$$

(b) Let $n_1 \geq \dots \geq n_m$ be a partition of n into m nonnegative parts. The $G_\theta(n, m)$ -orbit consisting of those triples $(E, A, B) \in C_\theta(n, m)$ for which $\text{HI}(E, A, B) = (n_m, \dots, n_1)$ contains the canonical triple $(\cos \theta I + \sin \theta A_c, -\sin \theta I + \cos \theta A_c, B_c)$ where (A_c, B_c) is the Brunovsky canonical form associated with the minimal column indices (n_m, \dots, n_1) .

Proof. (a) Let $(E_1, A_1, B_1), (E_2, A_2, B_2) \in C_\theta(n, m)$, and suppose there exist R, P, Q, F such that $(E_2, A_2, B_2) = g_\theta(R, P, Q, F)(E_1, A_1, B_1)$. Applying successively Proposition 2.2, Proposition 2.3 and Theorem 3.1(a) together with Remark 3.3, Proposition 2.3, Proposition 2.2, and Proposition 3.2, we have

$$\begin{aligned} \text{HI}(E_1, A_1, B_1) &= \text{HI}(R_{-\theta}(E_1, A_1, B_1)) \\ &= \text{CI}(R_{-\theta}(E_1, A_1, B_1)) = \text{CI}(g_0(R, P, Q, F) \circ R_{-\theta}(E_1, A_1, B_1)) \\ &= \text{HI}(g_0(R, P, Q, F) \circ R_{-\theta}(E_1, A_1, B_1)) \\ &= \text{HI}(R_\theta \circ g_0(R, P, Q, F) \circ R_{-\theta}(E_1, A_1, B_1)) \\ &= \text{HI}(g_\theta(R, P, Q, F)(E_1, A_1, B_1)) = \text{HI}(E_2, A_2, B_2). \end{aligned}$$

Conversely, suppose $(E_1, A_1, B_1), (E_2, A_2, B_2) \in C_\theta(n, m)$ with $\text{HI}(E_1, A_1, B_1) = \text{HI}(E_2, A_2, B_2)$. By Proposition 2.2, we have $\text{HI}(R_{-\theta}(E_1, A_1, B_1)) = \text{HI}(R_{-\theta}(E_2, A_2, B_2))$. By Proposition 2.3, it follows that $\text{CI}(R_{-\theta}(E_1, A_1, B_1)) = \text{CI}(R_{-\theta}(E_2, A_2, B_2))$. Consequently, by Theorem 3.1(a) together with Remark 3.3, there exist R, P, Q, F such that $R_{-\theta}(E_2, A_2, B_2) = g_0(R, P, Q, F) \circ R_{-\theta}(E_1, A_1, B_1)$. Thus, it follows from Proposition 3.2 that $(E_2, A_2, B_2) = g_\theta(R, P, Q, F)(E_1, A_1, B_1)$.

(b) Let $(E, A, B) \in C_\theta(n, m)$. It is proved in [2] that there is some partition of n into m nonnegative parts, $n_1 \geq n_2 \geq \dots \geq n_m$, such that the $G_\theta(n, m)$ -orbit of (E, A, B) contains the triple $(\cos \theta I + \sin \theta A_c, -\sin \theta I + \cos \theta A_c, B_c)$, where (A_c, B_c) is the Brunovsky canonical form associated with the minimal column indices (n_m, \dots, n_1) . It remains to show that (n_m, \dots, n_1) are the homogeneous indices of (E, A, B) . By successively using part (a), Proposition 2.2, and Proposition 2.3, we have $\text{HI}(E, A, B) = \text{HI}(\cos \theta I + \sin \theta A_c, -\sin \theta I + \cos \theta A_c, B_c) = \text{HI}(I, A_c, B_c) = \text{CI}(I, A_c, B_c) = (n_m, \dots, n_1)$. \square

Remark 3.4. Kalman has noted [9] that the action (3.5) of the state feedback group can be regarded as a special case of the strict equivalence action on matrix pencils. To see this, let (\hat{A}, \hat{B}) denote the right-hand side of (3.5). Then,

$$(3.14) \quad [\lambda I - \hat{A}, \hat{B}] = P^{-1}[\lambda I - A, B] \begin{bmatrix} P & 0 \\ -F & Q \end{bmatrix}.$$

Thus, the pencil $[\lambda I - \hat{A}, \hat{B}]$ is obtained from the pencil $[\lambda I - A, B]$ by a strict equivalence transformation. Consequently, Brunovsky canonical form is a special case of the Kronecker canonical form for matrix pencils.

An obvious question to ask is whether the action (3.13) of the CRPD feedback group $G_\theta(n, m)$ can somehow be regarded as a special case of the strict equivalence

action on matrix pencils. It is easy to show that the answer to this question is negative. One way to show this is to note that if (3.13) corresponded to a strict equivalence transformation, then the minimal column indices would be invariant under the action of $G_\theta(n, m)$. This is not the case. For example, let (E, A, B) be as in Remark 2.3, and let $\theta = \pi/2$. Then, $R_{-\theta}(E, A, B) = (-A, E, B)$, which is a controllable regular system, i.e., an element in $C_0(n, m)$. Thus, $(E, A, B) \in C_\theta(n, m)$. Let $F = [1 \ 0]$. Then $g_\theta(I, I, I, F)(E, A, B) = (E + BF, A, B)$ with

$$E + BF = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Thus, $(E + BF, A, B)$ is a controllable regular single-input system. Consequently, $CI(E + BF, A, B) = (2)$. On the other hand, it was noted in Remark 2.3 that $CI(E, A, B) = (1)$.

Another way to appreciate the distinction between the action of the CRPD feedback groups and the strict equivalence action is to express (3.13) as a transformation on matrix pencils. Let $(\tilde{E}, \tilde{A}, \tilde{B})$ denote the right-hand side of (3.13). Then,

$$(3.15) \quad [\lambda \tilde{E} - \mu \tilde{A}, \tilde{B}] = R^{-1}[\lambda E - \mu A, B] \begin{bmatrix} P & 0 \\ (\lambda \sin \theta - \mu \cos \theta)F & Q \end{bmatrix}.$$

In contrast to (3.14), the transformation (3.15) is not of the strict equivalence type since the right-multiplication is not by a constant matrix.

In the literature, other types of feedback groups have been considered in conjunction with generalized linear systems. Hayton [19] studies the action of the state feedback group $H(n, m)$ on generalized linear systems. Pandolfi [20] considers the transformation group generated by exponential rescaling, left-multiplication, change of basis in the state-space, change of basis in the input space, and static state feedback. A complete set of invariants for the action of this group on the set of controllable admissible generalized linear systems is determined.

The transformation groups $\{G_\theta(n, m)\}$ which we study differ considerably from those in [19] and [20]. The feedback in [19] and [20] is static state feedback, which implies that the rank of E is invariant. In contrast, the feedback in $G_\theta(n, m)$ is of the CRPD-type, and can modify the rank of E . Only when $\theta = 0$ does the CRPD feedback coincide with pure state feedback. If (E, A, B) is a singular system, then $(E, A, B) \notin \Sigma_0(n, m)$. Since the transformations in $G_\theta(n, m)$ are applied only to the systems in $\Sigma_\theta(n, m)$, pure state feedback (i.e., no derivative contribution) is never applied to a singular system.

4. Control structure theorem. The problem of pole-assignment by state feedback for singular systems has been studied by Cobb [21] and Pandolfi [22]. Armentano [23] and Lewis and Ozcaldiran [24] have investigated eigenvector-assignment by state feedback. Mukundan and Dayawansa [25] have studied pole-assignment by proportional and derivative state feedback. In this section, we consider a different problem, namely, the determination of which closed-loop invariant polynomials can be obtained using constant-ratio proportional and derivative feedback.

The Control Structure Theorem of Rosenbrock [6] is an important result which describes precisely which invariant polynomials can be assigned by application of state feedback to a controllable regular system:

THEOREM 4.1 [6]. *Let (A, B) be controllable. Let $r = \text{rank } B$, and let $n_1 \geq \dots \geq n_r$ be the nonzero minimal column indices of (I, A, B) . Let $q \leq r$, and let $\tilde{\psi}_1(\lambda), \dots, \tilde{\psi}_q(\lambda)$ be any set of nonunity monic polynomials such that $\tilde{\psi}_{i+1} | \tilde{\psi}_i$ ($i = 1, \dots, q-1$) and $\sum_{i=1}^q \deg \tilde{\psi}_i = n$. Then there is a state feedback gain F such that the given polynomials*

are the nonunity invariant polynomials of the closed loop system $A + BF$ if and only if

$$\sum_{i=1}^p \deg \tilde{\psi}_i \geq \sum_{i=1}^p n_i, \quad p = 1, \dots, q.$$

Let (E, A, B) be an admissible generalized linear system, i.e., $(E, A, B) \in \Sigma(n, m)$. By the invariant polynomials of (E, A, B) , we mean the invariant polynomials of the pencil $\lambda E - A$. In order to generalize Rosenbrock's Theorem, it is necessary to define the *homogeneous invariant polynomials* for (E, A, B) . In the Weierstrass treatment of infinite elementary divisors for a regular pencil [8, p. 26], (homogeneous) invariant polynomials are defined for the homogeneous pencil $\lambda E - \mu A$. Let $D_k(\lambda, \mu)$ be the greatest common divisor of the minors of order k ($k = 1, \dots, n$). The invariant polynomials of $\lambda E - \mu A$ are the quotients $i_1 = D_n/D_{n-1}$, $i_2 = D_{n-1}/D_{n-2}$, etc. Each D_k, i_j is homogeneous. We define the homogeneous invariant polynomials of (E, A, B) to be the invariant polynomials of $\lambda E - \mu A$. Note that these polynomials are defined modulo multiplication by nonzero real numbers.

Let $r_\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ with

$$(4.1) \quad r_\phi(\lambda, \mu) = ((\cos \phi)\lambda + (\sin \phi)\mu, (-\sin \phi)\lambda + (\cos \phi)\mu).$$

The following proposition describes how the homogeneous invariant polynomials transform under system rotation.

PROPOSITION 4.1. *Let ψ_1, \dots, ψ_n denote the homogeneous invariant polynomials of $(E, A, B) \in \Sigma(n, m)$. Then the homogeneous invariant polynomials of $R_\phi(E, A, B)$ are $\psi_1 \circ r_\phi, \dots, \psi_n \circ r_\phi$.*

Proof. Let $(\hat{E}, \hat{A}, \hat{B}) = R_\phi(E, A, B)$, and let $(\hat{\lambda}, \hat{\mu}) = r_{-\phi}(\lambda, \mu)$. Then,

$$(4.2) \quad \lambda E - \mu A = \hat{\lambda} \hat{E} - \hat{\mu} \hat{A}.$$

Since $(\lambda, \mu) = r_\phi(\hat{\lambda}, \hat{\mu})$, it follows immediately from (4.2) that the invariant polynomials of $\lambda \hat{E} - \mu \hat{A}$ are $\psi_1 \circ r_\phi, \dots, \psi_n \circ r_\phi$. \square

Remark 4.1. It follows from Proposition 4.1 that the degrees of the homogeneous invariant polynomials of $R_\phi(E, A, B)$ are equal to the degrees of the homogeneous invariant polynomials of (E, A, B) . However, the corresponding statement for the ordinary invariant polynomials is definitely not true. For example, let (E, A, B) be as in Remark 2.3, and let $\phi = \pi/2$. Then $R_\phi(E, A, B) = (A, -E, B)$. The homogeneous invariant polynomials of (E, A, B) are $\mu^2, 1$, which have the same degrees as $\lambda^2, 1$, the homogeneous invariant polynomials of $R_\phi(E, A, B)$. On the other hand, the (ordinary) invariant polynomials of (E, A, B) are $1, 1$, whereas those of $R_\phi(E, A, B)$ are $\lambda^2, 1$.

PROPOSITION 4.2. *$(E, A, B) \in \Sigma(n, m)$ is a regular system if and only if no homogeneous invariant polynomial is divisible by μ . In this case, there is a degree-preserving one-to-one correspondence between the homogeneous invariant polynomials and the invariant polynomials given by $\psi_i(\lambda, \mu) \leftrightarrow \psi_i(\lambda, 1)$.*

Proof. See [8, p. 27]. \square

Remark 4.2. The condition in Proposition 4.2 that no homogeneous invariant polynomial is divisible by μ is equivalent to the condition that $\det(\lambda E - \mu A)$ is not divisible by μ .

Remark 4.3. Roughly speaking, Propositions 4.1 and 4.2 are the analogues for the homogeneous invariant polynomials of Propositions 2.2 and 2.3 for the homogeneous indices.

We are now ready to state and prove the generalization of Rosenbrock's Theorem.

THEOREM 4.2 (Control Structure Theorem for Generalized Linear Systems). *Let $(E, A, B) \in C_\theta(n, m)$. Let $r = \text{rank } B$, and let $n_1 \geq \dots \geq n_r$ be the nonzero homogeneous*

indices of (E, A, B) . Let $q \leq r$, and let $\psi_1(\lambda, \mu), \dots, \psi_q(\lambda, \mu)$ be any set of nonconstant homogeneous polynomials such that $\psi_{i+1} | \psi_i$ ($i = 1, \dots, q-1$) and $\sum_{i=1}^q \deg \psi_i = n$. Then there is a CRPD state feedback gain F such that the given polynomials are the nonconstant homogeneous invariant polynomials of the closed loop system $g_\theta(F)(E, A, B)$ if and only if the following two conditions are satisfied:

- (i) ψ_i is not divisible by $(-\sin \theta)\lambda + (\cos \theta)\mu$, $i = 1, \dots, q$;
- (ii) $\sum_{i=1}^p \deg \psi_i \geq \sum_{i=1}^p n_i$, $p = 1, \dots, q$.

Proof. Let $(\hat{E}, \hat{A}, B) = R_{-\theta}(E, A, B) \in C_0(n, m)$. By Propositions 2.2 and 2.3, we have $\text{HI}(E, A, B) = \text{HI}(\hat{E}, \hat{A}, B) = \text{CI}(\hat{E}, \hat{A}, B)$. Thus, $n_1 \geq \dots \geq n_r$ are the nonzero minimal column indices of (\hat{E}, \hat{A}, B) . Let $\hat{\psi}_i = \psi_i \circ r_{-\theta}$, and let $\tilde{\psi}_i(\lambda) = \hat{\psi}_i(\lambda, 1)$. Clearly, $\deg \tilde{\psi}_i = \deg \psi_i$.

Suppose that ψ_1, \dots, ψ_q satisfy (i) and (ii). Since ψ_i is not divisible by $(-\sin \theta)\lambda + (\cos \theta)\mu$, $\hat{\psi}_i$ is not divisible by μ . Consequently, $\deg \tilde{\psi}_i = \deg \psi_i$. Thus,

$$\sum_{i=1}^p \deg \tilde{\psi}_i \geq \sum_{i=1}^p n_i, \quad p = 1, \dots, q.$$

Replacing $\tilde{\psi}_1, \dots, \tilde{\psi}_q$ with nonzero scalar multiples if necessary, we may assume $\tilde{\psi}_1, \dots, \tilde{\psi}_q$ are monic. Applying Rosenbrock's Theorem (Theorem 4.1) to the controllable pair $(\hat{E}^{-1}\hat{A}, \hat{E}^{-1}B)$, we conclude that there is a feedback gain F such that $\tilde{\psi}_1, \dots, \tilde{\psi}_q$ are the nonunity invariant polynomials of $\hat{E}^{-1}\hat{A} + \hat{E}^{-1}BF$, i.e., of the pencil $\lambda I - (\hat{E}^{-1}\hat{A} + \hat{E}^{-1}BF)$. Thus, $\tilde{\psi}_1, \dots, \tilde{\psi}_q$ are the nonunity invariant polynomials of the pencil $\lambda \hat{E} - (\hat{A} + BF)$, and hence of the regular system $(\hat{E}, \hat{A} + BF, B)$. By Proposition 4.2, $\tilde{\psi}_1, \dots, \tilde{\psi}_q$ are the nonconstant homogeneous invariant polynomials of $(\hat{E}, \hat{A} + BF, B)$. By Proposition 4.1, ψ_1, \dots, ψ_q are the nonconstant homogeneous invariant polynomials of $R_\theta(\hat{E}, \hat{A} + BF, B) = R_\theta \circ g_0(F) \circ R_{-\theta}(E, A, B) = g_\theta(F)(E, A, B)$, as required.

Conversely, suppose that there exists a feedback gain F for which the closed loop system $g_\theta(F)(E, A, B)$ has ψ_1, \dots, ψ_q as its nonconstant homogeneous polynomials. By Proposition 4.1, $\hat{\psi}_1, \dots, \hat{\psi}_q$ are the nonconstant homogeneous invariant polynomials of $R_{-\theta} \circ g_\theta(F)(E, A, B) = g_0(F)(\hat{E}, \hat{A}, B)$. By Proposition 4.2, $\hat{\psi}_1, \dots, \hat{\psi}_q$ are not divisible by μ , $\tilde{\psi}_1, \dots, \tilde{\psi}_q$ are the nonconstant invariant polynomials of $g_0(F)(\hat{E}, \hat{A}, B)$, and $\deg \tilde{\psi}_i = \deg \psi_i$. Since ψ_i is not divisible by μ and $\psi_i = \hat{\psi}_i \circ r_\theta$, it follows that ψ_i is not divisible by $(-\sin \theta)\lambda + (\cos \theta)\mu$. Thus, condition (i) is satisfied. Applying Theorem 4.1, we conclude that $\sum_{i=1}^p \deg \tilde{\psi}_i \geq \sum_{i=1}^p n_i$, $p = 1, \dots, q$. Since $\deg \tilde{\psi}_i = \deg \psi_i = \deg \hat{\psi}_i$, it follows immediately that condition (ii) is satisfied. \square

Remark 4.4. Theorem 4.2 says that for a generalized linear system $(E, A, B) \in C_\theta(n, m)$, the closed loop homogeneous invariant polynomials can be assigned arbitrarily by CRPD feedback $g_\theta(F)$ subject to two restrictions. Let $(E_F, A_F, B) = g_\theta(F)(E, A, B)$. The first restriction, that the homogeneous invariant polynomials of (E_F, A_F, B) cannot be divisible by $(-\sin \theta)\lambda + (\cos \theta)\mu$, is equivalent to the restriction that $\det(\lambda E_F - \mu A_F)$ is not divisible by $(-\sin \theta)\lambda + (\cos \theta)\mu$. This corresponds to the fact that $\Sigma_\theta(n, m)$ is invariant under the CRPD feedback $g_\theta(F)$, so we must have $\det(\cos \theta E_F - \sin \theta A_F) \neq 0$.

Remark 4.5. Rosenbrock's Theorem (Theorem 4.1) can be easily recovered by setting $\theta = 0$ in Theorem 4.2. Let (A, B) , r , $n_1 \geq \dots \geq n_r$, q and $\tilde{\psi}_1(\lambda), \dots, \tilde{\psi}_q(\lambda)$ be as in the hypotheses of Theorem 4.1. Then $(I, A, B) \in C_0(n, m)$, and by Proposition 2.3, $n_1 \geq \dots \geq n_r$ are the nonzero homogeneous indices of (I, A, B) . Let $\psi_i(\lambda, \mu)$ be the homogenization of $\tilde{\psi}_i(\lambda)$. That is, $\psi_i(\lambda, \mu) = \mu^{d_i} \tilde{\psi}_i(\lambda/\mu)$, where d_i denotes the degree of $\tilde{\psi}_i$. Since ψ_i is automatically not divisible by μ , it follows from setting $\theta = 0$ in Theorem 4.2 that there exists F such that $(I, A + BF, B)$ has nonconstant

homogeneous invariant polynomials ψ_1, \dots, ψ_q if and only if $\sum_{i=1}^p \deg \psi_i \equiv \sum_{i=1}^p n_i$, $p = 1, \dots, q$. By Proposition 4.2, $(I, A + BF, B)$ has nonconstant invariant polynomials $\tilde{\psi}_1, \dots, \tilde{\psi}_q$ if and only if it has nonconstant homogeneous invariant polynomials ψ_1, \dots, ψ_q . Since $\deg \tilde{\psi}_i = \deg \psi_i$, we conclude that there exists F such that $A + BF$ has nonconstant invariant polynomials $\tilde{\psi}_1, \dots, \tilde{\psi}_q$ if and only if $\sum_{i=1}^p \deg \tilde{\psi}_i \equiv \sum_{i=1}^p n_i$, $p = 1, \dots, q$.

Remark 4.6. The analogue of condition (i) in Theorem 4.2 is absent from Theorem 4.1 only because Theorem 4.1 is stated in terms of the invariant polynomials rather than the *homogeneous* invariant polynomials. If Theorem 4.1 were restated in terms of the homogeneous invariant polynomials, it would be necessary to include the requirement that they not be divisible by μ .

REFERENCES

- [1] Special Issue: Semistate Systems, Circuits, Systems, and Signal Processing, 5 (1986).
- [2] M. A. SHAYMAN AND Z. ZHOU, *Feedback control and classification of generalized linear systems*, IEEE Trans. Automat. Control, AC-32, (1987), pp. 483–494.
- [3] Z. ZHOU, *Feedback synthesis of singular systems—a geometric approach*, D.Sc. thesis, Washington University, St. Louis, MO, 1984.
- [4] Z. ZHOU, M. A. SHAYMAN AND T. J. TARN, *Singular systems: a new approach in the time domain*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 42–50.
- [5] M. A. CHRISTODOULOU, *Decoupling in the design and synthesis of singular systems*, Automatica, 22 (1986), pp. 245–249; Proc. IFAC World Congress, Budapest, 1984.
- [6] H. H. ROSENBROCK, *State-Space and Multivariable Theory*, Nelson, London, 1970.
- [7] L. KRONECKER, *Algebraische reduction der schaaren bilinearer formen*, S.-B. Akad. Berlin, 1890, pp. 763–776.
- [8] F. R. GANTMACHER, *Theory of Matrices, Volume II*, Chelsea, New York, 1959.
- [9] R. E. KALMAN, *Kronecker invariants and feedback*, in Ordinary Differential Equations, L. Weiss, ed., Academic Press, New York, 1972.
- [10] H. H. ROSENBROCK, *Structural properties of linear dynamical systems*, Internat. J. Control, 20 (1974), pp. 191–202.
- [11] E. L. YIP AND R. F. SINCOVEC, *Solvability, controllability, and observability of continuous descriptor systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 702–707.
- [12] P. BRUNOVSKY, *A classification of linear controllable systems*, Kybernetika, 6 (1970), pp. 173–188.
- [13] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1979.
- [14] R. W. BROCKETT, *The geometry of the set of controllable linear systems*, Research Reports Automatic Control Laboratory, Nagoya Univ., 24 (1977), pp. 1–7.
- [15] W. M. WONHAM AND A. S. MORSE, *Feedback invariants of linear multivariable systems*, Automatica, 8 (1972), pp. 93–100.
- [16] S. L. CAMPBELL, *Singular Systems of Differential Equations*, Pitman, London, 1980.
- [17] ———, *Singular Systems of Differential Equations II*, Pitman, London, 1982.
- [18] J. H. M. WEDDERBURN, *Lectures on Matrices*, Dover, New York, 1964.
- [19] G. E. HAYTON, *Properties of dynamical indices*, Internat. J. Control, 22 (1975), pp. 289–293.
- [20] L. PANDOLFI, *A canonical form for generalized linear systems*, Boll. Un. Mat. Ital., B(4), (1985), pp. 125–137.
- [21] J. D. COBB, *Feedback and pole placement in descriptor variable systems*, Internat. J. Control, 33 (1981), pp. 1135–1146.
- [22] L. PANDOLFI, *Coefficient assignment for generalized control systems*, Systems Sci., 8 (1982), pp. 195–203.
- [23] V. A. ARMENTANO, *Eigenvalue placement for generalized linear systems*, Systems Control Lett., 4 (1984), pp. 199–202.
- [24] F. L. LEWIS AND K. OZCALDIRAN, *On the eigenstructure assignment of singular systems*, Proc. 24th IEEE Conference on Decision and Control, Ft. Lauderdale, FL, December 1985, pp. 179–182.
- [25] R. MUKUNDAN AND W. DAYAWANSA, *Feedback control of singular systems—proportional and derivative feedback of the state*, Internat. J. Systems Sci., 14 (1983), pp. 615–632.

INVARIANT POLYNOMIAL CURVES OF PIECEWISE LINEAR MAPS*

B. CURTIS EAVES† AND URIEL G. ROTHBLUM‡

Abstract. Consider a system which evolves in n -space from x to $f(x)$ as time moves from t to $t+1$. Assuming f is piecewise linear our interest is in establishing conditions for existence of and an algorithm for computation of polynomials $y[\cdot]$ in time t with the property that $f(y[t]) = y[t+1]$, that is, the system moves from $y[t]$ to $y[t+1]$ in one unit of time.

Key words. piecewise linear functions, invariant curves, Markov decision chains, system evolution, ordered fields

AMS(MOS) subject classifications. primary 47H10, 47H09; secondary 90C47, 90C48

1. Introduction. Consider a system that evolves according to the piecewise linear function $f: R^n \rightarrow R^n$; that is, if the system is found in state x at time t , then it will be found in state $f(x)$ at time $t+1$. Our goal is to determine conditions under which there exists a vector polynomial $y[\cdot]$ in the variable t , such that

$$(1.1) \quad f(y[t]) = y[t+1] \quad \text{for all } t \geq 0.$$

Moreover, in the case where these conditions are satisfied we shall compute a vector polynomial $y[\cdot]$ satisfying (1.1). Evidently, if $y[\cdot]$ satisfies the system (1.1), then the set $\{y[t]: t \geq 0\}$ is mapped by f into itself. Accordingly we call the vector polynomial $y[\cdot]$ satisfying (1.1) an *invariant polynomial curve*.

The task of determining invariant polynomial curves arises in system theory, stochastic games, matrix theory, dynamic programming and branching Markov decision chains, e.g., Rothblum and Veinott [1987]. Note that if f has a *fixed point* y , i.e., $f(y) = y$, then $y[t] \equiv y$ is an invariant polynomial curve of f . As a trivial example we take the function $f: R^2 \rightarrow R^2$, defined by $f(x_1, x_2) = (x_1 + x_2, x_2)$. Then both $y[t] = (1, 0)$ and $y[t] = (t, 1)$, $t \geq 0$, are invariant polynomial curves of f , that is, they both satisfy the system (1.1).

Fixed points of *PL* maps have been studied extensively; see Eaves [1976], Eaves and Scarf [1976], Todd [1976], or Allgower and Georg [1980]. In the special case where the solution $y[\cdot]$ to (1.1) is linear, we refer to the set $\{y[t]: t \geq 0\}$ as an *invariant ray of f* . Kohlberg [1980] considered the invariant ray problem and showed that if f is nonexpansive, then f must have an invariant ray. Note that f is defined to be *nonexpansive* if for some norm $\|\cdot\|$ on R^n , we have $\|f(x) - f(y)\| \leq \|x - y\|$ for all x and y in R^n . Kohlberg's proof is based on the existence of a unique fixed point for contraction operators on R^n ; in particular, his analysis is not concerned with the computation of invariant rays. We extend Kohlberg's results in three directions. First we weaken the requirement that f is nonexpansive and conclude the existence of an invariant polynomial curve which is not necessarily a ray (the fact that our requirement is weaker than nonexpansiveness is demonstrated in Eaves and Rothblum [1986], [1987]). Second, our proofs are constructive—we obtain an explicit finite algorithm

* Received by the editors August 4, 1986; accepted for publication (in revised form) June 22, 1987. This research is based on work supported in part by the National Science Foundation under grants DMS-8404121 and DMS-8603232, by United States-Israel Binational Science Foundation grant 85-00295, and by U.S. Department of Energy contract DE-AA03-76SF000326, PA# DE-AS03-76ER72018.

† Department of Operations Research, Stanford University, Stanford, California 94305.

‡ Faculty of Industrial Engineering and Management, Technion-Israel Institute of Technology, Haifa 32000, Israel.

for computing invariant polynomial curves. Third, our results hold for all ordered fields whereas Kohlberg's results are restricted to the reals; however, we hasten to add that by Tarski's principle, his results extend immediately to any real closed field.

Preliminaries about ordered fields are listed in § 2. In § 3 we formally introduce the notion of piecewise linear maps and obtain the main theorem about existence of fixed points for such maps. The main result asserts the existence of a fixed point for a piecewise linear map f whenever the determinants of the Jacobians of $f - I$ preserve sign for points that are sufficiently far away from the origin. In § 4 we discuss extensions of piecewise linear maps to larger fields. Next, in § 5 we show how an ordered field can be augmented by an infinitesimal to obtain a larger ordered field and in § 6 we show how fixed points for the perturbed extension of piecewise linear maps to that larger field can be used to obtain invariant polynomial curves of the original maps. Finally, in § 7 we discuss an algebraic representation of the fixed point of the perturbed extension of a piecewise linear map, and obtain a simple upper bound on the degree of invariant polynomial curves, when they exist. In particular, we show that under Kohlberg's assumption of nonexpansiveness (over the reals) invariant polynomial curves must be invariant rays.

The present paper has been extracted from the technical report by Eaves and Rothblum [1986]; indeed, many ideas developed in the present paper are carried further in the technical report, especially the relationships of the sufficient conditions for existence of invariant polynomial curves. The latter will appear in Eaves and Rothblum [1987]. Also, occasionally we refer to Eaves and Rothblum [1984] which is unfinished and not readily available; our intent in doing so is merely to indicate that we have given the matter at hand full attention.

2. Preliminaries. Our main framework, throughout, consists of ordered fields, such as the set of rationals Q and the set of reals R . We begin by defining ordered fields. Let G be a set having two distinctive elements called "zero" and "one," denoted 0 and 1, respectively, where two operations called "addition" and "multiplication," denoted $+$ and \cdot , respectively, are defined on G . Also, a relation called "greater than," denoted $>$, is defined on G . The tuple $(G, 0, 1, +, \cdot, >)$ is defined to be an *ordered field* if 0 is the addition identity, 1 is the multiplication identity, addition and multiplication are both associative and commutative, multiplication is distributive over addition, all elements in G have an additive inverse, all nonzero elements have a multiplicative inverse, and greater than is a total order that is preserved under addition and under multiplication by positive elements, the latter means that for x, y and z in G with $x > y$ we have that $x + z > y + z$ and if $z > 0$ then $xz > yz$. If $(G, 0, 1, +, \cdot, >)$ satisfies the above axioms except for the existence of a multiplicative inverse, the tuple is defined to be an *ordered ring*. For convenience we shall usually refer to the ordered ring or field $(G, 0, 1, +, \cdot, >)$ by G .

In an ordered ring or field G let $-x$, x^{-1} , $x - y$, xy , $x/y = xy^{-1}$, x^n , $|x|$, and $x \geq y$ have the usual meanings of *additive inverse*, *multiplicative inverse*, *subtraction*, *product*, *division*, *power*, *absolute value*, and *greater than or equal*, respectively. We refer to addition, multiplication, subtraction, division, and comparison as the *five field operations*. If $x > 0$ we say x is *positive*, etc. Given two ordered rings or fields G and F , the words *subordered ring or field*, *extension*, *isomorphism*, *imbedding*, and *identification* are employed in the usual way.

Let G be an ordered field. For nonnegative integers m and n , let $G^{m \times n}$ denote the set of all $m \times n$ matrices whose elements are in G . We shall use subscripts to denote the rows of a matrix, superscripts to denote its columns and both sub- and superscripts

to denote its elements, e.g., A_i, A^j and A_i^j . For $A, B \in G^{m \times n}$, we write $A \leq B$ or $A \ll B$ if $A_i^j \leq B_i^j$ or if $A_i^j < B_i^j$ for all $i = 1, \dots, m$ and $j = 1, \dots, n$, respectively. Also, we write $A < B$ if $A \leq B$ and $A \neq B$. The zero matrix and the (square) identity matrix will be denoted 0 and I , respectively (their order will be clear from the context). As usual, $G^{n \times 1}$ will be denoted G^n and its elements will be called *vectors*. The l_∞ -norm on G^n is defined by $\|z\|_\infty = \max \{|z_i|: i = 1, \dots, n\}$. As usual vectors a^1, \dots, a^k in G^n are called *linearly independent* if β_1, \dots, β_k in G and $\sum_{i=1}^k \beta_i a^i = 0$ imply the β_i 's are all zero.

A set $\sigma \subseteq G^n$ is called a *cell* if it has the representation $\sigma = \{x \in G^n: Ax \leq a\}$ for some matrix $A \in G^{m \times n}$ and $a \in G^m$ where m is some nonnegative integer. In this case we say that (m, A, a) is a *representation* of σ . In particular, if $m = 0$, we have that $\sigma = G^n$. Let $\sigma \subseteq G^n$ be a cell. The *tangential hull* of σ , denoted $\text{tng } \sigma$, is the set $\{\alpha(x - y): x, y \in \sigma, \alpha \in G\}$. The *dimension* of σ is the maximal number of linearly independent vectors in $\text{tng } \sigma$. The *affine hull* of σ , denoted $\text{aff } \sigma$, is the set $\{x + z: z \in \text{tng } \sigma\}$ for some, or equivalently every, $x \in \sigma$. The cell σ is called *affine* if $\sigma = \text{aff } \sigma$. We say that σ is *unbounded* if for every $K \in G$ there exists some $z \in \sigma$ with $\|z\|_\infty > K$. If (m, A, a) is a representation of σ , then σ is unbounded if and only if $\sigma \neq \phi$ and $\{x \in G^n: Ax \leq 0, x \neq 0\} \neq \phi$. A function $f: \sigma \rightarrow G^k$ is called *affine* if for some $B \in G^{k \times n}$ and $b \in G^k$, $f(x) = Bx + b$ for all x in σ .

The *determinant* of a square matrix A will be defined as the usual sum of signed products of the elements of A and is denoted $\det A$. In any ordered field the following conditions remain equivalent: $\det A \neq 0$, A has a multiplicative inverse and $Ax \neq 0$ for all $x \in G^n \setminus \{0\}$, and, in this case, A is called *nonsingular*.

We say that a statement depending on a parameter ε holds *for sufficiently small positive ε* if for some $\delta > 0$ the statement holds for all ε with $0 < \varepsilon < \delta$; also, we use the phrase *for sufficiently large t* in a similar way.

3. Piecewise linear maps and fixed points. In this section we formally introduce piecewise linear maps, and determine conditions under which such maps have fixed points. Moreover, we identify an algorithm that can be used to compute corresponding fixed points under those conditions.

Throughout let G be a given ordered field. A function $f: G^n \rightarrow G^k$ is called *piecewise linear*, abbreviated PL, if G^n can be expressed as a finite union of cells having dimension n where the restriction of f to each of these sets is affine. If Σ is such a set of cells and for $\sigma \in \Sigma$,

$$(3.1) \quad \sigma = \{x \in G^n: A_\sigma x \leq a_\sigma\}$$

and

$$(3.2) \quad f(x) = B_\sigma x + b_\sigma \quad \text{for all } x \in \sigma,$$

where $A_\sigma \in G^{m_\sigma \times n}$, $a_\sigma \in G^{m_\sigma}$, $B_\sigma \in G^{k \times n}$ and $b_\sigma \in G^k$, we say that $\{(m_\sigma, A_\sigma, a_\sigma, B_\sigma, b_\sigma): \sigma \in \Sigma\}$ is a *representation of f* . Evidently, a finite set of quintuples $\{(m_\sigma, A_\sigma, a_\sigma, B_\sigma, b_\sigma): \sigma \in \Sigma\}$ where for each $\sigma \in \Sigma$, m_σ is a nonnegative integer, $A_\sigma \in G^{m_\sigma \times n}$, $a_\sigma \in G^{m_\sigma}$, $B_\sigma \in G^{k \times n}$ and $b_\sigma \in G^k$ form a representation of a PL function if and only if

$$(3.3) \quad \dim \{x \in G^n: A_\sigma x \leq a_\sigma\} = n \quad \text{for all } \sigma \in \Sigma,$$

$$(3.4) \quad \bigcup_{\sigma \in \Sigma} \{x \in G^n: A_\sigma x \leq a_\sigma\} = G^n$$

and

$$(3.5) \quad (A_\sigma x \leq a_\sigma) \wedge (A_\tau x \leq a_\tau) \Rightarrow (B_\sigma x + b_\sigma = B_\tau x + b_\tau) \quad \text{for all } x \in G^n.$$

In this case we identify each $\sigma \in \Sigma$ with the n -dimensional cell $\{x \in G^n: A_\sigma x \leq a_\sigma\}$. Of course, representations of two distinct PL functions cannot coincide, i.e., a representation of a PL function uniquely defines that function. However, each PL function will have many representation. A PL function $f: G^n \rightarrow G^n$ will be called a *PL map on G^n* , or simply a *PL map*.

In the following theorem and corollary we cite conditions which insure that a PL map is onto or has a fixed point. The proof is constructive and identifies an algorithm that can be used to compute fixed points under the corresponding conditions.

THEOREM 3.1. *Let $g: G^n \rightarrow G^n$ be a PL map having a representation $\{(m_\sigma, A_\sigma, a_\sigma, B_\sigma, b_\sigma): \sigma \in \Sigma\}$. Suppose that the signs of $\det B_\sigma$ for every unbounded $\sigma \in \Sigma$ are either all positive or all negative. Then g is onto.*

Proof. A constructive proof of this result when $G = R$ is the central issue of Chein and Kuh [1976]. The construction is a special case of the more general PL homotopy method; see Appendix A. \square

COROLLARY 3.2. *Let $f: G^n \rightarrow G^n$ be a PL map having the representation $\{(m_\sigma, A_\sigma, a_\sigma, B_\sigma, b_\sigma): \sigma \in \Sigma\}$. Suppose that the signs of $\det (B_\sigma - I)$ for every unbounded cell $\sigma \in \Sigma$ are either all positive or are all negative. Then there exists a vector $x \in G^n$ satisfying $f(x) = x$.*

Proof. Apply the Theorem 3.1 to $f - I$. \square

In Eaves and Rothblum [1986], [1987] necessary and sufficient conditions are given for the determinant conditions in the above theorem and corollary.

4. Extending PL functions. In this section we show how representations of PL functions can be used to extend such functions to larger ordered fields. We first demonstrate that representations of cells can be used to extend them to larger ordered fields in a unique way while preserving some of the algebraic properties.

LEMMA 4.1. *Let $\sigma \subseteq G^n$ be a cell having the representation (m, A, a) and let F be an extension of G . Let $\bar{\sigma} = \{x \in F^n: Ax \leq a\}$. Then:*

- (a) $\bar{\sigma}$ is independent of the representation of σ used to define it;
- (b) The dimension of $\bar{\sigma}$ (as a cell in F^n) coincides with the dimension of σ (as a cell in G^n); and
- (c) $\bar{\sigma}$ is unbounded if and only if σ is unbounded.

Proof. We present arguments for (a) and (c) here to give the flavor of the development; however, further details are given in Eaves and Rothblum [1984].

First assume that (m', A', a') is an alternative presentation of σ . Now, $\{x \in F^n: Ax \leq a\} \neq \{x \in F^n: A'x \leq a'\}$ if and only if either for some $i = 1, \dots, m'$ the system

$$(4.1) \quad Ax \leq a, \quad A'_i x > a'_i$$

has a solution in F^n , or for some $i = 1, \dots, m$, the system

$$(4.2) \quad A_i x > a_i, \quad A'x \leq a'$$

has a solution in F^n . By Corollary B.2 of Appendix B, this happens if and only if either system has a solution in G^n , respectively, implying that (m, A, a) and (m', A', a') cannot be representations of the same cell. So, part (a) follows.

Next, to establish (c), recall that σ is unbounded if and only if the system

$$(4.3) \quad Ax \leq a, \quad Ay \leq 0, \quad y \neq 0$$

has a solution in G^n . Similarly, the cell $\bar{\sigma}$ is unbounded if and only if (4.3) has a solution in F^n . By Lemma B.2 of Appendix B, the system (4.3) has a solution in F^n if and only if it has a solution in G^n . So, (c) follows. \square

Given a cell $\sigma \subseteq G^n$ and an extension F of G , we shall use the notation $\bar{\sigma}_F$ for the cell in F^n constructed in Lemma 4.1.

We next show how a representation of a PL function $f: G^n \rightarrow G^k$ can be used to extend f to a function on F^n whose range is F^k where F is any extension of G . Moreover, we show that the extension is independent of the representation of f .

LEMMA 4.2. *Suppose $f: G^n \rightarrow G^k$ is a PL function having the representation $\{(m_\sigma, A_\sigma, a_\sigma, B_\sigma, b_\sigma): \sigma \in \Sigma\}$ and let F be an extension of G . Consider the coordinates of $A_\sigma, a_\sigma, B_\sigma$ and b_σ as elements in F . Then $\{(m_\sigma, A_\sigma, a_\sigma, B_\sigma, b_\sigma): \sigma \in \Sigma\}$ is a representation of a PL function on F^n which coincides with f on G^n . Moreover, the resulting PL function on F^n is independent of the representation of f .*

Proof. Here, we only sketch the arguments of the proof; however, further details are given in Eaves and Rothblum [1984]. First, recall that in order to show that the quintuples $\{(m_\sigma, A_\sigma, a_\sigma, B_\sigma, b_\sigma): \sigma \in \Sigma\}$ form a representation of a PL function mapping F^n into F^k we have to establish (3.3)–(3.5) with F^n replacing G^n . As $\{(m_\sigma, A_\sigma, a_\sigma, B_\sigma, b_\sigma): \sigma \in \Sigma\}$ is a representation of f we have that (3.3)–(3.5) hold without any modification. Now (3.3) and part (b) of Lemma 4.1 imply the modified version of (3.3). Also, (3.4) asserts that for every set of integers $J \equiv \{i_\sigma: \sigma \in \Sigma\}$ with $i_\sigma \in \{1, \dots, m_\sigma\}$ for each $\sigma \in \Sigma$, the system

$$\bigwedge_{i \in J} (A_\sigma)_i x > (a_\sigma)_i$$

has no solution in G^n . Hence, by Corollary B.2 of Appendix B neither system has a solution in F^n , establishing the modified version of (3.4). Next, the modified version of (3.5) follows from similar arguments by examining the systems

$$A_\sigma x \leq a_\sigma, \quad A_\tau x \leq a_\tau, \quad (B_\sigma)_i x + (b_\sigma)_i < (B_\tau)_i x + (b_\tau)_i$$

for $\sigma \in \Sigma, \tau \in \Sigma$ and $i \in \{1, \dots, k\}$. Finally, assume that $\{(m'_\sigma, A'_\sigma, a'_\sigma, B'_\sigma, b'_\sigma): \sigma \in \Sigma'\}$ is an alternative representation of f for which the corresponding extension of f to F^n does not coincide with the one obtained from $\{(m_\sigma, A_\sigma, a_\sigma, B_\sigma, b_\sigma): \sigma \in \Sigma\}$. Then for some $\sigma \in \Sigma, \sigma' \in \Sigma$ and $i \in \{1, \dots, k\}$ the system

$$A_\sigma x \leq a_\sigma, \quad A'_{\sigma'} x \leq a'_{\sigma'}, \quad (B_\sigma)_i x + (b_\sigma)_i \neq (B'_{\sigma'})_i x + (b'_{\sigma'})_i$$

has a solution in F^n . Then, by Corollary B.2, this system has a solution in G^n contradicting the assumption that we had two representations of the same PL map. \square

Given a PL function $f: G^n \rightarrow G^k$ and an extension F of G , we shall denote by \bar{f}_F the extension of f to F^n constructed from any one representation of f in the way described in Lemma 4.2. In particular, we identify the set of cells in the representation of f and \bar{f}_F . Of course, f might have many other extensions to F^n . For example if $f: G \rightarrow G$ is defined by $f(x) = 0$ for all $x \in G$ and $G(\omega)$ is the ordered field of rational functions with an infinitesimal over G (see § 5 for formal definitions), then the function $h: G(\omega) \rightarrow G(\omega)$ defined by

$$h(x) = \begin{cases} 0 & \text{if } x \leq \omega \text{ or } x \geq 3\omega, \\ x - \omega, & \text{if } \omega \leq x \leq 2\omega, \\ 3\omega - x & \text{if } 2\omega \leq x \leq 3\omega \end{cases}$$

is a PL function and $f(x) = h(x)$ for all $x \in G$. In this example $\bar{f}_{G(\omega)}$ is the function mapping each $x \in G(\omega)$ into 0.

5. Extending an ordered field with an infinitesimal. We turn our attention to two important extensions of a given ordered field G . Let $G\{\omega\}$ be the set of all infinite sequences $a = (\dots, a_{-1}, a_0, a_1, \dots)$ of elements of G such that $a_i = 0$ for sufficiently small i (possibly negative). For $a \in G\{\omega\}$ we shall use subscripts to denote the elements

of the corresponding sequence. We next define addition, multiplication and order on $G\{\omega\}$. Let $a = (\cdots, a_{-1}, a_0, a_1, \cdots)$ and $b = (\cdots, b_{-1}, b_0, b_1, \cdots)$ be elements in $G\{\omega\}$. We define $a + b$ and $a \cdot b$ to be, respectively, the sequence of elements in G with $(a + b)_i = a_i + b_i$ and $(a \cdot b)_i = \sum_{k=-\infty}^{\infty} a_k b_{i-k}$, for all integers i ; of course, the infinite sum $\sum_{k=-\infty}^{\infty} a_k b_{i-k}$ has only finitely many nonzero elements and is a finite sum. We note that under the above definitions $a + b \in G\{\omega\}$ and $a \cdot b \in G\{\omega\}$. The order in $G\{\omega\}$ is taken to be the lexicographic order, namely, $a > b$ if the first nonzero elements in the sequence $(\cdots, a_{-1} - b_{-1}, a_0 - b_0, a_1 - b_1, \cdots)$ is positive; of course, as $a \in G\{\omega\}$ and $b \in G\{\omega\}$ the above sequence has a first nonzero element whenever $a \neq b$. The following lemma asserts that $G\{\omega\}$ is an ordered field under the above definitions of addition, multiplication and order.

LEMMA 5.1. *The set $G\{\omega\}$ with the above definitions of addition, multiplication and order is an ordered field. In particular, the element $x \in G\{\omega\}$ with $x_i = 0$ for all i is the "zero" in $G\{\omega\}$, and the element $x \in G\{\omega\}$ with $x_0 = 1$ and $x_i = 0$ for all integers $i \neq 0$ is the "one" in $G\{\omega\}$. Also, if $a \in G\{\omega\}$, then $(-a)_i = -a_i$ for all integers i , and if $a \neq 0$ and $a_i = 0$ for $i \leq q$ and $a_q \neq 0$, then $(a^{-1})_i = 0$ for all $i < -q$, $(a^{-1})_{-q} = (a_q)^{-1}$ and $(a^{-1})_i$ can be determined inductively for $i > -q$ by $(a^{-1})_i = -(a_q)^{-1}(\sum_{j=-q}^{i-1} (a^{-1})_j a_{i+q-j})$. \square*

Notice that G is imbedded in $G\{\omega\}$ by the imbedding $\alpha \rightarrow x$ where $x_0 = \alpha$ and $x_i = 0$ for integers $i \neq 0$. Thus, we consider G to be a subset of $G\{\omega\}$. In particular, $0 \in G$ and $1 \in G$ are clearly the "zero" and "one" elements of $G\{\omega\}$. Also, we denote the element $x \in G\{\omega\}$ where $x_1 = 1$ and $x_i = 0$ for integers $i \neq 1$ by ω . In particular, for every integer k , $(\omega^k)_k = 1$ and $(\omega^k)_i = 0$ for integers $i \neq k$. Thus, when $x \in G\{\omega\}$ we write $x = \sum_{i=-\infty}^{\infty} x_i \omega^i$. We emphasize that this is a formal notation and does not represent convergence. Also, we write $x = \sum_{i=q}^{\infty} x_i \omega^i$ for any integer q for which $x_i = 0$ for all $i < q$.

For any positive $\alpha \in G$ we have that $\alpha > \omega$ (in $G\{\omega\}$). We therefore speak of ω as being an *infinitesimal* relative to G . In particular, for every positive integer n , $n^{-1} > \omega$; so n^{-1} does not "tend" to zero in $G\{\omega\}$ as the integer n becomes large. However, ω^n does "tend" to zero in $G\{\omega\}$ as n becomes large.

Call an element $a \in G\{\omega\}$ a *polynomial form over G* if $a_i = 0$ for all negative i and all sufficiently large i . In particular, in this case, $a = \sum_{i=0}^m a_i \omega^i$. The set of all polynomial forms over G will be denoted $G[\omega]$. Of course, $G[\omega]$ is a subordered ring of $G\{\omega\}$. Next, call an element in $G\{\omega\}$ a *rational form over G* if it has the representation ab^{-1} where $a, b \in G[\omega]$ and $b \neq 0$. The set of all rational forms over G will be denoted $G(\omega)$. The next lemma characterizes this set.

LEMMA 5.2. *The set $G(\omega)$ is a subordered field of $G\{\omega\}$. Moreover, it is the smallest subordered field of $G\{\omega\}$ that includes G and contains ω . \square*

Each of the five field operations—addition, multiplication, subtraction, division and comparison—in $G\{\omega\}$ is defined via a countable number of field operations in G . However, our principal interest will be in $G[\omega]$ and $G(\omega)$ where it is next shown that each of the five field operations can be executed through a finite number of field operations in G . This is accomplished by representing elements in $G(\omega)$ by pairs of polynomial forms over G . Specifically, represent $x \in G(\omega)$ by any pair (a, b) where $a, b \in G[\omega]$, $b \neq 0$ and $x = ab^{-1}$. Evidently, if $a, b, c, d \in G[\omega]$ where $b \neq 0$ and $d \neq 0$, then $ad + bc$, $ad - bc$, bd , ac , bc and ad are polynomial forms over G and

$$(5.1) \quad ab^{-1} + cd^{-1} = (ad + bc)(bd)^{-1},$$

$$(5.2) \quad (ab^{-1})(cd^{-1}) = (ac)(bd)^{-1},$$

$$(5.3) \quad ab^{-1} - cd^{-1} = (ad - bc)(bd)^{-1},$$

$$(5.4) \quad ab^{-1}/cd^{-1} = (ad)(bc)^{-1},$$

the latter only if $c \neq 0$, and

$$(5.5) \quad ab^{-1} > cd^{-1} \quad \text{if and only if} \quad (ad - bc)(bd) > 0.$$

Hence, if x and y in $G(\omega)$ have the representation (a, b) and (c, d) , respectively, where $a, b, c, d \in G[\omega]$, $b \neq 0$ and $d \neq 0$, then $x + y$, xy , $x - y$ and x/y have the representation $(ad + bc, bd)$, (ac, bd) , $(ad - bc, bd)$ and (ad, bc) , respectively, the latter only if $x \neq 0$ which occurs if and only if $c \neq 0$. Of course, addition, multiplication and subtraction of polynomial forms over G can be executed by a finite number of operations in G . Also, (5.5) shows that $x > y$ if and only if $(ad - bc)(bd) > 0$, the latter being checkable by a finite number of comparisons in G . A bound on the finite number of field operations in G needed to execute the various field operations in $G(\omega)$ is easily obtained; such a bound, of course, depends upon the representations in $G(\omega)$.

We have seen the virtue of representing an element $x \in G\{\omega\}$ by a pair (a, b) where $a, b \in G[\omega]$, $b \neq 0$ and $x = ab^{-1}$. In the following (see Corollary 6.3) we shall have an element $x \in G(\omega)$ and will have to use the elements of the sequence $(\dots, x_{-1}, x_0, x_1, \dots)$. We shall next see how each element in this sequence can be computed by a *finite number* of field operations in G from any representation of x by a pair of corresponding polynomials. See Winograd [1980] for corresponding efficient computational procedures.

LEMMA 5.3. Suppose $x = (\dots, x_{-1}, x_0, x_1, \dots) = ab^{-1}$ where $a = \sum_{i=0}^m a_i \omega^i \in G[\omega]$, $b = \sum_{i=0}^p b_i \omega^i \in G[\omega]$ and $b \neq 0$. Let $q = \min \{i \geq 0: b_i \neq 0\}$. Then

$$(5.6) \quad x_i = \begin{cases} 0 & \text{if } i < -q, \\ \sum_{j=-q}^i c_j a_{i-j} & \text{if } i \geq -q, \end{cases}$$

where the c_j 's are defined inductively by

$$(5.7) \quad c_i = \begin{cases} 0 & \text{if } i < -q, \\ (b_q)^{-1} & \text{if } i = -q, \\ -(b_q)^{-1} \left(\sum_{j=-q}^{i-1} c_j b_{i+q-j} \right) & \text{if } i > -q. \end{cases}$$

Proof. Let $c = b^{-1}$. Then (5.6) follows from the representation of the inverse of an element in $G\{\omega\}$ given in Lemma 5.1. So, (5.7) follows from the rules of multiplication in $G\{\omega\}$. \square

For a positive integer p , let $p! = p(p-1) \cdots 1$ and let $p! = 1$ if $p = 0$. Next for $t \in G$ and integer p we define the *binomial coefficient t choose p* , denoted $\binom{t}{p}$, by

$$(5.8) \quad \binom{t}{p} = \begin{cases} [t(t-1) \cdots (t-p+1)]/p! & \text{for } p = 1, 2, \dots \text{ and all } t \in G, \\ 1 & \text{for } p = 0 \text{ and all } t \in G, \\ 0 & \text{for } p = -1, -2, \dots \text{ and all } t \in G. \end{cases}$$

It is easy to verify the (standard) identity that

$$(5.9) \quad \binom{t}{p} + \binom{t}{p-1} = \binom{t+1}{p} \quad \text{for all integers } p \text{ and all } t \in G.$$

LEMMA 5.4. Suppose $a \in G\{\omega\}$ where $a = \sum_{j=-k}^0 a_j \omega^j$. If $a \leq 0$ then for sufficiently large $t \in G$, $\sum_{i=0}^k a_{-i} \binom{t}{i} \leq 0$.

Proof. The result is trivial if $a = 0$. In particular, this happens if $k < 0$. Next assume that $a \neq 0$ and $k \geq 0$. Let $q = \max \{i \leq k: a_{-i} \neq 0\}$. As $a \leq 0$, clearly $a_{-q} < 0$. Now, if $q = 0$, then $\sum_{i=0}^k a_{-i} \binom{t}{i} = a_0 < 0$ for all $t \in G$, establishing the desired assertion. Next assume that $q > 0$. Now, if $t > 2q - 1$ then for $j = 1, \dots, q$, $(t - j + 1)j^{-1} > 1$, implying that for $i = 1, \dots, q - 1$

$$0 < \binom{t}{i} (t - q) q^{-1} \leq \binom{t}{i} \prod_{j=i+1}^q [(t - j + 1)j^{-1}] = \binom{t}{q}.$$

Hence, if in addition $(t - q)|a_{-q}| > (\sum_{i=0}^{q-1} |a_{-i}|)q$, then

$$\begin{aligned} \sum_{i=0}^k a_{-i} \binom{t}{i} &= -|a_{-q}| \binom{t}{q} + \sum_{i=0}^{q-1} a_{-i} \binom{t}{i} \\ &\leq -|a_{-q}| \binom{t}{q} + \sum_{i=0}^{q-1} |a_{-i}| (t - q)^{-1} q \binom{t}{q} \leq 0, \end{aligned}$$

completing our proof. \square

Each coordinate of a matrix $A \in G\{\omega\}^{m \times n}$ has representation as a power series in ω with coefficients in G . Thus, A can be expressed as a corresponding power series in ω with coefficients in $G^{m \times n}$, e.g., $A = \sum_{i=-\infty}^{\infty} A(t)\omega^i$ where each $A(t)$ is in $G^{m \times n}$. Addition and multiplication generalize to matrices in the natural way. In particular, we have the “infinite distributivity” asserted in the following lemma. The verification of the result is straightforward and is omitted.

LEMMA 5.5. *Let $A \in G^{m \times n}$ and $B \in G\{\omega\}^{n \times p}$ where $B = \sum_{i=-\infty}^{\infty} B(t)\omega^i$. Then $AB = \sum_{i=-\infty}^{\infty} [AB(t)]\omega^i$. \square*

6. Existence of invariant polynomial curves. In this section we show how an invariant polynomial curve of a PL map f can be computed from a fixed point of a perturbation of $\bar{f}_{G(\omega)}$. For notational convenience we will use the notation \bar{f} for $\bar{f}_{G(\omega)}$. We consider extensions of PL functions to $G(\omega)$ rather than $G\{\omega\}$ because the field operations in $G(\omega)$ are executable by a finite number of field operations in G (see § 5).

We first use Corollary 3.2 to determine conditions that assure that for a given PL map f on G^n and β in $G(\omega)$, $\beta\bar{f}$ has a (computable) fixed point (in $G(\omega)^n$).

PROPOSITION 6.1. *Let $f: G^n \rightarrow G^n$ be a PL map having the representation $\{(m_\sigma, A_\sigma, a_\sigma, B_\sigma, b_\sigma): \sigma \in \Sigma\}$ and let $\beta \in G(\omega)$. Suppose that for every unbounded cell $\sigma \in \Sigma$, the signs of $\det(\beta B_\sigma - I)$ (as elements in $G(\omega)$) are either all positive or are all negative. Then there exists a vector $x \in G(\omega)^n$ satisfying*

$$(6.1) \quad \beta\bar{f}(x) = x.$$

Proof. Evidently, $\beta\bar{f}$ is a PL map on $G(\omega)^n$ having the representation $\{(m_\sigma, A_\sigma, a_\sigma, \beta B_\sigma, \beta b_\sigma): \sigma \in \Sigma\}$ (see the proof of Lemma 4.2). By Lemma 4.1 $\{x \in G^n: A_\sigma x \leq a_\sigma\} \subseteq G^n$ is unbounded if and only if $\{x \in G(\omega)^n: A_\sigma x \leq a_\sigma\} \subseteq G(\omega)^n$ is unbounded. It follows that $\beta\bar{f}$ with the above representation satisfies the assumptions of Theorem 3.1 and therefore the existence of $x \in G(\omega)^n$ satisfying (6.1) follows from Corollary 3.2. \square

We remind the reader that the proof of Corollary 3.2 was constructive. In particular, the algorithm described in Appendix A can be used to compute fixed points of the map $\beta\bar{f}$ examined in Proposition 6.1.

We next show how an invariant polynomial curve of a PL map f can be obtained from a solution of (6.1) with $\beta = (1 + \omega)^{-1}$. In Appendix C of Eaves and Rothblum [1986] it is shown how invariant polynomial curves of f can be computed from solutions of (6.1) for other values of β ; however, we consider here only the (restricted) case of

$\beta = (1 + \omega)^{-1}$ for two reasons. First, the expressions representing the corresponding polynomials are simpler, and second, as is shown in Appendix C of Eaves and Rothblum [1986] solutions of (6.1) for other corresponding values of β can be obtained by the methods discussed earlier in this paper if and only if this can be done for $\beta = (1 + \omega)^{-1}$. That is, the wider range of corresponding β 's does not increase the applicability of the methods of this paper for computing invariant polynomial curves.

Given a PL map f for which (6.1) has a solution over $G(\omega)^n$ for $\beta = (1 + \omega)^{-1}$, we shall determine a function $z: G \rightarrow G^n$ for which

$$(6.2) \quad f(z[t]) = z[t+1] \quad \text{for all } t \text{ sufficiently large.}$$

Of course, if z satisfies (6.2) with the corresponding equality holding for all $t \geq t^*$, then $y: \{t \in G: t \geq 0\} \rightarrow G^n$ defined by $y[t] = z[t + t^*]$ satisfies (1.1). So, if z is a polynomial in its variable t , then y is an invariant polynomial curve of f .

Recall the binomial coefficients defined by (5.8).

THEOREM 6.2. *Let $f: G^n \rightarrow G^n$ be a PL map and let $x \in G(\omega)^n$ satisfy $(1 + \omega)^{-1}\bar{f}(x) = x$. Suppose $x = \sum_{i=-q}^{\infty} x_i \omega^i$ where q is a nonnegative integer. Consider the function $z: \{t \in G: t \geq 0\} \rightarrow G^n$ defined by*

$$(6.3) \quad z[t] = \sum_{i=0}^q \binom{t}{i} x_{-i}.$$

Then for sufficiently large t

$$(6.4) \quad f(z[t]) = z[t+1].$$

Proof. Let $\sigma \in \Sigma$ be a cell containing x , i.e., $A_\sigma x \leq a_\sigma$. Lemma 5.5 implies that

$$\sum_{i=-q}^{\infty} (A_\sigma x_i) \omega^i = A_\sigma \left(\sum_{i=-q}^{\infty} x_i \omega^i \right) = A_\sigma x \leq a_\sigma,$$

immediately implying that $\sum_{i=-q}^0 (A_\sigma x_i) \omega^i \leq a_\sigma$. It now follows from Lemma 5.4 that for sufficiently large t , $\sum_{i=0}^q \binom{t}{i} A_\sigma x_{-i} \leq a_\sigma$, and therefore, by the associativity of matrix multiplication, for such t , $A_\sigma \left(\sum_{i=0}^q \binom{t}{i} x_{-i} \right) \leq a_\sigma$, i.e., $z[t] \in \sigma$.

We next observe that as $\bar{f}(x) = (1 + \omega)x$ and $x \in \sigma$, we have from Lemma 5.5 that

$$\begin{aligned} \sum_{i=-q}^{\infty} (B_\sigma x_i) \omega^i + b_\sigma &= B_\sigma \left(\sum_{i=-q}^{\infty} x_i \omega^i \right) + b_\sigma \\ &= B_\sigma x + b_\sigma = \bar{f}(x) = (1 + \omega)x = (1 + \omega) \left(\sum_{i=-q}^{\infty} x_i \omega^i \right) \\ &= \sum_{i=-q}^{\infty} x_i \omega^i + \sum_{i=-q}^{\infty} x_i \omega^{i+1}. \end{aligned}$$

By equating corresponding coefficients of the left-hand side and right-hand side of the above equation we conclude that

$$(6.5) \quad B_\sigma x_i = x_i + x_{i-1}, \quad i = -q, \dots, -1$$

and

$$(6.6) \quad B_\sigma x_0 + b_\sigma = x_0 + x_{-1}.$$

(By convention, $x_{-q-1} = 0$.) Now for t for which $z[t] \in \sigma$ we have, from elementary arguments (6.5) and (6.6), that

$$\begin{aligned} f(z[t]) &= B_\sigma x[t] + b_\sigma = B_\sigma \left[\sum_{i=0}^q \binom{t}{i} x_{-i} \right] + b_\sigma \\ &= \sum_{i=0}^q \binom{t}{i} B_\sigma x_{-i} + b_\sigma = \sum_{i=0}^q \binom{t}{i} (x_{-i} + x_{-i-1}) \\ &= \sum_{i=0}^q \binom{t}{i} x_{-i} + \sum_{i=1}^{q+1} \binom{t}{i-1} x_{-i} \\ &= \sum_{i=1}^q \left[\binom{t}{i} + \binom{t}{i-1} \right] x_{-i} + x_0 = \sum_{i=0}^q \binom{t+1}{i} x_i = z[t+1], \end{aligned}$$

proving that for sufficiently large t , $z[t]$ satisfies (6.4). \square

We next combine Proposition 6.1 and Theorem 6.2 to obtain a sufficient condition for a PL map to have an invariant polynomial curve.

COROLLARY 6.3. *Let $f: G^n \rightarrow G^n$ be a PL map having the representation $\{(m_\sigma, A_\sigma, a_\sigma, B_\sigma, b_\sigma): \sigma \in \Sigma\}$. Suppose that for every unbounded cell σ , the signs of $\det((1+\omega)^{-1}B_\sigma - I)$ are either all positive or are all negative. Then f has an invariant polynomial curve.* \square

We note that a PL map f need not have an invariant polynomial curve when no assumptions are imposed on f . For example, let $G = \mathbb{R}$ and $f: \mathbb{R} \rightarrow \mathbb{R}$ be defined by

$$f(x) = \begin{cases} 2x+1 & \text{if } x \geq 0, \\ -2x+1 & \text{if } x \leq 0, \end{cases}$$

and suppose that $z[\cdot]$ satisfies (6.4) for $t \geq t^*$. Let $\beta \equiv f(z(t^*))$. As $\beta > 0$ we conclude from (6.4) and the explicit definition of f that for $k = 1, 2, \dots$

$$z(t^* + k) = f^k(z(t^*)) = f^{k-1}(\beta) = \sum_{i=0}^{k-2} 2^i + 2^{k-1}\beta = 2^k(1+\beta) - 1,$$

immediately implying that $z[t]$ is not a polynomial in t . Of course, by arbitrarily selecting $z[t]$ for $0 \leq t \leq 1$, one can inductively construct $z[t]$ for all $t \geq 0$ so that (6.4) is satisfied (when G is not Archimedean the induction is transfinite).

We next demonstrate that for every positive integer k there exists a PL map having an invariant polynomial curve of degree k only. Let k be a positive integer. Set $G = \mathbb{R}$ and consider the function $f: \mathbb{R}^k \rightarrow \mathbb{R}^k$ defined by $f(x) = Bx + b$ where

$$B = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & \cdots & 0 & 0 \\ & & & \ddots & & \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}.$$

The results of Eaves and Rothblum [1986], [1987] show that this map satisfies the assumptions of Theorem 6.2. We will next construct an invariant polynomial curve of the above map. For $j = 1, \dots, k$, let e^j be the j th unit vector in \mathbb{R}^k and let $e^{k+1} \equiv 0$. Evidently, $(B - I)^k = 0$ and for $j = 1, \dots, k$, $(B - I)^{j-1}b = e^j$ and $Be^j = e^{j+1} + e^j$. Let

$z[t] \equiv \sum_{i=1}^k \binom{t}{i} e^i$. Then $z[t]$ is a vector polynomial in t of degree k and for $t \geq 0$ we have from the above and (5.9) that

$$\begin{aligned} Bz[t] &= \sum_{i=1}^k \binom{t}{i} B e^i = \sum_{i=1}^k \binom{t}{i} (e^{i+1} + e^i) = \sum_{i=1}^k \left[\binom{t}{i} + \binom{t}{i-1} \right] e^i \\ &= \sum_{i=1}^k \binom{t+1}{i} e^i = z[t+1], \end{aligned}$$

establishing (6.4). We next argue that if $z[\cdot]$ is any polynomial satisfying (6.4) for sufficiently large t , then the degree of $z[\cdot]$ is k . Now suppose that $z[\cdot]$ is such a polynomial and that (6.4) holds for $t \geq t^*$. Then, for $p = 1, 2, \dots$ we get from the binomial formula that

$$\begin{aligned} z[t^* + p] &= f^p(z[t^*]) = \sum_{i=0}^{p-1} B^i b + B^p z[t^*] \\ &= \sum_{i=0}^{p-1} \sum_{j=0}^i \binom{i}{j} (B - I)^j b + \sum_{j=0}^p \binom{p}{j} (B - I)^j z[t^*]. \end{aligned}$$

As $(B - I)^k = 0$ and $(B - I)^j b = e^{j+1}$ for $j = 1, \dots, k-1$ we get that

$$\begin{aligned} z[t^* + p] &= \sum_{i=0}^{p-1} \sum_{j=0}^{k-1} \binom{i}{j} e^{j+1} + \sum_{j=0}^{p-1} \binom{p}{j} (B - I)^j z[t^*] \\ &= \sum_{j=0}^{k-1} \left(\sum_{i=0}^{p-1} \binom{i}{j} \right) e^{j+1} + \sum_{j=0}^{p-1} \binom{p}{j} (B - I)^j z[t^*] \\ &= \sum_{j=0}^{k-1} \binom{p}{j+1} e^{j+1} + \sum_{j=0}^{p-1} \binom{p}{j} (B - I)^j z[t^*] \end{aligned}$$

is a polynomial in p of degree k (with leading coefficient $(k!)^{-1} e^k$). It immediately follows that $z[t^* + p]$ is a polynomial in p of degree k (with leading coefficient $(k!)^{-1} e^k$). Standard arguments show that the degree of the polynomial $z[t]$ (in the variable t) is k .

We include the following theorem for its curiosity value; indeed, we do not know what to make of it.

THEOREM 6.4. *Let $f: G^n \rightarrow G^n$ be a PL map and let $x \in G(\omega)^n$ satisfy $(1 - \omega)^{-1} \bar{f}(x) = x$. Suppose $x = \sum_{i=-q}^{\infty} x_i \omega^i$ where q is a nonnegative integer. Consider the function $z: \{t \in G: t \geq 0\} \rightarrow G^n$ defined by*

$$(6.7) \quad z[t] = \sum_{i=0}^q \binom{t}{i} x_{-i}.$$

Then for sufficiently large t

$$(6.8) \quad f(z[t]) = z[t-1].$$

Proof. The proof of this theorem follows from arguments similar to those used in the proof of Theorem 6.2. We review only the modifications. First let $\{(m_\sigma, A_\sigma, a_\sigma, B_\sigma, b_\sigma)\}: \sigma \in \Sigma\}$ be a representation of f and let $\tau \in \Sigma$ satisfy $A_\tau x \leq a_\tau$. The arguments of the proof of Theorem 6.2 show that for sufficiently large t , $z[t] \in \tau$. Next, (6.5) and (6.6) have to be modified to

$$B_\tau x_i = x_i - x_{i-1}, \quad i = -q, \dots, -1$$

and

$$B_\tau x_0 + b_\tau = x_0 - x_{-1}.$$

So, for $t \in \sigma$ with $z[t] \in \tau$,

$$\begin{aligned} f(z[t]) &= b_\tau + B_\tau z[t] = b_\tau + B_\tau \left[\sum_{i=0}^q \binom{t}{i} x_{-i} \right] \\ &= x_0 - x_{-1} + \sum_{i=1}^q \binom{t}{i} (x_{-i} - x_{-i-1}) \\ &= x_0 + \sum_{i=1}^q \left[\binom{t}{i} - \binom{t}{i-1} \right] x_{-i} \\ &= x_0 + \sum_{i=1}^q \binom{t-1}{i} x_{-i} = z[t-1]. \end{aligned} \quad \square$$

Our final result in this section characterizes solvability of (6.1) with $\beta = (1 + \omega)^{-1}$ in terms of solvability of related systems over G .

THEOREM 6.5. *Let $f: G^n \rightarrow G^n$ be a PL map. Then there exists a vector $x \in G(\omega)^n$ satisfying*

$$(6.9) \quad (1 + \omega)^{-1} \bar{f}(x) = x$$

if and only if for sufficiently small positive ε the system

$$(6.10) \quad (1 + \varepsilon)^{-1} f(x) = x$$

has a solution in G^n .

Proof. The conclusion follows from Corollary B.2 and the arguments used in Kohlberg [1980, § 6] (see also Eaves and Rothblum [1985, § 5]). \square

The previous lemma was the major tool in the existence proof of fixed points of $(1 + \omega)^{-1} \bar{f}$ in Kohlberg [1980]. In that paper, Banach's contraction theorem was used to show that if f is nonexpansive, then $(1 + \varepsilon)^{-1} f$ has a fixed point for each $\varepsilon > 0$, and therefore so does $(1 + \omega)^{-1} \bar{f}$. We note, however, that the contraction arguments cannot be used in any other ordered field but the reals because of the lack of completeness.

7. Representations of invariant polynomial curves and bounds on their degrees. Our construction of an invariant polynomial curve of a PL map f given in Theorem 6.2 uses a fixed point $x \in G(\omega)^n$ of the PL map $(1 + \omega)^{-1} \bar{f}$. It follows directly from this construction that the degree of the corresponding invariant polynomial curve equals the integer q that is the largest integer i with $x_{-i} \neq 0$; moreover, the constructed invariant polynomial curve is determined by x_{-q}, \dots, x_0 . Now, suppose it is known that x lies in a specific cell $\sigma \in \Sigma$. Then $(1 + \omega)^{-1} (B_\sigma x + b_\sigma) = x$. As $\det [(1 + \omega)I - B_\sigma]$ is a polynomial in ω of degree n we have that it is nonzero. So, $(1 + \omega)I - B_\sigma$ is nonsingular and x has the representation

$$(7.1) \quad x = [(1 + \omega)I - B_\sigma]^{-1} b_\sigma.$$

It follows that representations of $[(1 + \omega)I - B_\sigma]^{-1}$ can be used to compute x .

Equation (7.1) suggests that representations of inverses of matrices of the form $\omega I - Q \in G(\omega)^{n \times n}$, where $Q \in G^{n \times n}$, can be useful. We will obtain such a representation after introducing some definitions and stating an auxiliary lemma.

Let $B \in G^{n \times n}$, $\lambda \in G$ and $Q = \lambda I - B$. The *index* of λ for B , denoted $\nu_\lambda(B)$, is the smallest integer $m \geq 0$ for which the null spaces of Q^m and Q^{m+1} coincide. In particular, we have that $\nu_\lambda(B) \leq n$. The following result is well known (see Rothblum [1981] where the case $G = R$ is considered).

LEMMA 7.1. Let $Q \in G^{n \times n}$, let ν be the index of zero for Q , and let M and N be, respectively, the range and null spaces of Q^ν . Then:

- (a) $G^n = M + N$,
- (b) there is a unique projection E on N along M ,
- (c) $Q^\nu E = 0$,
- (d) $Q^{\nu-1} E \neq 0$ if $\nu > 0$,
- (e) $(Q - E)$ is nonsingular,
- (f) $QD = I - E$ where $D \equiv (Q - E)^{-1}(I - E)$,
- (g) $QD^{j+1} = D^j$ for $j = 1, 2, \dots$ \square

For a given matrix $B \in G^{n \times n}$ and $\lambda \in G$, the matrices E and D defined in Lemma 7.1 for $Q \equiv B - \lambda I$ are called, respectively, the *eigenprojection* and *Drazin inverse* of B at λ . Methods to compute these matrices are well known (e.g., Rothblum [1976] or Campbell and Meyer [1979]).

We are now ready for the promised representation of inverses of matrices of the form $\omega I - Q \in G(\omega)^{n \times n}$ where $Q \in G(\omega)^{n \times n}$. The result resembles known expansions of resolvents of complex matrices (see Rothblum [1981, Thm. 3.1]).

THEOREM 7.2. Let $Q \in G^{n \times n}$ have eigenprojection E and Drazin inverse D and let ν be the index of zero for Q . Then $\omega I - Q$ is nonsingular and

$$(7.2) \quad (\omega I - Q)^{-1} = \sum_{j=1}^{\nu} Q^{j-1} E \omega^{-j} - \sum_{j=0}^{\infty} D^{j+1} \omega^j.$$

Proof. First observe that part (c) of Lemma 7.1 implies that

$$(7.3) \quad (\omega I - Q) \left(\sum_{j=1}^{\nu} Q^{j-1} E \omega^{-j} \right) = E - Q^\nu E = E,$$

and parts (f) and (g) of Lemma 7.1 imply that

$$(7.4) \quad (\omega I - Q) \left(\sum_{j=0}^{\infty} D^{j+1} \omega^j \right) = -QD + \sum_{j=1}^{\infty} (D^j - QD^{j+1}) \omega^j = -(I - E).$$

As noted earlier $\det(\omega I - Q)$ is a polynomial in ω of degree n and therefore $\det(\omega I - Q) \neq 0$. So, $\omega I - Q$ is nonsingular. Subtracting (7.4) from (7.3) and premultiplying the resulting equation by $(\omega I - Q)^{-1}$ yields (7.2). \square

Theorem 7.2 yields an immediate bound on the degree of invariant polynomial curves constructed by the methods of Theorem 6.2.

THEOREM 7.3. Let $f: G^n \rightarrow G^n$ be a PL map having the representation $\{(m_\sigma, A_\sigma, a_\sigma, B_\sigma, b_\sigma): \sigma \in \Sigma\}$. Let $x \in G(\omega)^n$ satisfy $(1 + \omega)^{-1} \tilde{f}(x) = x$ and suppose that $A_\tau x \leq a_\tau$, i.e., $x \in \tau$. Let $z[\cdot]$ be the polynomial obtained from x via the construction described in Theorem 6.2. Then the degree d of $z[\cdot]$ is bounded from above by $\nu_1(B_\tau)$; in particular, $d \leq \max \{\nu_1(B_\sigma): \sigma \in \Sigma\} \leq n$.

Proof. Theorem 6.2 shows that if $q \geq 0$ and $x_i = 0$ for all $i < -q$ then the degree of the constructed invariant polynomial curve is bounded by q . By (7.1) we have that $x = [\omega I - (B_\tau - I)]^{-1} b_\tau$ and therefore Theorem 6.2 ensures that $x_i = 0$ for all $i < -\nu_1(B_\tau)$, immediately implying our conclusions. \square

The example following the proof of Theorem 6.2 shows that the bound in Theorem 7.3 is tight and there exist PL maps mapping G^n into itself while satisfying the assumptions of Theorem 6.2 where the degree of any invariant polynomial curve is n .

A norm on R^n is a function $\|\cdot\|$, mapping each $x \in R^n$ into $\|x\| \in R$ such that $\|x\| = 0$ if and only if $x = 0$, $\|x + y\| \leq \|x\| + \|y\|$ and $\|\lambda x\| = |\lambda| \|x\|$ for all $x, y \in R^n$. For

example, the l_∞ -norm defined by $\|x\|_\infty = \max_i |x_i|$ for all $x \in R^n$ is a norm. A map $f: G^n \rightarrow G^n$ is called *nonexpansive* with respect to the norm $\|\cdot\|$ if

$$\|f(x) - f(y)\| \leq \|x - y\| \quad \text{for all } x, y \in G^n.$$

We conclude this section by demonstrating that the degree of invariant polynomial curves of nonexpansive PL maps over the reals are necessarily of degree one or less (see Kohlberg [1980]).

LEMMA 7.4. *Let $f: R^n \rightarrow R^n$ be a PL map that is nonexpansive with respect to some norm defined on R^n . Suppose $y[\cdot]$ is an invariant polynomial curve of f . Then the degree of $y[\cdot]$ is one or less.*

Proof. Suppose $y[t] = \sum_{j=0}^q y_j t^j$ for all $t \geq 0$ where $q > 1$ and $y_q \neq 0$ and we will establish a contradiction. For $t \geq 1$ we have that

$$\|y[t+1] - y[t]\| = \|f(y[t]) - f(y[t-1])\| \leq \|y[t] - y[t-1]\|.$$

A simple inductive argument shows that for each positive integer k ,

$$(7.5) \quad \|y(k) - y(k-1)\| \leq \|y(1) - y(0)\|.$$

But $y(k+1) - y(k)$ can be easily expressed as a polynomial in k of degree $q-1$ with leading coefficient qy_q , say $qy_q k^{q-1} + \sum_{j=0}^{q-2} \beta_j k^j$. As $\|qy_q k^{q-1} + \sum_{j=0}^{q-2} \beta_j k^j\| \geq q\|y_q\| k^{q-1} - \sum_{j=0}^{q-2} \|\beta_j\| k^j \rightarrow \infty$ as $k \rightarrow \infty$, we get a contradiction to (7.5). \square

Appendix A. The PL homotopy method. Let $g: G^n \rightarrow G^n$ be a PL map with representation $\{(m_\sigma, A_\sigma, a_\sigma, B_\sigma, b_\sigma): \sigma \in \Sigma\}$. Our purpose here is to briefly describe the PL homotopy method for solving $g(x) = y$ under the conditions of Theorem 3.1; we follow the formalism of Eaves [1976]. The general idea is captured in the homotopy principle (see Eaves [1972]).

Homotopy principle. To solve a given system of equations, the system is first deformed to one which is trivially solved. Beginning with the solution to the trivial problem a route of solutions is followed as the system is deformed, perhaps with retrogression, back to the given system. The route terminates with a solution to the given problem.

For our purposes herein we assume that the $\det B_\sigma$ for all unbounded σ of Σ have the same nonzero sign (see Chein and Kuh [1976]). First observe that $g(x)$ is bounded when x ranges over the finite union of bounded cells σ of Σ . Also, since $\det B_\sigma \neq 0$ for all unbounded σ in Σ , we have that $g(x) = B_\sigma x + b_\sigma$ is unbounded when x ranges over any given unbounded σ in Σ . Thus, for each unbounded σ in Σ there exists some $x^\sigma \in \sigma$ having the property that $g(x^\sigma) \neq g(x)$ for all x belonging to any bounded cell σ . Let $x^1 \in G^n$ be any vector having the property that $g(x^1) \neq g(x)$ for all x belonging to bounded cells σ in Σ .

We deform the target system $g(x) = y$ to the trivially solved system $g(x) = g(x^1)$. Specifically, define the PL homotopy $F: G^{n+1} \rightarrow G^n$ by $F(x, \theta) = g(x) + \theta(y - g(x^1))$ and consider the system

$$F(x, \theta) = y.$$

For $\theta = 1$ we have a system with solution x^1 and for $\theta = 0$ we have the target system.

In the full PL homotopy method the next step is to perturb y infinitesimally to avoid degeneracies, however, here we forego this technical matter and merely assume nondegeneracy (see Eaves [1976] for details).

A *path* is defined to be a *route* or *loop* which, in turn, is defined to be a finite union of closed line segments that is PL homeomorphic to the line G or to the PL

circle $\{x \in G^2: \|x\|_\infty = 1\}$, respectively. Assuming nondegeneracy, the set $F^{-1}(y) \equiv \{(x, \theta): F(x, \theta) = y\}$ is known to be a disjoint collection of routes and loops (see Eaves [1976] for details).

Beginning at the point $(x^1, 1)$ in $F^{-1}(y)$ we move along the path in $F^{-1}(y)$ that contains $(x^1, 1)$ in the direction that θ decreases. With such an orientation at the beginning it can be shown that the assumption concerning the determinants of the B_σ 's implies that θ will decrease on any unbounded cell $\sigma \in \Sigma$ (see Eaves [1976, § 12]). Consequently, θ cannot return to unit value on the selected path. To see this fact suppose that the first point (x, θ) along the path with $\theta = 1$ is the point $(w, 1)$. In particular, $g(w) = g(x^1)$ and therefore by choice of x^1 , w cannot belong to any bounded cell; hence, w must belong to some unbounded cell. We conclude that when moving along the selected path in the selected orientation, θ decreases as (x, θ) passes through $(w, 1)$. Since θ decreases as (x, θ) leaves the initial point $(x^1, 1)$, the mean value theorem for PL functions in ordered fields ensures that somewhere between $(x^1, 1)$ and $(w, 1)$ we pass through a point (x, θ) with $\theta = 1$. This conclusion contradicts the selection of $(w, 1)$. So, it follows that θ cannot return to unit value, and therefore the selected path must be a route. For $\sigma \in \Sigma$, let ρ_σ be the intersection of our route with the set $\{(x, \theta): x \in \sigma, \theta \in G\}$. It follows that if σ is bounded, ρ_σ is either empty or a line segment, and if σ is unbounded ρ_σ is either empty or is a line segment or half line. Since a route is PL homeomorphic to G we cannot just pass through finitely many line segments, but must reach a halfline. Since such a halfline corresponds to an unbounded cell σ we have that θ decreases on it indefinitely. It now follows from the mean value theorem for PL functions over ordered fields that we will encounter a point of the form $(x^0, \theta^0) = (x^0, 0)$. In particular, $F(x^0, 0) = y$, or equivalently, $g(x^0) = y$. So, x^0 is a solution to the target system.

The task of following the path in $F^{-1}(y)$ is essentially an oscillation between pivot steps as in the simplex method and movement from cell to adjacent cell in the covering Σ .

Appendix B. Linear inequalities over ordered fields. In this Appendix we review some useful facts about solvability of systems consisting of linear inequalities. Our discussion borrows from Kohlberg [1976].

The first result is a variant of Farkas' Lemma (e.g., Stoer and Witzgall [1970]).

LEMMA B.1. *A system*

$$Ax \leq a, \quad Bx \ll b, \quad Cx = c,$$

where $A \in G^{m \times n}$, $B \in G^{k \times n}$, $C \in G^{p \times n}$, $a \in G^m$, $b \in G^k$ and $c \in G^p$, has no solution $x \in G^n$ if and only if there exist vectors $\lambda \in G^{1 \times m}$, $\mu \in G^{1 \times k}$ and $\eta \in G^{1 \times p}$ such that

$$\lambda A + \mu B + \eta C = 0,$$

$$\lambda \geq 0, \quad \mu \geq 0,$$

$$\lambda a + \mu b + \eta c \leq 0,$$

$$(\mu, \lambda a + \mu b + \eta c) \neq 0.$$

□

The following corollary to Farkas' Lemma will be useful.

COROLLARY B.2. *Suppose a finite system of equalities, weak inequalities and strict inequalities with coefficients in the ordered field G does not have a solution in G . Then it also has no solution in any ordered field extension of G .*

Proof. The existence of Farkas' multipliers in G follows from our assumptions and Lemma B.1. Since these multipliers are also in F , Lemma B.1 implies that the system does not have solutions in F . □

Acknowledgment. The authors would like to express their appreciation to a referee for his careful reading and thoughtful suggestions.

REFERENCES

- E. ALLGOWER AND K. GEORG [1980], *Simplicial and continuation methods for approximating fixed points and solutions to systems of equations*, SIAM Rev., 22, pp. 28–85.
- S. L. CAMPBELL AND C. D. MEYER [1979], *Generalized Inverses of Linear Transformations*, Pitman, London.
- M. J. CHEIN AND E. S. KUH [1976], *Solving piecewise linear equations for resistive networks*, Circuit Theory and Applications, 4, pp. 3–24.
- B. C. EAVES [1972], *Homotopies for the computation of fixed points*, Math. Programming, 3, pp. 1–22.
- [1976], *A short course for solving equations with LP homotopies*, SIAM-AMS Proc., 9, pp. 73–143.
- B. C. EAVES AND U. G. ROTHBLUM [1984], *Computation in linear and piecewise linear structures*, unpublished manuscript.
- [1985], *A theory on extending algorithms for parametric problems*, manuscript.
- [1986], *Invariant Polynomial Curves of Piecewise Linear Maps*, SOL 86–13 Department of Operations Research, Stanford University, Stanford, CA.
- [1987], *More on properties of piecewise linear functions*, in preparation.
- B. C. EAVES AND H. SCARF [1976], *The solution of systems with piecewise linear equations*, Math. Oper. Res., 1, pp. 1–27.
- E. KOHLBERG [1980], *Invariant half-lines of nonexpansive piecewise linear transformations*, Math. Oper. Res., 5, pp. 366–372.
- U. G. ROTHBLUM [1976], *Computation of the eigenprojection of a nonnegative matrix at its spectral radius*, Math. Programming Stud., 6, pp. 188–201.
- [1981], *Resolvent expansions of matrices and applications*, Linear Algebra Appl., 38, pp. 33–49.
- U. G. ROTHBLUM AND A. F. VEINOTT, JR. [1987], *Branching Markov decision chains*, in progress.
- J. STOER AND C. WITZGALL [1970], *Convexity and Optimization in Finite Dimensions I*, Springer-Verlag, New York.
- M. J. TODD [1976], *The Computation of Fixed Points and Applications*, Lecture Notes in Economics and Mathematical Systems 124, Springer-Verlag, New York.
- S. WINOGRAD [1980], *Arithmetic Complexity of Computation*, CBMS-NSF Regional Conference Series in Applied Mathematics 33, Society for Industrial and Applied Mathematics, Philadelphia, PA.

THE EXPANDED LAGRANGIAN SYSTEM FOR CONSTRAINED OPTIMIZATION PROBLEMS*

A. B. POORE† AND Q. AL-HASSAN†

Abstract. Smooth penalty functions can be combined with numerical continuation and bifurcation techniques to produce a class of robust and potentially fast algorithms for constrained optimization problems. The key to the development of these algorithms is an expanded Lagrangian system which is derived and analyzed in this work. This parametrized system of nonlinear equations contains the penalty path as a solution, provides a smooth path into the first-order necessary conditions, and sometimes yields a method for finding multiple optima. Furthermore, the inevitable ill-conditioning present in a sequential optimization algorithm is removed for three and essentially only three smooth penalty functions: the quadratic penalty function for equality constraints, the logarithmic barrier function (an interior method), and the quadratic loss function (an exterior method) for inequality constraints. Preliminary test results for several nontrivial test problems are included to demonstrate the robustness, efficiency and potential of the methodology.

Key words. smooth penalty functions, logarithmic barrier, quadratic loss and quadratic penalty functions, expanded Lagrangian system, constrained optimization, nonlinear programming, numerical bifurcation and continuation techniques

AMS(MOS) subject classifications. 49, 65, 90

1. Introduction. The use of smooth penalty functions in a sequential optimization algorithm to solve constrained optimization problems has long been regarded as unfashionable and numerically defective, principally because of an inevitable ill-conditioning that occurs as the penalty parameter tends to the prescribed limit. This view, which is prevalent in most texts and review articles [6], [7], [9], [15], [16], is quite valid and has certainly motivated development of other numerical optimization techniques such as exact penalty functions and multiplier methods. Our objective herein is to re-examine the use of these smooth penalty functions, not in a sequential optimization algorithm, but in a continuation-based algorithm for following the penalty function path defined as a solution of a parametrized system of nonlinear equations. This methodology now becomes viable because of the extensive development of numerical continuation/bifurcation techniques during the last decade [2], [12], [13], [18] and because we can remove the ill-conditioning for three fundamentally important smooth penalty functions. These continuation methods are capable of producing robust and potentially fast algorithms, and some times provide a method for finding multiple optima.

Although the techniques discussed here apply to constrained optimization in general and to linear and nonlinear programming, calculus of variations, optimal control and parameter identification in particular, the main focus in this work is on the general nonlinear programming problem

$$(1.1) \quad \min \{f(x) | h(x) = 0, g(x) \geq 0\}$$

where $f: \mathbb{R}^n \rightarrow \mathbb{R}^1$, $h: \mathbb{R}^n \rightarrow \mathbb{R}^q$ and $g: \mathbb{R}^n \rightarrow \mathbb{R}^p$ are assumed to be twice continuously differentiable. The essence of the idea is as follows: We first convert the constrained

* Received by the editors July 7, 1986; accepted for publication (in revised form) August 31, 1987. The research of the first author was supported by National Science Foundation grant DMS-85-10201, by Air Force Office of Scientific Research grant AFOSR-ISSA-85-00079 and by National Aeronautics and Space Administration contract NAS1-18107 while he was in residence at the Institute for Computer Applications in Science and Engineering, NASA-Langley Research Center, Hampton, Virginia 23665-5225.

† Department of Mathematics, Colorado State University, Fort Collins, Colorado 80523.

optimization problem to an unconstrained optimization problem $\min P(x, r)$ where r is the penalty parameter arranged so that the desired limit is $r=0$ and where the equality and inequality constraints are incorporated into the penalized objective function $P(x, r)$ through the use of smooth penalty functions. Then the penalty path(s) is described as the solution set of $\min P(x, r)$ as r varies and must, by the first-order necessary conditions, be a solution of $\nabla P=0$. (The gradient operator ∇ is with respect to x and ∇P denotes a column vector.) Then the gradient of the penalty function ∇P is formally identified with that of the Lagrangian $\nabla \mathcal{L}$ by defining the multipliers appropriately. Using the definition of these multipliers as additional equations, we obtain an expanded Lagrangian system (ELS) of nonlinear equations which, with an additional modification, becomes a perturbation of the Fritz John first-order necessary conditions. This modification prevents unbounded multipliers. Two different systems, one based on the quadratic penalty-logarithmic barrier function and the other on the quadratic penalty-quadratic loss function, are derived and analyzed in §§ 2 and 3.

Returning to the question of ill-conditioning, we shall show in § 4 that there are only three smooth penalty functions that yield well-conditioned expanded Lagrangian systems (ELS) and each is a method of order one in the sense of Lootsma [15]. The canonical examples of these three classes are the quadratic penalty function for equality constraints, logarithmic barrier function (an interior method) and the quadratic loss function (an exterior method) for inequality constraints. The remaining smooth penalty functions introduce artificial singularities and ill-conditioning into the ELS and thus are not used numerically.

The salient features of this class of algorithms can now be described. We first use an unconstrained optimization technique to get on the penalty path at a value of r , say r^0 , where the problem is reasonably well conditioned. Predictor-corrector continuation techniques [12], [13], [18] are then used to follow the penalty path as a solution of the expanded Lagrangian system to optimality at $r=0$. Section 5 contains a discussion of one such algorithm based on the quadratic penalty-logarithmic barrier function. Numerical test results are then presented to demonstrate the robustness, efficiency and potential of the methodology.

2. The mixed quadratic penalty-logarithmic barrier function. In this section we derive and analyze the expanded Lagrangian system (ELS) for the general nonlinear programming problem when the quadratic penalty function is used for the equality constraints and the logarithmic barrier function, for the inequality constraints. This mixed quadratic penalty-logarithmic barrier function leads to an expanded system which we modify to bound and normalize the multipliers. The resulting equations, which represent a perturbation of the Fritz John first-order necessary conditions, are summarized in Theorem 2.1. The existence and regularity of the penalty path is given in Theorem 2.2. Throughout this section we assume that f , g and h are at least \mathcal{C}^1 .

The mixed quadratic penalty-logarithmic barrier function is

$$(2.1) \quad P(x, r) = f(x) + \frac{1}{2r} h^T(x)h(x) - r \sum_{i=1}^p \ln(g_i(x)).$$

Assuming $\{x: g(x) > 0\}$ to be nonempty and robust [16], the first-order necessary condition for a minimum of P is that $\nabla P=0$, which is a parametrized system of nonlinear equations. However, this system suffers from the numerical problem that the Jacobian of ∇P , the Hessian of P , becomes increasingly ill-conditioned as $r \rightarrow 0^+$. (The l_2 condition $\kappa_2(\nabla^2 P) = \mathcal{O}(1/r)$ as $r \rightarrow 0^+$.) To remove the ill-conditioning, we expand the equations $\nabla P=0$ as follows: Assuming $g(x) > 0$ and $r > 0$, (x, r) solves

$\nabla P = \nabla f + \nabla h^T(h/r) - \sum_{i=1}^p \nabla g_i(r/g_i) = 0$ if and only if (x, λ, μ, r) solves the expanded system

$$\begin{aligned} \nabla f - \nabla h^T \lambda - \nabla g^T \mu &= 0, \\ h + r\lambda &= 0, \\ Mg - re &= 0, \end{aligned} \quad (2.2)$$

where $M = \text{diag}(\mu_1, \dots, \mu_p)$ and $e = (1, 1, \dots, 1)^T \in \mathbb{R}^p$. (The last two equations are derived from the definitions $\lambda = -h/r$ and $\mu_i = r/g_i(x)$.) This expanded system (2.2) is not new; it has been in the literature for some time and was used by Fiacco and McCormick [5] to investigate the behavior of the penalty path near $r=0$. The system (2.2) still has a numerical deficiency in that a multiplier may tend to infinity when either a constraint cannot be satisfied or a constraint qualification fails. Also the use of shifts in the barrier function can sometimes cause a multiplier μ_i to tend to infinity at a finite value of r . The modification that we now introduce trades this problem for limit point singularities for which the arclength or pseudo-arclength predictor-corrector continuation methods are particularly effective.

THEOREM 2.1. *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^1$, $h: \mathbb{R}^n \rightarrow \mathbb{R}^q$, and $g: \mathbb{R}^n \rightarrow \mathbb{R}^p$ be \mathcal{C}^1 functions, and let $\beta_0, r \in \mathbb{R}$ and $x \in \mathbb{R}^n$ be such that $\beta_0 > 0$, $r \neq 0$ and $g(x) > 0$, componentwise. The vector $(x; r)$ solves*

$$\nabla P = \nabla f + \nabla h^T \left(\frac{h}{r} \right) - \sum_{i=1}^p \nabla g_i \left(\frac{r}{g_i} \right) = 0 \quad (2.3)$$

if and only if there exist $\lambda \in \mathbb{R}^q$, $\mu \in \mathbb{R}^p$, $\mu_0 \in \mathbb{R}$ with $\mu_0 > 0$ such that the following equations hold:

$$F(x, \lambda, \mu, \mu_0; \bar{r}) = \begin{cases} \nabla \mathcal{L} = 0, \\ h + \bar{r}\lambda = 0, \\ Mg - \mu_0^2 \bar{r}e = 0, \\ \mu_0^2 + \|\mu\|_2^2 + \|\lambda\|_2^2 - \beta_0^2 = 0 \end{cases} \quad (2.4)$$

where $\mathcal{L} = \mu_0 f - h^T \lambda - g^T \mu$, $M = \text{diag}(\mu_1, \dots, \mu_p)$, $\mu = (\mu_1, \dots, \mu_p)^T$, $\lambda = (\lambda_1, \dots, \lambda_q)$, and $\bar{r} = r/\mu_0$.

Proof. Suppose (2.3) is valid. Multiply through $\nabla P = 0$ by $\mu_0 = \beta_0[1 + \sum(h_j/r)^2 + \sum(r/g_i)^2]^{-1/2}$ and define $\bar{r} = r/\mu_0$, $\lambda = -h/\bar{r}$, $\mu_i = (\mu_0^2 \bar{r})/g_i$, so that $(x, \lambda, \mu, \mu_0, \bar{r})$ now solves the expanded system (2.4). When $\mu_0 > 0$ and $(x, \lambda, \mu, \mu_0, \bar{r})$ solves (2.4), the reduction to (2.3) also follows directly from these definitions once the normalization has been dropped. \square

Note that if we drop the normalization and set $\mu_0 = 1$, we are back to the system (2.2). The use of the parameter β_0 allows for the efficient transfer from the solution of the minimization problem $\min P(x, r)$ to the system $F = 0$. For example, if x^0 is a solution of $\min P(x, r^0)$, we could define $\lambda^0 = -h(x^0)/r^0$, $\mu_i^0 = r^0/g_i(x^0)$, and $\beta_0 = [1 + \|\mu^0\|_2^2 + \|\lambda^0\|_2^2]^{1/2}$. Then $(x, \lambda, \mu, \mu_0; \bar{r}) = (x^0, \lambda^0, \mu^0, 1; r^0)$ is a solution of $F = 0$; and, assuming no singularities are encountered, we can follow the penalty path ($\min P(x, r)$) to $r = 0$ by following the solution of $F = 0$ to $\bar{r} = 0$. Given this formulation we now give necessary and sufficient conditions for the system (2.4) be regular at $\bar{r} = 0$.

THEOREM 2.2. *Let the system (2.4) be denoted by $F(z; \bar{r}) = 0$ with $z = (x, \lambda, \mu, \mu_0)$. Let $(z^0; 0)$ be a solution of $F = 0$ and assume f, h and g are twice continuously differentiable in a neighborhood of x^0 . Define two index sets \mathcal{A} and \mathcal{A} and a corresponding tangent*

space \bar{T} by

$$\begin{aligned}\bar{\mathcal{A}} &= \{i: 1 \leq i \leq p, g_i(x^0) = 0\}, & \mathcal{A} &= \{i \in \bar{\mathcal{A}}: \mu_i^0 \neq 0\}, \\ \bar{T} &= \{y \in \mathbb{R}^n: y^T \nabla h_j(x^0) = 0 (j = 1, \dots, q), y^T \nabla g_i(x^0) = 0 (i \in \bar{\mathcal{A}})\}.\end{aligned}$$

A necessary and sufficient condition that the Jacobian $D_z F(z^0; 0)$ be nonsingular is that each of the following three conditions hold:

- (a) $\bar{\mathcal{A}} = \mathcal{A}$;
- (b) $S := \{\{\nabla g_i(x^0)\}_{i \in \bar{\mathcal{A}}} \cup \{\nabla h_j(x^0)\}_{j=1}^q\}$ is a linearly independent collection of $q + |\bar{\mathcal{A}}|$ vectors, where $|\bar{\mathcal{A}}|$ denotes the cardinality of $\bar{\mathcal{A}}$;
- (c) The Hessian of the Lagrangian $\nabla^2 \mathcal{L}$ is nonsingular on the tangent space \bar{T} at z^0 .

If $D_z F(z^0; 0)$ is nonsingular, there exist open neighborhoods \mathcal{B}_1 of $\bar{r} = 0$ and \mathcal{B}_2 of $(z^0; 0)$ and a function $\phi \in \mathcal{C}^1(\mathcal{B}_1)$ such that $F(\phi(\bar{r}), \bar{r}) = 0$ for all $\bar{r} \in \mathcal{B}_1$ and $\phi(0) = z^0$. This solution is locally unique in the sense that if $(z; \bar{r}) \in \mathcal{B}_2$ and $F(z; \bar{r}) = 0$, then z belongs to the manifold defined by ϕ , i.e., $z = \phi(\bar{r})$. Furthermore, if f, g and h are \mathcal{C}^k (\mathcal{C}^∞ or real analytic) then ϕ is \mathcal{C}^{k-1} (\mathcal{C}^∞ or real analytic, respectively) on \mathcal{B}_1 .

Proof. The necessary and sufficient conditions for the nonsingularity of the Jacobian $D_z F(z^0; 0)$ have been established in the work of Poore and Tiaht [17] in the context of the parametric programming problem and will not be repeated here. The remaining part of the theorem follows from the implicit function theorem [3]. \square

Several remarks are in order: If x^0 is a Fritz John or Karush–Kuhn–Tucker point, condition (a) is called strict complementarity ($g_i(x^0) = 0$ implies μ_i^0 is nonzero) while condition (b) is the linear independence constraint qualification. Furthermore, if conditions (a) and (b) are satisfied and condition (c) is strengthened to the Hessian of the Lagrangian being positive definite on the tangent space \bar{T} , then we have a second-order sufficient condition for x^0 to be a local minimum provided $\mu^0 \geq 0$ and $g \geq 0$. Finally we note that the theorem is valid regardless of the type of extremal: local minimum, saddle point, local maximum, feasible or nonfeasible critical point.

3. The quadratic penalty-quadratic loss function. In this section we will use the quadratic loss function, an exterior method, to handle the inequality constraints. Since the virtues of exterior methods have been discussed previously [5], we forego an extensive comparison of interior and exterior methods. The essence of their advantage is that an initial strictly feasible point is not required and that when many inactive inequality constraints are present, the resulting problem size is much smaller than that for an interior method. The price to be paid for these advantages is a lack of higher-order differentiability; however, when the three conditions (a)–(c) of Theorem 2.2 hold, second-order differentiability is present along the solution path for r positive and sufficiently small, as will be shown in Theorem 3.2.

The quadratic penalty-quadratic loss function used in this section is

$$(3.1) \quad P(x, r) = f(x) + \frac{1}{2r} h^T(x) h(x) + \frac{1}{2r} (g^-(x))^T (g^-(x))$$

where $g^- = (g_1^-, \dots, g_p^-)^T$ and $g_i^-(x) = \min(g_i(x), 0)$. A difficulty with this penalty function is that the Hessian becomes discontinuous across an inequality constraint boundary $g_i = 0$. Although this discontinuity is a simple finite jump discontinuity, the jump in $\nabla^2 P$ tends to infinity as r tends to zero, and the additional and inevitable ill-conditioning is still present in $\nabla^2 P$. In the expanded Lagrangian system the jump discontinuity tends to zero as r tends to zero and the ill-conditioning is removed! In the presence of a second-order sufficient condition, Fiacco and McCormick [5] have

shown that there is no jump along the path for sufficiently small r . We extend this result in Theorem 3.2, but first we derive an appropriate expanded Lagrangian system.

THEOREM 3.1. *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^1$, $h: \mathbb{R}^n \rightarrow \mathbb{R}^q$, and $g: \mathbb{R}^n \rightarrow \mathbb{R}^p$ be \mathcal{C}^1 functions, and let $\beta_0, r \in \mathbb{R}$, $x \in \mathbb{R}^n$ be such that $\beta_0 > 0$ and $r \neq 0$. Then $(x; r)$ solves*

$$(3.2) \quad \nabla P = \nabla f + \nabla h^T \left(\frac{h}{r} \right) + \nabla g^T \left(\frac{g^-}{r} \right) = 0,$$

where g^- is defined following (3.1), if and only if there exist $\lambda \in \mathbb{R}^q$, $\mu \in \mathbb{R}^p$, $\mu_0 \in \mathbb{R}$ with $\mu_0 > 0$ such that the following equations hold:

$$(3.3) \quad G(x, \lambda, \mu, \mu_0; \bar{r}) := \begin{cases} \nabla \mathcal{L} = 0, \\ h + \bar{r}\lambda = 0, \\ g_i + \mu_i \bar{r} = 0 & (i \in V(x)), \\ \mu_i = 0 & (i \notin V(x)), \\ \mu_0^2 + \|\mu\|_2^2 + \|\lambda\|_2^2 - \beta_0^2 = 0 \end{cases}$$

where $\mathcal{L} = \mu_0 f - h^T \lambda - g^T \mu$, $\bar{r} = r/\mu_0$, and $V(x) = \{i: 1 \leq i \leq p, g_i(x) \leq 0\}$.

Except for the manipulation of g^- and its derivatives, which is explained in Luenberger [16], the proof is similar to that of Theorem 2.1. The corresponding trajectory analysis around $\bar{r} = 0$ is given next.

THEOREM 3.2. *Let the system (3.3) be denoted by $G(z, \bar{r}) = 0$ with $z = (x, \lambda, \mu, \mu_0)$, and assume f, h and g are twice continuously differentiable in a neighborhood of x^0 . Then $(z^0; 0)$ is a solution of (2.4) ($F = 0$) if and only if it is a solution of (3.3) ($G = 0$). Let $(z^0; 0)$ be a solution of either system and let x^0 denote the corresponding x component. Modify the equations $G = 0$ by fixing the index set $V(x)$ to be $V(x^0)$, and denote the resulting equations by $G^0(z; \bar{r}) = 0$. Then if $D_z F(z^0; 0)$ is nonsingular so is $D_z^0 G(z^0; 0)$ and thus the three conditions (a)–(c) in Theorem 2.2 guarantee the nonsingularity of $D_z G^0(z^0; 0)$.*

Let $(z^0; 0)$ be a solution of $G = 0$, suppose $D_z G^0(z^0; 0)$ is nonsingular. Then the system $G^0(z; \bar{r}) = 0$, has a solution $(z, \bar{r}) = (\phi(\bar{r}), \bar{r})$ with the following properties: There exist open neighborhoods, \mathcal{B}_1 of $\bar{r} = 0$ and \mathcal{B}_2 of $(z^0; 0)$ such that $\phi \in \mathcal{C}^1(\mathcal{B}_1)$, $G^0(\phi(\bar{r}), \bar{r}) = 0$ for all $\bar{r} \in \mathcal{B}_1$ and $\phi(0) = z^0$. This solution is locally unique in the sense that $(z; \bar{r}) \in \mathcal{B}_2$ and $G^0(z; \bar{r}) = 0$ implies $z = \phi(\bar{r})$. Furthermore, if f, g and h are in \mathcal{C}^k (\mathcal{C}^∞ or real analytic) then ϕ is \mathcal{C}^{k-1} (\mathcal{C}^∞ or real analytic).

The proof follows directly from the implicit function theorem [3] and our previous work [17] and is thus omitted. This theorem gives a smooth path through the given solution $(z^0; 0)$ only by fixing $V(x^0)$ and thus by modifying G . In practice we start with a positive \bar{r} so that (3.3) implies $\mu_i \geq 0$. If in addition conditions (a)–(c) of Theorem 2.2 are satisfied, then $\mu_i^0 > 0$ for $i \in V(x^0)$ and $\mu_i^0 = 0$ otherwise. Thus the eventual active inequality constraints approach from the exterior of the feasible region, which implies that the solution $(z; \bar{r}) = (\phi(\bar{r}), \bar{r})$ of $G = 0$ has the stated smoothness only for $\bar{r} \in (0, \infty) \cap \mathcal{B}_1$. In fact, the “solution” of $G = 0$ may not be continuous across $\bar{r} = 0$. On the other hand, we can still continue in the positive r direction in hopes of hitting a turning point and returning to $\bar{r} = 0$. In this sense this method still maintains a capability of determining multiple optima.

4. Other penalty functions. In this section we show that the three penalty methods described in the previous section are essentially the only smooth penalty methods that do not introduce artificial singularities and thus ill-conditioning into the penalty path for nonsingular nonlinear programming problems. Our classification of interior and

exterior methods for inequality constraints and of equality penalty methods is based on the work of Lootsma [15].

We shall use the customary term “barrier function” instead of “interior penalty function.” These functions generally have the following properties:

$$(4.1) \quad \begin{aligned} & \text{(i) } \phi: \mathbb{R}^+ \rightarrow \mathbb{R}, \lim_{u \rightarrow 0^+} \phi(u) = +\infty, \\ & \text{(ii) } \phi'(u) < 0 \quad \text{and} \quad \phi''(u) > 0 \quad \text{for } u > 0. \end{aligned}$$

For our purposes we need a refinement of these properties and use a classification given by Lootsma [15].

DEFINITION 4.1. A barrier function $\phi: \mathbb{R}^+ \rightarrow \mathbb{R}$ satisfying properties (i) and (ii) in (4.1) is said to be a barrier function of order α if ϕ' is real analytic on \mathbb{R}^+ and has a pole of order α at the origin.

The three most popular barrier functions are $\phi = -\ln u$ ($\alpha = 1$) due to Frisch (1955), $\phi = u^{-1}$ ($\alpha = 2$) due to Carroll (1961), and $\phi = u^{-2}$ ($\alpha = 3$) due to Kowalik (1966), Box (1969) and Fletcher and McCann (1969) [15]. The principal result is contained in Theorem 4.1.

THEOREM 4.1. Let ϕ be a barrier function of order α and let f , g and h be \mathcal{C}^2 functions. Then $\alpha \geq 1$ and if $\alpha > 1$, the expanded Lagrangian system is singular at the penalty parameter value $r = 0$.

Proof. We first note that $\alpha < 1$ implies that $\lim_{u \rightarrow 0^+} \phi(u) = +\infty$ is violated. Thus we restrict ourselves to the case $\alpha \geq 1$. We consider, without loss of generality, the nonlinear programming problem $\min \{f(x): g(x) \geq 0\}$, which has the penalized objective function

$$P(x, r) = f(x) + r^\alpha \sum_{i=1}^p \phi(g_i(x))$$

where the power of r is included so that the minimizer $x(r)$ will be smooth in r (Lootsma [15]). A minimizer $x(r)$ must satisfy

$$(4.2) \quad \nabla P = \nabla f + r^\alpha \sum_{i=1}^p \phi'(g_i) \nabla g_i = 0$$

for which the expanded system is

$$(4.3) \quad \nabla \mathcal{L} = 0, \quad MG + r^\alpha e = 0$$

where $M = \text{diag}(\mu_1, \dots, \mu_p)$, $G = (-1/\phi'(g_1), \dots, -1/\phi'(g_p))^T$ and $\mathcal{L} = f - g^T \mu$. Now since ϕ has a pole of order α at the origin, $\phi''(u)/[\phi'(u)]^2 = \mathcal{O}(u^{\alpha-1})$ as $u \rightarrow 0$. Thus $\alpha > 1$ implies $\phi''(u)/[\phi'(u)]^2 \rightarrow 0$ as $u \rightarrow 0^+$. To complete the proof for $\alpha > 1$, it suffices to consider the case $g_i(x) \rightarrow 0$ as $r \rightarrow 0$ either smoothly or as a sequence. Now the $(n+1)$ th row of the Jacobian of the expanded system (4.3) has as the only potential nonzero entries $1/\phi'(g_i)$ and $(-\mu_i \phi''(g_i)/[\phi'(g_i)]^2)(\nabla g_i)^T$, both of which tend to zero as $g_i \rightarrow 0^+$. Thus, the $(n+i)$ th row of the Jacobian approaches zero at $r \rightarrow 0^+$. (Notice that we could have arrived at this same conclusion by extending the definition of $\phi''/[\phi']^2$ to $u = 0$ by continuity.) \square

In case $\alpha = 1$, the quotient $\phi''(u)/[\phi'(u)]^2$ tends to a finite nonzero limit as $u \rightarrow 0^+$ and we can establish the nonsingularity of the expanded Lagrangian system just as in Theorem 2.2. Of course, the canonical example from the class of barrier functions of order one is the logarithmic barrier function which was examined in § 2.

The next class of penalty methods to be considered is the class of exterior penalty functions which we call loss functions [5]. We generally require that a loss function satisfy the properties

$$(4.4) \quad \begin{aligned} & \text{(iii)} \quad \psi(u) = 0 \text{ for } u \geq 0, \psi(u) > 0 \text{ for } u < 0, \\ & \text{(iv)} \quad \psi'(u) < 0 \text{ and } \psi''(u) > 0 \text{ for } u < 0, \\ & \text{(v)} \quad \psi \text{ is continuous across } u = 0. \end{aligned}$$

Following Lootsma [15], we further restrict this class of exterior methods as follows.

DEFINITION 4.2. A loss function ψ satisfying (4.4) is said to be of order $\gamma > 0$ provided $\psi'(u)$ is a real analytic for $u < 0$ with a zero of order γ at $u = 0$, i.e., $\psi'(u) = \mathcal{O}((-u)^\gamma)$ as $u \rightarrow 0^-$.

Given this definition we can state the principal results for exterior methods.

THEOREM 4.2. Let f, g, h be \mathcal{C}^2 functions and suppose $\psi(u)$ is a loss function of order γ . If either $\gamma < 1$ or $\gamma > 1$, the corresponding expanded Lagrangian system is singular at $r = 0$.

Proof. It suffices again to consider the problem $\min \{f(x): g(x) \geq 0\}$ with the corresponding penalized objective function

$$P(x, r) = f(x) + r^{-\gamma} \sum_{i=1}^p \psi(g_i(x))$$

where ψ is a loss function of order γ and the power γ of r is included to ensure a smooth dependence of x on r [15]. Then a minimizer of P satisfies

$$\nabla P = \nabla f + r^{-\gamma} (\nabla g^T)(\psi'(g_1), \dots, \psi'(g_p))^T = 0,$$

which has the expanded Lagrangian system

$$(4.5) \quad \begin{aligned} & \nabla \mathcal{L} = 0, \\ & (\psi'(g_1), \dots, \psi'(g_p))^T - r^\gamma \mu = 0 \end{aligned}$$

where the Lagrangian $\mathcal{L} = f - g^T \mu$. If $\gamma < 1$, then $\psi'(u)$ becomes unbounded as $u \rightarrow 0^-$. Thus (4.5) is singular in that $\psi'(u)$ becomes unbounded as $u \rightarrow 0^-$. If $\gamma > 1$, then $g_i \rightarrow 0$ as $r \rightarrow 0^+$ implies that the $(n+i)$ th row of the Jacobian tends to zero so that the system (4.5) becomes singular. \square

For the case $\gamma = 1$, we can again prove a result similar to that obtained in § 3 for the canonical quadratic loss function. Finally we come to the penalty functions for equality constraints. In analogy with our two previous definitions we define penalty functions for equality constraints in the following definition.

DEFINITION 4.3. Let $\theta: \mathbb{R} \rightarrow \mathbb{R}^+ \cup \{0\}$. We say that θ is a penalty function of order β , provided $\theta'' > 0$, $\theta' < 0$ for $u < 0$ and $\theta' > 0$ for $u > 0$, θ is analytic on $\mathbb{R} - \{0\}$ and θ' has a zero of order β at $u = 0$, i.e., $\theta = \mathcal{O}(|u|^\beta)$ as $u \rightarrow 0$.

For this class of penalty methods, the corresponding result is contained in Theorem 4.3.

THEOREM 4.3. Let θ be a penalty function for equality constraints of order β . Then the use of θ to incorporate equality constraints into the penalized objective function yields a corresponding singular expanded Lagrangian system if $\beta < 1$ or $\beta > 1$.

We omit the proof of this theorem since it closely parallels the previous proof. If $\beta = 1$, we can prove a theorem corresponding to those in the previous two sections. The canonical order-one method is the quadratic penalty function $\theta(u) = u^2$.

In conclusion, any of the order-one methods lead to well-conditioned expanded Lagrangian systems, and the canonical examples of these order-one penalty functions are the quadratic penalty function for equality constraints, the logarithmic barrier function, an interior method, and the quadratic-loss function, an exterior method, for inequality constraints.

5. An algorithm and numerical examples. The alternatives for dealing with the constraints in the general nonlinear programming problem (1.1) can be summarized as follows:

Equality constraints	Inequality constraints
(a) Quadratic penalty	(c) Logarithmic barrier
(b) Not penalized	(d) Quadratic loss

where the “Not penalized” means that the equality constraints are handled directly and are not incorporated into the penalized objective function. This appears to be particularly advantageous for dealing with linear (affine) constraints. In this section we present the salient features of a class of algorithms based on the use of the combination (a)–(c) and the corresponding expanded Lagrangian system (2.3). We conclude this section with the preliminary results of our testing on some nontrivial test problems [10] to demonstrate the robustness, efficiency, and potential for the methodology.

Although the algorithms have many variants, we separate the different phases as follows: In the first phase we perform an initialization to produce an initial point \hat{x} satisfying $g(\hat{x}) > 0$. An alternative is to introduce a shift into the log function by choosing a δ so that at an initial point \hat{x} the inequality $g(\hat{x}) + \delta > 0$ is satisfied. Then we use the shifted penalty function $P(x, r) = f(x) + h^T(x)h(x)/(2r) - r\sum \ln(g_i(x) + r\delta_i/r^0)$ which is well defined at $r = r^0$ and $x = \hat{x}$. If the point \hat{x} is not close to the feasible region, the path can be long and the resulting algorithm degrades in time and the number of iterations. Although we have not experienced it numerically, an additional difficulty is that the introduction of the shift can introduce limit point singularities into the solution path of the normalized expanded Lagrangian system (5.1). In the numerical examples to follow we start with the given initial value of x , [10], minimize the loss function $g^-(x)^T g^-(x)$ to produce a point at least close to the feasible region, and then introduce the shift to complete the initialization.

In the second phase we use an unconstrained minimization technique such as quasi-Newton with a BFGS update or a preconditioned conjugate gradient method to minimize the penalty function P at a value of r , say r^0 , at which the problem is reasonably well conditioned. In the examples to follow we have used a quasi-Newton method with a quadratic-cubic line search and an Armijo stopping criterion [4], modified to maintain feasibility ($g(x) + \delta > 0$). Since a highly accurate answer of $\min P(x, r^0)$ may have only a digit or two in common with the final answer, we have found it to be more efficient to compute the answer to within a moderate to low accuracy and then add a homotopy to the equation $\nabla \mathcal{L} = 0$ to account for a moderate value of ∇P . Thus given an approximate answer x^0 to $\min P(x, r^0)$ we convert to the expanded system

$$\begin{aligned}
 (5.1) \quad & \nabla \mathcal{L} - (\mu_0^2 \bar{r}/r^0) \nabla P(x^0, r^0) = 0, \\
 & h + \bar{r}\lambda = 0, \\
 & M(g + \mu_0 \bar{r}\delta/r^0) - \bar{r}\mu_0^2 e = 0, \\
 & \mu_0^2 + \|\mu\|_2^2 + \|\lambda\|_2^2 - \beta_0^2 = 0
 \end{aligned}$$

where $\mathcal{L} = \mu_0 f - h^T \lambda - g^T \mu$, $M = \text{diag}(\mu)$, $\mu = (\mu_1, \dots, \mu_p)^T$, $\bar{r} = r/\mu_0$. At $x = x^0$ and $r = r^0$ we have a solution of (5.1) given by $(x, \lambda, \mu, \mu_0, \bar{r}) = (x^0, \lambda^0, \mu^0, 1, r^0)$, where $\lambda^0 = -h(x^0)/r^0$, $\mu_i^0 = r^0/g_i(x^0)$, and where β_0 is defined and fixed by $\beta_0 = (1 + \|\lambda^0\|_2^2 + \|\mu^0\|_2^2)^{1/2}$. Notice that Theorem 2.2 remains valid for this system even though it is a perturbation of the one discussed in § 2. Continuation is performed on system (5.1) in the scaled penalty parameter \bar{r} with β_0 fixed.

In the path following phase we use a predictor-corrector continuation technique, which is closely patterned after the work of Keller [12], to follow the solution of the parameterized system of (5.1) to optimality at $\bar{r} = 0$. However, we use higher-order Adams-Bashforth predictors [1], [18] to increase the efficiency. In this methodology we start with an Euler method (an Adams-Bashforth one-step method), and then increase the order of the method as more steps are taken. The error-stepsize control is based on the use of consecutive order methods wherein the higher order method is used to estimate the error in the lower order method [1]. The stepsize is then adjusted to achieve the desired error in the predictor. Typically, two to three steps are needed to get to $r = 0$ where we complete the convergence with Newton's method. Thus the method is locally quadratically convergent.

To get some estimation of the relative performance of this algorithm, we have solved several test problems from the book by W. Hock and K. Schittkowski [10] and give a comparative summary of the number of function evaluations in the table below. The function evaluation for the codes other than PENCON are taken from [10]. To be consistent with those function evaluation counts, we count the evaluation of a p -dimensional vector as p function evaluations; however, we do not count upper and lower bounds on variables, since they are handled directly in the code and gradient evaluations of linear constraints are counted only once. The approximation of the Hessian of the Lagrangian in the continuation phase is based on finite differences [4]. Several values of r^0 ranging from 1.0 to 0.0001 have been tested and indeed the initial minimization is sensitive to this choice. In the test results in Table 1 we have used the value $r^0 = 0.1$; however, in some cases we found the number of function evaluations to be less for smaller values of r^0 .

TABLE 1

Code	Author	Method
VFO2AD	Powell	Quadratic approximation
OPRQP	Bartholomew-Biggs	Quadratic approximation
GRGA	Abadie	Generalized reduced gradient
VFO1A	Fletcher	Multiplier
FUNMIN	Kraft	Multiplier
FMIN	Kraft, Lootsma	Penalty
PENCON	Al-Hassan, Poore	Penalty-continuation

Comparison in Table 2 illustrates that this penalty-continuation algorithm as coded in PENCON [1] currently comes in second or third and occasionally first and fourth; however, the methodology is quite robust, primarily because of the robustness of penalty paths leading to optimality and the robustness of the continuation methodology. Thus this algorithm offers an alternative to sequential quadratic programming illustrated by the codes of Powell and Bartholomew-Biggs above. Furthermore, these test results are even more promising because of what we have not done in the implementation of the algorithm. Between 50 percent and 90 percent of the function evaluations occur

TABLE 2

Code								
Prob no.	VFO2AD	OPRQP	GRGA	VFO1A	FUNMIN	FMIN	PENCON	Our Rank
10	48	126	678	280	554	687	88	[2]
12	48	132	277	300	492	306	130	[2]
13	180	300	192	565	928	4,178	269	[3]
29	52	206	646	421	482	414	155	[2]
34	48	291	346	619	**	2,483	221	[2]
38	218	213	185	212	95	739	83	[1]
43	96	220	1,580	448	1,032	1,949	509	[4]
44	84	245	203	1,470	2,394	8,547	127	[2]
55	14	84	**	**	2,737	4,767	523	[3]
63	54	153	758	276	555	1,065	125	[2]
64	28	65,238	532	416	990	2,113	365	[2]
65	44	3,942	249	308	2,520	374	148	[2]
71	30	201	411	557	1,575	6,165	350	[3]
118	**	1,800	3,857	**	**	**	2,804	[2]

Function evaluation count

** indicates failure.

in the first and second phases wherein we perform the unconstrained optimization. However, the efficiency could be considerably improved through the use of scaling of the constraints, weights and special line searches for the log barrier function [8]. In the continuation phase, efficiency could be improved through the use of updating procedures after a prediction along the path and during the correction back to the path. Finally, our answers are computed to approximately twelve digits of accuracy on a 14+ digit CDC machine, whereas the remaining answers listed above are computed on a 10 digit machine [10]. It is also interesting to note the value of the objective function in problem numbers 64 and 65 are off by two orders of magnitude and 300 percent, respectively, for Powell's VFO2AD, whereas the values in our tests agree to at least ten digits.

6. Summary and conclusions. One of the objectives in this work has been to re-examine the use of smooth penalty functions as an effective computational technique in light of recent developments in numerical continuation techniques. The expanded Lagrangian system, which has been a focal point of this work, is the key link between these methods as well as the key to the removal of the ill-conditioning traditionally associated with smooth penalty methods. This parametrized system of nonlinear equations is well conditioned only for three smooth penalty methods as explained in § 4. The canonical examples are the quadratic penalty function for equality constraints, logarithmic barrier function, an interior method, and the quadratic loss function, an exterior method, for inequality constraints. An additional technique is to leave linear (affine) equality constraints out of the penalized objective function and treat them directly. In linear programming this technique leads to a class of algorithms which are qualitatively similar to the interior point methods [8], [11]. Finally, this methodology generalizes in a natural way to optimal control, calculus of variations, and parameter identification.

The robustness and efficiency of this class of algorithms based on these smooth penalty functions, corresponding expanded Lagrangian system, and predictor-corrector

continuation techniques have been illustrated in § 5. The methodology shows considerable promise and potential for solving constrained optimization problems; however, as with any method, a word of caution is appropriate. We can construct simple examples illustrating the following situations for penalty paths: given $\varepsilon > 0$ a penalty path may exist only for the penalty parameter $\bar{r} \geq \varepsilon$ or only for $\bar{r} \leq \varepsilon$, may not exist at all, may diverge, or may exist for $\bar{r} > 0$ but the limit point at $\bar{r} = 0$ is not a local minimum of the original problem. (When this last situation occurs, the expanded Lagrangian system is singular.) Despite these examples, penalty path following is a mathematically robust method of solving constrained optimization problems as is illustrated by the examples in Hock and Schittkowski [10], and we hope that these penalty-continuation algorithms based on the expanded Lagrangian system make the methodology numerically robust and efficient.

REFERENCES

- [1] Q. AL-HASSAN, *Continuation methodology for solving constrained optimization problems*, Ph.D. thesis, Colorado State University, Fort Collins, CO, 1987.
- [2] E. L. ALLGOWER AND K. GEORG, *Predictor-corrector and simplicial methods for approximating fixed points and zero points of nonlinear mappings*, in *Mathematical Programming: The State of the Art*, A. Bachem, M. Grottschel and B. Korte, eds., Springer-Verlag, New York, 1983.
- [3] S.-N. CHOW AND J. K. HALE, *Methods of Bifurcation Theory*, Springer-Verlag, New York, 1982.
- [4] J. E. DENNIS AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
- [5] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.
- [6] R. FLETCHER, *Practical Methods of Optimization*, Vol. 2, John Wiley, New York, 1981.
- [7] ———, *Penalty functions*, in *Mathematical Programming: The State of the Art*, A. Bachem, M. Grottschel and B. Korte, eds., Springer-Verlag, New York, 1983.
- [8] P. E. GILL, W. MURRAY, M. A. SAUNDERS, J. A. TOMLIN AND M. H. WRIGHT, *On projected Newton barrier methods for linear programming and an equivalence to Karmarkar's projective method*, *Math. Programming*, 36 (1986), pp. 183–209.
- [9] P. E. GILL, W. MURRAY AND M. H. WRIGHT, *Practical Optimization*, Academic Press, New York, 1981.
- [10] W. HOCK AND K. SCHITTKOWSKI, *Test Examples for Nonlinear Programming Code*, Springer-Verlag, New York, 1981.
- [11] N. KARMARKAR, *A new polynomial-time algorithm for linear programming*, *Proc. 16th Annual ACM Symposium on the Theory of Computing*, 1984, pp. 302–311.
- [12] H. B. KELLER, *Numerical solution of bifurcation and nonlinear eigenvalue problems*, in *Applications of Bifurcation Theory*, P. Rabinowitz, ed., Academic Press, New York, 1977, 359–384.
- [13] T. KUPPER, H. D. MITTELMANN AND H. WEBER, *Numerical Methods for Bifurcation Problems*, Birkhauser-Verlag, Boston, 1984.
- [14] F. A. LOOTSMA, ED., *Numerical Methods for Nonlinear Optimization*, Academic Press, New York, 1972.
- [15] F. A. LOOTSMA, *A survey of methods for solving constrained minimization problems via unconstrained minimization*, in *Numerical Methods for Nonlinear Optimization*, F. A. Lootsma, ed., Academic Press, New York, 1972.
- [16] D. G. LUENBERGER, *Linear and Nonlinear Programming*, 2nd ed., Addison-Wesley, Reading, MA, 1984.
- [17] A. B. POORE AND C. TIAHRT, *Bifurcation problems in nonlinear parametric programming*, *Math. Programming* (1987).
- [18] W. C. RHEINOLDT, *Numerical Analysis of Parametrized Nonlinear Equations*, John Wiley, New York, 1986.
- [19] L. F. SHAMPINE AND M. K. GORDON, *Computer Solution of Ordinary Differential Equations: The Initial Value Problem*, W. H. Freeman, San Francisco, 1975.

NUMERICAL APPROXIMATION FOR THE INFINITE-DIMENSIONAL DISCRETE-TIME OPTIMAL LINEAR-QUADRATIC REGULATOR PROBLEM*

J. S. GIBSON† AND I. G. ROSEN‡

Abstract. An abstract approximation framework is developed for the finite and infinite horizon discrete-time linear-quadratic regulator problems for systems whose state dynamics are described by a linear semigroup of operators on an infinite-dimensional Hilbert space. The schemes included in the framework yield finite-dimensional approximations to the linear state feedback gains which determine the optimal control law. Convergence arguments are given. Examples involving hereditary and parabolic systems and the vibration of a flexible beam are considered. Spline-based finite element schemes for these classes of problems, together with numerical results, are presented and discussed.

Key words. linear-quadratic regulator, discrete-time, distributed parameter system

AMS(MOS) subject classifications. 93C25, 93C55, 65J10

1. Introduction. Recent advances in microprocessor technology have led to increased interest in digital, or discrete-time, control systems. In addition, because many current application areas involve complex systems which are most appropriately modeled using functional and/or partial differential equations, it has become important to study digital control techniques in the context of infinite-dimensional or distributed systems.

A great deal of attention has been given to the continuous-time infinite-dimensional linear-quadratic regulator problem. The general theory and characterization of the linear state feedback form of the optimal control are discussed in [5], [6], [9], [10], [23] and [24], while its application to hereditary, parabolic and hyperbolic systems with emphasis on approximation is treated in [2], [3], [8], [11], [13], [15], [16] and [19], to mention just some of the work that has been done.

On the other hand, relatively little can be found in the literature concerning the corresponding discrete-time problem. The major contributions in this area can be found in the papers by Lee, Chow and Barr [22] and Zabczyk [30]. In these studies the Riccati difference equations that characterize the linear feedback form of the optimal control for the finite-time problem are given and limiting properties as the length of the time horizon tends to infinity are discussed. However, the issue of approximation is not considered.

In the present paper, we develop numerical approximation schemes that yield finite-dimensional approximations to the feedback gain operators which determine the discrete-time optimal control law. We consider control systems whose dynamics can be described in terms of a linear semigroup of operators on an infinite-dimensional Hilbert space. The basis of our approach is the construction of a sequence of finite-dimensional (presumably finite-element based) state approximations which in turn

* Received by the editors December 30, 1985; accepted for publication (in revised form) May 1, 1987. Part of this research was carried out while the authors were visiting scientists at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, Virginia 23665 which is operated under NASA contracts NAS1-17070 and NAS1-18107.

† Department of Mechanical Aerospace and Nuclear Engineering, University of California, Los Angeles, California 90024. This research was supported in part by the Air Force Office of Scientific Research under contract AFOSR-84-0309.

‡ Department of Mathematics, University of Southern California, Los Angeles, California 90089. This research was supported in part by the Air Force Office of Scientific Research under contract AFOSR-84-0393.

leads to a sequence of finite-dimensional discrete-time linear-quadratic regulator problems each of which can be solved using standard techniques.

Under appropriate assumptions on the nature of the original problem and the convergence of the state approximation, we are able to prove that the approximating optimal controls and feedback gains converge to the true optimal control sequences and feedback laws for the original infinite-dimensional system. Depending upon the convergence properties of the state approximation, we are able to establish strong or uniform norm convergence of the approximating gain operators and the corresponding weak or strong convergence of the approximating feedback kernels which are used in the implementation of the optimal control. We treat both the finite and infinite horizon problems.

We have tested our schemes on a wide variety of examples. This paper includes numerical results for problems with state dynamics given by hereditary and parabolic (heat/diffusion) differential equations and a hybrid system of partial and ordinary differential equations for the vibration of an Euler-Bernoulli beam connected to a rigid rotating hub and a lumped mass at the tip. We implemented and tested the methods on an IBM Personal Computer.

We provide a brief outline of the remainder of the paper. In § 2 we briefly outline previous results concerning the characterization of the optimal control and feedback gains for both the finite and infinite horizon discrete-time regulator problems for distributed systems. The Riccati difference and algebraic equations whose solutions determine the optimal feedback control law are discussed. In § 3 we develop the abstract approximation framework and convergence arguments. Section 4 contains a discussion of particular schemes for the classes of problems mentioned above together with the results of our numerical studies. Some concluding remarks are given in § 5.

2. The optimal control problem.

2.1. Optimal control on a finite interval. Let Z and U be Hilbert spaces with inner products $\langle \cdot, \cdot \rangle_Z$ and $\langle \cdot, \cdot \rangle_U$, respectively, with U finite dimensional. For $\{H, \langle \cdot, \cdot \rangle_H\}$ a Hilbert space, let $l^2(t_0, t_f; H)$ denote the usual Hilbert space of sequences $x = \{x(t)\}_{t=t_0}^{t_f} = \{x(t_0), x(t_0+1), \dots, x(t_f)\}$ with t, t_0 and t_f integers and $x(t) \in H$ together with the inner product

$$(2.1) \quad \langle x, y \rangle_{l^2} = \sum_{t=t_0}^{t_f} \langle x(t), y(t) \rangle_H.$$

The discrete-time linear-quadratic regulator problem on the finite time interval $[t_0, t_f]$ is:

(P1) Choose $u \in l^2(t_0, t_f; U)$ to minimize the quadratic performance index

$$(2.2) \quad \begin{aligned} J(G; t_0, t_f, z(t_0), u) \\ = \sum_{t=t_0}^{t_f-1} [\langle Qz(t), z(t) \rangle_Z + \langle Ru(t), u(t) \rangle_U] + \langle Gz(t_f), z(t_f) \rangle_Z \end{aligned}$$

subject to the discrete-time control system

$$(2.3) \quad z(t+1) = Tz(t) + Bu(t), \quad t \geq t_0 \quad (t \text{ and } t_0 \text{ integers}), \quad z(t_0) \in Z,$$

where T and B are bounded linear operators from Z into Z and U into Z , respectively, Q and G are bounded, nonnegative self-adjoint operators on Z , and R is a positive definite self-adjoint operator on U .

The solution to problem ($\mathcal{P}1$) has been given for infinite-dimensional control systems in [22], [30], and the equations representing the solution have the same form as in the finite-dimensional case. We will now give the version of the solution that is most useful for our purposes.

For given $z(t_0)$, $J(G; t_0, t_f, z(t_0), u)$ is a bounded linear-quadratic functional on $l^2(t_0, t_f; U)$ with coercive quadratic part. Therefore, for each $z(t_0)$, there exists a unique optimal control sequence in $l^2(t_0, t_f; U)$. Also, the minimum value of the performance index is a quadratic functional of $z(t_0)$, so that there exists a unique nonnegative, self-adjoint $\Pi(t_0) \in \mathcal{L}(Z)$ such that

$$(2.4) \quad J_* = \min J(G; t_0, t_f, z(t_0), u) = \langle \Pi(t_0)z(t_0), z(t_0) \rangle_Z.$$

Application of the principle of dynamic optimality establishes that the optimal control has the feedback form

$$(2.5) \quad u_*(t) = -F(t)z_*(t), \quad t_0 \leq t \leq t_f - 1,$$

where $z_*(t)$ is the optimal trajectory,

$$(2.6) \quad F(t) = \tilde{R}(t)^{-1}B^*\Pi(t+1)T,$$

$$(2.7) \quad \tilde{R}(t) = R + B^*(t+1)B$$

and $\Pi(t)$ satisfies the Riccati difference equation

$$(2.8) \quad \Pi(t) = T^*[\Pi(t+1) - \Pi(t+1)B\tilde{R}(t)^{-1}B^*\Pi(t+1)]T + Q, \quad t \leq t_f - 1, \quad \Pi(t_f) = G.$$

The optimal trajectory z_* is given by $z_*(t+1) = S(t)z_*(t)$, for $t \geq t_0$ where

$$(2.9) \quad S(t) = T - BF(t).$$

2.2. Control on the infinite interval. Here, $t_f = \infty$ and $G = 0$. To simplify notation, we will write $J(t_0, \infty, z(t_0), u)$ instead of $J(0; t_0, \infty, z(t_0), u)$.

DEFINITION 2.1. A control sequence $u \in l^2(0, \infty; U)$ is an *admissible control for the initial condition* z if $J(0, \infty, z, u) < \infty$.

The discrete-time linear-quadratic regulator problem on the infinite interval is

$$(\mathcal{P}2) \quad \text{Choose an admissible control } u_* \text{ to minimize } J(0, \infty, z, u), \text{ if an admissible control exists for the initial condition } z.$$

That a unique control u_* exists whenever at least one admissible control exists follows from the fact that the quadratic part of $J(0, \infty, z, u)$ is coercive on a subspace of $l^2(0, \infty; U)$. See the discussion following Definition 4.1 of [10].

DEFINITION 2.2. A bounded linear operator Π on Z is a *solution to the algebraic Riccati equation* if

$$(2.10) \quad \Pi = T^*[\Pi - \Pi B(R + B^*\Pi B)^{-1}B^*\Pi]T + Q.$$

The following theorem summarizes results from Zabczyk [30].

THEOREM 2.3. *The following are equivalent:*

- (i) *There exists an admissible control for each $z \in Z$;*
- (ii) *For each $z \in Z$, $\sup_{t < t_f} \langle \Pi(t)z, z \rangle_Z < \infty$, where $\Pi(t)$ is the Riccati operator in (2.10) and $\Pi(t_f) = 0$ for fixed t_f ;*
- (iii) *As $t \rightarrow -\infty$, $\Pi(t)$ converges strongly to a nonnegative self-adjoint solution to the algebraic Riccati equation;*
- (iv) *There exists a nonnegative self-adjoint solution to the algebraic Riccati equation.*

For uniqueness of the solution to the algebraic Riccati equation and characterization of the optimal control, Zabczyk treated two cases: when Q is coercive and when

the spectral radius of T is less than 1 (i.e., the open-loop system is uniformly exponentially stable). Since neither is the case in the example we discuss in § 4.2 and other applications in which we are interested, we will need the following hypothesis and theorem.

Hypothesis 2.4. The operators T , B and Q are such that, if $z(0) \in Z$ and u is an admissible control for $z(0)$, then

$$(2.11) \quad \lim_{t \rightarrow \infty} |z(t)|_Z = 0.$$

In finite dimensions, Hypothesis 2.4 is equivalent to saying that the pair (Q, T) is detectable. In infinite dimensions, Hypothesis 2.4 implies that any unstable finite-dimensional eigenspace of T is observable by output Qz , and this is sufficient for Hypothesis 2.4 in most applications, including those discussed in § 4.

THEOREM 2.5. *When Hypothesis 2.4 holds, there exists at most one nonnegative self-adjoint solution to the algebraic Riccati equation. If such a solution Π exists, then there exists a unique solution to problem $(\mathcal{P}2)$ for each initial condition $z(0) \in Z$, the minimum value of the performance index is*

$$(2.12) \quad J_* = \min_{u \text{ admissible}} J(0, \infty, z(0), u) = \langle \Pi z(0), z(0) \rangle_Z,$$

the optimal control has the feedback form

$$(2.13) \quad u_*(t) = -Fz_*(t), \quad t \geq 0,$$

$$(2.14) \quad F = \tilde{R}^{-1} B^* \Pi T,$$

$$(2.15) \quad \tilde{R} = R + B^* \Pi B$$

and the optimal trajectory $z_(t)$ satisfies $z_*(t+1) = Sz_*(t)$, $t \geq 0$, with*

$$(2.16) \quad S = T - BF.$$

Proof. Let Π be such a solution and note that, for any finite t_f , Π is a constant solution to (2.8) with $G = \Pi$. Then the corresponding $F(t)$ and $\tilde{R}(t)$ defined by (2.6) and (2.7) are the constant operators in (2.14) and (2.15). For $z(0) \in Z$, define $\bar{z}(0) = z(0)$, $\bar{z}(t+1) = (T - BF)\bar{z}(t)$, $t \geq 0$, and $\bar{u}(t) = -F\bar{z}(t)$, $t \geq 0$. Now suppose that u is an admissible control for $z(0)$ and that $z(t)$ is the corresponding solution to (2.3). For $t_f > 0$, the preceding results about the solution to problem $(\mathcal{P}1)$ with $G = \Pi$ imply

$$(2.17) \quad \begin{aligned} J(\Pi; 0, t_f, z(0), \bar{u}) &\leq J(0; 0, t_f, z(0), u) + \langle \Pi z(t_f), z(t_f) \rangle_Z \\ &\leq J(0; 0, \infty, z(0), u) + \langle \Pi z(t_f), z(t_f) \rangle_Z. \end{aligned}$$

Also,

$$(2.18) \quad \begin{aligned} J(\Pi; 0, t_f, z(0), \bar{u}) &= \langle \Pi z(0), z(0) \rangle_Z \\ &= J(0; 0, t_f, z(0), \bar{u}) + \langle \Pi \bar{z}(t_f), \bar{z}(t_f) \rangle_Z. \end{aligned}$$

Since $z(t_f) \rightarrow 0$ as $t_f \rightarrow \infty$, (2.17) shows that \bar{u} is both admissible and optimal for problem $(\mathcal{P}2)$. Since $\bar{z}(t_f) \rightarrow 0$ as $t_f \rightarrow \infty$, (2.18) shows (2.12). As we see now, (2.12) must hold for any nonnegative self-adjoint solution of the algebraic Riccati equation; therefore, such a solution is unique.

Remark 2.6. When Hypothesis 2.4 does not hold, the algebraic Riccati equation may have more than one nonnegative self-adjoint solution. In this case, the minimal such solution—there will be one—gives the solution to problem $(\mathcal{P}2)$ as in Theorem 2.5. Throughout this paper, we assume that Hypothesis 2.4 holds.

LEMMA 2.7. Suppose that $Q \cong m$ for some positive constant m , and set $C_n = \sum_{t=0}^n (T^*)'QT'$, for $n = 1, 2, \dots$. Then $|C_n z|_Z$ is bounded in n for each $z \in Z$ if and only if C_n converges in norm to the operator $C = \sum_{t=0}^{\infty} (T^*)'QT'$ and

$$(2.19) \quad |T'| \leq (|C|/m)(1 - m/|C|)', \quad t = 1, 2, \dots$$

Proof. Since C_n is an increasing sequence of bounded self-adjoint linear operators, C_n converges strongly to some bounded self-adjoint C if and only if $\langle C_n z, z \rangle_Z$ is bounded in n for each z , if and only if $|C_n z|_Z$ is bounded in n for each z . This is a standard result. The proof of the lemma is then a standard exercise using the Lyapunov functional $\langle Cz(t), z(t) \rangle_Z$ for the homogeneous part of (2.3).

COROLLARY 2.8. If $Q \cong m > 0$ and the algebraic Riccati equation has a nonnegative self-adjoint solution Π , then the spectral radius of the operator S in (2.16) is less than 1, and

$$(2.20) \quad |S^t| \leq (|\Pi|/m)(1 - m/|\Pi|)', \quad t = 1, 2, \dots$$

Proof. This follows from Lemma 2.7 and

$$(2.21) \quad \langle \Pi z, z \rangle_Z = \sum_{t=0}^{\infty} \langle (S^*)' [Q + F^* R F] S^t z, z \rangle_Z.$$

For Q coercive, Zabczyk proved a stronger result than part (iii) of Theorem 2.3: if a nonnegative self-adjoint solution to the Riccati algebraic equation exists, then $|\Pi(t) - \Pi| \rightarrow 0$ geometrically fast as $t \rightarrow -\infty$. (Also, see [14].) We will need such a result, along with an explicit convergence rate, for the approximation theory in § 3.2. Since Zabczyk's proof does not yield an explicit convergence rate, we give the following theorem.

THEOREM 2.9. Suppose that there exists a nonnegative self-adjoint solution Π to (2.10) and that

$$(2.22) \quad |S^t| \leq M r^t, \quad t = 1, 2, \dots,$$

where M and r are positive constants with $r < 1$ and S is the optimal closed-loop operator in Theorem 2.5. If $\Pi(\cdot)$ is the operator in (2.8) with $t_f = 0$ and $\Pi(0) \geq \Pi$, then

$$(2.23) \quad \langle \Pi z, z \rangle_Z \leq \langle \Pi(-t)z, z \rangle_Z \leq \langle \Pi z, z \rangle_Z + (M r^t) |\Pi(0)|, \quad t = 1, 2, \dots$$

Proof. For t_0 a negative integer, let u_0 be the optimal control sequence for the finite-time problem ($\mathcal{P}1$) on the interval $[t_0, 0]$ with initial condition $z(t_0) \in Z$, with z_0 the corresponding optimal trajectory. Also, let u_* be the optimal control sequence on the infinite interval for problem ($\mathcal{P}2$) with initial condition $z(t_0)$, with z_* the corresponding optimal trajectory. Since Π is a constant solution to (2.8) for the final condition $G = \Pi$, we have

$$(2.24) \quad \begin{aligned} \langle \Pi z(t_0), z(t_0) \rangle_Z &= J(\Pi; 0, -t_0, z(t_0), u_*(\cdot - t_0)) \\ &\leq J(0; t_0, 0, z(t_0), u_0) + \langle \Pi z_0(0), z_0(0) \rangle_Z \\ &\leq J(0; t_0, 0, z(t_0), u_0) + \langle \Pi(0) z_0(0), z_0(0) \rangle_Z \\ &= \langle \Pi(t_0) z(t_0), z(t_0) \rangle_Z. \end{aligned}$$

On the other hand (note that $z_*(-t_0) = S^{-t_0} z(t_0)$),

$$(2.25) \quad \begin{aligned} \langle \Pi(t_0) z(t_0), z(t_0) \rangle_Z &\leq J(0; 0, -t_0, z(t_0), u_*) + \langle \Pi(0) z_*(-t_0), z_*(-t_0) \rangle_Z \\ &\leq J(0; 0, \infty, z(t_0), u_*) + \langle \Pi(0) z_*(-t_0), z_*(-t_0) \rangle_Z \\ &\leq \langle \Pi z(t_0), z(t_0) \rangle_Z + |\Pi(0)| \|S^{-t_0} z(t_0)\|_Z^2. \end{aligned}$$

3. Approximation theory.

3.1. The finite time interval problem. In this section we develop a general approximation framework for the finite time interval problem ($\mathcal{P}1$) and describe associated convergence results.

For each $N = 1, 2, \dots$, let $Z_N \subset Z$ be a finite-dimensional subspace of Z and let $P_N: Z \rightarrow Z_N$ denote projection-like mappings of Z onto Z_N . We require the following hypotheses.

Hypothesis 3.1. There exist operators $T_N: Z_N \rightarrow Z_N$, $B_N: U \rightarrow Z_N$, $Q_N: Z_N \rightarrow Z_N$ and $G_N: Z_N \rightarrow Z_N$ which satisfy $T_N P_N \rightarrow T$, $T_N^* P_N \rightarrow T^*$, $B_N \rightarrow B$, $Q_N P_N \rightarrow Q$ and $G_N P_N \rightarrow G$ strongly, as $N \rightarrow \infty$ with Q_N and G_N self-adjoint and nonnegative.

Hypothesis 3.2. The spaces Z_N are approximating subspaces in the sense that the P_N satisfy $P_N \rightarrow I$ strongly on Z as $N \rightarrow \infty$.

We note that since U has been assumed to be finite-dimensional, Hypothesis 3.1 above necessarily implies that $B_N \rightarrow B$ and $B_N^* P_N \rightarrow B^*$ in the uniform norm topology on $\mathcal{L}(U, Z)$ and $\mathcal{L}(Z, U)$, respectively.

We define a sequence of approximating discrete-time linear quadratic regulator problems on the finite time interval $[t_0, t_f]$ as follows:

($\mathcal{P}1_N$) Find $u_*^N \in l^2(t_0, t_f - 1; U)$ which minimizes

$$(3.1) \quad J_N(G_N; t_0, t_f, z(t_0), u) = \sum_{t=t_0}^{t_f-1} [\langle Q_N z_N(t), z_N(t) \rangle_Z + \langle Ru(t), u(t) \rangle_U] \\ + \langle G_N z_N(t_f), z_N(t_f) \rangle_Z$$

subject to

$$(3.2) \quad z_N(t+1) = T_N z_N(t) + B_N u(t), \quad t \geq t_0, \quad z_N(t_0) = P_N z(t_0).$$

The results stated in § 2.1 concerning the existence and uniqueness of solutions to problem ($\mathcal{P}1$) apply to the problems ($\mathcal{P}1_N$) as well. Indeed, there exists a unique solution $u_*^N \in l^2(t_0, t_f - 1; U)$ to problem ($\mathcal{P}1_N$) which is given in feedback form by

$$(3.3) \quad u_*^N(t) = -F_N(t) z_*^N(t), \quad t_0 \leq t \leq t_f - 1,$$

$$(3.4) \quad F_N(t) = \tilde{R}_N(t)^{-1} B_N^* \Pi_N(t+1) T_N,$$

$$(3.5) \quad \tilde{R}_N(t) = R + B_N^* \Pi_N(t+1) B_N$$

and the operators $\{\Pi_N(t)\}_{t=t_0}^{t_f}$ on Z_N satisfying the Riccati difference equation

$$(3.6) \quad \Pi_N(t) = T_N^* [\Pi_N(t+1) - \Pi_N(t+1) B_N \tilde{R}_N(t)^{-1} B_N^* \Pi_N(t+1)] T_N \\ + Q_N \Pi_N(t_f) = G_N.$$

The optimal trajectory z_*^N is given by $z_*^N(t+1) = S_N(t) z_*^N(t)$, $t \geq t_0$, $z_*^N(t_0) = P_N z(t_0)$ where

$$(3.7) \quad S_N(t) = T_N - B_N F_N(t), \quad t \geq t_0.$$

The operators $\{\Pi_N(t)\}_{t=t_0}^{t_f}$ are bounded, self-adjoint and nonnegative. The minimum value of the performance index (3.1) is given by

$$(3.8) \quad J_*^N = J_N(G_N; t_0, t_f, z(t_0), u_*^N) = \langle \Pi_N(t_0) z_*^N(t_0), z_*^N(t_0) \rangle_Z.$$

The fundamental convergence result is given in the following theorem.

THEOREM 3.3. Let u_*^N and u_* be the unique solutions to problems ($\mathcal{P}1_N$) and ($\mathcal{P}1$), respectively, with z_*^N and z_* the corresponding optimal trajectories. Let J_N , Π_N and F_N

and J , Π and F be given by (3.1), (3.6) and (3.4) and by (2.2), (2.8) and (2.6). Then, if Hypotheses 3.1 and 3.2 hold, we have

- (i) $\lim_{N \rightarrow \infty} \|u_*^N - u_*\|_{l^2}^2 = 0,$
- (ii) $\lim_{N \rightarrow \infty} \|z_*^N - z_*\|_{l^2}^2 = 0,$
- (iii) $\lim_{N \rightarrow \infty} \|J_*^N - J_*\| = 0,$
- (iv) $\lim_{N \rightarrow \infty} \|\Pi_N(t)P_N z - \Pi(t)z\|_Z = 0, \quad z \in Z, \quad t_0 \leq t \leq t_f \quad \text{and}$
- (v) $\lim_{N \rightarrow \infty} \|F_N(t)P_N - F(t)\| = 0, \quad t \leq t \leq t_f - 1.$

Proof. We first note that $\Pi_N(t)$ being nonnegative implies that $|\tilde{R}_N(t)| \geq |R|$ and consequently that $|\tilde{R}_N(t)^{-1}| \leq |R|^{-1}$. It follows therefore that for $u \in U$.

$$(3.9) \quad \begin{aligned} |(\tilde{R}_N(t)^{-1} - \tilde{R}(t)^{-1})u|_U &= |\tilde{R}_N(t)^{-1}(\tilde{R}(t) - \tilde{R}_N(t))\tilde{R}(t)^{-1}u|_U \\ &\leq |R|^{-1}|(\tilde{R}(t) - \tilde{R}_N(t))\tilde{R}(t)^{-1}u|_U. \end{aligned}$$

The above estimate together with (2.7), (3.5), the fact that $\Pi(t_f) = G$ and $\Pi_N(t_f) = G_N$, and Hypothesis 3.1 imply that $\tilde{R}_N(t_f - 1)^{-1} \rightarrow \tilde{R}(t_f - 1)^{-1}$ as $N \rightarrow \infty$ strongly on U . Since U is finite dimensional the convergence is in fact uniform. It then follows immediately from (2.6), (3.4) and Hypothesis 3.1 that $F_N(t_f - 1)P_N \rightarrow F(t_f - 1)$, uniformly as $N \rightarrow \infty$, and from (2.8) and (3.6), that $\Pi_N(t_f - 1)P_N \rightarrow \Pi(t_f - 1)$ strongly on Z as $N \rightarrow \infty$. A simple induction yields (iv) and (v) from which (i)–(iii) then follow trivially.

Remark. It will, on occasion, be the case that in constructing a particular approximation scheme $T_N P_N \rightarrow T$ strongly but $T_N^* P_N \rightarrow T^*$ only weakly (see, for example, [3]). However, by using the fact that $(T_N^* \Pi_N(t+1))^* = \Pi_N(t+1)T_N$ implies that $T_N^* \Pi_N(t+1) \rightarrow T^* \Pi(t+1)$ weakly if $\Pi_N(t+1) \rightarrow \Pi(t+1)$ weakly, we conclude that Theorem 3.3 continues to hold under these somewhat weaker hypotheses with the strong convergence in (iv) replaced by weak and the uniform convergence in (v) replaced by strong.

Under certain additional hypotheses it can be shown that the operators $\Pi(t)$, $t_0 \leq t \leq t_f$ given by (2.8) are trace class (see [17]) and that $\lim_{N \rightarrow \infty} \|\Pi_N(t)P_N - \Pi(t)\|_1 = 0$, $t_0 \leq t \leq t_f$, where $\|\cdot\|_1$ denotes the trace norm, the strongest of all common operator norms. To show this we require the following lemmas.

LEMMA 3.4. *If $\{a_i\}_{i=1}^\infty$ is an absolutely summable sequence of real numbers then there exist sequences $\{b_i\}_{i=1}^\infty$ and $\{c_i\}_{i=1}^\infty$ such that $\lim_{i \rightarrow \infty} b_i = 0$, $\{c_i\}_{i=1}^\infty$ is absolutely summable and $a_i = b_i c_i$.*

Proof. Let $\alpha = \sum_{i=1}^\infty |a_i|$ and for $j = 0, 1, 2, \dots$, define nonnegative integers n_j as follows. Let $n_0 = 0$ and let n_j denote the first index for which

$$(3.10) \quad \sum_{i=1}^{n_j} |a_i| > \alpha - \frac{1}{j^3},$$

$j = 1, 2, \dots$. Set $b_i = 1/j$, $c_i = ja_i$, $i = n_{j-1} + 1, \dots, n_j$, $j = 1, 2, \dots$. Then $b_i c_i = a_i$, $i = 1, 2, \dots$, $\lim_{i \rightarrow \infty} b_i = 0$ and

$$(3.11) \quad \sum_{i=1}^\infty |c_i| = \sum_{j=1}^\infty j \sum_{k=n_{j-1}+1}^{n_j} |a_k| \leq \alpha + \sum_{j=1}^\infty \frac{1}{j^2} < \infty.$$

LEMMA 3.5. *If L is a self-adjoint trace class operator on a separable Hilbert space H , then L can be written as $L^1 L^2$ where L^1 is compact and L^2 is trace class.*

Proof. Let $\{\lambda_i\}_{i=1}^\infty$ denote the eigenvalues of L repeated according to multiplicity and let $\{\phi_i\}_{i=1}^\infty$ denote the corresponding eigenvectors. Then $\{\lambda_i\}_{i=1}^\infty$ is a sequence of real numbers, each of finite multiplicity, and

$$(3.12) \quad \sum_{i=1}^{\infty} |\lambda_i| = \|L\|_1 < \infty.$$

According to Lemma 3.4, there exist sequences $\{\mu_i\}_{i=1}^\infty$ and $\{\nu_i\}_{i=1}^\infty$ with $\lim_{i \rightarrow \infty} \mu_i = 0$, $\sum_{i=1}^{\infty} |\nu_i| < \infty$ and $\lambda_i = \mu_i \nu_i$. With L^1 and L^2 defined by $L^1 \phi = \sum_{i=1}^{\infty} \mu_i \langle \phi, \phi_i \rangle_H \phi_i$ and $L^2 \phi = \sum_{i=1}^{\infty} \nu_i \langle \phi, \phi_i \rangle_H \phi_i$ for $\phi \in H$, respectively, the lemma immediately follows.

LEMMA 3.6. Let $\{S_N\}_{N=1}^\infty$ be a sequence of bounded linear operators on a separable Hilbert space H that converges strongly to a bounded linear operator S . Let $\{L_N\}_{N=1}^\infty$ be a sequence of trace class operators on H that converges in trace norm to an operator L . If L can be written as $L = L^1 L^2$ with L^1 compact and L^2 trace class then the sequence $\{S_N L_N\}_{N=1}^\infty$ converges in trace norm to SL .

Proof. The result follows immediately from the estimate:

$$(3.13) \quad \begin{aligned} \|S_N L_N - SL\|_1 &\leq \|S_N(L_N - L)\|_1 + \|(S_N - S)L^1 L^2\|_1 \\ &\leq \|S_N\| \|L_N - L\|_1 + \|(S_N - S)L^1\| \|L^2\|_1. \end{aligned}$$

THEOREM 3.7. If Q and G are trace class operators then the operators $\{\Pi(t)\}_{t=t_0}^{t_f}$ given by (2.8) are trace class. Moreover, if Hypotheses 3.1 and 3.2 hold and $Q_N P_N \rightarrow Q$ and $G_N P_N \rightarrow G$ in trace norm as $N \rightarrow \infty$ then we have

$$(3.14) \quad \lim_{N \rightarrow \infty} \|\Pi_N(t) P_N - \Pi(t)\|_1 = 0, \quad t_0 \leq t \leq t_f.$$

Proof. That the operators $\Pi(t)$, $t_0 \leq t \leq t_f$ are trace class is an immediate consequence of the hypotheses of the theorem, (2.8) and the fact that the trace class operators form a two-sided ideal of $\mathcal{L}(Z)$, the space of bounded linear operators on Z (see [17]). The trace norm convergence stated in (3.14) will follow once we have shown that $\lim_{N \rightarrow \infty} \|\Pi_N(t+1) P_N - \Pi(t+1)\|_1 = 0$ implies

$$(i) \quad \lim_{N \rightarrow \infty} \|T_N^* \Pi_N(t+1) T_N P_N - T^* \Pi(t+1) T\|_1 = 0, \quad \text{and}$$

$$(ii) \quad \lim_{N \rightarrow \infty} \|T_N^* \Pi_N(t+1) B_N \tilde{R}_N(t)^{-1} B_N^* \Pi_N(t+1) T_N P_N - T^* \Pi(t+1) B \tilde{R}(t)^{-1} B^* \Pi(t+1) T\|_1 = 0.$$

To argue (i) we first note that Hypothesis 3.1 and Lemmas 3.5 and 3.6 imply $\lim_{N \rightarrow \infty} \|T_N^* \Pi_N(t+1) P_N - T^* \Pi(t+1)\|_1 = 0$. Taking adjoints we obtain $\lim_{N \rightarrow \infty} \|\Pi_N(t+1) T_N P_N - \Pi(t+1) T\|_1 = 0$. Another application of the previous two lemmas yields

$$(3.15) \quad \begin{aligned} &\lim_{N \rightarrow \infty} \|T_N^* \Pi_N(t+1) T_N P_N - T^* \Pi(t+1) T\|_1 \\ &\leq \lim_{N \rightarrow \infty} \|T_N^* \Pi_N(t+1) T_N P_N - \Pi(t+1) T\|_1 \\ &\quad + \lim_{N \rightarrow \infty} \|(\Pi(t+1) T - T^* \Pi(t+1) T)\|_1 \end{aligned}$$

where $\Pi(t+1) = \Pi^1(t+1) \Pi^2(t+1)$ is the factorization of $\Pi(t+1)$ described in Lemma 3.6. The verification of (ii) is analogous and the theorem is proven.

We note that if Hypotheses 3.1 and 3.2 hold and if the operators Q and G are trace class with Q_N and G_N defined by $Q_N = P_N Q$ and $G_N = P_N G$, then Lemmas 3.5 and 3.6 imply that the trace norm convergence hypotheses in Theorem 3.7 hold. This is the case in the time delay problem considered in § 4.2 below.

3.2. Approximation on the infinite interval. Problem (\mathcal{P}_{2N}) is problem $(\mathcal{P}2)$ for the control system in (3.2) and the performance index

$$(3.16) \quad J_N(0, \infty, z_N(0), u) = \sum_{t=0}^{\infty} [\langle Q_N z_N(t), z_N(t) \rangle_Z + \langle Ru(t), u(t) \rangle_U].$$

Hypothesis 3.8. For each N , there exists exactly one nonnegative self-adjoint solution to the Riccati algebraic equation

$$(3.17) \quad \Pi_N = T_N^*[\Pi_N - \Pi_N B_N (R + B_N^* \Pi_N B_N)^{-1} B_N^* \Pi_N] T_N + Q_N.$$

One may suspect that solvability of the approximating finite-dimensional Riccati equations should follow from solvability of the infinite-dimensional Riccati equation, but we have no such result. Even in the corresponding continuous time case, which has been studied extensively, the solvability of the approximating problems must be established for particular approximation schemes and does not follow from solvability of the infinite-dimensional problem.

By Theorem 2.3, Hypothesis 3.8 implies that $\lim_{t \rightarrow -\infty} |\Pi_N - \Pi_N(t)| = 0$ for each N , since $\dim(Z_N) < \infty$.

As in Theorem 2.5, we write

$$(3.18) \quad F_N = \tilde{R}_N^{-1} B_N^* \Pi_N T_N,$$

$$(3.19) \quad \tilde{R}_N = R + B_N^* \Pi_N B_N,$$

$$(3.20) \quad S_N = T_N - B_N F_N.$$

From here on, Π will be the nonnegative self-adjoint solution to the infinite-dimensional Riccati algebraic equation (2.10)—when it exists— F will be the corresponding feedback operator in (2.14) and S will be the corresponding closed-loop operator in (2.16).

THEOREM 3.9. *If $\Pi_N P_N$ converges strongly to some bounded linear operator Π , then Π is a nonnegative self-adjoint solution to (2.10), $F_N P_N$ converges in norm to F and $S_N P_N$ converges strongly to S .*

Proof. This follows from Hypotheses 3.1 and 3.2, and the fact that the control space U has fixed finite dimension.

THEOREM 3.10. *Suppose that there exist positive constants M and r , independent of N , with $r < 1$, such that*

$$(3.21) \quad \Pi_N \leq M, \quad N = 1, 2, \dots, \quad \text{and}$$

$$(3.22) \quad |S_N^t| \leq M r^t, \quad t = 1, 2, \dots, \quad N = 1, 2, \dots.$$

Then a nonnegative self-adjoint solution Π to (2.10) exists, and $\Pi_N P_N \rightarrow \Pi$ strongly as $N \rightarrow \infty$. If there exists a positive m , independent of N , such that $Q_N \geq m$, $N = 1, 2, \dots$, then (3.21) implies the existence of an r less than one and independent of N for which (3.22) holds.

Proof. For each N , let $\Pi_N(\cdot)$ satisfy (3.6) with $t_f = 0$ and $\Pi_N(0) = MI$, where I denotes the identity operator on Z_N . From (2.23), $|\Pi_N - \Pi_N(-t)| \rightarrow 0$ as $t \rightarrow \infty$, uniformly in N . Now, for $z \in Z$, write

$$(3.23) \quad \begin{aligned} \langle (\Pi_N - \Pi_{N'})z, z \rangle_Z &= \langle (\Pi_N - \Pi_N(-t))z, z \rangle_Z + \langle (\Pi_N(-t) - \Pi_{N'}(-t))z, z \rangle_Z \\ &\quad + \langle (\Pi_{N'}(-t) - \Pi_{N'})z, z \rangle_Z. \end{aligned}$$

For $\varepsilon > 0$ choose $t > 0$ such that $|\langle (\Pi_N - \Pi_N(-t))z, z \rangle_Z| < \varepsilon$ and $|\langle (\Pi_{N'} - \Pi_{N'}(-t))z, z \rangle_Z| < \varepsilon$. Then, for N and N' large enough, $|\langle (\Pi_N(-t) - \Pi_{N'}(-t))z, z \rangle_Z| < \varepsilon$. This shows that Π_{Nz}

is a Cauchy sequence in Z for each z . Therefore, Π_N converges strongly to a nonnegative self-adjoint solution to (2.10). The last sentence in the theorem follows from Lemma 2.7.

An important application of this theorem is the case when the approximating open-loop operators T_N have an exponential decay rate independent of N , Q is coercive and $Q_N = P_N Q|_{Z_N}$. In this case, the zero control gives an upper bound, independent of N , on Π_N . Such is the case in the example discussed in § 4.1 and in applications to flexible structures with no rigid-body modes and coercive structural damping.

THEOREM 3.11. *Suppose that $\Pi_N P_N$ converges strongly to Π , $Q_N P_N$ converges in trace norm to Q (hence Q is trace class), and (3.22) holds for positive M and r independent of N with r less than one. Then $\Pi_N P_N$ converges in trace norm to Π .*

Proof. From (3.18)–(3.20), it follows that F_N^* and hence F_N converge in trace norm and that S_N converges strongly. Then for each N , it follows from Lemmas 3.5 and 3.6 that $(S_N^*)'[Q_N + F_N^* R F_N] S_N^t$ converges in trace norm for each t . Writing (2.21) for the N th approximating problem yields $\Pi_N = \sum_{t=0}^{\infty} (S_N^*)'[Q_N + F_N^* R F_N] S_N^t$ where the series converges absolutely in trace norm, uniformly in N , because $\|(S_N^*)'[Q_N + F_N^* R F_N] S_N^t\|_1 \leq \|S_N\|^{2t} \|Q_N + F_N^* R F_N\|_1$. Therefore, since each term in the series converges in trace norm, Π_N converges in trace norm to Π .

Note that $Q_N P_N$ converges in trace norm to Q if Q is trace class and $Q = P_N Q P_N|_{Z_N}$.

THEOREM 3.12. *If $|\Pi_N|$ is bounded in N , then a nonnegative self-adjoint solution Π to (2.10) exists, $\Pi_N P_N$ converges weakly to Π , and $F_N P_N$ and $S_N P_N$ converge strongly to F and S , respectively.*

Proof. According to Theorem 6 of [12], $\Pi_N P_N$ converges weakly to some nonnegative self-adjoint bounded Π . It follows from (3.17) and Hypotheses 3.1 and 3.2 that Π satisfies (2.10) and that F_N and S_N converge as indicated.

Note that Theorem 3.12 holds if $S_N P_N$ converges strongly but $S_N^* P_N$ converges only weakly.

3.3. Implementation of the approximation schemes. The expressions given by (3.3)–(3.6) are operator equations and although they are finite-dimensional, they are not appropriate for computations. To make use of our approximation framework, we must first determine equivalent matrix formulations. Toward this end we assume, without loss of generality, that $U = R^m$ with the standard basis and inner product and let $\{\phi_N^i\}_{i=1}^{K_N}$ be a basis for Z_N . Define the $K_N \times K_N$ Gram matrix M_N by $[M_N]_{ij} = \langle \phi_N^i, \phi_N^j \rangle_Z$. For an operator A we denote its matrix representation with respect to the bases defined above by $[A]$. Similarly, for an element $z \in Z$ or $u \in U$, we let its vector representation be given by $[z]$ or $[u]$ respectively. Standard calculations yield $[T_N^*] = M_N^{-1} [T_N]^T M_N$ and $[B_N^*] = [B_N]^T M_N$. Defining $\hat{\Pi}_N(t) = M_N [\Pi_N(t)]$, $\hat{Q}_N = M_N [Q_N]$, and $\hat{G}_N = M_N [G_N]$ we obtain

$$(3.24) \quad [u_*^N(t)] = -[F_N(t)][z_*^N(t)], \quad t_0 \leq t \leq t_f - 1,$$

$$(3.25) \quad [F_N(t)] = [\tilde{R}_N(t)]^{-1} [B_N]^T \hat{\Pi}_N(t+1) [T_N],$$

$$(3.26) \quad [\tilde{R}_N(t)] = [R] + [B_N]^T \hat{\Pi}_N(t+1) [B_N],$$

$$(3.27)$$

$$\hat{\Pi}_N(t) - [T_N]^T (\hat{\Pi}_N(t+1) - \hat{\Pi}_N(t+1) [B_N] [\tilde{R}_N(t)]^{-1} [B_N]^T \hat{\Pi}_N(t+1)) [T_N] + \hat{Q}_N,$$

$$t_0 \leq t \leq t_f - 1,$$

$$(3.28) \quad \hat{\Pi}_N(t_f) = \hat{G}_N.$$

Note that since Q_N and G_N are self-adjoint and nonnegative, \hat{Q}_N and \hat{G}_N are also. Equations (3.24)–(3.28) are therefore in the form of the standard ones obtained for the feedback law for a discrete-time linear-quadratic regulator problem in R^{K_N} . Consequently they can be solved using conventional techniques. The minimum value of the performance index is given by $J_*^N = [z_*^N(t_0)]^T \hat{\Pi}_N(t_0) [z_*^N(t_0)]$.

Analogously, for the infinite horizon problem, we have

$$(3.29) \quad [u_*^N(t)] = -[F_N][z_*^N(t)], \quad t \geq t_0,$$

$$(3.30) \quad [F_N] = [\tilde{R}_N]^{-1} [B_N]^T \hat{\Pi}_N [T_N],$$

$$(3.31) \quad [\tilde{R}_N] = [R] + [B_N]^T \hat{\Pi}_N [B_N],$$

where $\hat{\Pi}_N$ is the solution to the matrix algebraic Riccati equation

$$(3.32) \quad \hat{\Pi}_N = [T_N]^T (\hat{\Pi}_N - \hat{\Pi}_N [B_N] [\tilde{R}_N]^{-1} [B_N]^T \hat{\Pi}_N) [T_N] + \hat{Q}_N.$$

The minimum value of the performance index is given by $J_*^N = [z_*^N(t_0)]^T \hat{\Pi}_N [z_*^N(t_0)]$.

4. Examples and numerical results. In this section we describe the application of the general approximation framework developed above to a variety of examples. In addition to theoretical considerations, in each of the examples below, we discuss some numerical results for an infinite horizon problem of the form given in problem (P2).

Of primary concern to us will be applications in which one considers piecewise constant controls in the sampled form of the continuous-time control system

$$(4.1) \quad \dot{z}(s) = \mathcal{A}z(s) + \mathcal{B}u(s), \quad s > 0, \quad z(0) = z_0,$$

where \mathcal{A} is the infinitesimal generator of a \mathcal{C}_0 -semigroup of bounded linear operators $\mathcal{T}(s)$, $s \geq 0$, on Z , \mathcal{B} is a (possibly unbounded) linear operator from U into Z and $z_0 \in Z$. In this case we have

$$(4.2) \quad T = \mathcal{T}(\tau) \quad \text{and} \quad B = \int_0^\tau \mathcal{T}(s) \mathcal{B} ds,$$

where τ is the length of the sampling interval. If, as in the example in § 4.1, where u is a boundary control in a heat equation, \mathcal{B} is unbounded (more precisely, \mathcal{B} maps U not into Z but into some larger space), then the integral in (4.2) is not interpreted literally (see, for example, [7]).

In constructing the approximating operators T_N , B_N , Q_N and G_N a standard Galerkin approach is often taken; that is, $T_N = P_N T$, $B_N = P_N B$, $Q_N = P_N Q$ and $G_N = P_N G$. We note however that explicit representations for the operators T and B are frequently not available. In particular, this can occur when the discrete-time system (2.3) arises from the sampling of an infinite-dimensional continuous-time system as was described above. In this case it is the operators \mathcal{A} and \mathcal{B} which are approximated by a sequence of finite-dimensional operators \mathcal{A}_N and \mathcal{B}_N on Z_N , from which an approximation to the semigroup $\{\mathcal{T}(s): s \geq 0\}$ is obtained as $\mathcal{T}_N(s) = \exp(\mathcal{A}_N s)$, $s \geq 0$. The operators T_N and B_N are then $T_N = \mathcal{T}_N(\tau)$ and $B_N = \int_0^\tau \mathcal{T}_N(s) \mathcal{B}_N ds$, respectively. The strong convergence $T_N P_N \rightarrow T$ and $B_N \rightarrow B$ is then usually argued using an appropriate formulation of the Trotter-Kato theorem, a well-known semigroup approximation result (see [17], [25]).

Matrix exponentials were computed from eigenvalue-eigenvector decompositions obtained using the QR algorithm. The matrix Riccati equations (3.32) were solved using a Schur-vector decomposition of the Hamiltonian matrix (see [20], [26]). It should be noted that if the eigenvalue pairs of the Hamiltonian matrix for a continuous-time linear-quadratic regulator problem are asymptotic to $\pm \gamma(n)$ as $n \rightarrow \infty$, then the

eigenvalue pairs of the Hamiltonian matrix for the corresponding discrete-time problem will be asymptotic to $e^{\pm \gamma(n)\tau}$ as $n \rightarrow \infty$. Consequently, for all but very small τ , conditioning problems arise more quickly than in the continuous-time case when the approximating matrix algebraic Riccati equations are solved.

All numerical studies were performed on an IBM Personal Computer. The machine we used was equipped with an Intel 8087 math co-processor chip and 640K bytes of random access memory (of which less than 384K was required).

4.1. A heat equation with boundary control. In this example we consider the scalar parabolic system with boundary control given by

$$(4.3) \quad \frac{\partial}{\partial s} w(s, x) = \frac{\partial}{\partial x} \left(a(x) \frac{\partial}{\partial x} w(s, x) \right), \quad s > 0, \quad x \in (0, 1),$$

$$(4.4) \quad w(s, 0) = 0, \quad w(s, 1) = v(s),$$

$$(4.5) \quad w(0, x) = \phi(x)$$

with $a \in H^1(0, 1)$, $a(x) \geq \alpha > 0$, $x \in [0, 1]$, $\phi \in L_2(0, 1)$ and $v \in L_2(0, \infty)$.

To formulate the discrete-time state equation for this system we let τ denote the sampling interval and consider piecewise constant controls v given by $v(s) = u(t)$, $s \in [t\tau, (t+1)\tau)$, $t = 0, 1, 2, \dots$. We choose our state space Z to be $L_2(0, 1)$ with the usual inner product denoted by $\langle \cdot, \cdot \rangle$. The state $z(t) \in Z$ is

$$(4.6) \quad z(t) = \lim_{s \rightarrow t\tau} w(s, \cdot), \quad t = 1, 2, \dots, \quad z(0) = \phi.$$

For $t \in \{0, 1, 2, \dots\}$, we define $y(s) \in Z$ by

$$(4.7) \quad y(s) = w(s, \cdot) - \psi_0 u(t), \quad s \in (t\tau, (t+1)\tau), \quad y(t\tau) = z(t) - \psi_0 u(t),$$

where $\psi_0 \in Z$ is the ramp function given by $\psi_0(x) = x$, $x \in [0, 1]$. A straightforward calculation reveals that $y(s) = y(s, \cdot)$ satisfies

$$(4.8) \quad \dot{y}(s) = D(aDy(s)) + a'u(t), \quad s \in (t\tau, (t+1)\tau),$$

$$(4.9) \quad y(s)|_0 = 0, \quad y(s)|_1 = 0, \quad s \in (t\tau, (t+1)\tau),$$

$$(4.10) \quad y(t\tau) = z(t) - \psi_0 u(t),$$

where D denotes the differentiation operator on $H^1(0, 1)$. Let $\mathcal{A}: \text{Dom}(\mathcal{A}) \subset Z \rightarrow Z$ be given by $\mathcal{A}\psi = D(aD\psi)$ for $\psi \in \text{Dom}(\mathcal{A}) = H^2(0, 1) \cap H_0^1(0, 1)$. The operator \mathcal{A} is densely defined and self-adjoint. It satisfies

$$(4.11) \quad \langle \mathcal{A}z, z \rangle \leq -\omega |z|^2, \quad z \in \text{Dom}(\mathcal{A})$$

for some $\omega > 0$ and has compact resolvent. Also, \mathcal{A} is the infinitesimal generator of an analytic semigroup of contractions $\{\mathcal{T}(s): s \geq 0\}$ on Z which, in light of (4.11), satisfies $|\mathcal{T}(s)| \leq e^{-\omega s}$, $s \geq 0$. It follows, therefore, that

$$(4.12) \quad y(s) = \mathcal{T}(s - t\tau)y(t\tau) + \int_{t\tau}^s \mathcal{T}(s - \sigma)a' d\sigma u(t), \quad s \in [t\tau, (t+1)\tau).$$

The continuity of y , (4.6) and (4.7) imply $z(t) = y(t\tau) + \psi_0 u(t)$ and $z(t+1) = y((t+1)\tau) + \psi_0 u(t)$, and hence that

$$(4.13) \quad \begin{aligned} z(t+1) &= y((t+1)\tau) + \psi_0 u(t) \\ &= \mathcal{T}(\tau)(z(t) - \psi_0 u(t)) + \int_{t\tau}^{(t+1)\tau} \mathcal{T}((t+1)\tau - \sigma)a' d\sigma u(t) + \psi_0 u(t). \end{aligned}$$

Defining the operators $T \in \mathcal{L}(Z)$ and $B \in \mathcal{L}(R^1, Z)$ by $Tz = \mathcal{T}(\tau)z$ for $z \in Z$ and $Bu = [(I - \mathcal{T}(\tau))\psi_0 + \int_0^\tau \mathcal{T}(\sigma)a' d\sigma]u$, for $u \in R^1$, we obtain a discrete time control system of the form (2.3).

We take the performance index to be

$$(4.14) \quad J(g_0; 0, t_f, \phi, u) = \sum_{t=0}^{t_f-1} \{q_0|z(t)|_0^2 + ru(t)^2\} + g_0|z(t_f)|_0^2$$

with $q_0, g_0 \geq 0$ and $r > 0$.

Applying the theory developed in § 2.1, we have, for the finite time interval problem, that the optimal control u_* is given by $u_*(t) = -F(t)z_*(t)$, $t = 0, 1, 2, \dots, t_f - 1$, where for each t , $F(t)$ is the continuous linear functional on Z given by (2.6)–(2.8). It follows that $F(t)$ has a representation $f(t, \cdot) \in L_2(0, 1)$ and that $F(t)\psi = \int_0^1 f(t, \cdot)\psi(\theta) d\theta$, for $\psi \in L_2(0, 1)$, $t = 0, 1, 2, \dots, t_f - 1$.

For the infinite interval problem it is immediately clear that Hypothesis 2.4 is satisfied. It is also clear that (4.11) implies that for each $z(0) \in L_2(0, 1)$, $u(t) = 0$, $t = 0, 1, 2, \dots$, is an admissible control, and hence that there exists a unique nonnegative self-adjoint solution of the algebraic Riccati equation (2.10). From (2.13)–(2.15) we obtain $u_*(t) = -Fz_*(t)$, $t = 0, 1, 2, \dots$, where F is a continuous linear functional on Z and $F\psi = \int_0^1 f(\theta)\psi(\theta) d\theta$ for $\psi \in L_2(0, 1)$ with $f \in L_2(0, 1)$.

We define a standard Ritz–Galerkin approximation scheme. We define the space $V = H_0^1(0, 1)$ together with the inner product $\langle \phi, \psi \rangle_V = \langle A D\phi, D\psi \rangle$. We note that V is the energy space associated with the coercive operator $-A$, $V = \text{Dom}((-A)^{1/2})$, and it is the closure of $\text{Dom}(A)$ with respect to the V -norm defined above.

For each $N = 2, 3, \dots$, let Δ_N denote the uniform partition of the interval $[0, 1]$ given by $\{0, 1/N, 2/N, \dots, (N-1)/N, 1\}$. Let $\{e_N^j\}_{j=1}^{N-1}$ denote the usual linear B-splines (“hat” functions) on $[0, 1]$ corresponding to the partition Δ_N which satisfy $e_N^j(0) = e_N^j(1) = 0$, $j = 1, 2, \dots, N-1$. Let $Z_N = \text{span}\{e_N^j\}_{j=1}^{N-1} \subset V$. Define $P_N: Z \rightarrow Z_N$ to be the orthogonal projection of Z onto Z_N with respect to the L_2 inner product and $\mathcal{P}_N: V \rightarrow Z_N$ to be the orthogonal projection of V onto Z_N with respect to the V inner product. Define the operator $\mathcal{A}_N: Z_N \rightarrow Z_N$ as the inverse of the operator $\mathcal{A}_N^{-1} = \mathcal{P}_N \mathcal{A}^{-1}|_{Z_N}$. The invertibility of \mathcal{A} of course follows from the coercivity of $-A$ (see (4.11)). On the other hand, a straightforward calculation yields that

$$(4.15) \quad \langle \mathcal{A}_N^{-1} z_N, z_N \rangle_V = |z_N|^2$$

for $z_N \in Z_N$ and consequently that the operator \mathcal{A}_N^{-1} is invertible and the operator \mathcal{A}_N is well defined. It is also self-adjoint. Indeed, for $z_N, y_N \in Z_N$, $\langle \mathcal{A}_N z_N, y_N \rangle = -\langle z_N, y_N \rangle_V$. Also $\langle \mathcal{A}_N z_N, z_N \rangle \leq -\omega |z_N|^2$ for $z_N \in Z_N$ and therefore \mathcal{A}_N is the infinitesimal generator of a semigroup of bounded linear operators on Z_N , $\{\mathcal{T}_N(s): s \geq 0\}$ with $\mathcal{T}_N(s) = \exp(\mathcal{A}_N s)$, $s \geq 0$ and $|\mathcal{T}_N(s)| \leq e^{-\omega s}$, $s \geq 0$.

Elementary properties of spline functions (see [29]) imply $P_N \rightarrow I$ strongly on Z and $\mathcal{P}_N \rightarrow I$ strongly on V as $N \rightarrow \infty$. Furthermore, \mathcal{A}^{-1} compact and the estimate

$$(4.16) \quad |\mathcal{P}_N \mathcal{A}^{-1} z - \mathcal{A}^{-1} z| \leq |\mathcal{P}_N \mathcal{A}^{-1} z - \mathcal{A}^{-1} z|_V = |(\mathcal{P}_N - I) \mathcal{A}^{-1} z|_V$$

imply that $\mathcal{P}_N \mathcal{A}^{-1} \rightarrow \mathcal{A}^{-1}$ as $N \rightarrow \infty$ in the uniform operator topology on $\mathcal{L}(Z)$. We have, therefore, that $|\mathcal{A}_N^{-1} P_N - \mathcal{A}^{-1}|_Z \rightarrow 0$ as $N \rightarrow \infty$ and we conclude (see [4], [15]) that $\mathcal{T}_N(s) P_N z \rightarrow \mathcal{T}(s) z$ and $\mathcal{T}_N^*(s) P_N z \rightarrow \mathcal{T}^*(s) z$ as $N \rightarrow \infty$ for each $z \in Z$, uniformly in s for s in compact subintervals.

With $T_N = \mathcal{T}_N(\tau)$, $Q_N = q_0 P_N$, $G_N = g_0 P_N$ and $B_N = (I - T_N) P_N \psi + \int_0^\tau \mathcal{T}_N(\sigma) P_N a' d\sigma$, Hypotheses 3.1 and 3.2 hold and hence the convergence results for the finite time interval problem given in Theorem 3.3 apply.

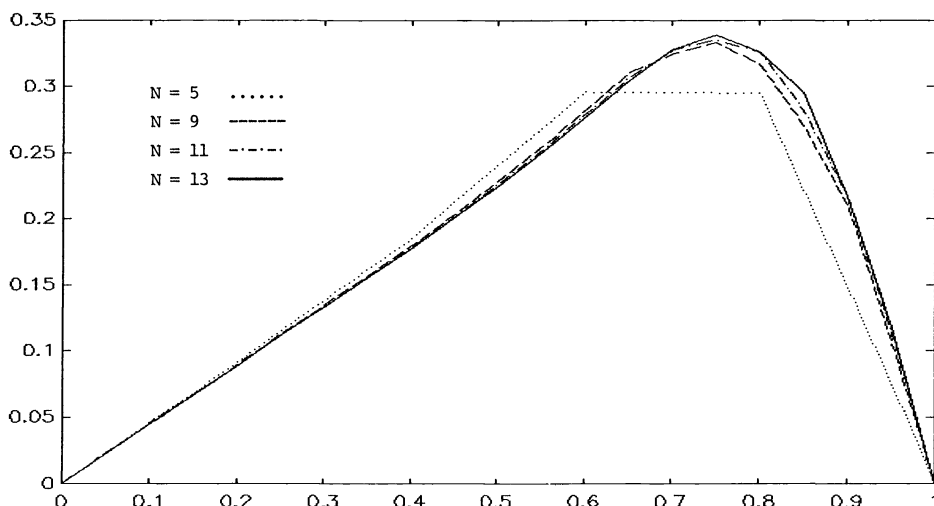


FIG. 4.1

For the infinite interval problem, the uniform coercivity of the $-\mathcal{A}_N$ implies that Hypothesis 3.8 is satisfied. Moreover, if $q_0 > 0$ the conditions given in the statement of Theorem 3.10 are satisfied (see the remark following the proof of Theorem 3.10) and consequently the convergence results for the infinite horizon problem given in Theorem 3.9 hold.

Define the R^{N-1} vector valued function E_N on $[0, 1]$ by $E_N(\theta)^T = (e_N^1(\theta), e_N^2(\theta), \dots, e_N^{N-1}(\theta))$, $\theta \in [0, 1]$ and the $(N-1) \times (N-1)$ Gram matrix $M_N = \langle E_N, E_N^T \rangle$. Let the matrix H_N be given by $H_N = \langle E_N, \mathcal{A}_N E_N^T \rangle = -\langle a D e_N^1, D e_N^1 \rangle$. The matrix representation $[\mathcal{A}_N]$ for the operator \mathcal{A}_N is given by $[\mathcal{A}_N] = M_N^{-1} H_N$, while, for the operators T_N , Q_N and G_N , we have $[T_N] = \exp([\mathcal{A}_N]\tau)$, $[Q_N] = q_0 I_N$ and $[G_N] = g_0 I_N$ where I_N denotes the $(N-1) \times (N-1)$ identity matrix. If we define ψ_{0N} , $a'_N \in R^{N-1}$ by $\psi_{0N} = \langle E_N, \psi_0 \rangle$ and $a'_N = \langle E_N, a' \rangle$ we obtain

$$\begin{aligned} [B_N] &= (I_N - [T_N])M_N^{-1}\psi_{0N} + \int_0^\tau \exp([\mathcal{A}_N]\sigma)M_N^{-1}a'_N d\sigma \\ (4.17) \quad &= (I_N - [T_N])M_N^{-1}\psi_{0N} + [\mathcal{A}_N]^{-1}([T_N] - I_N)M_N^{-1}a'_N. \end{aligned}$$

When the finite-dimensional approximating gain matrices $[F_N(t)]$, $t = 0, 1, 2, \dots, t_f - 1$ for the finite interval problem have been computed an approximation $f_N(t, \cdot)$ to the feedback kernel $f(t, \cdot)$ can be obtained from $f_N(t, \theta) = E_N(\theta)^T M_N^{-1} [F_N(t)]^T$, $t = 0, 1, 2, \dots, t_f - 1$, $\theta \in [0, 1]$. We have $f_N(t, \cdot) \rightarrow f(t, \cdot)$ in $L_2(0, 1)$ as $N \rightarrow \infty$ for each $t = 0, 1, 2, \dots, t_f - 1$. Similarly, for the infinite interval problem, an approximation f_N to f is given by $f_N(\theta) = E_N(\theta)^T M_N^{-1} [F_N]^T$, $\theta \in [0, 1]$ where the matrix $[F_N]$ is computed from (3.30)–(3.32). We have $f_N \rightarrow f$ in $L_2(0, 1)$ as $N \rightarrow \infty$.

We demonstrate the feasibility of our schemes on an infinite interval problem. Taking $q_0 = 1.0$, $r = 1.0$, $a(x) = a = 1.0$, $x \in [0, 1]$ and $\tau = .01$ we obtained the approximating feedback kernels shown in Fig. 4.1.

4.2. A time delay system. In this example we consider linear hereditary control systems of the form

$$(4.18) \quad \dot{x}(s) = A_0 x(s) + A_1 x(s-r) + B_0 v(s), \quad s \geq 0, \quad x(0) = \eta, \quad x_0 = \phi,$$

where $r > 0$, $x(s) \in R^n$, $u(s) \in R^m$, $\eta \in R^n$, $\phi \in L_2(-r, 0; R^n)$ and A_0, A_1 and B_0 are real matrices of appropriate dimensions. For each $s \geq 0$ the function x_s represents the history of the state x on the interval $[s-r, s]$; that is $x_s(\theta) = x(s+\theta)$, $\theta \in [-r, 0]$. The extension of the results outlined below to more general linear hereditary systems (i.e., ones involving multiple discrete delays and integral terms with distributed history kernels) is straightforward. In the state space $Z = R^n \times L_2((-r, 0); R^n)$, the hereditary system (4.18) can be reformulated as an abstract evolution system of the form (4.1) with $z(s) = (x(s), x_s)$, $\mathcal{A}: \text{Dom}(\mathcal{A}) \subset Z \rightarrow Z$ given by $\mathcal{A}(\phi(0), \phi) = (A_0\phi(0) + A_1\phi(-r), D\phi)$ for $(\phi(0), \phi) \in \text{Dom}(\mathcal{A}) = \{(\xi, \psi) \in Z: \psi \in H^1((-r, 0); R^n), \xi = \psi(0)\}$, $\mathcal{B}: R^m \rightarrow Z$ defined by $\mathcal{B}u = (B_0u, 0)$, $u \in R^m$, and $z_0 = (\eta, \phi)$ (see [1]). The discrete time performance index is taken to be

$$(4.19) \quad J(G; 0, t_f, z_0, u) = \sum_{t=0}^{t_f-1} \{ \langle Qz(t), z(t) \rangle + u(t)^T R u(t) \} + \langle Gz(t_f), z(t_f) \rangle$$

where R is a positive-definite symmetric $m \times m$ matrix and the operators $Q: Z \rightarrow Z$ and $G: Z \rightarrow Z$ are given by $Q(\xi, \psi) = (Q_0\xi, 0)$ and $G(\xi, \psi) = (G_0\xi, 0)$ respectively with Q_0 and G_0 nonnegative, symmetric $n \times n$ matrices. For the infinite time problem we of course have $t_f = \infty$ and $G_0 = 0$.

For the finite time problem, the optimal control is given by (2.5)–(2.8). The operator $F(t)$ can be represented by a matrix of operators, $[F^0(t), F^1(t)]$ with $F^0(t) \in \mathcal{L}(R^n, R^m)$ and $F^1(t) \in \mathcal{L}(L_2((-r, 0); R^n); R^m)$. It follows therefore that

$$(4.20) \quad u_*(t) = -f^0(t)x_*(t\tau) - \int_{-r}^0 f^1(t, \theta)(x_*)_{t\tau}(\theta) d\theta, \quad t = 0, 1, 2, \dots, t_f - 1,$$

where $f^0(t)$ is an $m \times n$ matrix, $f^1(t, \cdot)$ is a square integrable $m \times n$ matrix valued function on $(-r, 0)$ and $z_*(t) = (x_*(t\tau), (x_*)_{t\tau})$, $t = 0, 1, 2, \dots, t_f$ is the optimal trajectory.

For the infinite time problem we assume that our original hereditary system and Q_0 are such that there exists an admissible control for each $z(0) = (\eta, \phi) \in Z$ and that Hypothesis 2.4 is satisfied. (This is equivalent to assuming that any unstable modes of the original hereditary system, of which there are at most a finite number, are stabilizable and detectable.) Then Theorems 2.3 and 2.5 imply that there exists a unique nonnegative self-adjoint solution Π to the Riccati algebraic equation (2.10) with the optimal control u_* given by (2.13)–(2.15). The feedback gain F can be represented by a matrix of operators $[F^0, F^1]$ with $F^0 \in \mathcal{L}(R^n, R^m)$ and $F^1 \in \mathcal{L}(L_2((-r, 0); R^n), R^m)$. We have

$$(4.21) \quad u_*(t) = -f^0 x_*(t\tau) - \int_{-r}^0 f^1(\theta)(x_*)_{t\tau}(\theta) d\theta, \quad t = 0, 1, 2, \dots,$$

where f^0 is an $m \times n$ matrix and f^1 is a square integrable $m \times n$ matrix valued function on $(-r, 0)$.

Approximation methods for the solution of the continuous time linear quadratic regulator problem for hereditary systems in closed-loop form have been studied extensively (see, for example, [3], [8], [12], [18], [19], [21], [27], [28]). Recently, a new spline-based state approximation for hereditary systems has been proposed in [15]. This new method appears to exhibit those characteristics which are most desirable in an approximation scheme when it is used to solve linear-quadratic control problems. It also performs at or above the level of any of the approximation schemes for delay systems which have been described in the literature to date. We have chosen this scheme to describe and implement here for the discrete-time problem.

For each $N = 1, 2, \dots$, define $\hat{e}_N^0 = (I_n, 0)$ and $\hat{e}_N^j = (0, e_N^j I_n)$, $j = 1, 2, \dots, N+1$, where I_n denotes the $n \times n$ identity matrix and the e_N^j are the “hat” functions with respect to the uniform mesh $\{-r, -(N-1)r/N, \dots, -r/N, 0\}$ on the interval $[-r, 0]$. Let

$$(4.22) \quad Z_N = \left\{ z \in Z : z = \sum_{j=0}^{N+1} \alpha_j \hat{e}_N^j, \alpha_j \in \mathbb{R}^n \right\}.$$

We shall refer to the collection $\{\hat{e}_N^j\}_{j=0}^{N+1}$ as a “basis” for Z_N and a vector $\alpha \in \times_{j=0}^{N+1} \mathbb{R}^n$ as being a “coordinate vector” for an element in Z_N . Defining $\hat{E}_N^T = (\hat{e}_N^0, \hat{e}_N^1, \dots, \hat{e}_N^{N+1})$ we have $Z_N = \{z \in Z : z = \hat{E}_N^T \alpha, \alpha \in \times_{j=0}^{N+1} \mathbb{R}^n\}$. Let M_N denote the Gram matrix corresponding to the basis $\{\hat{e}_N^j\}_{j=0}^{N+1}$; $M_N = \langle \hat{E}_N, \hat{E}_N^T \rangle$. If P_N denotes the orthogonal projection of Z on to Z_N then $P_N(\xi, \psi) = (\xi, p_N \psi)$ where p_N is the orthogonal projection of $L_2((-r, 0); \mathbb{R}^n)$ on to $\text{span} \{e_N^j\}_{j=1}^{N+1}$.

Noting that $Z_N \not\subset \text{Dom}(\mathcal{A})$, we motivate the definition of \mathcal{A}_N by first formally extending the operator \mathcal{A} to an operator defined on Z_N . For $z_N = (\xi_N, \psi_N) \in Z_N$ define $\mathcal{A}^N z_N = (A_0 \xi_N + A_1 \psi_N(-r), D^+ \psi_N + \delta(\xi_N - \lim_{\theta \rightarrow 0^-} \psi_N(\theta)))$ where δ is the Dirac delta impulse concentrated at zero and $D^+ \psi$ denotes the right-hand derivative of ψ . For each $N = 1, 2, \dots$, let $\mathcal{A}_N : Z_N \rightarrow Z_N$ be given by $\mathcal{A}_N z_N = (A_0 \xi_N + A_1 \psi_N(-r), p_N D^+ \psi_N) + \delta_N(\xi_N - \lim_{\theta \rightarrow 0^-} \psi_N(\theta))$ where $\delta_N = \hat{E}_N^T \gamma_N$ with $\gamma_N = M_N^{-1} \text{col}(0, \lim_{\theta \rightarrow 0^-} e_N^1(\theta), \dots, \lim_{\theta \rightarrow 0^-} e_N^{N+1}(\theta))$.

We define the operators $\mathcal{B}_N : \mathbb{R}^m \rightarrow Z_N$, $Q_N : Z_N \rightarrow Z_N$ and $G_N : Z_N \rightarrow Z_N$ by $\mathcal{B}_N = P_N \mathcal{B}$, $Q_N = P_N Q$ and $G_N = P_N G$.

Once the matrix representations for the approximating feedback gains have been computed, $[F_N(t)]$, $t = 0, 1, 2, \dots, t_f - 1$ from (3.25)–(3.28) for the finite time interval problem and $[F_N]$ from (3.30)–(3.32) (assuming, for the moment, that solutions to (3.17) exist) for the infinite interval problem, approximations for $f^0, f^1(t, \cdot)$, $t = 0, 1, 2, \dots, t_f - 1$ and $f^0, f^1(\cdot)$ can be computed from $((f_N^0(t))^T, (f_N^1(t, \cdot))^T)^T = \hat{E}_N^T M_N^{-1} [F_N(t)]^T$, $t = 0, 1, 2, \dots, t_f - 1$, and $((f_N^0)^T, (f_N^1(\cdot))^T)^T = \hat{E}_N^T M_N^{-1} [F_N]^T$ respectively.

For the approximation scheme defined above, it is shown in [15] that $P_N \rightarrow I$ strongly on Z and that $\exp(\mathcal{A}_N s) P_N \rightarrow \mathcal{T}(s)$ and $\exp(\mathcal{A}_N^* s) P_N \rightarrow \mathcal{T}^*(s)$ strongly on Z and uniformly in s for s in compact intervals. Hypothesis 3.1 is a simple consequence of these results. The present scheme, therefore, satisfies all of the hypotheses of Theorem 3.3 and we may conclude that the convergence results for the finite time interval problem given in the statement of the theorem hold. In particular, we have $f_N^0(t) \rightarrow f^0(t)$ in $\mathbb{R}^{m \times n}$ and $f_N^1(t, \cdot) \rightarrow f^1(t, \cdot)$ in $L_2((-r, 0); \mathbb{R}^{m \times n})$ for each $t = 0, 1, 2, \dots, t_f - 1$.

With the operators Q and G and the operators Q_N and G_N as they have been defined above it is clear that the hypotheses given in the statement of Theorem 3.7 are satisfied. We have therefore that for the present example the operators $\{\Pi(t)\}_{t=0}^{t_f}$ are trace class and that $\lim_{N \rightarrow \infty} \|\Pi_N(t) P_N - \Pi(t)\|_1 = 0$, $t = 0, 1, 2, \dots, t_f$.

For the infinite time problem and the approximation scheme discussed here, we believe the situation with regard to Hypothesis 3.8 and convergence is much the same as it is for the continuous time problem (see [15], [16]). It is shown in [16] that for the present scheme, in the continuous time setting, the existence of a unique nonnegative self-adjoint solution to the infinite dimensional algebraic Riccati equation (i.e., the stabilizability and detectability of the underlying hereditary system) implies the existence of unique, nonnegative self-adjoint solutions to the approximating finite dimensional algebraic Riccati equations for all N sufficiently large. We expect that the situation is much the same but rigorous verification of this would be very tedious and probably unenlightening—and outside the scope of this paper. In general we are

unable to demonstrate the existence of an M and an $r < 1$ for which (3.19) holds nor do we expect to be able to. Indeed, our numerical studies yield sequences of closed-loop eigenvalues of the approximating discrete-time control problems $(\mathcal{P}2_N)$ which tend toward the unit circle as $N \rightarrow \infty$. This agrees with what is observed in the continuous time case; see [16]. Of course Theorem 3.10 provides only sufficient conditions for strong convergence of Π_N to Π . The existence of the uniform bounds (3.18) and (3.19) and the uniform coercivity of Q_N notwithstanding, our numerical studies appear to indicate that we do in fact have strong convergence. We note that in [16], for the continuous time problem, Kappel and Salamon show that the present approximation scheme does satisfy a somewhat weaker uniform stability condition—what they call uniform output stability—and indicate that this is sufficient to conclude strong convergence of the Riccati operators for the infinite time problem. Once again we suspect that an analogous result holds in the discrete-time case.

Upon solving the approximating problems it can often be observed that $|\Pi_N|$ is uniformly bounded in N . Consequently Theorem 3.12 may be applied to conclude that a solution Π to (2.10) exists, $\Pi_N P_N \rightarrow \Pi$ weakly and $F_N P_N \rightarrow F$ strongly as $N \rightarrow \infty$. It would then follow that $f_N^0 \rightarrow f^0$ in $R^{m \times n}$ and $f_N^1 \rightarrow f^1$ weakly in $L_2((-r, 0); R^{m \times n})$ as $N \rightarrow \infty$. When Π_N does converge to Π strongly, $f_N^1 \rightarrow f^1$ strongly in $L_2(-r, 0); R^{m \times n}$.

We applied the scheme to an infinite-time problem with $n = 2$, $r = 1$, $m = 1$,

$$A_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}, \quad B_0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad Q_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

This system is the first order form of a second order linear harmonic oscillator with delayed restoring force. It is not difficult to show that the stabilizability and detectability conditions which guarantee the existence of a solution to the infinite-dimensional algebraic Riccati equation are satisfied for this system. The nature of this system and the approximation scheme is such that we must have $[f^1(\theta)]_2 = [f_N^1(\theta)]_2 = 0$, $-1 \leq \theta \leq 0$, $N \geq 1$, where $[v]_2$ denotes the second component of the vector $v \in R^2$ (see [12]).

Setting $\tau = .01$ we obtained the results shown in Table 4.2 and in Fig. 4.3 below with $R = .05$. With $R = 1.0$, the results shown in Table 4.4 and Fig. 4.5 were obtained. As the cost of control increases the effect that the optimal control for the infinite-dimensional problem has on higher modes decreases. Consequently, the finite-dimensional approximations converge more rapidly.

4.3. Control of a flexible structure. We consider an Euler-Bernoulli beam cantilevered to a rigid hub which is free to rotate about its fixed center, point 0. Also, a point mass m_1 is attached to the other end of the beam. The control is a torque u applied to the hub, and all motion is in the plane. See Fig. 4.6 and Table 4.7.

The angle θ represents the rotation of the hub (the rigid-body mode), $w(t, \eta)$ is the elastic deflection of the beam from the rigid-body position, and $w_1(t)$ is the displacement of m_1 from the rigid-body position. For technical reasons, we do not yet impose the condition $w_1(t) = w(t, l)$.

TABLE 4.2

N	2	4	8	10
$[f_N^0]_1$	4.5483	4.5452	4.5451	4.5451
$[f_N^0]_2$	5.2954	5.2948	5.2948	5.2948

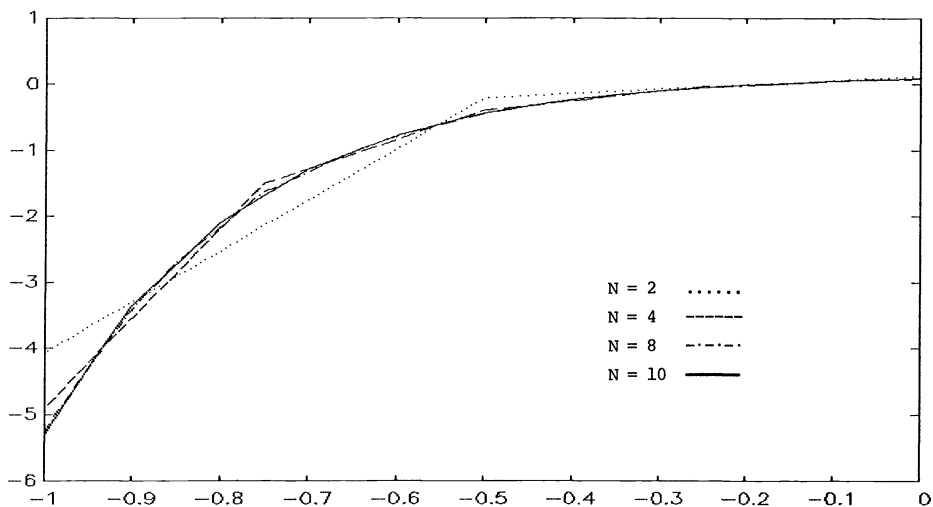


FIG. 4.3

TABLE 4.4

N	2	4	8	10
$[f_N^0]_1$	1.4050	1.4054	1.4054	1.4054
$[f_N^0]_2$	1.9477	1.9479	1.9479	1.9479

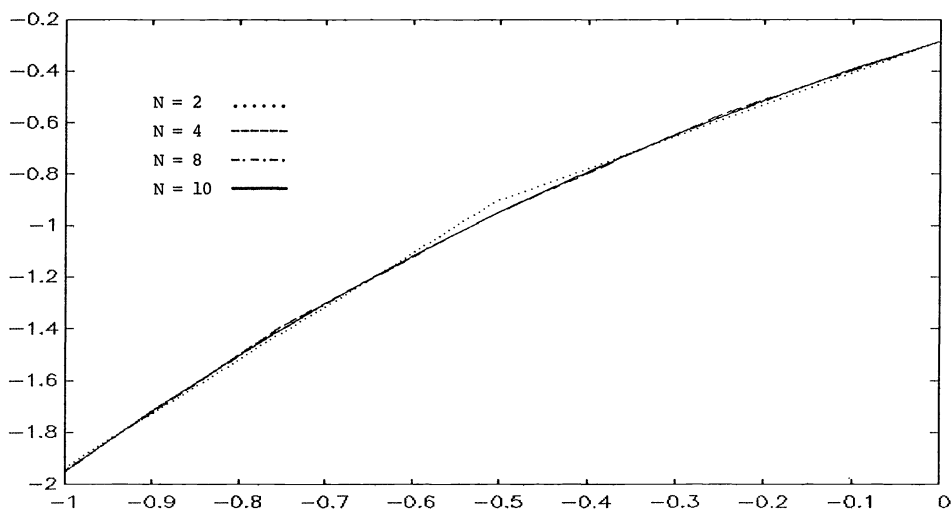


FIG. 4.5

The control problem is to stabilize rigid-body motions and linear (small) transverse elastic vibrations about the state $\theta = 0$ and $w = 0$. Our linear model assumes not only that the elastic deflection of the beam is linear but also that the axial inertial force produced by the rigid-body angular velocity has negligible effect on the bending stiffness of the beam. The rigid-body angle need not be small.

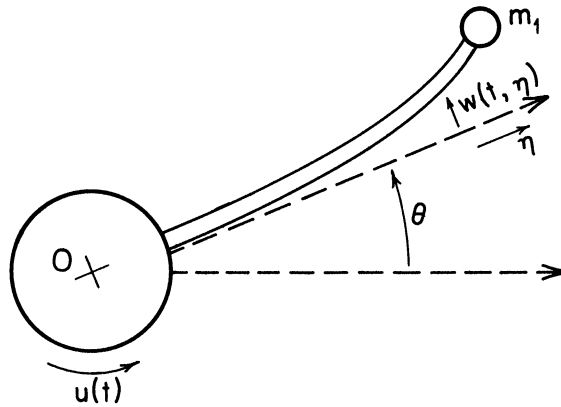


FIG. 4.6

TABLE 4.7

r = hub radius	10 in
l = beam length	100 in
I_0 = hub moment of inertia about axis perpendicular to page through O	100 slug in ²
m_b = beam mass per unit length	.01 slug/in
m_1 = tip mass	1 slug
EI = product of elastic modulus and second moment of cross section for beam	13,333 slg in ³ /sec ²
fundamental frequency of undamped structure	.9672 rad/sec

For this example, it is a straightforward exercise to derive the coupled ordinary and partial differential equations of motion in θ , w and w_1 . However, rather than writing these equations explicitly, it is easier and more useful for our purposes to derive an abstract second order evolution equation for the structure. To do this, we define the generalized displacement vector $x = (\theta, w, w_1) \in H = R^1 \times L_2(0, l) \times R^1$. The kinetic energy in the system is then $\frac{1}{2} \langle M_0 \dot{x}, \dot{x} \rangle_H$ where M_0 is the unique bounded self-adjoint linear mass operator M_0 on H such that

$$(4.23) \quad \langle M_0 x, \hat{x} \rangle_H = I_0 \theta \hat{\theta} + m_b \langle w + \psi_0 \theta, \hat{w} + \psi_0 \hat{\theta} \rangle_{L_2} + m_1 (w_1 + \psi_0(l) \theta) (\hat{w}_1 + \psi_0(l) \hat{\theta}),$$

where $\psi_0 \in L_2(0, l)$ is given by $\psi_0(\eta) = r + \eta$. It is easy to show that M_0 is also coercive. The elastic strain energy is $\frac{1}{2} a(x, x)$ with $a(x, \hat{x}) = EI \langle D^2 w, D^2 \hat{w} \rangle_{L_2}$. We make $a(\cdot, \cdot)$ into an inner product by setting $\langle x, \hat{x} \rangle_V = a(x, \hat{x}) + \theta \hat{\theta}$ and define the strain-energy space $V = \{x = (\theta, \phi, \phi(l)) : \phi \in H^2(0, l), \phi(0) = D\phi(0) = 0\}$. The additional term in the definition of $\langle \cdot, \cdot \rangle_V$ is necessary because there is no strain energy associated with the rotation of the hub.

We define the stiffness operator $A_0: \text{Dom}(A_0) \subset H \rightarrow H$ by $\text{Dom}(A_0) = \{x = (\theta, \phi, \phi(l)) \in V : \phi \in H^4(0, l), D^2 \phi(l) = 0\}$ and

$$(4.24) \quad A_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & EI D^4 & 0 \\ 0 & -EI D^3 & 0 \end{bmatrix}.$$

This operator is self-adjoint with compact resolvent and all positive eigenvalues except the one zero eigenvalue corresponding to the rigid-body mode. Note that V is the domain of the square root of A_0 .

With these mass and stiffness operators, we can write the equations of motion as

$$(4.25) \quad M_0 \ddot{x}(s) + c_0 A_0 \dot{x}(s) + A_0 x(s) = B_0 u(s), \quad s \geq 0,$$

where c_0 is a positive constant and the term $c_0 A_0 \dot{x}$ represents viscoelastic damping in the beam. The input operator is $B_0 = (1, 0, 0)$.

Letting $Z = V \times H$ with inner product $\langle (v, h), (\hat{v}, \hat{h}) \rangle = \langle v, \hat{v} \rangle_V + \langle M_0 h, \hat{h} \rangle_H$, this system can be written in the form (4.1) where $z = (x, \dot{x}) \in Z$ and \mathcal{A} is the unique extension of the operator

$$(4.26) \quad \mathcal{A} = \begin{bmatrix} 0 & I \\ -M_0^{-1} A_0 & -c_0 M_0^{-1} A_0 \end{bmatrix}, \quad \text{Dom}(\mathcal{A}) = \text{Dom}(A_0) \times \text{Dom}(A_0),$$

that generates a \mathcal{C}_0 -semigroup on the space Z . Of course, \mathcal{B} is

$$(4.27) \quad \mathcal{B} = \begin{bmatrix} 0 \\ M_0^{-1} B_0 \end{bmatrix}.$$

(See [11] and [13].) The hub-beam-tip mass structure here is discussed in more detail in [13], along with the continuous-time control problem.

As in the previous examples, we will solve a discrete-time optimal control problem on the infinite interval. In the performance index, we take the state weighting operator Q to be the identity on Z . This means that $\langle Qz, z \rangle_Z$ is twice the total energy in the structure plus the square of the rigid-body rotation. Since there is one input, the control weighting R is a scalar. The optimal control has the feedback form

$$(4.28) \quad u_*(t) = -\langle f, x(t) \rangle_V - \langle M_0 g, \dot{x}(t) \rangle_H$$

where f and g have the form $f = (f^1, f^2, f^3) \in V$ and $g = (g^1, g^2, g^3) \in H$.

To see that there exists a solution to the infinite-dimensional discrete-time regulator problem for this example, we note that all open-loop modes except the rigid body mode are exponentially stable and that any control which stabilizes the controllable rigid body mode is admissible. Existence of a solution to the infinite-dimensional Riccati equation then follows from Theorem 2.3. Uniqueness follows from Theorem 2.5, since Hypothesis 2.4 holds because $Q = I$ is coercive.

Our approximation of the structure is based on a finite element approximation of the beam which uses Hermite cubic splines as basis functions ([29]). We define the sequence of spaces $V_N = \text{span} \{e_N^j\}_{j=1}^{J_N}$ with $e_N^1 = (1, 0, 0)$, $e_N^j = (0, \phi_N^j, \phi_N^j(l))$, $j = 2, 3, \dots, J_N$, where the ϕ_N^j 's are the cubic splines. Each V_N is a subspace of V , and our approximation scheme is a Ritz-Galerkin approximation obtained by projecting (4.25) onto V_N . See [13] for details. Writing

$$(4.29) \quad x_N(s) = \sum_{j=1}^{J_N} [x_N(s)]_j e_N^j,$$

we have

$$(4.30) \quad M_N[\ddot{x}_N(s)] + c_0 K_N[\dot{x}_N(s)] + K_N[x_N(s)] = B_{0N}u(s)$$

to solve for the vector $[x_N(s)]$ of time-dependent coefficients $[x_N(s)]_j$. The mass matrix M_N and the stiffness matrix K_N are given by $[M_N]_{ij} = \langle M_0 e_N^i, e_N^j \rangle_H$, and $[K_N]_{ij} = \langle e_N^i, e_N^j \rangle_V$ and the input matrix is $B_{0N} = [1 \ 0 \ 0 \ \dots \ 0]^T$. With $z_N = (x_N, \dot{x}_N) \in V_N \times V_N$, (4.30) is the matrix representation of an evolution equation

$$(4.31) \quad \dot{z}_N(s) = \mathcal{A}_N z_N(s) + \mathcal{B}_N u(s)$$

where \mathcal{A}_N and \mathcal{B}_N approximate \mathcal{A} and \mathcal{B} . It is shown in [13] that, as N increases, the semigroup $\{\mathcal{T}_N(s): s \geq 0\}$ generated by \mathcal{A}_N converges strongly to the semigroup $\{\mathcal{T}(s): s \geq 0\}$ and that the adjoint semigroup $\{\mathcal{T}_N^*(s): s \geq 0\}$ converges strongly as well. Since \mathcal{B}_N is the Z -projection of \mathcal{B} onto $V_N \times V_N$, it converges strongly to \mathcal{B} .

For each N , the solution to the infinite-time optimal control problem is based on the N th Riccati operator equation (3.17). As in the previous examples, we solve the Riccati matrix equation (3.32) for $\hat{\Pi}_N$, which is related to $[\Pi_N]$ (the matrix representation of the operator Π_N) as in § 3.3, except here we have $\hat{\Pi}_N = W_N[\Pi_N]$, where

(4.32)
$$W_N = \begin{bmatrix} \tilde{K}_N & 0 \\ 0 & M_N \end{bmatrix}$$

and \tilde{K}_N is the stiffness matrix with 1 added to the first element. Since $Q = I$ in the infinite-dimensional problem, Q_N is the identity on $V_N \times V_N$ and it follows that the matrix \hat{Q}_N for (3.32) is W_N .

The optimal feedback control for the N th problem is given by (3.29) with the matrices in (3.30) and (3.31), and it has the equivalent representation

(4.33)
$$u_*^N(t) = -\langle f_N, x_N(t) \rangle_V - \langle M_0 g_N, \dot{x}_N(t) \rangle_H,$$

where $f_N = (f_N^1, f_N^2, f_N^3) \in V$, $g_N = (g_N^1, g_N^2, g_N^3) \in H$. From (3.29), (4.29) and (4.33), it follows that

(4.34)
$$\begin{bmatrix} f_N \\ g_N \end{bmatrix} = \begin{bmatrix} E_N^T & 0 \\ 0 & E_N^T \end{bmatrix} W_N^{-1} [F_N]^T,$$

where $E_N^T = (e_N^1, e_N^2, \dots, e_N^{J_N})$.

To see that Hypothesis 3.8 holds, we note that for each N , there are $N - 1$ exponentially stable modes and one controllable rigid body mode. If there were no rigid body mode, we would have a uniform positive lower bound on the stiffness matrices K_N and this would imply a uniform decay rate for the open-loop systems which would yield the uniform bounds (3.18) and (3.19). With the rigid body mode, the argument becomes more tedious but essentially similar. For more detail, see the continuous time case in [13].

For the sampling interval $\tau = .01$, the damping coefficient $c_0 = .001$ and the control weighting $R = 1$, Tables 4.8 and 4.9 give the values of the corresponding scalar gains, $f_N^i, g_N^i, i = 1$ and 3, for various values of N . The values of the functional gains $D^2 f_N^2$ and g_N^2 along the length of the beam also are plotted in Figs. 4.10 and 4.11. We plotted

TABLE 4.8

N	3	4	5	7
f_N^1	.9991	.9992	.9990	.9992
g_N^1	.1030	.1040	.1043	.1044

TABLE 4.9

N	3	4	5	7
f_N^3	.1750	.1769	.1774	.1777
g_N^3	-18.1231	-18.3385	-18.3902	-18.4158

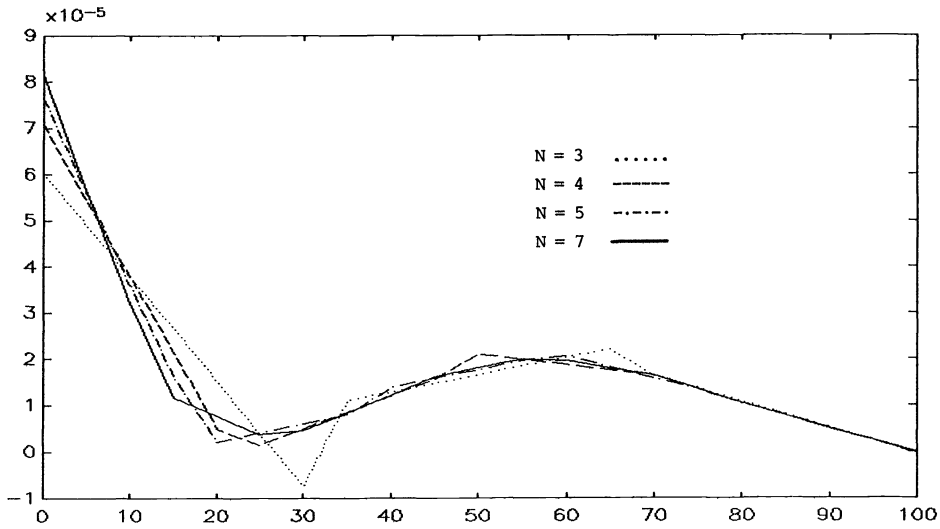


FIG. 4.10

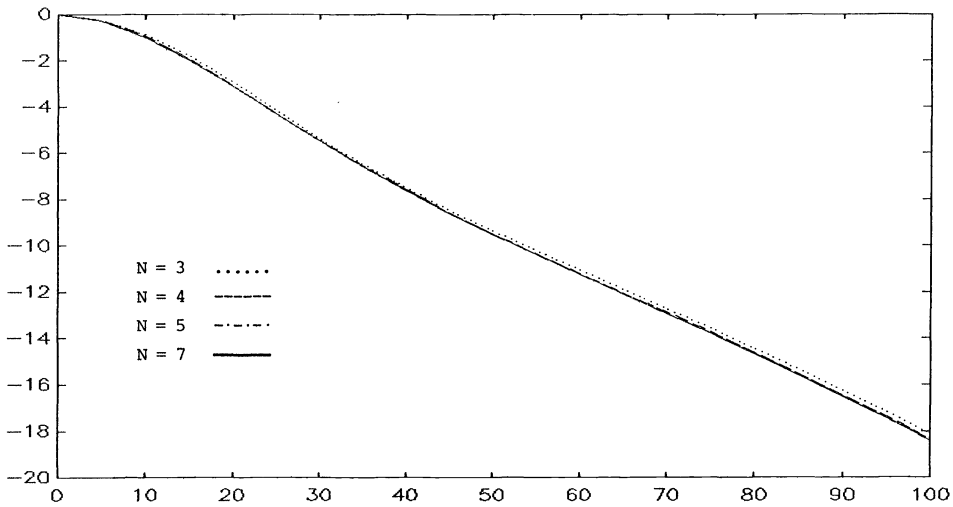


FIG. 4.11

$D^2 f_N^2$ because this is what appears in the V inner product in (4.33) and also to show the H^2 convergence. We note that the form of the V inner product is such that f_N^3 does not appear in the feedback law.

5. Concluding remarks. We have presented an approximation theory for numerical solution of the discrete-time optimal linear regulator problem in Hilbert space, on both finite and infinite time intervals. The motivation for this theory comes from optimal control problems for systems involving diffusion equations, hereditary differential equations and distributed models of flexible structures. We have demonstrated the application of the theory to examples from all three areas.

The solution to the infinite-dimensional optimal control problem is based on an infinite-dimensional Riccati operator equation—a difference equation in the finite-time problem and an algebraic equation in the infinite-time problem. We have shown that

the solution to the infinite-dimensional problem can be approximated by the solutions to a sequence of finite-dimensional problems each of which involves a finite-dimensional Riccati matrix equation to be solved numerically. The finite-dimensional problems are just the corresponding optimal control problems for finite element approximations to the infinite-dimensional control system. For the infinite-time problem, the finite-dimensional Riccati equations usually are solved via eigenspace decomposition of the Hamiltonian matrix.

In both continuous and discrete-time optimal regulator problems for distributed systems, the numerical solution often involves solution of large Riccati matrix equations. As we observed at the beginning of § 4, the asymptotic relationship between the eigenvalues of a continuous-time Hamiltonian system and the eigenvalues of the corresponding discrete-time Hamiltonian system is exponential. This means that the approximating finite-dimensional discrete-time Riccati equations for a given distributed system invariably are not as well conditioned as the corresponding continuous-time Riccati equations. Nonetheless, as our examples should illustrate, the numerical solution of such problems is well within the reach of current computing. To emphasize this, we obtained all of the numerical results in this paper on an IBM Personal Computer (not an XT or AT) with 640K of random access memory and an Intel 8087 math co-processor chip. The largest Riccati matrix equation that we solved here was a 30×30 steady state equation for the hub-beam-tip mass example. This solution takes 15–20 minutes on the PC. We have solved much larger Riccati equations easily on larger mainframe computers.

REFERENCES

- [1] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: Numerical methods based on averaging approximations*, this Journal, 16 (1978), pp. 169–208.
- [2] H. T. BANKS AND K. KUNISCH, *The linear regulator problem for parabolic systems*, this Journal, 22 (1984), pp. 684–698.
- [3] H. T. BANKS, I. G. ROSEN AND K. ITO, *A spline based technique for computing Riccati operators and feedback controls in regulator problems for delay equations*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 830–855.
- [4] A. BELLENI-MORANTE, *Applied Semigroups and Evolution Equations*, Clarendon Press, Oxford, 1979.
- [5] R. F. CURTAIN, *The infinite-dimensional Riccati equations with application to affine hereditary differential systems*, this Journal, 13 (1975), pp. 1130–1143.
- [6] 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.
- [7] R. F. CURTAIN AND D. SALAMON, *Finite-dimensional compensators for infinite-dimensional systems with unbounded input operators*, this Journal, 24 (1986), pp. 797–816.
- [8] M. C. DELFOUR, *The linear quadratic optimal control problem for hereditary differential systems: Theory and numerical solution*, Appl. Math. Optim., 3 (1977), pp. 101–162.
- [9] M. C. DELFOUR, C. MCCALLA AND S. K. MITTER, *Stability and the infinite-time quadratic cost problem for linear hereditary differential systems*, this Journal, 13 (1975), pp. 48–88.
- [10] J. S. GIBSON, *The Riccati integral equations for optimal control problems in Hilbert spaces*, this Journal, 17 (1979), pp. 537–565.
- [11] ———, *An analysis of optimal modal regulation: convergence and stability*, this Journal, 19 (1981), pp. 686–707.
- [12] ———, *Linear quadratic optimal control of hereditary differential systems: infinite-dimensional Riccati equations and numerical approximations*, this Journal, 21 (1983), pp. 95–139.
- [13] J. S. GIBSON AND A. ADAMIAN, *Approximation theory for optimal LQG control of flexible structures*, Report, Department of Mechanical, Aerospace and Nuclear Engineering, University of California—Los Angeles, Los Angeles, CA, 1986.
- [14] W. GREEN AND T. D. MORLEY, *Operator means, fixed points and the norm convergence of monotone approximates*, Math. Scand., to appear.

- [15] F. KAPPEL AND D. SALAMON, *Spline approximation for retarded systems and the Riccati equation*, this Journal, 24 (1986), pp. 1082-1117.
- [16] ———, *On the stability properties of spline approximations for retarded systems*, this Journal, submitted.
- [17] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
- [18] N. N. KRAVSOVSKIĬ, *The approximation of a problem of analytic design of controls in a system with time-lag*, J. Appl. Math. Mech., 28 (1964), pp. 876-885.
- [19] K. KUNISCH, *Approximation schemes for the linear quadratic optimal control problem associated with delay equations*, this Journal, 20 (1982), pp. 506-540.
- [20] A. J. LAUB, *A Shur method for solving algebraic Riccati equations*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 914-921.
- [21] E. B. LEE AND A. MANITIUS, *Computational approaches to synthesis of feedback controllers for multivariable systems with delays*, Proc. IEEE Conference on Decision and Control, November 20-22, 1974, Phoenix, AZ, pp. 791-792.
- [22] K. Y. LEE, S. CHOW AND R. O. BARR, *On the control of discrete-time distributed parameter systems*, this Journal, 10 (1972), pp. 361-376.
- [23] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
- [24] D. L. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101-121.
- [25] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
- [26] T. PAPPAS, A. J. LAUB AND N. R. SANDELL, JR., *On the numerical solution of the discrete-time algebraic Riccati equation*, IEEE Trans. Automat. Control., AC-25 (1980), pp. 631-641.
- [27] D. W. ROSS, *Controller design for time lag systems via a quadratic criterion*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 644-672.
- [28] D. W. ROSS AND I. FLUGGE-LOTZ, *An optimal control problem for systems with differential-difference equation dynamics*, this Journal, 7 (1969), pp. 609-623.
- [29] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [30] J. ZABCZYK, *Remarks on the control of discrete-time distributed parameter systems*, this Journal, 12 (1974), pp. 721-735.

OPTIMAL PERIODIC CONTROL: A SCENARIO FOR LOCAL PROPERNESS*

FRITZ COLONIUS†

Abstract. A fundamental problem in optimal periodic control is to decide whether proper periodic controls and trajectories yield better average performance than constant steady-state solutions. The present paper describes a situation where this holds true, because “nearby” the linearized system equation has a pair of eigenvalues on the imaginary axis. An example involving a retarded Liénard equation is discussed in detail.

Key words. optimal periodic control, local properness, functional differential equations, Hopf bifurcation

AMS(MOS) subject classifications. 49B34, 49D50

1. Introduction. In optimal periodic control theory, one looks for periodic controls and corresponding periodic trajectories of a control system described, for example, by a functional differential equation such that a certain average performance criterion is minimized. Suppose that a constant control u^0 and a corresponding steady state x^0 of the system are given, which are optimal among all such pairs (x, u) . If it is possible to obtain better average performance in every neighborhood of (x^0, u^0) by allowing proper periodic controls and corresponding periodic trajectories x , then the pair (x^0, u^0) is called locally proper. It is the purpose of the present paper to explore a situation where one may expect local properness because of structural properties of the system equation. In particular these properties are related to those of a Hopf bifurcation. The guiding idea is that local properness will occur, if the considered system has “nearby” a “natural” periodic motion giving better performance.

A connection between Hopf bifurcation and optimal periodic control theory has already been observed by Russell [20]. He was interested in coupled nonlinear oscillators, where a Hopf bifurcation causes periodic motions which he wanted to dampen. Since this was not possible by linear regulator theory, he considered this problem as an optimal periodic control problem where the performance criterion is constructed in such a way as to minimize the amplitude of the oscillations.

Observe, however, that the spirit of the present paper is quite different: Instead of trying to dampen periodic motions we are willing to introduce them in order to get better performance. This is motivated by problems from chemical engineering (output maximization of chemical reactors [19], [24], [25]) and aircraft flight performance optimization (fuel optimal flight [22], [23]). Further references are given in [8], [17], [18].

In § 2 the optimal periodic control problem is formally defined for systems described by retarded functional differential equations. Furthermore, among other preliminaries, relevant information on necessary optimality conditions is cited from [8].

Section 3 exhibits a scenario for local properness. Theorem 3.6 contains the main result of this paper. In conclusion, § 4 discusses an example which, in fact, was the

* Received by the editors February 5, 1985; accepted for publication (in revised form) May 6, 1987. This work was partially supported by a grant from Deutsche Forschungsgemeinschaft and by the U.S. Air Force Office of Scientific Research under grant AFOSR-84-0398. This research was performed during a visit to the Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

† “Institut für Dynamische Systeme,” Universität Bremen, FB 3, D-2800 Bremen 33, Federal Republic of Germany.

starting point of the present analysis. The importance of the result in § 3 is twofold: (i) It explains a mechanism by which local properness may occur and thus gives some insight into this phenomenon. (ii) It gives a hint, where to look for local properness, namely near equilibria, where the linearized system equation has a pair of eigenvalues on the imaginary axis.

It is worthwhile emphasizing that these results, which may be viewed as a contribution to the qualitative theory of optimal control, are also new for the special case of systems governed by ordinary differential equations.

Notation. The transpose of an element $x \in \mathbb{R}^n$ is denoted by x^T ; similarly for matrices. For a map F between Banach spaces X and Y , $\mathcal{D}F(x^0)$ denotes its Fréchet-derivative at $x^0 \in X$. For maps between finite dimensional space we also use a subscript x in order to denote the partial derivative with respect to x . The second Fréchet-derivative at $x^0 \in X$ is denoted by $\mathcal{D}\mathcal{D}F(x^0)$. For an element $x \in \mathbb{R}^n$, \bar{x} denotes the constant function $\bar{x}(s) \equiv x$ (in various function spaces).

2. Problem formulation and optimality conditions. In this section, a parameter dependent optimal periodic control problem (OPC) $^\alpha$ and the corresponding optimal steady-state problem (OSS) $^\alpha$ are formulated. Furthermore optimality conditions and results on smooth dependence of optimal solutions are cited, slightly modified for our purposes, from [8].

Consider the following optimal periodic control problem.

$$\begin{aligned} \text{(OPC)}^\alpha \quad & \text{Minimize } 1/\tau \int_0^\tau g(x(s), u(s)) \, ds \\ & \text{over } (x, u) \in C(-r, \tau; \mathbb{R}^n) \times L^\infty(0, \tau; \mathbb{R}^m) \\ & \text{subject to} \end{aligned}$$

$$(2.1) \quad \dot{x}(t) = f(x_t, u(t), \alpha) \quad \text{a.e. } t \in [0, \tau],$$

$$(2.2) \quad x_0 = x_\tau,$$

$$(2.3) \quad h(u(t)) \in \mathbb{R}_-^1 \quad \text{a.e. } t \in [0, \tau],$$

$$(2.4) \quad \int_0^\tau k(x(t), u(t)) \, dt = 0;$$

here $x_t(s) = x(t+s) \in \mathbb{R}^n$, $s \in [-r, 0]$, $r > 0$ is the length of the delay, $\alpha \in A$ is a parameter, $A \subset \mathbb{R}$ open, $f = (f^i): C(-r, 0; \mathbb{R}^n) \times \mathbb{R}^m \times A \rightarrow \mathbb{R}^n$, $g: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}$, $h = (h^i): \mathbb{R}^m \rightarrow \mathbb{R}^l$, $k = (k^i): \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^{n_1}$.

The period length $\tau > 0$ is considered fixed here (we also allow $\tau < r$). The requirement (2.2) is imposed in order to allow periodic extensions of x and u to periodic solutions of (2.1) on $\mathbb{R}_+ := [0, \infty)$.

Abbreviate

$$\begin{aligned} \Omega &:= \{u \in \mathbb{R}^m: h(u) \in \mathbb{R}_-^l\}, \\ (2.5) \quad \mathcal{U}_{\text{ad}} &:= \{u \in L^\infty(0, \tau; \mathbb{R}^m): u(t) \in \Omega \text{ a.e. } t \in [0, \tau]\}. \end{aligned}$$

The corresponding steady-state problem has the following form:

$$\begin{aligned} \text{(OSS)}^\alpha \quad & \text{Minimize } g(x, u) \text{ over } (x, u) \in \mathbb{R}^n \times \mathbb{R}^m \\ & \text{subject to} \end{aligned}$$

$$(2.6) \quad 0 = f(\bar{x}, u, \alpha),$$

$$(2.7) \quad h(u) \in \mathbb{R}_-^l,$$

$$(2.8) \quad 0 = k(x, u);$$

here f , g , h and k are as in $(\text{OPC})^\alpha$.

We are interested in the behavior of $(\text{OPC})^\alpha$ near an optimal solution $(x^0, u^0) \in \mathbb{R}^n \times \mathbb{R}^m$ of $(\text{OSS})^{\alpha_0}$ (i.e., near the constant pair $(\bar{x}^0, \bar{u}^0) \in C(-r, \tau; \mathbb{R}^n) \times L^\infty(0, \tau; \mathbb{R}^m)$).

DEFINITION. A local solution (x^α, u^α) of problem $(\text{OSS})^\alpha$ is called locally proper, if for all $\varepsilon > 0$ there exist $(x, u) \in C(-r, \tau; \mathbb{R}^n) \times L^\infty(0, \tau; \mathbb{R}^m)$ with $\sup_{t \in [0, \tau]} |x^0 - x(t)| < \varepsilon$ satisfying (2.1)–(2.4) and

$$1/\tau \int_0^\tau g(x(t), u(t)) dt < g(x^\alpha, u^\alpha).$$

As is well known, first order necessary optimality conditions (based on weak variations) do not allow one to decide the question of local properness. Hence we will give below second order necessary optimality conditions for $(\text{OPC})^\alpha$.

Let the Pontryagin function $H: C(-r, 0; \mathbb{R}^n) \times \mathbb{R}^m \times \mathbb{R}^{n+n_1} \times A \rightarrow \mathbb{R}$ for $(\text{OPC})^\alpha$ be

$$(2.9) \quad H(\varphi, u, y, \alpha) := g(\varphi(0), u) + y^T \begin{pmatrix} f(\varphi, u, \alpha) \\ k(\varphi(0), u) \end{pmatrix}$$

and let the Lagrange function $\mathcal{L}: \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{n+n_1} \times \mathbb{R}^l \times A \rightarrow \mathbb{R}$ for $(\text{OSS})^\alpha$ be

$$(2.10) \quad \mathcal{L}(x, u, y, z, \alpha) := g(x, u) + y^T \begin{pmatrix} f(\bar{x}, u, \alpha) \\ k(x, u) \end{pmatrix} + z^T h(u).$$

The following hypotheses will be used.

Hypothesis 2.1. The functions f , g , h and k are twice continuously Fréchet differentiable in a neighborhood of $(\bar{x}^0, u^0, \alpha_0)$ (respectively, (x^0, u^0) , u^0 , (x^0, u^0)); the function f and its first and second derivatives are bounded for bounded arguments; the set Ω is convex.

Hypothesis 2.2. There exist $(y^0, z^0) \in \mathbb{R}^{n+n_1} \times \mathbb{R}^l$ such that

$$(2.11) \quad z^{0T} h(u^0) = 0,$$

$$(2.12) \quad \mathcal{D}_{1,2} \mathcal{L}(x^0, u^0, y^0, z^0, \alpha_0) = 0,$$

$$(2.13) \quad \mathcal{D}_{1,2} \mathcal{D}_{1,2} \mathcal{L}(x^0, u^0, y^0, z^0, \alpha_0)((x, u), (x, u)) > 0$$

for all $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$ with $\mathcal{D}_{1,2} f(\bar{x}^0, u^0, \alpha_0)(\bar{x}, u) = 0$, $k_{x,u}(x^0, u^0)(x, u) = 0$, $h_u^i(u^0)u < 0$ if $h^i(u^0) = 0$, $i \in \{1, \dots, l\}$.

Hypothesis 2.3. The gradients in \mathbb{R}^{n+m}

$$(2.14) \quad \begin{aligned} & \mathcal{D}_{1,2} f^i(\bar{x}^0, u^0, \alpha_0), \quad i = 1, \dots, n, \\ & (0, h_u^i(u^0)) \quad \text{with } h^i(u^0) = 0, \quad i \in \{1, \dots, l\}, \\ & k_{x,u}^i(x^0, u^0), \quad i = 1, \dots, n_1 \end{aligned}$$

are linearly independent and the multiplier $z^0 = (z^{0,i})$ from Hypothesis 2.2 satisfies $z^{0,i} > 0$ if $h^i(u^0) = 0$, $i \in \{1, \dots, l\}$.

Hypothesis 2.4. For all $\alpha \neq \alpha^0$ in a neighborhood of α_0 , the linearized equation

$$(2.15) \quad \dot{x}(t) = \mathcal{D}_1 f(\bar{x}^\alpha, u^\alpha, \alpha) x_t, \quad t \geq 0$$

has only the trivial τ -periodic solution; here (x^α, u^α) are elements in $\mathbb{R}^n \times \mathbb{R}^m$ to be determined in Theorem 2.7, below.

Next we comment on these hypotheses.

Remark 2.5. Hypothesis 2.4 is equivalent to

$$(2.16) \quad \text{rank } \Delta(j\omega, \alpha) = n \quad \text{for } \omega = 2k\pi/\tau, \quad k \in \mathbb{Z},$$

where $\Delta(z, \alpha)$ is the characteristic function of (2.15),

$$\Delta(z, \alpha) = zI - \mathcal{D}_1 f(\bar{x}^\alpha, u^\alpha, \alpha)(e^{z \cdot} I), \quad z \in \mathbb{C}.$$

This hypothesis will guarantee that all Lagrange multipliers for $(\text{OPC})^\alpha$ can be obtained from Lagrange multipliers for $(\text{OSS})^\alpha$.

Remark 2.6. Hypotheses 2.3 and 2.2 are a constraint qualification and a second order sufficient optimality condition, respectively, for the steady-state problem $(\text{OSS})^{\alpha_0}$. Note that in (2.14) $f(\cdot, u^0, \alpha_0)$ is considered as a map $\mathbb{R}^n \rightarrow \mathbb{R}^n$, $x \mapsto f(\bar{x}, u^0, \alpha_0)$.

First we analyze the steady-state problem $(\text{OSS})^\alpha$.

THEOREM 2.7. *Suppose that $(x^0, u^0) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfy Hypotheses 2.1–2.3. Then*

(i) *The pair (x^0, u^0) is an isolated local minimum of Problem $(\text{OSS})^{\alpha_0}$, and the Lagrange multipliers $(y^0, z^0) \in \mathbb{R}^{n+n_1} \times \mathbb{R}^l$ are uniquely determined by (2.11) and (2.12).*

(ii) *There exists a continuously differentiable function $\alpha \rightarrow (x^\alpha, u^\alpha, y^\alpha, z^\alpha) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{n+n_1} \times \mathbb{R}^l$ defined on a neighborhood of α_0 such that (x^α, u^α) is an isolated minimum of $(\text{OSS})^\alpha$, $(x^{\alpha_0}, u^{\alpha_0}, y^{\alpha_0}, z^{\alpha_0}) = (x^0, u^0, y^0, z^0)$ and $(x^\alpha, u^\alpha, y^\alpha, z^\alpha)$ satisfy conditions (2.11)–(2.13) with α_0 replaced by α .*

Proof. This follows from a result in Fiacco [9, § 3.2]. \square

Next we state second order necessary optimality conditions for $(\text{OPC})^\alpha$.

THEOREM 2.8. *Let $(x^0, u^0) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfy Hypotheses 2.1–2.3 and suppose that (x^α, u^α) determined by Theorem 2.7 satisfy Hypothesis 2.4. There exists a neighborhood A_0 of α_0 with the following property. Let $\alpha \in A_0$, $\alpha \neq \alpha_0$ and assume that the constant functions $(\bar{x}^\alpha, \bar{u}^\alpha) \in C(-r, \tau; \mathbb{R}^n) \times L^\infty(0, \tau; \mathbb{R}^m)$ are a local minimum of $(\text{OPC})^\alpha$.*

Then for all $(x, u) \in C(-r, \tau; \mathbb{R}^n) \times L^\infty(0, \tau; \mathbb{R}^m)$ with

$$(2.17) \quad \int_0^\tau [g_x(x^\alpha, u^\alpha)x(t) + g_u(x^\alpha, u^\alpha)u(t)] dt \leq 0$$

and

$$(2.18) \quad \begin{aligned} x_0 &= x_\tau, \quad \dot{x}(t) = \mathcal{D}_1 f(\bar{x}^\alpha, u^\alpha, \alpha)x_t + f_u(\bar{x}^\alpha, u^\alpha, \alpha)u(t) \quad \text{a.e. } t \in [0, \tau], \\ \bar{u}^0 + u &\in \text{int } \mathcal{U}_{\text{ad}} \end{aligned}$$

it follows that

$$(2.19) \quad \begin{aligned} \int_0^\tau [\mathcal{D}_1 \mathcal{D}_1 H(\bar{x}^\alpha, u^\alpha, y^\alpha, \alpha)(x_t, x_t) + 2\mathcal{D}_1 \mathcal{D}_2 H(\bar{x}^\alpha, u^\alpha, y^\alpha, \alpha)(x_t, u(t)) \\ + \mathcal{D}_2 \mathcal{D}_2 H(\bar{x}^\alpha, u^\alpha, y^\alpha, \alpha)(u(t), u(t))] dt \geq 0. \end{aligned}$$

Sketch of proof. By continuity Hypotheses 2.1–2.3 hold for α in a neighborhood of α_0 . Problem $(\text{OPC})^\alpha$ can be reformulated as an optimization problem over $(\varphi, u) \in C(-r, 0; \mathbb{R}^n) \times L^\infty(0, \tau; \mathbb{R}^m)$ (with $\varphi := x_0$) using the implicit function theorem near $(\bar{x}^\alpha, \bar{u}^\alpha) \in C(-r, \tau; \mathbb{R}^n) \times L^\infty(0, \tau; \mathbb{R}^m)$ (cf. [8, Chap. 5]). Application of optimization theory in Banach spaces (cf. [8, Chap. 2] or [16]) yields second order necessary optimality conditions for $(\varphi^\alpha, \bar{u}^\alpha)$ with $\varphi^\alpha := \bar{x}^\alpha$, involving Lagrange multipliers $(l^\alpha, y^\alpha, z^\alpha) \in C(-r, 0; \mathbb{R}^n)^* \times \mathbb{R}^{n+n_1} \times \mathbb{R}^l$. Since by assumption $\bar{u}^0 + u \in \text{int } \mathcal{U}_{\text{ad}}$, the term with z vanishes. Hypothesis 2.4 yields that the Lagrange multipliers for $(\text{OPC})^\alpha$ can be obtained from Lagrange multipliers for $(\text{OSS})^\alpha$ (cf. [8, Prop. VII.2.7]); these, however, are unique by Theorem 2.7.

For more details see Theorem VII.3.1 of [8].

Theorem 2.8 furnishes a test for local properness: If there are (x, u) satisfying (2.17) and (2.18) but violating (2.19), then $(\bar{x}^\alpha, \bar{u}^\alpha)$ cannot be a local optimal solution

of $(\text{OPC})^\alpha$. Using Hypothesis 2.4 it is advantageous to consider special (sinusoidal) test functions (x, u) . First we introduce the following abbreviations ($\omega \in \mathbb{R}_+$, $\alpha \in A_0$):

$$\begin{aligned} P(\omega, \alpha) &:= \mathcal{D}_1 \mathcal{D}_1 H(\bar{x}^\alpha, u^\alpha, y^\alpha, \alpha)(e^{j\omega \cdot} I, e^{-j\omega \cdot} I), \\ Q(\omega, \alpha) &:= \mathcal{D}_2 \mathcal{D}_1 H(\bar{x}^\alpha, u^\alpha, y^\alpha, \alpha)(e^{j\omega \cdot} I), \\ R(\alpha) &:= \mathcal{D}_2 \mathcal{D}_2 H(\bar{x}^\alpha, u^\alpha, y^\alpha, \alpha), \\ B(\alpha) &:= \mathcal{D}_2 f(\bar{x}^\alpha, u^\alpha, \alpha) \end{aligned} \quad (2.20)$$

and for later purposes

$$L(\alpha) := \mathcal{D}_1 f(\bar{x}^\alpha, u^\alpha, \alpha).$$

Identify $P(\omega, \alpha)$, $Q(\omega, \alpha)$ and $R(\alpha)$ with elements in $\mathbb{C}^{n \times n}$, $\mathbb{C}^{n \times m}$ and $\mathbb{R}^{m \times m}$, respectively. Define

$$\begin{aligned} \Pi(\omega, \alpha) &= B(\alpha)^T \Delta(-j\omega, \alpha)^T P(\omega, \alpha) \Delta(j\omega, \alpha) B(\alpha) \\ &\quad + Q(-\omega, \alpha)^T \Delta(j\omega, \alpha) B(\alpha) \\ &\quad + B(\alpha)^T \Delta(-j\omega, \alpha)^T Q(\omega, \alpha) + R(\alpha). \end{aligned} \quad (2.21)$$

COROLLARY 2.9 (II-Test). *Let the assumptions of Theorem 2.8 be satisfied. Then (x^α, u^α) is locally proper, if there exist $\nu_0, \nu_1 \in \mathbb{C}^m$ with $(\omega = 2\pi/\tau)$*

$$\begin{aligned} &[g_x(x^\alpha, u^\alpha) \Delta(0, \alpha) B(\alpha) + g_u(x^\alpha, u^\alpha)] \nu_0 \\ &\quad + [g_x(x^\alpha, u^\alpha) \Delta(j\omega, \alpha) B(\alpha) + g_u(x^\alpha, u^\alpha)] \nu_1 \leq 0, \\ &h(u^\alpha + \nu_0 + \text{Re}(\nu_1 e^{j\omega t})) \in \text{int } \mathbb{R}_-^l \quad \text{for all } t \in [0, \tau], \end{aligned} \quad (2.22)$$

$$\nu_0^T \Pi(0, \alpha^0) \nu_0 + 2\nu_1^T \Pi(\omega, \alpha) \nu_1 < 0. \quad (2.23)$$

Sketch of proof. Choose $u(t) := \nu_0 + \text{Re}(\nu_1 e^{j\omega t})$, $t \in [0, \tau]$. Then (2.22) ensures that (2.17) and (2.18) are satisfied. Computation of the expression in (2.19) yields the one in (2.23) (cf. [8, Thm. VII.3.3]).

3. A scenario for local properness. Now we will relate local properness to structural changes in the system equation. The analysis is motivated by the following consideration. Suppose a Hopf bifurcation occurs at $\alpha = \alpha_0$. See Hale [12] or Hassard, Kazarinoff and Wang [13] for an exposition of Hopf bifurcation theory of functional differential equations. If the generated periodic solution is “better” than the steady-state solution, one will expect local properness at $\alpha = \alpha_0$. It turns out that under a controllability condition this is true for all α close to α_0 . The controllability condition guarantees that the free periodic motion can be approximated by forced periodic motions for $\alpha \neq \alpha_0$. In fact it is not necessary that a Hopf bifurcation actually occur; instead some weaker properties stated below are sufficient.

Throughout this section we assume that Hypotheses 2.1–2.4 hold and hence the assertions of Theorems 2.7, 2.8, and Corollary 2.9 hold. Recall that the characteristic function of the linearized equation (with $L(\alpha) := \mathcal{D}_1 f(\bar{x}^\alpha, u^\alpha, \alpha)$)

$$\dot{x}(t) = L(\alpha)x_t, \quad t \geq 0 \quad (3.1)$$

is given by

$$\Delta(z, \alpha) = zI - L(\alpha)(e^{z \cdot} I), \quad z \in \mathbb{C}. \quad (3.2)$$

LEMMA 3.1. *Suppose that for a pair $(\omega_0, \alpha_0) \in (0, \infty) \times A$*

$$\begin{aligned} &\text{rank } \Delta(j\omega_0, \alpha_0) = n - 1, \\ &\text{rank } \Delta(j\omega, \alpha) = n \quad \text{for all } (\omega, \alpha) \neq (\omega_0, \alpha_0) \text{ close to } (\omega_0, \alpha_0). \end{aligned} \quad (3.3)$$

Then for all α in a neighborhood of α_0 , (3.1) has a simple eigenvalue $z(\alpha)$ and $z(\alpha)$ has a continuous derivative $z'(\alpha_0)$ at $\alpha = \alpha_0$.

Proof. By Theorem 2.7, the map $\alpha \rightarrow L(\alpha)$ is continuously Fréchet differentiable, and Hale [12, Lemma 2.2, p. 171] implies the assertion. \square

Remark 3.2. Condition (3.3) does not require that an eigenvalue actually cross the imaginary axis at $\alpha = \alpha_0$.

LEMMA 3.3. Condition (3.3) implies that there exists a nontrivial τ -periodic solution of (3.1) with $\alpha = \alpha_0$, $\tau := 2\pi/\omega_0$; furthermore, there exists $p_1 \in \mathbb{C}^n$ such that for every such τ -periodic solution p

$$(3.4) \quad p(t) = 2\gamma \operatorname{Re}(e^{j\omega t} p_1), \quad t \geq 0,$$

for some $\gamma \in \mathbb{R}$.

Proof. By assumption the eigenspace corresponding to $z = j\omega_0$ is one-dimensional and the assertion follows (cf. Hale [12]). \square

LEMMA 3.4. Suppose that condition (3.3) is satisfied. Then the following two conditions are equivalent:

$$(3.5) \quad \text{There exists } \nu_1 \in \mathbb{C}^m \text{ with } p_1 = [\operatorname{Adj} \Delta(j\omega_0, \alpha_0)]B(\alpha_0)\nu_1$$

where p_1 is given by Lemma 3.3 and Adj denotes the adjunct;

$$(3.6) \quad [\operatorname{Adj} \Delta(j\omega_0, \alpha_0)]B(\alpha_0) \neq 0.$$

Proof. Recall that

$$\Delta(j\omega_0, \alpha_0)[\operatorname{Adj} \Delta(j\omega_0, \alpha_0)] = \det \Delta(j\omega_0, \alpha_0) \cdot I$$

(see, e.g., Kowalsky [15, Kap. 4]). Thus the range of

$$[\operatorname{Adj} \Delta(j\omega_0, \alpha_0)]B(\alpha_0)$$

is contained in the kernel of $\Delta(j\omega_0, \alpha_0)$ which is spanned by p_1 . \square

Condition (3.5) may be viewed as a “controllability condition” for the periodic solution (3.4).

LEMMA 3.5. Let condition (3.3) be satisfied. Then

$$\bar{p}_1^T P(\omega_0, \alpha_0) p_1 = \int_0^\tau \mathcal{D}_{1,2} \mathcal{D}_{1,2} H(\bar{x}^0, u^0, y^0, \alpha^0)((p_t, 0), (p_t, 0)) dt$$

where $p_1, p(\cdot)$ are as in Lemma 3.3.

Proof. Obvious from the definitions and Lemma 3.3. \square

The next theorem establishes the connection to local properness.

THEOREM 3.6. Let $(x^0, u^0) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfy the constraints of Problem (OSS) $^{\alpha_0}$ and suppose Hypotheses 2.1–2.4 hold. Furthermore, assume that conditions (3.3) and (3.5) are satisfied, and that there exists $\nu_0 \in \mathbb{C}^m$ such that ν_0 and ν_1 satisfy (2.22) with $\alpha = \alpha_0$. Let $p_0 := [\operatorname{Adj} \Delta(0, \alpha_0)]B(\alpha_0)\nu_0$ and assume for P given by (2.20)

$$(3.7) \quad \bar{p}_0^T P(0, \alpha_0) p_0 + \bar{p}_1^T P(\omega_0, \alpha_0) p_1 < 0.$$

Then there exists a neighborhood \mathcal{N} of (ω_0, α_0) such that the steady states (x^α, u^α) being isolated local minima of (OSS) $^\alpha$ are locally proper and (2.22), (2.23) hold for all $(\omega, \alpha) \in \mathcal{N}$, $(\omega, \alpha) \neq (\omega_0, \alpha_0)$.

Proof. In view of Theorem 2.7 and Corollary 2.9 it only remains to establish (2.22) and (2.23). By continuity, (2.22) is satisfied for (ω, α) near (ω_0, α_0) and ν_0, ν_1 replaced

by some elements $\nu_0^\alpha, \nu_1^\alpha$, which depend continuously on α . Furthermore $B(\alpha)$, $\text{Adj } \Delta(j\omega, \alpha)$ and $P(\omega, \alpha)$ are continuous with respect to (ω, α) and

$$[\det \Delta(j\omega, \alpha)]^2 > 0$$

for $(\omega, \alpha) \neq (\omega_0, \alpha_0)$ in a neighborhood of (ω_0, α_0) . We have

$$\begin{aligned} & \bar{\nu}_0^{\alpha T} B(\alpha)^T \Delta^{-1}(0, \alpha)^T P(0, \alpha) \Delta^{-1}(0, \alpha) B(\alpha) \nu_0^\alpha \\ & + \bar{\nu}_1^{\alpha T} B(\alpha)^T \Delta^{-1}(-j\omega, \alpha)^T P(\omega, \alpha) \Delta^{-1}(j\omega, \alpha) B(\alpha) \nu_1^\alpha \\ & = [\det \Delta(0, \alpha)]^{-2} \{ \bar{\nu}_0^\alpha B(\alpha)^T [\text{Adj } \Delta(0, \alpha)^T] P(0, \alpha) [\text{Adj } \Delta(0, \alpha)] B(\alpha) \nu_0^\alpha \} \\ & + [\det \Delta(j\omega, \alpha)]^{-2} \{ \bar{\nu}_1^\alpha B(\alpha)^T [\text{Adj } \Delta(-j\omega, \alpha)^T] P(\omega, \alpha) [\text{Adj } \Delta(j\omega, \alpha)] B(\alpha) \nu_1^\alpha \}. \end{aligned}$$

For $(\omega, \alpha) \rightarrow (\omega_0, \alpha_0)$ we have that $\det \Delta(j\omega, \alpha)$ tends to zero, while the second factor $\{\cdot \cdot \cdot\}$ in the second summand converges to $\bar{p}_1^T P(\omega_0, \alpha_0) p_1 < 0$.

Now consider the definition (2.21) of $\Pi(\omega, \alpha)$: For $(\omega, \alpha) \rightarrow (\omega_0, \alpha_0)$ the first summand tends to minus infinity with $[\det \Delta(j\omega, \alpha)]^{-2}$, the others tend to infinity with at most $|\det \Delta(j\omega, \alpha)|^{-1}$.

Thus the first summand becomes dominant and hence

$$(3.8) \quad \bar{\nu}^{0T} \Pi(0, \alpha) \nu^0 + \bar{\nu}^{1T} \Pi(j\omega, \alpha) \nu^1 < 0$$

for all $(\omega, \alpha) \neq (\omega_0, \alpha_0)$ in a neighborhood of (ω_0, α_0) . \square

COROLLARY 3.7. *Let $(x^0, u^0) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfy the constraints of problems $(\text{OPC})^{\alpha_0}$ and $(\text{OSS})^{\alpha_0}$ without control constraints (i.e., $h \equiv 0$) and assume that Hypotheses 2.1–2.4 hold and conditions (3.3) and (3.5) are satisfied. If*

$$(3.9) \quad \bar{p}_1^T P(\omega_0, \alpha_0) p_1 < 0$$

where p_1 is given by Lemma 3.3, then there exists a neighborhood \mathcal{N} of (ω_0, α_0) such that the steady states (x^α, u^α) being isolated local minima of $(\text{OSS})^\alpha$ are locally proper and

$$(3.10) \quad \bar{\nu}_1^T \Pi(\omega, \alpha) \nu_1 < 0 \quad \text{for all } (\omega, \alpha) \in \mathcal{N}, \quad (\omega, \alpha) \neq (\omega_0, \alpha_0),$$

where ν_1 is given by Lemma 3.4.

Proof. Follows from Lemma 3.4 and Theorem 3.6.

Remark 3.8. In Corollary 3.7, Condition (3.5) may be replaced by (3.6).

Remark 3.9. The second order sufficient optimality condition for the steady-state problem (i.e., Hypothesis 2.2) and the “complementary slackness” condition in Hypothesis 2.3 are needed in order to guarantee smooth dependence of $(x^\alpha, u^\alpha, y^\alpha)$ on α . If this can be guaranteed by other arguments (e.g., if the steady-state problem is independent of α as in the example of § 4, below) we can replace Hypothesis 2.2 by the assumption that (x^α, u^α) are a local minimum of $(\text{OSS})^\alpha$.

The following result is a partial converse of Corollary 3.7.

THEOREM 3.10. *Let the assumptions of Corollary 3.7 be satisfied. If there exists a sequence $(\omega_n, \alpha_n) \rightarrow (\omega_0, \alpha_0)$, $(\omega_n, \alpha_n) \neq (\omega_0, \alpha_0)$ with*

$$(3.11) \quad \bar{\nu}^T \Pi(\omega_n, \alpha_n) \nu > 0$$

where $\nu = \nu_1$ is given by Lemma 3.4, then

$$(3.12) \quad \bar{p}_1^T P(\omega_0, \alpha_0) p_1 \geq 0$$

where p_1 is given by Lemma 3.3.

Proof. Condition (3.12) and (2.21) imply

$$\begin{aligned} 0 &< \bar{\nu}^T \Pi(\omega_n, \alpha_n) \nu \\ &= \bar{\nu}^T B(\alpha_n)^T \Delta^{-1}(-j\omega_n, \alpha_n)^T P(\omega_n, \alpha_n) \Delta^{-1}(j\omega_n, \alpha_n) B(\alpha_n) \nu \\ &\quad + \bar{\nu}^T \{B(\alpha_n)^T \Delta^{-1}(-j\omega_n, \alpha_n)^T Q(\omega_n, \alpha_n) \\ &\quad + Q(-\omega_n, \alpha_n)^T \Delta^{-1}(j\omega_n, \alpha_n) B(\alpha_n) + R(\alpha_n)\} \nu. \end{aligned}$$

The first summand equals

$$[\det \Delta^{-1}(-j\omega_n, \alpha_n)]^{-2} \{\bar{\nu}^T B(\alpha_n) [\text{Adj } \Delta(-j\omega_n, \alpha_n)]^T P(\omega_n, \alpha_n) \text{Adj } \Delta(j\omega_n, \alpha_n) B(\alpha_n) \nu\}.$$

Again $[\det \Delta(-j\omega_n, \alpha_n)]^2 > 0$, and the second factor converges to

$$\begin{aligned} &\bar{\nu}^T B(\alpha_0) [\text{Adj } \Delta(-j\omega_0, \alpha_0)]^T P(\omega_0, \alpha_0) \text{Adj } \Delta(j\omega_0, \alpha_0) B(\alpha_0) \nu \\ &= \bar{p}_1^T P(\omega_0, \alpha_0) p_1. \end{aligned}$$

Arguing as in the proof of Theorem 3.6, we obtain (3.12). \square

Remark 3.11. Suppose that a Hopf bifurcation occurs at $\alpha = \alpha_0$ (cf. Hale [12, Thm. 1.1, p. 246]). Then Theorem 3.6 may be interpreted as follows: At $\alpha = \alpha_0$, a “natural” periodic solution of $\dot{x}(t) = f(x_t, u^\alpha, \alpha)$ bifurcates from the steady state x^α , $\alpha = \alpha_0$. By (3.7), this periodic motion shows better average performance than the steady state. Condition (3.3) is satisfied and the controllability condition (3.5) guarantees (by continuity) that for all α near α_0 the periodic trajectory can be approximated by trajectories corresponding to a sinusoidal control. Hence, for α near α_0 , the points (x^α, u^α) are locally proper. Suppose nontrivial periodic trajectories exist for $\alpha > \alpha_0$. Then, also for $\alpha < \alpha_0$, where no free periodic trajectory exists, we can generate periodic trajectories by appropriate sinusoidal controls. Thus it is not surprising that the assumption can be weakened by requiring only the assumptions of Theorem 3.6: it is not necessary that the nonlinear equation actually has a free periodic trajectory. In view of this discussion, it seems feasible to me to use the expression “Controlled Hopf Bifurcation” if conditions (3.3) and (3.5) are satisfied.

Remark 3.12. The stability properties of the periodically forced equations near $\alpha = \alpha_0$ may be very complicated; cf. Gambaudo [10] for a classification in the case of two-dimensional ordinary differential equations.

4. An example. In this section we consider an optimal periodic control problem for a retarded Lienard equation where Corollary 3.7 applies. First results for this problem were obtained in [6]. The problem is the following:

$$\text{Minimize} \quad -\frac{1}{\tau} \int_0^\tau x(s) ds + \frac{1}{2\tau} \int_0^\tau u(s)^2 ds$$

subject to

$$(4.1) \quad \ddot{x}(t) + f(x(t))\dot{x}(t) + g(x(t-r)) = u(t) \quad \text{a.e. } t \in [0, \tau],$$

$$(4.2) \quad x_0 = x_\tau, \quad (\dot{x})_0 = (\dot{x})_\tau,$$

$$(4.3) \quad \int_0^\tau u(t) dt = 0;$$

here f and $g: \mathbb{R} \rightarrow \mathbb{R}$; $x(t), u(t) \in \mathbb{R}$, and $r, \tau > 0$. We require that

$$(4.4) \quad f \text{ and } g \text{ are } C^2\text{-functions in a neighborhood of zero with } f(0) = g(0) = 0, \\ g'(0) = 1, g''(0) = -1, \text{ and } f(x) \neq 0 \text{ for } x > 0.$$

Writing (4.1) as a system of first order equations and applying the time transformation $t := tr$, we get

$$(4.5) \quad \begin{aligned} \dot{x}_1(t) &= \frac{1}{r} x_2(t), \\ \dot{x}_2(t) &= \frac{1}{r} [-f(x_1(t))x_2(t) - g(x(t-1)) + u(t)]. \end{aligned}$$

Consider $\alpha = 1/r > 0$ as bifurcation parameter.

The corresponding steady-state problem is

$$(4.6) \quad \begin{aligned} \text{(OSS)} \quad & \text{Minimize } -x_1 + \frac{1}{2}u^2 \quad \text{subject to} \\ & 0 = x_2, \\ & 0 = -f(x_1)x_2 - g(x_1) + u, \\ & 0 = u. \end{aligned}$$

The assumptions in (4.4) guarantee that $(x^0, u^0) = (0, 0)$ is the unique optimal solution of (OSS). Observe that (OSS) is independent of α ; hence Remark 3.9 applies and we can omit Hypothesis 2.2.

Furthermore Hypothesis 2.3 is satisfied and the corresponding Lagrange multipliers are

$$(4.7) \quad \begin{aligned} y_1 &= -g'(0)^{-1}f(0) = 0, \\ y_2 &= -g'(0)^{-1} = -1, \\ y_3 &= g'(0) = 1. \end{aligned}$$

The linearized system equation is

$$(4.8) \quad \dot{x}(t) = \alpha \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x(t) + \alpha \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix} x(t-1) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(t)$$

or

$$(4.9) \quad \ddot{x}(t) + \alpha^2 x(t-1) = u(t).$$

Thus the characteristic equation is

$$(4.10) \quad \det \Delta(z, \alpha) = z^2 + \alpha^2 e^{-z} = 0.$$

LEMMA 4.1. (i) *There exists an eigenvalue z of (4.8) on the imaginary axis if and only if $\alpha = \alpha_n = 1/(2n\pi)$, $n \in \mathbb{N}$.*

(ii) *If $\alpha = \alpha_n$ for some $n \in \mathbb{N}$, then the eigenvalues z on the imaginary axes are $z = \pm j$.*

(iii) *For $\alpha \rightarrow 0$ all eigenvalues in the right half-plane tend to the origin.*

(iv) *For α close to α_n , there exists a C^1 -function $\alpha \rightarrow z(\alpha)$ such that $z(\alpha)$ is a simple eigenvalue of (4.8) and $z(\alpha_n) = j$, $z'(\alpha_n) > 0$ and $\operatorname{Re} z(\alpha) < 0$ for $\alpha < \alpha_n$.*

Proof. The proof follows by an elementary analysis of (4.10).

The lemma shows that for $\alpha = \alpha_n$, $n \in \mathbb{N}$, a Hopf bifurcation occurs with frequency $\omega_0 = 1$.

A nontrivial periodic solution of

$$\dot{x}(t) = \alpha_n \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x(t) + \alpha_n \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix} x(t-1)$$

with period $\tau = 2\pi$ is given by

$$p(t) = 2 \begin{pmatrix} \cos t \\ -\sin t \end{pmatrix}.$$

The Fourier coefficients of p are

$$p_1 = \hat{p}(1) = \begin{pmatrix} 1 \\ j \end{pmatrix}, \quad \bar{p}_1 = \hat{p}(-1) = \begin{pmatrix} 1 \\ -j \end{pmatrix},$$

$$\hat{p}(k) = 0 \quad \text{for } k \neq \pm 1.$$

The function $H: C(-1, 0; \mathbb{R}^2) \times \mathbb{R} \times \mathbb{R}^3 \times (0, \infty) \rightarrow \mathbb{R}$ is given by

$$H(\phi, u, y, \alpha) := -\phi_1(0) + \frac{1}{2}u^2 + \alpha y^T \begin{pmatrix} \phi_2(0) \\ -f(\phi_1(0))\phi_2(0) - g(\phi_1(-1)) + u \\ u \end{pmatrix}.$$

We compute

$$\begin{aligned} \bar{p}_1^T P(\omega_0, \alpha_0) p_1 &= (1 \quad -j) \begin{pmatrix} -1 & f'(0) \\ f'(0) & 0 \end{pmatrix} \begin{pmatrix} 1 \\ j \end{pmatrix} \\ &= -1 < 0; \end{aligned}$$

thus (3.9) holds.

It only remains to show the controllability condition (3.6) (cf. Remark 3.8). We easily compute

$$\text{Adj} [\Delta(j\omega, \alpha)] B_0(\alpha) = \begin{pmatrix} j\omega & 1 \\ -\exp(-j\omega, \alpha) & j\omega \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ j\omega \end{pmatrix}.$$

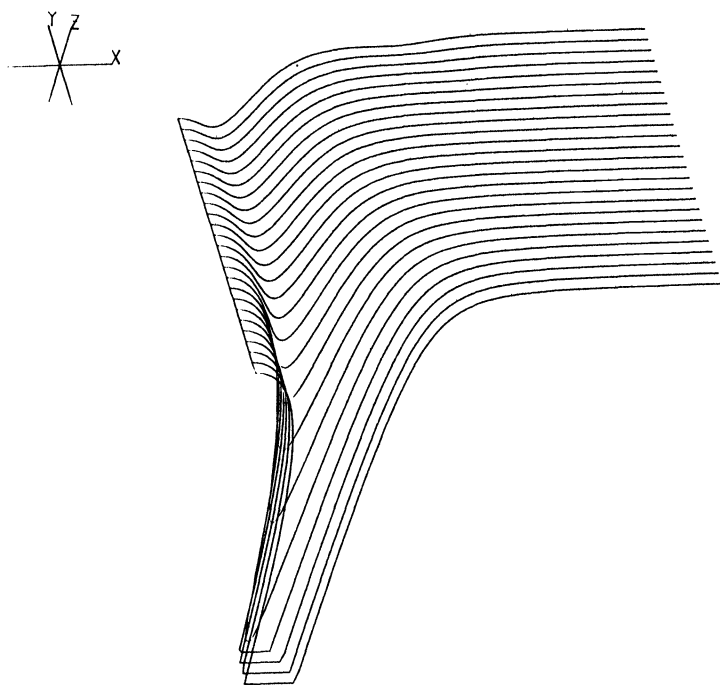


FIG. 1. Shows $\Pi(\omega, r)$, $0 \leq \omega \leq 4$, for different values of r between $r=0$ and $r=3$ ($X=\omega$, $Y=r$, $Z=\Pi(\omega, r)$). The function values are cut off for $z < -3$.

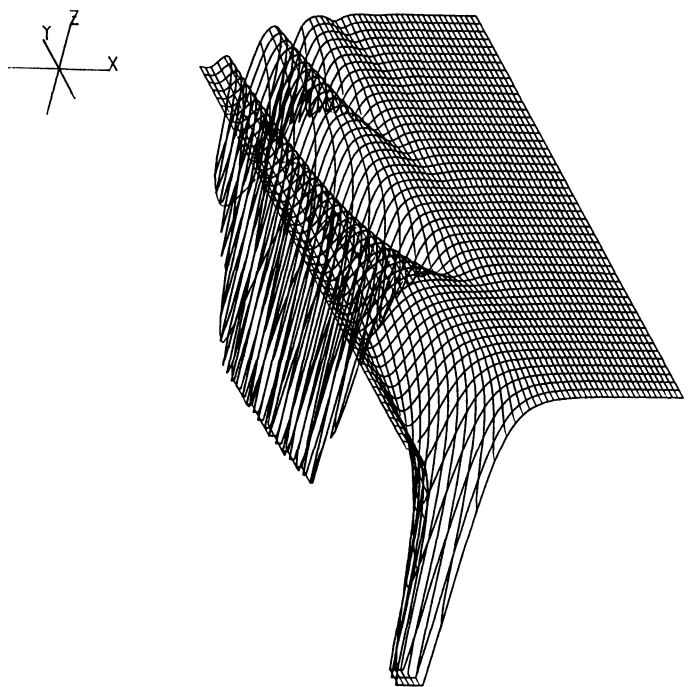


FIG. 2. Shows $\Pi(\omega, r)$, $0 \leq \omega \leq 4$, for different values of r between $r = 0$ and $r = 10$.

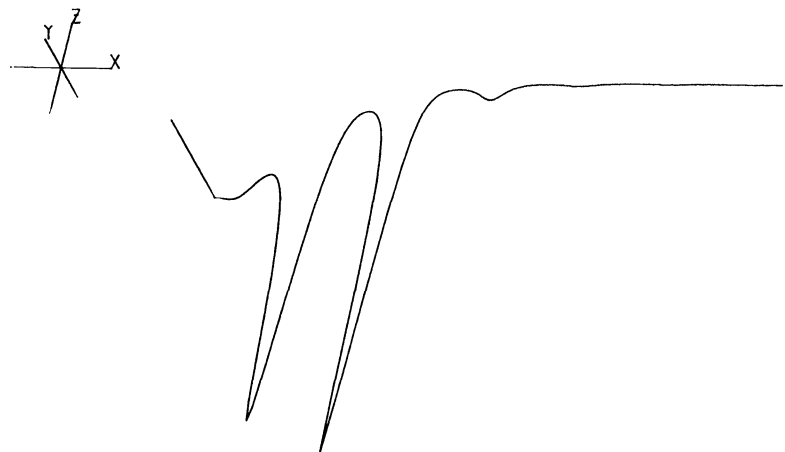


FIG. 3. Shows $\Pi(\omega, r)$, $0 \leq \omega \leq 4$, for $r = 10$.

Clearly

$$p_1 = \begin{pmatrix} 1 \\ j\omega_0 \end{pmatrix} \in \mathcal{R}[\text{Adj}[\Delta(\alpha_0, j\omega_0)]B_0(\alpha_0)].$$

Thus all the assumptions of Corollary 3.7 are verified. It is advantageous to write Π as a function of the delay r and the frequency ω . Then a simple computation yields

(4.11)
$$\Pi(\omega, r) = 1 - 1/[\omega^4 - 2\omega^2 \cos(\omega r) + 1]$$

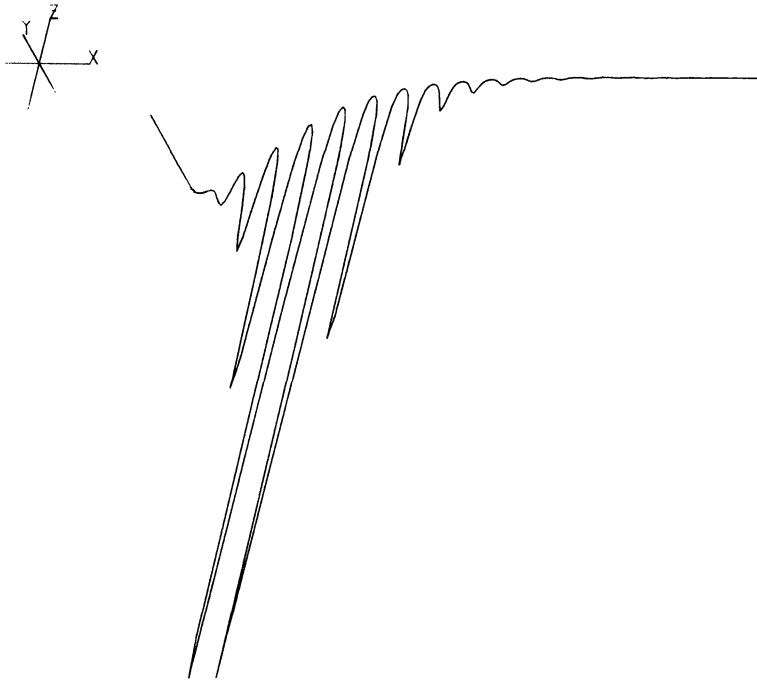


FIG. 4. Shows $\Pi(\omega, r)$, $0 \leq \omega \leq 4$, for $r = 30$.

for $(\omega, r) \neq (1, r_n)$, where $r_0 = 1$, $r_n := 1/\alpha_n$, $n \in \mathbb{N}$. Zones of local properness, indicated by $\Pi(\omega, r) < 0$ occur for (ω, r) close to $(1, r_n)$.

An analysis of the function Π given by (4.11) yields that for all $\omega, r \in \mathbb{R}_+$

$$1 - 1/[\omega^2 - 1]^2 \leq \Pi(\omega, r) \leq 1 - 1/[\omega^2 + 1]^2,$$

$$\Pi(0, r) = 0, \quad \lim_{\omega \rightarrow \infty} \Pi(\omega, r) = 1.$$

Figures 1–4 show $\Pi(\omega, r)$ for different values of r (here $X = \omega$, $Y = r$, $Z = \Pi(\omega, r)$). A significant feature of this example is that the zones of properness (i.e., the ω -intervals where $\Pi(\omega, r) < 0$) which occur at a Hopf bifurcation at $r = r_n$ (indicated by a negative pole of $\Pi(\omega, r)$ at $\omega = 1$) do not vanish for increasing r . Thus for large r , $\Pi(\omega, r)$ becomes very oscillatory (see Fig. 4).

Remark 4.2. It is easy to check that the function $\Pi: \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R} \cup \{-\infty\}$ has no local minima besides $(\omega, r) = (1, r_n)$. This may be interpreted in the following way: Local properness in this problem occurs *only* via the mechanism described by Corollary 3.7. Naturally, this may not be true for other problems (e.g., local properness may be due to nonlinearities in the performance criterion).

Acknowledgments. I thank Professor H. T. Banks for the invitation to work at Brown University. Furthermore I would like to acknowledge the use of programs by Dr. M. Pratt for the plotting of the diagrams.

REFERENCES

- [1] W. ALT, *Lipschitzian perturbations of infinite optimization problems*, in *Mathematical Programming with Data Perturbations II*, A. V. Fiacco, ed., Marcel Dekker, New York, 1983, pp. 7–21.

- [2] D. S. BERNSTEIN, *Control constraints, abnormality and improved performance by periodic control*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 367–376.
- [3] D. S. BERNSTEIN AND E. G. GILBERT, *Optimal periodic control: the Π test revisited*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 673–684.
- [4] S. BITTANTI, G. FRONZA AND G. GUARBADASSI, *Periodic control: A frequency domain approach*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 33–38.
- [5] P. L. BUTZER AND R. J. NESSEL, *Fourier Analysis and Approximation*, Vol. 1, Birkhäuser, Basel, 1971.
- [6] F. COLONIUS, *Optimal periodic control of retarded Liénard equations*, in Control of Distributed Parameter Systems and Applications, F. Kappel, K. Kunisch and W. Schappacher, eds., Springer-Verlag, Berlin, Heidelberg, New York, 1986.
- [7] ———, *Optimality for periodic control of functional differential systems*, J. Math. Anal. Appl., 120 (1986), pp. 119–149.
- [8] ———, *Optimal Periodic Control*, FS “Dynamische Systeme” Report No. 140, Universität Bremen, Bremen, Federal Republic of Germany, 1985.
- [9] A. V. FIACCO, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, New York, 1983.
- [10] J. M. GAMBAUDO, *Perturbation of a Hopf bifurcation by an external time-periodic forcing*, J. Differential Equations, 57 (1985), pp. 172–199.
- [11] E. G. GILBERT, *Optimal Periodic Control, A general theory of necessary conditions*, this Journal, 15 (1977), pp. 717–746.
- [12] J. K. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, Berlin, 1977.
- [13] B. D. HASSARD, N. D. KAZARINOFF AND Y. H. WANG, *Theory and Application of Hopf Bifurcation Theory*, London Mathematical Society Lecture Notes 41, Cambridge University Press, Cambridge, 1981.
- [14] A. D. IOFFE AND V. M. TIKHOMIROV, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.
- [15] H. J. KOWALSKY, *Lineare Algebra*, De Gruyter, Berlin, 1963.
- [16] A. LINNEMANN, *Higher order necessary conditions for infinite and semi-infinite optimization*, J. Optim. Theory Appl., 28 (1982), pp. 483–512.
- [17] M. MATSUBARA, Y. NISHIMURA, N. WATANABE AND K. ONOGI, *Periodic Control Theory and Applications*, Research Reports of Automatic Control Laboratory, Vol. 28, Faculty of Engineering, Nagoya University, 1981.
- [18] E. NOLDUS, *A survey of optimal periodic control of continuous systems*, Journal A., 16 (1975), pp. 11–16.
- [19] K. ONOGI AND M. MATSUBARA, *Structure analysis of periodically controlled chemical processes*, Chem. Engrg. Sci., 34 (1980), pp. 1009–1019.
- [20] D. L. RUSSELL, *Optimal orbital regulation in dynamical systems subject to Hopf bifurcation*, J. Differential Equations, 44 (1982), pp. 188–223.
- [21] D. SINCIC AND J. E. BAILEY, *Optimal periodic control of variable time-delay systems*, Internat. J. Control, 27 (1978), pp. 547–555.
- [22] J. L. SPEYER, *Non-optimality of steady-state cruise for aircraft*, AIAA J., 14 (1976), pp. 1604–1610.
- [23] J. L. SPEYER AND R. T. EVANS, *A second variational theory of optimal periodic processes*, IEEE Trans. Automat. Control, 29 (1984), pp. 138–148.
- [24] N. WATANABE, K. ONOGI AND M. MATSUBARA, *Periodic control of continuous stirred tank reactors—I*, Chem. Engrg. Sci., 36 (1981), pp. 809–818.
- [25] ———, *Periodic control of continuous stirred tank reactors—II*, Chem. Engrg. Sci., 37 (1982), pp. 745–752.

REMARKS ON A PAPER BY R. M. HIRSCHORN*

A. KUMPERA†

Abstract. We indicate some imprecisions appearing in a paper by R. M. Hirschorn and suggest modifications.

Key words. nonlinear control system, reachable set, Lie algebra

AMS(MOS) subject classification. 93B05

1. Introduction. In 1976, R. M. Hirschorn published in this journal [1] far-reaching results concerning the relationship between the structure of the reachable sets of a nonlinear control system of the form

$$\frac{dx}{dt} = A(x) + u_1 B_1(x) + \cdots + u_m B_m(x)$$

and the properties of the associated, eventually infinite-dimensional, Lie algebra of vector fields

$$\mathcal{L} = \{A, B_1, \dots, B_m\}_{LA}.$$

Unfortunately, the underlying hypotheses are revealed to be insufficient. In § 2, we provide a counterexample to Theorem 3.6 of [1] and, consequently, to the main Theorem 3.2, and point out the gap in the proof. In § 3, we discuss weaker forms for the above-mentioned theorems. The basic definitions as well as the notation are borrowed from Hirschorn's paper.

2. A counterexample. We consider on the manifold $M = \mathbb{R}^2$ the system

$$(*) \quad \frac{dX}{dt} = A(X) + uB(X), \quad X = (x, y)$$

where $A = \partial/\partial x + \partial/\partial y$ and $B = \sin^2(x) \partial/\partial x$. Let $C = \sin(2x) \partial/\partial x$, $D = \cos(2x) \partial/\partial x$ and $E = \partial/\partial x$; then

$$\mathcal{L} = \{A, B, C, D, E\}_{LS},$$

$$\mathcal{B} = \{B\}_{LS},$$

$$\mathcal{L}_0 = \{B, C, D, E\}_{LS}$$

and the commutation relations read as follows:

$$[A, B] = C, \quad [B, C] = -2B, \quad [C, D] = -2E, \quad [D, E] = 2C.$$

$$[A, C] = 2D, \quad [B, D] = -C, \quad [C, E] = -2D,$$

$$[A, D] = -2C, \quad [B, E] = -C,$$

$$[A, E] = 0,$$

The algebra \mathcal{L} is composed of analytic complete vector fields and the hypothesis $[\mathcal{L}_0, \mathcal{B}(a)] \subset \mathcal{B}(a)$ (cf. [1, Thm. 3.6]) is satisfied since $[C, B] = 2B \in \mathcal{B}$ and $[D, B] = [E, B] = C$ vanishes whenever B does.

* Received by the editors May 13, 1987; accepted for publication July 20, 1987.

† Instituto de Matemática, Universidade Estadual de Campinas, 13.081 Campinas SP, Brazil.

Let us now examine the reachable sets for the system (*). Since, for any constant c , $A + cB = \partial/\partial y + f_c \partial/\partial x$, where $f_c(x) = 1 + c \sin^2 x$, and since

$$\left[\frac{\partial}{\partial y}, f_c \frac{\partial}{\partial x} \right] = 0,$$

we infer that any point $Q = (x, y)$ reachable from $P = (x_0, y_0)$ at time t (i.e., $Q \in \mathcal{R}_t(P)$) via the finite succession of constant controls c_1, c_2, \dots, c_l , during respective times t_1, \dots, t_l with $\sum t_i = t$ (i.e., the control function u in (*) takes the value $u(s) = c_i$ when $s \in [t_1 + \dots + t_{i-1}, t_1 + \dots + t_i]$) can, in fact, be reached from P by steering initially along the vector field $\partial/\partial y$, during time t , thus attaining the point $P_1 = (x_0, y_0 + t)$ and subsequently, steering along the succession of vector fields $f_{c_1} \partial/\partial x, \dots, f_{c_l} \partial/\partial x$ during the corresponding times t_1, \dots, t_l . Hence, the reachable set $\mathcal{R}_t(P)$ for the system (*) is equal to the reachable set $\hat{\mathcal{R}}_t(P_1)$ relative to the family of vector fields $\{f_c \partial/\partial x\}$. Observing that $f_c(0) = 1 > 0$ whatever the value c , we infer that any solution of (*) with initial data $x(0) = 0$ necessarily enters into the domain $\Delta = \{(x, y): x > 0\}$ since $x'(t) > 0$ as long as control $u = c_1$ is maintained. Let us switch now to control $u = c_2$ and initial data $x(t_1) > 0$. Then, either $c_2 \geq -1$ —hence $f_{c_2}(x) \geq 0$, and consequently $x'(t) \geq 0$; or $c_2 < -1$, and the following two possibilities arise: (a) $f_{c_2}(x(t_1)) \geq 0$ or (b) $f_{c_2}(x(t_1)) < 0$. In case (a), $x'(t) \geq 0$ as before and in case (b), $x'(t) < 0$. However, in the latter case, the solution $x(t)$, though decreasing, never attains the first zero point of f_{c_2} to the left of $x(t_1)$, as shown by the graph of f_{c_2} (see Fig. 1).

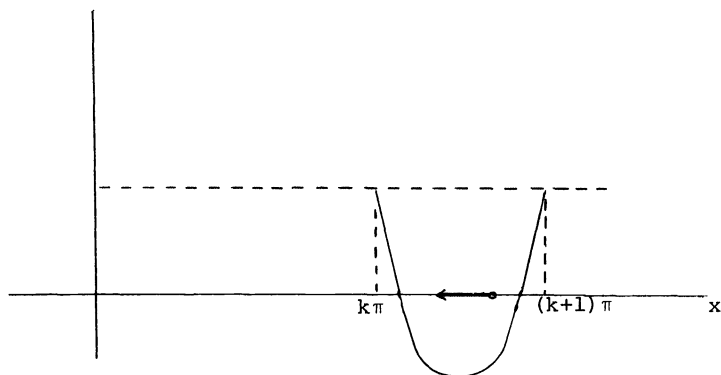


FIG. 1

We infer that $X(t)$ remains in the domain Δ as long as control c_2 is maintained and, iterating the above argument to the finite succession of controls c_1, \dots, c_l , we eventually conclude that $\mathcal{R}_t(0) \subset \Delta$. On the other hand, since the constant controls c_i can be arbitrarily chosen, it follows that $\mathcal{R}_t(0) = \{(x, t): x > 0\}$ for any $t > 0$. A similar argument shows, in general, that $\mathcal{R}_t(P) = \{(x, y_0 + t): x > k\pi\}$ where $P = (x_0, y_0)$, $t > 0$ and $k\pi \leq x_0 < (k+1)\pi$. In conclusion,

$$\mathcal{R}_t(P) \neq I'(\mathcal{L}_0, P) = \{(x, y_0 + t): x \in \mathbb{R}\}.$$

The gap in the proof of Theorem 3.6 of [1, p. 707, 1–19] appears in the claim: “A straightforward computation \dots .” In fact, according to our example, the vector field $\xi = \sin(x) \partial/\partial x$ satisfies the property $\xi(a) \in \mathcal{B}(a)$, for all $a \in M$, but $[E, \xi] = \cos(x) \partial/\partial x$ fails to do so.

3. The stronger hypothesis. If we assume that the distribution $x \mapsto \mathcal{B}(x)$ on the manifold M is regular, then the above-mentioned claim is true and Theorem 3.6 holds.

However, regularity conditions are strange and inconvenient to control theory. Moreover, in the context of [1], such a condition would make any attempt to define a suitable A -radical extremely difficult. It seems, therefore, that the simplest and most convenient remedy consists in precisely assuming whatever the proof requires, namely, replacing the hypothesis $[\mathcal{L}_0, \mathcal{B}](x) \subset \mathcal{B}(x)$ by the stronger version

$$(\text{ad}_{\mathcal{L}_0}^k \mathcal{B})(x) \subset \mathcal{B}(x), \quad k \geq 1.$$

Under this assumption, Theorem 3.6 holds and, if we redefine accordingly the A -radical for \mathcal{B} , Theorem 3.2 also holds. Needless to say, the above assumption is, for practical purposes, rather discouraging.

REFERENCE

- [1] R. M. HIRSCHORN, *Global controllability of nonlinear systems*, this Journal, 14 (1976), pp. 700-711.

GAMES WITH REPEATED DECISIONS*

STEVE ALPERN†

Abstract. Extensive form games are traditionally required to obey the axiom of “nonrepetition”: *No information set may contain more than one node from any directed path in the tree.* However, in order to model games where the players are teams or automata, it is necessary to consider extensive forms which do not satisfy the nonrepetition axiom. This paper develops a noncooperative strategic theory for such games. Mixed or behavioral strategies are inadequate for such games, but there always exist Nash equilibria in finite combinations of behavioral strategies. It can also be shown that any given pair of such strategies form the unique Nash equilibrium for some two person zero-sum game. Mathematically, the theory is closely related to that of polynomial and separable games as developed by Dresher, Karlin and Shapley (Ann. Math. Stud., 24 (1950), pp. 161–180) and by Gale and Gross (Pacific J. Math., 8 (1958), pp. 735–741).

Key words. games, extensive form, polynomial games

AMS(MOS) subject classification. 90D

1. Introduction. Game theory has traditionally been studied under the assumption that a player is never confronted with the same decision problem more than once in any play of a game. This assumption arises, in the conventional definition of an extensive form game (tree), through the following “nonrepetition” hypothesis placed on information sets: *No information set may contain more than one node from any directed path in the tree.* We will call games which satisfy this hypothesis “nonrepetitive,” and those which do not “repetitive.” It is clear that games where the players are individuals and have perfect memories can be modeled as nonrepetitive games and even as games of perfect recall. However, the contention of this paper is that repetitive games may be required to model situations where the players are teams (and the decisions are not agent-specific) or automata with limited memories. We outline a noncooperative strategic theory for repetitive games which establishes that there are always equilibria in finite mixtures of behavioral strategies. We rely extensively on ideas from the theory of polynomial and separable games, as developed in [DKS], [GG], and [K].

Before going further into the mathematical analysis, some interpretive material may be useful to justify the whole enterprise. When games of imperfect recall are modeled by considering players to be represented by teams of agents, all decisions made by a team are assumed to be agent-specific. This means that there is an implicit map from information sets of player i to agents of player i . For example, in bridge, every information set of player North-South is uniquely assigned to either agent North or agent South. Such team games, where decisions are agent-specific, clearly satisfy the usual nonrepetition hypothesis because the individual agents are assumed to have perfect memories. Furthermore, such games can alternatively be modeled as games of perfect recall, with agents as players. However the situation is entirely different for team games where decisions of teams are not agent-specific. In many large organizations (teams) it is more a matter of chance which agent deals with a given decision. Sometimes this allocation of agents is formally randomized, as when incoming calls are handled on switchboards and given to the next available agent. More often it is informal. It is clear, for example, that if in two separate instances a policeman had given me a warning for speeding (as opposed to a fine) on the same day, both were confronted with the

* Received by the editors August 18, 1986; accepted for publication (in revised form) August 21, 1987.

† Department of Mathematics, London School of Economics, London, WC2A 2AE, England.

same decision problem. That is, neither one knew that he was the first or second to warn me. The situation would have been different (worse for me in this case, I'm afraid) if an individual agent had been permanently assigned to me. In summary of these arguments, repetitive games are needed to model team games in which decisions are not agent-specific.

As an example, consider "Team Chess" played by two large teams called White and Black. At every move for, say, White, a referee randomly selects a White agent, shows him the current position and requests a move. Since the agent is not told the moves which led to this position, he may select a move (such as castling when the king has already moved) which is illegal in standard chess—in Team Chess such moves are legal but lose. In this model positions are information sets, and thus the same information set may be entered more than once in a play of the game.

The second application of repetitive games is to cases when players are modeled as automata (for example, [Ru]). When a node of player i is reached, the referee inputs the code for the information set of that node into an automaton which is playing for player i (or *is* player i). Depending on the number of internal states (memory size) it is possible that the automaton "cannot distinguish" whether or not that code has been previously inputted. To carry out the types of strategies outlined in this paper the automata must be probabilistic.

To illustrate the extensive and normal forms to be studied in this paper we consider the two-person zero-sum game shown in Fig. 1. If player I goes right at his single information set with probability x and player II goes right with probability y then the payoff (to maximizer I) is given by the behavioral-normal form

$$A(x, y) = 3(1 - x) + 4x(1 - y)(1 - x) + 5x(1 - y)x + 6xy.$$

While this has the form of a polynomial game, with both x and y chosen from the unit interval, it is clear that as soon as some player has more than one information set the expression for A will be somewhat more general. We will call this form a "multinomial game." While the theory of polynomial games is not sufficient for our purposes, it turns out that an extension called "separable games" developed by Karlin et al. [DKS], [KS], [K] can be adapted to analyze the solutions of games with repeated decisions.

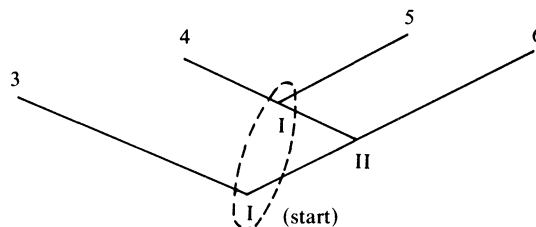


FIG. 1

The organization of the paper is as follows. In § 2 we give the definition of an extensive form game and of some related strategic notions. This is as usual (in fact we follow Owen [O]) except for the omission of the restriction on information sets and the inclusion of randomized strategies. We outline the easy argument which shows that there is always a Nash equilibrium in randomized strategies. Section 3 consists of examples of two-person zero-sum games with repeated decisions. These examples

demonstrate the inadequacy of mixed and behavioral strategies. Section 4 employs an analysis of the generalized moment space to establish that randomized strategies concentrated on a finite number of behavioral strategies are sufficient. Theorem 2 relates the required complexity of strategies to the informational complexity of the extensive form game. The results on strategic complexity given here are not sharp and can undoubtedly be improved by a better description of the moment space. In § 5 we use an algorithm of Gale and Gross [GG] to find an extensive game with a given pair of unique optimal strategies.

2. Extensive form games. In this section we define a game in extensive form. Our definition is standard (see [O]) except that the assumption of “no repeated decisions” is omitted. We define various notions of strategy and the resulting behavioral-normal form. We prove that all games possess a Nash equilibrium in terms of randomized strategies.

DEFINITION. By an n -person game in extensive form the following is meant:

(i) A tree T with a distinguished vertex v^* called the “starting point,” with all the edges of T directed away from v^* .

(ii) A function U , called the “payoff function,” which assigns an n -vector $U(v)$ to each terminal vertex v of T .

(iii) A partition of the nonterminal vertices of T into $n+1$ sets S_0, S_1, \dots, S_n , called the “player sets.”

(iv) A probability distribution, defined at each (chance) vertex of S_0 , among the immediate followers of the vertex.

(v) For each $i = 1, \dots, n$, a subpartition of S_i into subsets S_i^j , called “information sets,” such that two vertices in the same information set have the same number of immediate followers.

(vi) For each information set S_i^j , an index set I_i^j (called choices), together with 1:1 mappings of the set I_i^j onto the set of immediate followers of each vertex of S_i^j .

If no two vertices of S_i^j , for any j , lie on a single directed path, then player i is said to have “no repeated decisions.” If all players have no repeated decisions we call the game “nonrepetitive”; otherwise we call it repetitive. In our terminology, traditional game theory is the study of nonrepetitive games.

The primitive strategic notion for repetitive games is the behavioral strategy. A behavioral strategy b_i for player i is a function which assigns to each information set S_i^j of player i a probability distribution over the choices I_i^j . The set of all behavioral strategies for player i is therefore the Cartesian product of a finite number of simplices, which we denote by B_i . We observe for future use that each set B_i is compact and convex, and hence so is their Cartesian product B . The extreme points of B_i are those (finitely many) b_i which assign only the probabilities 0 or 1 to all choices, and we call these “pure strategies.” A Borel probability measure μ_i on B_i is called a randomized strategy, and these are denoted collectively by B_i^* . If μ_i has finite support we will call it a “finite randomized strategy” and if additionally the support consists entirely of pure strategies then we call μ_i a “mixed strategy.”

Corresponding to any n -tuple of behavioral strategies $b = (b_1, \dots, b_n)$ there is a distribution $D = D(b_1, \dots, b_n)$ over the set of terminal vertices. For each terminal vertex v , $D_b(v)$ is the probability of the game ending at v if the strategies b are employed. The number $D_b(v)$ is computed by multiplying all the probabilities assigned by b to the edges (choices) leading from the starting vertex v^* to v . It is important to note that $D(v)$ is a continuous function of b . Since $U_i(v)$ is the von Neumann–Morgenstern utility to player i of the play ending at v , the expected utility for i

corresponding to b is given by $U_i(b) = \sum_v D_b(v) U_i(v)$ (summation over terminal vertices v) which is therefore also continuous in b . We call these maps $U_i: B \rightarrow R$ the behavioral-normal form of the game. Finally, if $\mu = (\mu_1, \dots, \mu_n)$ is an n -tuple of randomized strategies, then we define the expected utility to player i by $U_i(\mu) = \int \dots \int U_i(b_1, \dots, b_n) d\mu_1(b_1) \dots d\mu_n(b_n)$.

There are many equilibrium concepts which have proved useful in the study of nonrepetitive games. However in this first study of repetitive games we will consider only the most basic concept, that of a Nash equilibrium. An n -tuple $\bar{\mu} = (\bar{\mu}_1, \dots, \bar{\mu}_n)$ of randomized strategies is called a Nash equilibrium if for each $i = 1, \dots, n$ and all μ_i in B_i^* , $U_i(\bar{\mu}_1, \bar{\mu}_2, \dots, \bar{\mu}_{i-1}, \mu_i, \bar{\mu}_{i+1}, \dots, \bar{\mu}_n) \leq U_i(\bar{\mu})$.

Since each of the n utility functions U_i is continuous on the compact set B , a well-known result (see [O, p. 78] or [BO, p. 168]) asserts that there is a Nash equilibrium in Borel probability measures (usually called mixed strategies, but that term would be misleading here) on the sets B_i . Reinterpreting this result in our setting, we obtain the following.

THEOREM 1. *Every finite extensive form game has a Nash equilibrium in randomized strategies.*

3. Examples. In this section we solve three repetitive two-person zero-sum games. The first is given in extensive form, while the latter two are given in behavioral-normal form together with an algorithm (Theorem 2) for constructing an inducing extensive form. We will simplify an otherwise notationally complex problem by choosing only "binary" games in which all choices made by the players are binary, or equivalently, in which all nonterminal vertices have outdegree two. In the formal notation of § 2, this requires that $\#(I_i) = 2$ for all i and j . Since any choice among a finite number of alternatives can be reduced (nonuniquely) to a sequence of binary choices, every extensive form game is equivalent to a binary game. For binary games the probability distributions over the (two) alternatives at any information set may be identified with the unit interval $[0, 1]$ and hence the set of behavioral strategies B_i for each player i is a finite-dimensional cube, as is the product B of these sets.

For binary games there is a simple notation both for behavioral strategies and for labeling information sets. Let x_i (y_i) denote the probability of going to the right at the i th information set of player I (II) according to the behavioral strategy x (y). Label the corresponding edges of the tree x_i , and the complementary edges with " $1 - x_i$." Such a labeling of the edges of a binary tree indicates all the game theoretic information (player sets, information sets and identified alternatives) and also makes clear the corresponding Markov chain with absorbing barriers.

Example 1. Consider the game with extensive form shown in Fig. 2.

The behavioral-normal form corresponding to this extensive form is

$$A(x_1, x_2; y_1) = 6(1 - x_1)y_1 + 5x_1(1 - y_1) + 4x_1y_1(1 - x_2) + 6x_1y_1x_2(1 - x_1).$$

We can reduce the dimension of player I's strategies by the following observation. If

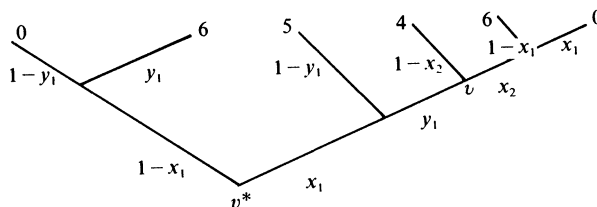


FIG. 2

x_1 is known to the agent choosing x_2 at the vertex labeled v , then clearly a maximizing agent should choose $x_2 = 1$ if $x_1 < \frac{1}{3}$ and $x_2 = 0$ if $x_1 > \frac{1}{3}$ (with any x_2 if $x_1 = \frac{1}{3}$). More precisely, $(x_1, 1)$ dominates (x_1, x_2) for $x_1 < \frac{1}{3}$ and $x_2 < 1$, and similarly for $(x_1, 0)$ if $x_1 > \frac{1}{3}$. Hence the game with normal form $A(x_1, x_2; y_1)$ is equivalent (with $x = x_1$ and $y = y_1$) to the game on the square with kernel

$$K(x, y) = \begin{cases} 6(1-x)y + 5x(1-y) + 6x(1-x)y & \text{if } x \leq \frac{1}{3}, \\ 6(1-x)y + 5x(1-y) + 4xy & \text{if } x \geq \frac{1}{3}. \end{cases}$$

It is not hard to calculate that $\inf_y \sup_x K(x, y) = 5 - q$, which is attained uniquely for $y = q = (170 + 40\sqrt{6})/386 \doteq 0.694$. This mixed (or behavioral) strategy for player II is optimal. We next observe that the $\sup_x K(x, q)$ is attained only at the two points $x = a = 5(1 - q)/12q \doteq 0.184$ and $x = 1$. It follows that player I has an optimal randomized strategy involving only the two behavioral strategies $(a, 1)$ and $(1, 0)$. To calculate the probability p such that the randomized strategy $p(a, 1) + (1 - p)(1, 0)$ is optimal we consider the 2×2 matrix game shown below.

	$y = 0$	$y = 1$
$(a, 1)$	$5a$	$6(1 - a^2)$
$(1, 0)$	5	4

The minimax solution to this matrix game is given by $y = 1$ with probability q (which we already knew) and $(a, 1)$ with probability $p = -1/(6a^2 + 5a - 7) \doteq 0.17$. The value of this 2×2 game, and hence also of the original extensive form game, is $5 - q$. It is worth noting that in contrast to nonrepetitive games, neither the value nor the probabilities used in the optimal strategies lie in the field generated by the payoffs (all integers in this example). This is of course due to the nonlinearity of the normal form. We wish to make one further remark concerning the nature of the optimal strategies found for this game. Observe that player I had to randomize over only two behavioral strategies. We could have anticipated this qualitative feature by observing that $\partial^2/\partial y_1^2 A(x_1, x_2; y_1) = 0$ and referring to Karlin's result [K, Thm. 4.3.1] that if $\partial^n/\partial y_1^n A(x_1, \dots, x_m; y_1) \geq 0$, then player I has an optimal strategy involving at most n points. This is part of the theory of "generalized convex games." See also Owen [O, Prob. 2, p. 83]. In the next section we will see how weaker results in this direction can be obtained even if both strategy spaces are multidimensional.

Observe that the behavioral-normal form of the game in Example 1, and indeed of any binary two-person zero-sum game, has the form

$$A(x, y) = \sum_{k=1}^L a_k f_k(x, y) \quad \text{where } f_k(x, y) = x_1^{i_1} \cdots x_m^{i_m} y_1^{j_1} \cdots y_n^{j_n};$$

the exponents depend on k . We will call such games, where the maximizer I chooses x in $[0, 1]^m$ and the minimizer II chooses y in $[0, 1]^n$, "multinomial games." This is an obvious generalization of the class of polynomial games, where $m = n = 1$. It is clear that the behavioral-normal form of a binary two person zero-sum game is always a multinomial game. Our next two examples are given most naturally as multinomial games, so it is useful to know that they arise from a (repetitive) game in extensive form. This follows from Theorem 2.

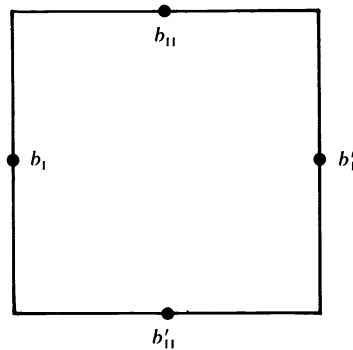
THEOREM 2. *Every multinomial game is the behavioral-normal form of some extensive form binary two-person zero-sum game. If the degree (highest exponent) of the multinomial is one, then the inducing extensive form game may be taken to be nonrepetitive.*

Proof. Unfortunately we know of no efficient algorithm for constructing the extensive form, so we give only the following trivial one: Suppose the multinomial $A(x, y)$ is in the form written above. At a starting vertex v^* we begin a tree by drawing L out-edges which are chance moves with probability $1/L$. The vertex following the k th of these edges is followed by $i_1 = i_1(k)$ consecutive vertices of information set 1 of player I, with every node following a $(1 - x_1)$ left choice being a terminal vertex with utility 0. At the vertex following $i_1(k)$ choices of x_1 (right) place a similar tree based on $i_2(k)$ choices of information set 2 of player I. Finally, at the terminal vertex v_k following chance move k , i_1 choices to the right (x_1) at information set 1 of player I, \dots , j_m choices to the right (y_n) at information set n of player II, let the utility be given by $U(v_k) = La_k$. Since the probability of arriving at a vertex of type v_k is $(1/L)f_k(x, y)$ and the utility of all vertices not of this type is 0, the behavioral-normal form of this game is the given multinomial $A(x, y)$. This construction clearly has the property claimed in the second part of the theorem.

Example 2. Consider the multinomial game given by $A(x, y) = (x_1 - y_1)^2 - (x_2 - y_2)^2$. The associated game $A_1(x_1, y_1) = (x_1 - y_1)^2$ with x_1 and y_1 in $[0, 1]$ has the unique optimal strategies x_1 equal to 0 or with 1 equiprobability and $y_1 = \frac{1}{2}$. Since the game $A_2(x_2, y_2) = -(x_2 - y_2)^2$ is the same with the players' roles reversed, and A is in a natural sense the sum of these games, the solution to the original game A is given by the randomized strategies $x = (0, \frac{1}{2})$ or $(1, \frac{1}{2})$ equiprobability and $y = (\frac{1}{2}, 0)$ or $(\frac{1}{2}, 1)$ equiprobability. The value of this symmetric game is of course zero (see Fig. 3).

In the solutions given in Fig. 3, no pure strategies (corners of the square) are used. We now compute player I's security level if he is restricted to using mixed strategies (concentrated on the four corners). Suppose player I uses the generic mixed strategy denoted by (p, q, r, s) , where these entries are the respective probabilities of the extreme points $(0, 0)$, $(1, 0)$, $(1, 1)$ and $(0, 1)$. If player II counters with the behavioral strategy (z, w) (that is, $y_1 = z$ and $y_2 = w$) then the expected payoff is computed as

$$\begin{aligned} & p(z^2 - w^2) + q((1 - z)^2 - w^2) + r((1 - z)^2 - (1 - w)^2) + s(z^2 - (1 - w)^2) \\ &= [pz^2 + q(1 - z)^2 + r(1 - z)^2 + sz^2] - [pw^2 + qw^2 + r(1 - w)^2 + s(1 - w)^2] \\ &= f(z) - g(w). \end{aligned}$$



$$\begin{aligned} \bar{\mu} &= \frac{1}{2}b_1 + \frac{1}{2}b'_1 \\ \bar{\mu}'' &= \frac{1}{2}b_1 + \frac{1}{2}b''_1 \\ B_1 &= B_1 = [0, 1]^2 \end{aligned}$$

FIG. 3

It follows that the minimizer II chooses z to minimize f and w to maximize g . Hence player II takes $z = q + r$ and w equal to 1 or 0 depending on whether $p + q$ is greater than or less than $r + s$. Assuming the former, we obtain without loss of generality that the expected value of A resulting from an optimal response to (p, q, r, s) is $(p + s)(q + r) - (p + q)$ which has maximal value $-\frac{1}{4}$ when all four corners are equally likely. Thus no mixed strategy is optimal.

Example 3. Our last example is a geometric game in the spirit of Ruckle [R]. A new town is to be built around a circular expressway. To minimize the cost of additional spur roads, new residents are fined according to the square of the distance that they choose to locate from the expressway. We consider the early stages where only two players, I and II, decide to live in this town. We assume that a player's fine is distributed among the remaining players. The final rub is that I hates II but II likes I, feelings which are related to utilities according to the square of the distance between them. More precisely, the expressway is centered at $(0, 0)$ and has radius r , and the two players must independently choose their respective positions x and y in the plane. The payoff to maximizer I is $A(x, y) = 2[d(y, S_r)]^2 - 2[d(x, S_r)]^2 + [d(x, y)]^2$. Here $d(y, S_r)$ denotes the Euclidean distance from the point y to the circle S_r representing the expressway. The squares are used to ensure that the kernel is a multinomial and the example was created so that the radial symmetry would produce optimal randomized strategies symmetrically distributed about the origin (and, in particular, with infinite nonconvex support).

Let $G(a, b)$ denote the average square of the distance from a point on S_a to a point on the concentric circle S_b . It is easily calculated that $G(a, b) = a^2 + b^2$. From the symmetry of the problem it is easy to see that the only real decision of each player is how far from the origin to locate. Let $F(a, b)$ denote the expected value of $A(x, y)$ if x and y are chosen uniformly over the circles of radius a and b , respectively. Then

$$\begin{aligned} F(a, b) &= 2(b - r)^2 - 2(a - r)^2 + G(a, b) \\ &= [3b^2 - 4rb] - [a^2 - 4ra]. \end{aligned}$$

Since $d/db[3b^2 - 4rb] = 6b - 4r$ and $d/da[a^2 - 4ra] = 2a - 4r$ it follows that the values $a = 2r$ and $b = 2r/3$ constitute the unique saddle point for $F(a, b)$. Hence for the original game the uniform distributions μ_1 and μ_2 over the concentric circles of radius $2r$ and $2r/3$, respectively, are optimal randomized strategies for players I and II. Consequently the value of the game is $F(2r, 2r/3) = 2r^2/9 - 2r^2 + 4r^2 + 4r^2/9 = 8r^2/3$.

The above analysis assumes that the strategies available to the players are all points in the plane, whereas for a multinomial game they must be restricted to the unit square. However if the circle is chosen with center at $(\frac{1}{2}, \frac{1}{2})$ and radius $r < \frac{1}{4}$ so that the concentric circle of radius $2r$ lies in the unit square, then the resulting multinomial game with strategies restricted to the square will have as optimal strategies scaled versions of the μ_1 and μ_2 given above. In Fig. 4 these optimal strategies are indicated by dotted circles.

We have not proven uniqueness so the question remains whether the players have optimal finite randomized strategies. To answer this we must use the following geometric lemma: *If y_0, \dots, y_{m-1} are $m > 1$ points equally spaced along the circle S_b then the average squared distance from any point x on the concentric circle S_a to the points y_0, \dots, y_{m-1} is $G(a, b)$.* It follows that any finite randomized strategy which gives equal weight to a set of equally spaced points along S_b is optimal for player II. In particular, any pair of equiprobable antipodal points on S_b constitutes an optimal randomized strategy for II, with a corresponding result holding for player I.

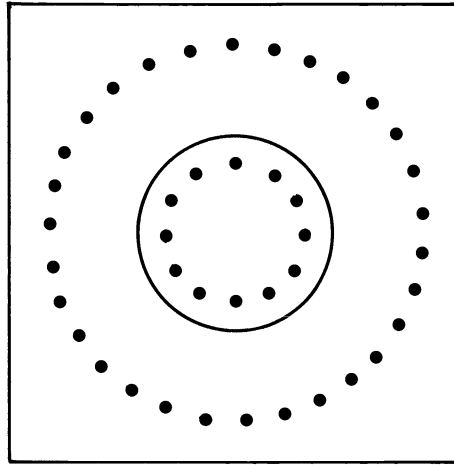


FIG. 4

The proof of the geometric lemma used in the above analysis is easy. Let $x = ae^{it}$ and for $j = 0, \dots, m-1$ let $y_j = be^{jq_i}$ where $q = 2\pi/m$ ($i^2 = -1$). The sum of the squared distances from x to y_j is given by

$$f(t) = \sum_{j=0}^{m-1} |be^{jq_i} - ae^{it}|^2 = \sum_{j=0}^{m-1} (be^{jq_i} - ae^{it})(be^{-jq_i} - ae^{-it}).$$

Hence

$$f'(t) = \sum_{j=0}^{m-1} [(be^{jq_i} - ae^{it})(ai e^{-it}) + (be^{-jq_i} - ae^{-it})(-ai e^{it})] = 0$$

because $\sum_{j=0}^{m-1} e^{jq_i} = \sum_{j=0}^{m-1} e^{-jq_i} = 0$. Thus the sum, $f(t)$, is constant. The average must also be a constant, which upon examination must be $G(a, b)$.

4. Sufficiency of finite randomized strategies. Let G be a two-person zero-sum game. According to Theorem 1 there are optimal strategies for each player consisting of probability distributions over their behavioral strategy spaces. In this section we further demonstrate that the players need to average over only a finite number of behavioral strategies, and give an upper bound for that finite number in terms of the complexity of the information structure of the game.

First suppose that G is a binary game. We define the multiplicity of an information set as the maximum number of times a game path may pass through it. (Thus a game is nonrepetitive if all information sets have multiplicity one.) Suppose player I has N information sets of multiplicities k_1, k_2, \dots, k_N and player II has M of multiplicities q_1, \dots, q_M . For each play of the game we consider the N -tuple (i_1, \dots, i_N) , where i_n is the number of times the n th information set of player I has been entered. Since each i_n satisfies $0 \leq i_n \leq k_n$ an upper bound on the number of possible N -tuples (i_1, \dots, i_N) is given by the number $K = \prod_{n=1}^N (k_n + 1)$ which we call the "informational multiplicity of player I." Similarly let $Q = \prod_{m=1}^M (q_m + 1)$ be the informational multiplicity of player II. If the game G is not in binary form, define its informational multiplicity, $K(G)$, as the minimum of $K(G')$ as G' varies over all binary games equivalent to G . Define $Q(G)$ similarly. The behavioral-normal form of the binary game G can be written in the form

$$A(x, y) = \sum_{i=1}^K \sum_{j=1}^Q a_{ij} r_i(x) s_j(y)$$

where $r_i(x)$ is of the form $x_1^{i_1} x_2^{i_2} \cdots x_N^{i_N}$ and $s_j(y)$ is of the form $y_1^{j_1} \cdots y_M^{j_M}$. This notation is explained in the previous section. The strategies x and y are chosen from $B_I = [0, 1]^N$ and $B_{II} = [0, 1]^M$, respectively.

THEOREM 3. *Let G be a two-person zero-sum game, where player I has informational multiplicity K . Then he has an optimal randomized strategy which uses at most $K + 1$ behavioral strategies.*

Proof. By Theorem 1 player I has an optimal randomized strategy, which we denote by $\bar{\mu}$. This means that

$$A(\bar{\mu}, y) = \int_{B_I} A(x, y) d\bar{\mu}(x) \geq V \quad \text{for all } y \text{ in } B_{II}$$

where V denotes the value of the game G . Actually we use the version of Theorem 1 appropriate for two-person zero-sum games. Let R denote the set of all points $r = (r_1, \cdots, r_K)$ of the form $r_i = \int_{B_I} r_i(x) d\mu(x)$, $i = 1, \cdots, K$, for some randomized strategy μ in B_I^* . Let C denote the subset of R determined by the point measures (behavioral strategies) μ_x in B_I^* . That is, $C = \{r: r_i = r_i(x), i = 1, \cdots, K, \text{ for some } x \text{ in } B_I\}$. It follows in a straightforward manner from the separating hyperplane theorem that R is the convex hull of C . This is stated in [K, Thm. 3.1.1], where a proof for $B_I = [0, 1]$ is given which applies equally well to all compact convex sets B_I .

Let \bar{r} be the point in R represented by the optimal randomized strategy $\bar{\mu}$: $\bar{r}_i = \int_{B_I} r_i(x) d\bar{\mu}(x)$, $i = 1, \cdots, K$. Since R is a subset of K -dimensional Euclidean space, every point in R is a convex sum of at most $K + 1$ extreme points. Hence we may write

$$\bar{r} = \sum_{p=1}^{K+1} z_p C^p$$

where the z_p are nonnegative numbers summing to one and each C^p in the set C has the form $C_i^p = r_i(x^p)$ for some behavioral strategy x^p in B_I . Define a finite randomized strategy μ' by the convex sum

$$\mu' = \sum_{p=1}^{K+1} z_p [x^p]$$

where $[x^p]$ denotes the point measure whose support is x^p . For $i = 1, \cdots, K$

$$\begin{aligned} \int_{B_I} r_i(x) d\mu'(x) &= \sum_{p=1}^{K+1} z_p r_i(x^p) = \bar{r}_i \\ &= \int_{B_I} r_i(x) d\bar{\mu}(x). \end{aligned}$$

In other words, the finite randomized strategy μ' has the same $K + 1$ generalized moments with respect to the r_i as the given optimal randomized strategy μ . It follows that for any behavioral strategy y in B_{II} ,

$$\begin{aligned} A(\mu', y) &= \int_{B_I} A(x, y) d\mu'(x) \\ &= \int_{B_I} \sum_{i=1}^K \sum_{j=1}^Q a_{ij} r_i(x) s_j(y) d\mu'(x) \\ &= \sum_{i=1}^K \sum_{j=1}^Q a_{ij} s_j(y) \int_{B_I} r_i(x) d\mu'(x) \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^K \sum_{j=1}^Q a_{ij} s_j(y) \int_{B_1} r_i(x) d\bar{\mu}(x) \\
&= \int_{B_1} (A(x, y) d\bar{\mu}(x) \\
&= A(\bar{u}, y).
\end{aligned}$$

This shows that any randomized strategy is in fact strategically equivalent to a finite randomized strategy, which in particular establishes the theorem.

For those interested in sharper bounds on the number of behavioral strategies one needs to average over, the theory of separable games as developed in [DKS] and [K] will prove useful. In particular it can be shown (using [K, Cor. 3.5.1]) that each player has a finite randomized strategy which involves at most $\min(K, Q)$ behavioral strategies. However, since our estimates which initially gave the numbers K and Q were themselves far from sharp, we have decided not to go into these details here.

5. Games with given optimal solutions. In the previous section we established (Theorem 3) that zero-sum two-person extensive form games have solution pairs consisting of finite randomized strategies. In this section we determine which such pairs are solutions of some game. To make this question more precise, define $\Omega(m, n)$ to be the class of all binary extensive form games (zero-sum, two person) in which the two players have m and n information sets, respectively. Let $F(j)$ denote the set of all finite convex combinations of point measures on $[0, 1]^j$. Then we can prove the following.

THEOREM 4. *For any pair (μ, ρ) of measures in $F(m) \times F(n)$ there is an extensive form game in $\Omega(m, n)$ such that μ and ρ are the unique optimal (randomized) strategies.*

Proof. Gale and Gross [GG, Thm. 2] showed how a multinomial function $A(x, y)$ can be explicitly constructed on $[0, 1]^m \times [0, 1]^n$ such that μ and ρ are the unique optimal (mixed) strategies for the abstract multinomial game. The explicit construction of Theorem 2 (of this paper) may then be applied to obtain an extensive form game in $\Omega(m, n)$ which has $A(x, y)$ as its behavioral-normal form. It now follows that μ and ρ , interpreted as finite randomized strategies, are uniquely optimal for the extensive form game.

REFERENCES

- [BO] T. BASAR AND G. J. OLSDER, *Dynamic Noncooperative Game Theory*, Academic Press, London, 1982.
- [DKS] M. DRESHER, S. KARLIN AND L. S. SHAPLEY, *Polynomial games*, in Contributions to the Theory of Games I, Ann. of Math. Stud. 24 (1950), pp. 161–180.
- [GG] D. GALE AND O. GROSS, *A note on polynomial and separable games*, Pacific J. Math., 8 (1958), pp. 735–741.
- [K] S. KARLIN, *Mathematical Methods and Theory in Games, Programming and Economics*, Vol. II, Pergamon Press, London, 1959.
- [KS] S. KARLIN AND L. S. SHAPLEY, *Geometry of moment spaces*, Mem. Amer. Math. Soc. (12), 1953.
- [O] G. OWEN, *Game Theory*, Academic Press, New York, 1982.
- [Ru] A. RUBINSTEIN, *Finite automata play the repeated prisoner's dilemma*, J. Econ. Theory, 39 (1986), pp. 83–96.
- [R] W. H. RUCKLE, *Geometric Games and Their Applications*, Pitman, Boston, 1983.

ON DIFFERENTIAL PURSUIT GAMES*

JIONGMIN YONG†

Abstract. A nonlinear differential pursuit game with a general closed terminal set is studied. Notions of various terminabilities are introduced. These notions include the notions of asymptotic stability, stabilizability and reachability of control systems. By using Lyapunov's direct method, some sufficient conditions for these terminabilities of the game are obtained. As a special case, a linear differential pursuit game with the terminal set as a subspace is discussed as well.

Key words. differential pursuit games, terminability, asymptotic terminability

AMS(MOS) subject classifications. 90D25, 90D26

1. Introduction. Consider a differential game governed by the following system:

$$(1.1) \quad \begin{aligned} \dot{z} &= f(z, u, v), \\ z(0) &= z_0, \end{aligned}$$

where $z \in \mathbb{R}^n$ is the state of the game and $z_0 \in \mathbb{R}^n$ is the initial state, $u \in U \subseteq \mathbb{R}^p$ is the pursuit control and $v \in V \subseteq \mathbb{R}^q$ is the evasion control, $f: \mathbb{R}^n \times U \times V \rightarrow \mathbb{R}^n$ is a given map and the dot above z means the derivative with respect to time t . A nonempty subset M of \mathbb{R}^n is also given. The game is terminated, if at some time $t_0 < \infty$, the state of the game hits M , namely, the trajectory $z(\cdot)$ of (1.1) corresponding to some pair of controls $u(\cdot)$ and $v(\cdot)$ satisfies

$$z(t_0) \in M.$$

Thus, M is called the terminal set of the game.

We denote, for $0 \leq a < b \leq +\infty$, that

$$\begin{aligned} \mathcal{U}[a, b] &= \{u: [a, b] \rightarrow U \mid u \text{ is measurable}\}, \\ \mathcal{V}[a, b] &= \{v: [a, b] \rightarrow V \mid v \text{ is measurable}\}, \\ \mathcal{Z}[a, b] &= \{z: [a, b] \rightarrow \mathbb{R}^n \mid z \text{ is absolutely continuous}\}. \end{aligned}$$

For $b < \infty$, we can define $\mathcal{U}[a, b]$, $\mathcal{V}[a, b]$ and $\mathcal{Z}[a, b]$ in a similar way.

In the game, the pursuer wants to terminate the game by choosing a suitable control $u(\cdot) \in \mathcal{U}[0, \infty)$, while the evader tries to prevent the game from terminating by selecting a proper control $v(\cdot) \in \mathcal{V}[0, \infty)$.

The pursuer is regarded as the "primary" player in a pursuit game, namely, the pursuer has some advantages in choosing his controls. To make it precise, let us introduce the following definition.

DEFINITION 1.1. Let $0 \leq a < b \leq +\infty$, and $U \subseteq \mathbb{R}^p$, $V \subseteq \mathbb{R}^q$ are given. A strategy S on $[a, b]$ is a map from $\mathcal{Z}[a, b] \times \mathcal{V}[a, b]$ to $\mathcal{U}[a, b]$, such that for any $t \in [a, b]$, $z(\cdot) \in \mathcal{Z}[a, b]$ and $v(\cdot) \in \mathcal{V}[a, b]$,

$$(1.2) \quad S(z(\cdot), v(\cdot))(\tau) = S(\hat{z}(\cdot), \hat{v}(\cdot))(\tau) \quad \text{a.e. } \tau \in [a, t],$$

whenever

$$(1.3) \quad \begin{aligned} z(\tau) &= \hat{z}(\tau) \quad \forall \tau \in [a, t], \\ v(\tau) &= \hat{v}(\tau) \quad \text{a.e. } \tau \in [a, t]. \end{aligned}$$

* Received by the editors March 12, 1986; accepted for publication (in revised form) July 17, 1987.

† Department of Mathematics, University of Texas, Austin, Texas 78712. Present address: Department of Mathematics, Fudan University, Shanghai, People's Republic of China.

Moreover, S is called admissible if for any $z_0 \in \mathbb{R}^n$, $v(\cdot) \in \mathcal{V}[a, b]$, there exist $z(\cdot) \in \mathcal{Z}[a, b]$, $u(\cdot) \in \mathcal{U}[a, b]$, such that

$$(1.4) \quad \begin{aligned} \dot{z}(t) &= f(z(t), u(t), v(t)) \quad \text{a.e. } t \in [a, b], \\ z(a) &= z_0, \end{aligned}$$

$$(1.5) \quad u(t) = S(z(\cdot), v(\cdot))(t) \quad \text{a.e. } t \in [a, b].$$

We call the control $u(\cdot)$ determined by (1.5) an outcome of the strategy S . We denote

$$\hat{\mathcal{U}}[a, b] = \{u(\cdot) \in \mathcal{U}[a, b] \mid u(\cdot) \text{ is an outcome of some admissible strategy } S \text{ on } [a, b]\}.$$

The pursuer will take his control from $\hat{\mathcal{U}}[0, \infty)$. We see that, roughly speaking, the pursuer can use the information that $\{z(s), v(s) \mid 0 \leq s \leq t\}$ when choosing the value $u(t)$ of the pursuit control $u(\cdot)$ at time t .

DEFINITION 1.2. A point $z_0 \in \mathbb{R}^n$ is said to be asymptotically terminable, if for any given $v(\cdot) \in \mathcal{V}[0, \infty)$, there exists a $u(\cdot) \in \hat{\mathcal{U}}[0, \infty)$, such that the corresponding trajectory $z(\cdot)$ of (1.1) satisfies

$$(1.6) \quad \lim_{t \rightarrow \infty} d(M, z(t)) = 0,$$

where $d(\cdot, \cdot)$ is the Euclidean distance in \mathbb{R}^n . If (1.6) is replaced by the following:

$$(1.7) \quad z(t_0) \in M \quad \text{for some } t_0 \in [0, \infty),$$

we call z_0 terminable.

We denote

$$C_a(M) = \{z_0 \in \mathbb{R}^n \mid z_0 \text{ is asymptotically terminable}\},$$

$$C(M) = \{z_0 \in \mathbb{R}^n \mid z_0 \text{ is terminable}\}.$$

Trivially, we have

$$(1.8) \quad M \subseteq C(M) \subseteq C_a(M).$$

DEFINITION 1.3. The game (1.1) with terminal set M is said to be:

(i) Locally asymptotically terminable, if there exists a neighborhood \mathcal{N} of M , such that

$$(1.9) \quad \mathcal{N} \subset C_a(M).$$

(ii) s -locally asymptotically terminable, if there exists a $\delta > 0$, such that

$$(1.10) \quad G_0^\delta(M) = \{z \in \mathbb{R}^n \mid 0 \leq d(M, z) < \delta\} \subset C_a(M).$$

(iii) Locally terminable, if there exists a neighborhood \mathcal{N} of M , such that

$$(1.11) \quad \mathcal{N} \subset C(M).$$

(iv) s -locally terminable, if there exists a $\delta > 0$, such that

$$(1.12) \quad G_0^\delta(M) \subset C(M).$$

(v) (Globally) asymptotically terminable, if

$$(1.13) \quad C_a(M) = \mathbb{R}^n.$$

(vi) (Globally) terminable, if

$$(1.14) \quad C(M) = \mathbb{R}^n.$$

Remark 1.4. The following implications about the above definitions are clear:

$$(1.15) \quad \begin{array}{ccccc} (vi) & \Rightarrow & (iv) & \Rightarrow & (iii) \\ \Downarrow & & \Downarrow & & \Downarrow \\ (v) & \Rightarrow & (ii) & \Rightarrow & (i) \end{array}$$

Also, if M is bounded and closed, then

$$(1.16) \quad (i) \Leftrightarrow (ii), \quad (iii) \Leftrightarrow (iv).$$

The problem of various terminabilities introduced above for differential pursuit game (1.1) with general terminal set M is closely related to many control problems and stability problems. First, we can easily see that if both U and V are singletons, then we can refer (i), (ii) and (v) to be the local, s -local and (global) asymptotic stabilities, respectively, of the system to set M . See [2] for similar notions. Second, if V is a singleton, then our problem is just the usual control problem. We can refer (i), (ii) and (v) to be the local, s -local and (global) asymptotic stabilizabilities, respectively, of the system to set M , and (iii), (iv) and (vi) to be the local, s -local and (global) reachabilities, respectively, of the system to set M . See [1], [4], [9] for some similar notions. Finally, regarding $v(\cdot)$ as the uncertain parameter, our problem is related to the problem of stabilizability or reachability of dynamic systems with uncertainty. For relevant results, see [3], [6], [16], for example, and the references cited therein.

The problem of differential pursuit games, especially the linear differential games with terminal set M being a subspace of \mathbb{R}^n has been substantially studied (see [5], [11]–[14] and [18], for example). The purpose of this paper is to give a (unified) way of treating the nonlinear differential pursuit games with general closed terminal set M using Lyapunov's direct method. Thus, our approach is different from those used in [5], [11]–[14]. We will also discuss the case that M is a convex and closed set, in particular the case that M is a subspace and the system is linear.

2. Preliminaries. We denote the Euclidean norm and inner product in \mathbb{R}^n by $|\cdot|$ and (\cdot, \cdot) , respectively. For any smooth function $\Phi(z)$ defined on a domain of \mathbb{R}^n , we denote its gradient by $(\partial\Phi/\partial z)(z)$.

Now, let us make some assumptions.

(A1) The sets $U \subseteq \mathbb{R}^p$ and $V \subseteq \mathbb{R}^q$ are compact. The function $f(z, u, v)$ is continuous in $(z, u, v) \in \mathbb{R}^n \times U \times V$, and there exists a constant $K > 0$, such that

$$(2.1) \quad |f(z, u, v) - f(\hat{z}, u, v)| \leq K|z - \hat{z}| \quad \forall z, \hat{z} \in \mathbb{R}^n, \quad (u, v) \in U \times V.$$

(A2) The set $M \subseteq \mathbb{R}^n$ is nonempty, not equal to \mathbb{R}^n and closed.

(A3) The set M is convex.

Remark 2.1. If (A1) holds and we set hereafter,

$$(2.2) \quad K_1 = K + \max_{(u,v) \in U \times V} |f(0, u, v)|,$$

then, by (2.1), we have

$$(2.3) \quad |f(z, u, v)| \leq K_1(1 + |z|) \quad \forall (z, u, v) \in \mathbb{R}^n \times U \times V.$$

LEMMA 2.2. Suppose (A1) holds. Then for any $u(\cdot) \in \mathcal{U}[0, \infty)$, $v(\cdot) \in \mathcal{V}[0, \infty)$ and $z_0 \in \mathbb{R}^n$, the trajectory $z(\cdot)$ of (1.1) satisfies

$$(2.4) \quad |z(t)| \leq (1 + |z_0|) e^{K_1 t} - 1 \quad \forall t \geq 0,$$

$$(2.5) \quad \max_{(u,v) \in U \times V} |f(z(t), u, v)| \leq K_1(1 + |z_0|) e^{K_1 t} \quad \forall t \geq 0.$$

The proof is straightforward.

Next, let M be a convex and closed subset of \mathbb{R}^n , nonempty, not equal to \mathbb{R}^n . We let ∂M be the boundary of M . For any $x \in \partial M$, we say that a vector

$$\eta \in S^n \equiv \{\eta \in \mathbb{R}^n \mid \|\eta\| = 1\}$$

is an outer normal of M at x if

$$(y - x, \eta) \leq 0 \quad \forall y \in M.$$

We define for any $x \in \partial M$ that

$$N(x) = \{\eta \in S^n \mid \eta \text{ is an outer normal of } M \text{ at } x\}.$$

Then, by the supporting hyperplane theorem, we have that $N(x) \neq \emptyset$ for all $x \in \partial M$. Also, it is clear that, in general, $N(x)$ may contain more than one point. On the other hand, we know that for any $z \in \mathbb{R}^n \setminus M$, there exists a unique $x \in \partial M$, such that

$$(2.6) \quad |z - x| = d(M, z).$$

Thus, the above defines a map $x: \mathbb{R}^n \setminus M \rightarrow \partial M$. About this map, we have the following lemmas.

LEMMA 2.3. For all $z, \hat{z} \in \mathbb{R}^n \setminus M$,

$$(2.7) \quad |x(z) - x(\hat{z})| \leq |z - \hat{z}|.$$

For the proof, see § 32 of [7].

LEMMA 2.4. If we define $\Psi(z) = |z - x(z)|^2$, for $z \in \mathbb{R}^n$, then

$$(2.8) \quad \frac{\partial \Psi}{\partial z}(z) = 2(z - x(z)) \quad \forall z \in \mathbb{R}^n.$$

LEMMA 2.5. For any $z \in \mathbb{R}^n \setminus M$, we define

$$(2.9) \quad \theta(z) = \frac{z - x(z)}{|z - x(z)|}.$$

Then, for any $z, \hat{z} \in G_s^\infty(M) \equiv \{z \in \mathbb{R}^n \mid s \leq d(M, z) < \infty\}$ ($s > 0$)

$$(2.10) \quad |\theta(z) - \theta(\hat{z})| \leq \frac{3}{s} |z - \hat{z}|.$$

The proof of Lemmas 2.4 and 2.5 are straightforward (see [18]).

Now, we return to the case with terminal set M simply being closed. Let $z_0 \in C(M)$. We define

$$\begin{aligned} \mathcal{T}_M(z_0) = \{t \geq 0 \mid \forall v(\cdot) \in \mathcal{V}[0, \infty), \exists u(\cdot) \in \hat{\mathcal{U}}[0, \infty) \\ \text{such that } z(s) \in M, \text{ for some } s \leq t\}, \end{aligned}$$

$$T_M(z_0) = \inf \mathcal{T}_M(z_0).$$

In the case where there is no confusion, we simply denote $\mathcal{T}(z_0) = \mathcal{T}_M(z_0)$ and $T(z_0) = T_M(z_0)$. We call $T(z_0)$ the terminal time of z_0 .

3. Various terminabilities. In this section, we prove the main results of this paper. We divide this section into several subsections.

3.1. Local and s -local terminabilities. We define (cf. (1.6)), for $0 \leq s < \delta \leq +\infty$,

$$G_s^\delta(M) \triangleq \{z \in \mathbb{R}^n \mid s \leq d(M, z) < \delta\}.$$

Our first result is concerned with the s -local terminability of the game.

THEOREM 3.1. Suppose (A1) and (A2) hold. Let there exist constants $\delta, \delta_0, C > 0$ and continuous functions $\omega: (0, \delta] \rightarrow \mathbb{R}^+$, $\Phi: G_0^\delta(M) \rightarrow \mathbb{R}^+$, such that

$$(3.1) \quad \begin{aligned} &\Phi(z) \text{ is nondecreasing with respect to } d(M, z), z \in G_0^\delta(M), \text{ i.e., } \Phi(z) \geq \Phi(\hat{z}), \\ &\text{whenever } d(M, z) \geq d(M, \hat{z}), \text{ for all } z, \hat{z} \in G_0^\delta(M), \end{aligned}$$

$$(3.2) \quad \Phi|_M = 0, \quad \Phi(z) > 0 \quad \forall z \in G_0^\delta(M) \setminus M,$$

$$(3.3) \quad \left| \frac{\partial \Phi}{\partial z}(z) \right| \leq C \quad \forall z \in G_0^\delta(M) \setminus M,$$

$$(3.4) \quad \left| \frac{\partial \Phi}{\partial z}(z) - \frac{\partial \Phi}{\partial z}(\hat{z}) \right| \leq \omega(s)|z - \hat{z}| \quad \forall z, \hat{z} \in G_s^\delta(M), \quad 0 < s < \delta,$$

$$(3.5) \quad \sup_{v \in V} \inf_{u \in U} \left(\frac{\partial \Phi}{\partial z}(z), f(z, u, v) \right) \leq -\delta_0 \quad \forall z \in G_0^\delta(M) \setminus M.$$

Then, the game (1.1) with terminal set M is s -locally terminable with

$$(3.6) \quad G_0^\delta(M) \subset C(M).$$

Moreover, for any $z \in G_0^\delta(M)$, the terminal time

$$(3.7) \quad T(z) \leq \frac{\Phi(z)}{\delta_0}.$$

Proof. Let $z_0 \in G_0^\delta(M) \setminus M$, $0 < \lambda < 1$ and $v(\cdot) \in \mathcal{V}[0, \infty)$ be given. We define the following functions:

$$\begin{aligned} \tau(d) &= \max \{1, d\} \quad \forall d \in \mathbb{R}^+ \equiv [0, \infty), \\ F(z, \tau) &= K_1(1 + |z|) e^{K_1 \tau} \quad \forall (z, \tau) \in \mathbb{R}^n \times \mathbb{R}^+, \\ t(d, F) &= \min \left\{ \frac{(1-\lambda)d}{1+F}, \frac{(1-\lambda)\delta_0}{\omega(\lambda d)F^2 + CKF} \right\} \quad \forall (d, F) \in \mathbb{R}^+ \times \mathbb{R}^+, \end{aligned}$$

where K_1 is determined by (2.2). We denote $d_0 = d(M, z_0)$, $F_0 = F(z_0, \tau(d_0))$, $t_0 = t(d_0, F_0)$. Let $u(\cdot) \in \hat{\mathcal{U}}[0, t_0]$ be such that

$$(3.8) \quad \left(\frac{\partial \Phi}{\partial z}(z_0), f(z_0, u(t), v(t)) \right) \leq -\delta_0 \quad \text{a.e. } t \in (0, t_0).$$

Then, we claim that for the trajectory $z(\cdot)$ of (1.1) corresponding to $u(\cdot)$ and $v(\cdot)$, we have

$$(3.9) \quad z(t) \in G_{\lambda d_0}^\delta(M), \quad t \in [0, t_0].$$

In fact, since $t_0 \leq (1-\lambda)d_0/(1+F_0) \leq d_0 \leq \tau(d_0)$, by (2.5), we have

$$(3.10) \quad d(M, z(t)) \geq d(M, z_0) - F_0 t_0 \geq \lambda d_0 \quad \forall t \in [0, t_0].$$

On the other hand, if (3.9) were not true, then there would exist $\hat{t} \in [0, t_0]$, such that

$$(3.11) \quad d(M, z(\hat{t})) = \delta, \quad d(M, z(t)) < \delta \quad \forall t \in [0, \hat{t}).$$

Then, by (3.8), (3.10) and our assumptions, we have

$$\begin{aligned} \frac{d}{dt} \Phi(z(t)) &= \left(\frac{\partial \Phi}{\partial z}(z_0), f(z_0, u(t), v(t)) \right) \\ &\quad + \left(\frac{\partial \Phi}{\partial z}(z(t)) - \frac{\partial \Phi}{\partial z}(z_0), f(z(t), u(t), v(t)) \right) \\ &\quad + \left(\frac{\partial \Phi}{\partial z}(z_0), f(z(t), u(t), v(t)) - f(z_0, u(t), v(t)) \right) \\ &\leq -\delta_0 + \omega(\lambda d_0) F_0^2 t + CK F_0 t \\ &\leq -\lambda \delta_0 \quad \forall t \in [0, \hat{t}). \end{aligned}$$

Hence, it follows that

$$(3.12) \quad \Phi(z(t)) \leq \Phi(z_0) - \lambda \delta_0 t, \quad t \in [0, \hat{t}).$$

By (3.1), we then have that

$$d(M, z(t)) \leq d(M, z_0) < \delta \quad \forall t \in [0, \hat{t}).$$

Thus, we end up with a contradiction to (3.11). Hence (3.9) holds. Then, from the above argument, we obtain

$$\Phi(z(t_0)) \leq \Phi(z_0) - \lambda \delta_0 t_0.$$

Set $z_1 = z(t_0)$, $d_1 = d(M, z_1)$. Repeating the above procedure, we have

$$z(t) \in G_{\lambda d_1}^\delta(M), \quad t \in [t_0, t_1],$$

$$\Phi(z(t_1)) \leq \Phi(z_1) - \lambda \delta_0(t_1 - t_0)$$

for $t_1 = t_0 + t(d_1, F_1)$, $F_1 = F(z_1, \tau(d_1))$. In general, we have, for some $u(\cdot) \in \hat{\mathcal{U}}[0, \infty)$, that for $k = 0, 1, 2, \dots$

$$z_k = z(t_{k-1}), \quad z_0 \text{ given,}$$

$$(3.13) \quad d_k = d(M, z_k), \quad F_k = F(z_k, \tau(d_k)), \quad t_k = t_{k-1} + t(d_k, F_k),$$

$$z(t) \in G_{\lambda d_k}^\delta(M), \quad t \in [t_{k-1}, t_k],$$

$$(3.14) \quad \Phi(z(t_k)) \leq \Phi(z_k) - \lambda \delta_0(t_k - t_{k-1}) \leq \dots \leq \Phi(z_0) - \lambda \delta_0 t_k.$$

Then, from (3.14), we see that (note $\Phi \geq 0$)

$$t_k \leq \frac{\Phi(z_0)}{\lambda \delta_0} \quad \forall k \geq 0.$$

Thus, we can assume that

$$(3.15) \quad \lim_{k \rightarrow \infty} t_k = \tilde{t} \in \left[0, \frac{\Phi(z_0)}{\lambda \delta_0}\right].$$

By (3.1), we see that $\{d_k\}$ is nonincreasing. Thus,

$$\lim_{k \rightarrow \infty} d_k = \tilde{d} \geq 0$$

exists. We claim that $\tilde{d} = 0$. Suppose the contrary, i.e., $\tilde{d} > 0$. Then, we let

$$\tilde{\omega} = \max \{ \omega(\gamma) \mid \gamma \in [\lambda \tilde{d}, \delta] \} < \infty,$$

$$\tilde{F} = F(z_0, \tilde{t}) < \infty.$$

Then, it is clear that for any $k \geq 0$,

$$t(d_k, F_k) \geq \min \left\{ \frac{(1-\lambda)\tilde{d}}{1+\tilde{F}}, \frac{(1-\lambda)\delta_0}{\tilde{\omega}\tilde{F}^2 + CK\tilde{F}} \right\} > 0.$$

This is a contraction to (3.15). Thus $\tilde{d} = 0$, and

$$d(M, z(\tilde{t})) = \lim_{k \rightarrow \infty} d(M, z(t_k)) = 0,$$

i.e., $z(\hat{t}) \in M$. Thus obtain (3.6). Finally, from (3.15), we have

$$T(z_0) \leq \frac{\Phi(z_0)}{\lambda \delta_0}.$$

Sending $\lambda \rightarrow 1$, we get (3.7). \square

Remark 3.2. It is easy to see that in the above proof, only some local arguments have been used. Thus, some of the conditions in Theorem 3.1 can be slightly relaxed. For example, the constant C in (3.3) can be replaced by a continuous function $C: G_0^\delta(M) \rightarrow \mathbb{R}^+$. The relaxation of condition (3.5) is a little interesting. To state the relaxed condition, let us set

$$(3.16) \quad \sigma(R) = - \sup_{\substack{z \in G_0^\delta(M) \setminus M \\ |z| \leq R}} \sup_{v \in V} \inf_{u \in U} \left(\frac{\partial \Phi}{\partial z}(z), f(z, u, v) \right), \quad R > 0.$$

Then, the relaxed conditions states as follows: There exists a continuous function $\sigma_0: G_0^\delta(M) \rightarrow [0, \infty)$ such that

$$(3.5') \quad -\sigma((1+|z|) e^{K_1(\Phi(z)/\sigma_0(z))} - 1) \leq -\sigma_0(z) \quad \forall z \in G_0^\delta(M)$$

where K_1 is determined by (2.2). It is clear that the proof of Theorem 3.1 applies if we replace δ_0 by $\sigma_0(z_0)$, and (3.5) implies (3.5') with $\sigma_0(z) \equiv \delta_0$. The above idea leads to the following result.

THEOREM 3.3. *Let \mathcal{N} be a neighborhood of M . Suppose all the assumptions of Theorem 3.1, except (3.1) and (3.5), hold with $G_0^\delta(M)$ replaced by \mathcal{N} . Let $\sigma(R)$ be defined as in (3.16) with $G_0^\delta(M)$ replaced by \mathcal{N} and*

$$(3.5'') \quad \sigma(R) < 0 \quad \forall R > 0.$$

Then the game (1.1) with terminal set M is locally terminable.

Proof. We define

$$(3.17) \quad \rho(z) = \sup \{ \rho > 0 \mid \mathcal{O}(z, \rho) \subset \mathcal{N} \}, \quad \mathcal{O}(z, \rho) \equiv \{ \hat{z} \in \mathbb{R}^n \mid |z - \hat{z}| < \rho \},$$

$$(3.18) \quad \mathcal{N}_0 = \left\{ z \in \mathcal{N} \mid \rho(z) > \frac{2\Phi(z)}{\sigma(|z|)}, -\sigma((1+|z|) e^{K_1(2\Phi(z)/\sigma(|z|))} - 1) \leq -\frac{\sigma(|z|)}{2} \right\}.$$

Then, it is not hard to see that \mathcal{N}_0 contains an open neighborhood of M . We claim that $\mathcal{N}_0 \subset C(M)$. In fact, for any $z_0 \in \mathcal{N}_0 \setminus M$, by arguments similar to those used in the proof of Theorem 3.1, we can get $z_0 \in C(M)$ and

$$T(z_0) \leq \frac{2\Phi(z_0)}{\sigma(|z_0|)}.$$

Hence, by definition, the game is locally terminable. \square

We should note that in the proof of $d_k \rightarrow 0$, we do not need the monotonicity of $\Phi(\cdot)$. The monotonicity of $\Phi(\cdot)$ in Theorem 3.1 is only needed to guarantee (3.9). While in the proof of Theorem 3.3, by letting $z_0 \in \mathcal{N}_0$, we automatically have $z(t) \in \mathcal{N}$ for $t \leq 2\Phi(z_0)/\sigma(|z_0|)$. Thus, we do not need the monotonicity of $\Phi(\cdot)$ in Theorem 3.3.

Now, let us give some consequences of the above.

COROLLARY 3.4. *Suppose (A1)–(A3) hold. Suppose there exists a $\delta_0 > 0$, such that*

$$(3.19) \quad \sup_{\substack{\eta \in N(x) \\ x \in \partial M}} \sup_{v \in V} \inf_{u \in U} (\eta, f(x, u, v)) \leq -\delta_0.$$

Then, the game (1.1) with terminal set M is s -locally terminable, with

$$(3.20) \quad G_0^{\delta_0/K}(M) \subset C(M).$$

Moreover, for any $z \in G_0^{\delta_0/K}(M)$,

$$(3.21) \quad T(z) \leq \frac{Kd(M, z)}{\delta_0 - Kd(M, z)}.$$

Proof. Let

$$\Phi(z) = d(M, z) \equiv |z - x(z)| \quad \forall z \in \mathbb{R}^n.$$

Then, by Lemma 2.4,

$$\frac{\partial \Phi}{\partial z}(z) = \theta(z) \equiv \frac{z - x(z)}{|z - x(z)|} \in N(x(z)).$$

Thus, (3.2)–(3.4) hold with $C = 1$ and $\omega(s) = 3/s$ (note Lemma 2.5). Next,

$$\begin{aligned} \sup_{v \in V} \inf_{u \in U} \left(\frac{\partial \Phi}{\partial z}(z), f(z, u, v) \right) &= \sup_{v \in V} \inf_{u \in U} \{ (\theta(z), f(x(z), u, v)) \\ &\quad + (\theta(z), f(z, u, v) - f(x(z), u, v)) \} \\ &\leq -\delta_0 + Kd(M, z). \end{aligned}$$

Hence, for any $\varepsilon \in (0, \delta_0/K)$, we have

$$\sup_{v \in V} \inf_{u \in U} \left(\frac{\partial \Phi}{\partial z}(z), f(z, u, v) \right) \leq -(\delta_0 - \varepsilon) < 0 \quad \forall z \in G_0^\varepsilon(M).$$

Thus, by Theorem 3.1, $G_0^\varepsilon(M) \subset C(M)$. Sending $\varepsilon \rightarrow \delta_0/K$, we get (3.20). Finally, (3.21) follows from (3.7) and the above arguments. \square

COROLLARY 3.5. *Suppose (A1)–(A3) hold. Suppose*

$$(3.22) \quad \sup_{\eta \in N(x)} \sup_{v \in V} \inf_{u \in U} (\eta, f(x, u, v)) < 0 \quad \forall x \in \partial M.$$

Then, the game (1.1) with terminal set M is locally terminable.

Proof. Since $(\eta, f(x, u, v))$ is Lipschitz continuous in x , uniformly in η, u, v , we obtain that

$$\sup_{\substack{|x| \leq R \\ x \in \partial M}} \sup_{\eta \in N(x)} \sup_{v \in V} \inf_{u \in U} (\eta, f(x, u, v)) < 0 \quad \forall R > 0.$$

Then, as in the proof of Corollary 3.4, by using Theorem 3.3, we can get the conclusion.

3.2. Local and s -local asymptotic terminability. We note that if $f(z, u, v) = 0$ for $z \in M$, then all the results of § 3.1 do not apply. In this section, we will discuss the situation allowing $f(z, u, v) = 0$ for $z \in M$. The Lyapunov functions we are going to use in this section will be different from those in § 3.1.

THEOREM 3.6. *Suppose (A1) and (A2) hold. Let there exist a constant $\varepsilon > 0$, a neighborhood \mathcal{N} of M and continuous functions $\mu: \mathbb{R}^+ \rightarrow \mathbb{R}^+$, $\Psi: \mathcal{N} \rightarrow \mathbb{R}^+$, such that*

$$(3.23) \quad \mu \text{ is nondecreasing, } \mu(0) = 0, \quad \mu(\gamma) > 0 \quad \forall \gamma > 0,$$

$$(3.24) \quad \Psi(z) > 0 \quad \forall z \in \mathcal{N} \setminus M, \quad \Psi|_M = 0,$$

$$(3.25) \quad \mathcal{N}_\varepsilon \equiv \{z \in \mathbb{R}^n \mid \Psi(z) \leq \varepsilon\} \subset \mathcal{N},$$

$$(3.26) \quad \frac{\partial \Psi}{\partial z}(\cdot) \in C(\overline{\mathcal{N} \setminus M}),$$

$$(3.27) \quad \sup_{v \in V} \inf_{u \in U} \left(\frac{\partial \Psi}{\partial z}(z), f(z, u, v) \right) \leq -\mu(d(M, z)) \quad \forall z \in \mathcal{N} \setminus M.$$

Then, the game (1.1) with terminal set M is locally asymptotically terminable with $\mathcal{N}_\varepsilon \subset C_a(M)$. Moreover, suppose instead of (3.25); we have, for some $\delta > 0$, that

$$(3.25') \quad G_0^\delta(M) \subset \mathcal{N}_\varepsilon \subset \mathcal{N}.$$

Then, the game (1.1) with terminal set M is s -locally asymptotically terminable with $G_0^\delta(M) \subset C_a(M)$.

Proof. Let $z_0 \in \mathcal{N}_\varepsilon \setminus M$ and $v(\cdot) \in \mathcal{V}[0, \infty)$ be given. Denote

$$d_0 = d(M, z_0), \quad \tau_0 = \frac{2\Psi(z_0)}{\mu(d_0/2)}.$$

We claim that there exist a $t_0 \in [0, \tau_0]$ and a $u(\cdot) \in \hat{\mathcal{U}}[0, t_0]$, such that the trajectory $z(\cdot)$ of (1.1) corresponding to $u(\cdot)$ and $v(\cdot)$ satisfies

$$(3.28) \quad d(M, z(t_0)) \leq \frac{d_0}{2}.$$

Suppose it is not the case. Then, we have for any $u(\cdot) \in \hat{\mathcal{U}}[0, \infty)$,

$$(3.29) \quad d(M, z(t)) > \frac{d_0}{2} \quad \forall t \in [0, \tau_0].$$

Now, we set

$$F_0 = K_1(1 + |z_0|) e^{K_1 \tau_0}.$$

Let $\omega_0(\cdot)$ be the modulus of continuity of $(\partial\Psi/\partial z)(z)$ on the set $\{z \in \overline{\mathcal{N} \setminus M} \mid |z| \leq (1 + |z_0|) e^{K_1 \tau_0} - 1\}$. Then, we define

$$(3.30) \quad \hat{t} = \sup \left\{ t \leq \tau_0 \mid \omega_0\left(\frac{d_0 F_0 t}{2}\right) F_0 + CK F_0 t \leq \frac{1}{2} \mu\left(\frac{d_0}{2}\right) \right\} > 0.$$

Then, similar to the proof of Theorem 3.1, we can find a $u(\cdot) \in \hat{\mathcal{U}}[0, \hat{t}]$, such that $z(t) \in \mathcal{N}$ for $t \in [0, \hat{t}]$ due to (3.25) and

$$\Psi(z(\hat{t})) \leq \Psi(z_0) - \frac{1}{2} \mu\left(\frac{d_0}{2}\right) \hat{t}.$$

Thus, the procedure can be repeated. Since \hat{t} only depends on τ_0 and z_0 , noting (3.29), we see that by repeating finitely many times of the above procedure, we will have the following:

$$\Psi(z(\tau_0)) \leq \Psi(z_0) - \frac{1}{2} \mu\left(\frac{d_0}{2}\right) \tau_0 = 0.$$

Thus, $z(\tau_0) \in M$ which contradicts (3.29). Hence (3.28) holds. Then, replacing z_0 by $z_1 \equiv z(t_0)$, we can repeat the above arguments. By induction, we see that $z_0 \in C_a(M)$. Hence, we obtain the local asymptotic terminability. It is clear that if we have (3.25'), then the above argument gives us that $G_0^\delta(M) \subset C_a(M)$. Hence, we have the s -local asymptotic terminability of the game. \square

The following result is well known (see [2, p. 135], for example).

COROLLARY 3.7. *Suppose (A1) and (A2) hold with U and V being singletons and M being bounded. Let there exist a neighborhood \mathcal{N} of M and a continuous function $\Psi: \mathcal{N} \rightarrow \mathbb{R}^+$ satisfying (3.24) and (3.26). Also, we assume that (let $f(z) \equiv f(z, U, V)$)*

$$(3.27') \quad \left(\frac{\partial\Psi}{\partial z}(z), f(z) \right) < 0 \quad \forall z \in \mathcal{N} \setminus M.$$

Then, the system (1.1) with terminal set M is (locally) asymptotically stable.

Proof. Since M is compact, we can easily define a continuous function $\mu(\cdot)$ satisfying (3.23) and

$$(3.27'') \quad \left(\frac{\partial \Psi}{\partial z}(z), f(z) \right) \leq -\mu(d(M, z)) \quad \forall z \in \mathcal{N} \setminus M.$$

Also, we can find $\varepsilon, \delta > 0$, such that (3.25') holds. Then Theorem 3.6 applies. \square

Remark 3.8. In the case where V is a singleton, Theorem 3.6 gives a sufficient condition for stabilizing system (1.1) to set M . In particular, for M being compact, we have a result similar to Corollary 3.7 for the stabilization of the system (1.1) to M .

The next result says that if in Theorem 3.6, $\Psi(z)$ goes to zero (as $d(M, z) \rightarrow 0$) much faster than $\mu(d(M, z))$ does, then we can have local (s -local) terminability of the game.

COROLLARY 3.9. *Suppose (A1) and (A2) hold. Let there exist a constant $\varepsilon > 0$, a neighborhood \mathcal{N} of M and continuous functions $\beta, \mu: \mathbb{R}^+ \rightarrow \mathbb{R}^+, \Psi: \mathcal{N} \rightarrow \mathbb{R}^+$, such that (3.23)–(3.27) hold. $\beta(\cdot)$ also satisfies (3.23) and*

$$(3.31) \quad \Psi(z) \leq \beta(d(M, z)) \quad \forall z \in \mathcal{N}.$$

In addition, we assume that there exists a sequence of decreasing positive numbers $\{d_k\}$ such that

$$(3.32) \quad \lim_{k \rightarrow \infty} d_k = 0, \quad d_0 = \varepsilon,$$

$$(3.33) \quad \sum_{k=0}^{\infty} \frac{\beta(d_k)}{\mu(d_{k+1})} < \infty.$$

Then, the game (1.1) with terminal set M is locally terminable with $\mathcal{N}_\varepsilon \subset C(M)$, and for any $z_0 \in \mathcal{N}_\varepsilon$, if $d_{k+1} < d(M, z_0) \leq d_k$, then

$$(3.34) \quad T(z_0) \leq 2 \sum_{i=k}^{\infty} \frac{\beta(d_i)}{\mu(d_{i+1})}.$$

Furthermore, if there exists a $\delta > 0$, such that (3.25') holds, then the game (1.1) with terminal set M is s -locally terminable with $G_0^\delta(M) \subset C(M)$.

Proof. Let $z_0 \in \mathcal{N}_\varepsilon \setminus M$, $v(\cdot) \in \mathcal{V}[0, \infty)$ be given. Let

$$\tau_0 = \frac{2\Psi(z_0)}{\mu(d_1)}.$$

Then, as in the proof of Theorem 3.6, we have that there exist $t_0 \in [0, \tau_0]$ and $u(\cdot) \in \hat{\mathcal{U}}[0, t_0]$, such that

$$d(M, z(t_0)) \leq d_1.$$

Set $z_1 = z(t_0)$, and repeat the procedure. In general, we have ($t_{-1} \triangleq 0$)

$$z_k = z(t_{k-1}), \quad \tau_k = \frac{2\Psi(z_k)}{\mu(d_{k+1})}, \quad t_k = t_{k-1} + \tau_k,$$

$$d(M, z_k) \leq d_k, \quad k = 0, 1, 2, \dots$$

Then, by (3.31) and (3.33), we have

$$t_k = 2 \sum_{i=0}^k \frac{\Psi(z_i)}{\mu(d_{i+1})} \leq 2 \sum_{i=0}^{\infty} \frac{\beta(d_i)}{\mu(d_{i+1})} < \infty.$$

Hence, $\lim_{k \rightarrow \infty} t_k$ exists and it is easy to see that $z_0 \in C(M)$. The remaining assertions follow immediately. \square

To close this section, we give some consequences of the above results for the case that M is convex.

COROLLARY 3.10. *Suppose (A1)–(A3) hold. Let there exist a constant $\delta > 0$, a continuous function $\hat{\mu} : \mathbb{R}^+ \rightarrow \mathbb{R}^+$, satisfying (3.23), such that*

$$(3.35) \quad \sup_{v \in V} \inf_{u \in U} (\theta(z), f(z, u, v)) \leq -\hat{\mu}(d(M, z)) \quad \forall z \in G_0^\delta(M) \setminus M,$$

where $\theta(z) \equiv (z - x(z))/(|z - x(z)|)$. (See § 2.) Then, the game (1.1) with terminal set M is s -locally asymptotically terminable with $G_0^\delta(M) \subset C(M)$. Moreover, if there exist $c > 0$, $0 < \alpha < 1$, such that

$$(3.36) \quad \hat{\mu}(\gamma) \geq c\gamma^\alpha \quad \forall \gamma \in [0, \delta],$$

then the game (1.1) with terminal set M is s -locally terminable with $G_0^\delta(M) \subset C(M)$.

Proof. Let $\Psi(z) = d(M, z)^2 \equiv |z - x(z)|^2$. Then, (3.24) and (3.26) are satisfied. By (3.35), we have (note Lemma 2.4)

$$\sup_{v \in V} \inf_{u \in U} \left(\frac{\partial \Psi}{\partial z}(z), f(z, u, v) \right) \leq -d(M, z)\hat{\mu}(d(M, z)) \quad \forall z \in G_0^\delta(M).$$

Thus, by Theorem 3.6, we have $G_0^\delta(M) \subset C_a(M)$.

Now, if (3.36) holds, then we let

$$d_k = \frac{1}{k^m}, \quad m > \frac{1}{1-\alpha}, \quad k \geq k_0.$$

Then, with $\beta(d) = d^2$, $\mu(d) = \hat{\mu}(d)d$, we have

$$\frac{\beta(d_k)}{\mu(d_{k+1})} \leq \frac{1}{c} \frac{d_k^2}{d_{k+1}^{1+\alpha}} = \frac{1}{c} \left(\frac{k+1}{k} \right)^{m(1+\alpha)} \frac{1}{k^{m(1-\alpha)}} \leq \frac{2^{m(1+\alpha)}}{c} \frac{1}{k^{m(1-\alpha)}}.$$

Hence, $\sum_{k \geq k_0} \beta(d_k)/\mu(d_{k+1}) < \infty$ and Corollary 3.9 applies. \square

3.3. Global terminabilities and remarks. In this section, we will briefly discuss the global (asymptotic) terminability of the game. The results are essentially the consequences of those in §§ 3.1 and 3.2.

PROPOSITION 3.11. *Suppose all the assumptions of Theorem 3.1 hold for a sequence of $\delta = \delta_k > 0$, with $\lim_{k \rightarrow \infty} \delta_k = +\infty$. Then, the game (1.1) with terminal set M is terminable and for any $z \in \mathbb{R}^n$, (3.7) holds.*

PROPOSITION 3.12. *Suppose all the assumptions of Theorem 3.6 hold with $\mathcal{N} = \mathbb{R}^n$. Then the game (1.1) with terminal set M is asymptotically terminable. Furthermore, suppose (3.31) holds in $\mathcal{N} = \mathbb{R}^n$ for some $\beta(\cdot)$ satisfying (3.23) and there exists a sequence $\{d_k\}_{k=-\infty}^\infty$, such that*

$$(3.32') \quad \begin{aligned} d_{k+1} &< d_k \quad \forall k = 0, \pm 1, \pm 2, \dots, \\ \lim_{k \rightarrow +\infty} d_k &= 0, \quad \lim_{k \rightarrow -\infty} d_k = +\infty, \end{aligned}$$

$$(3.33') \quad \sum_{k=-\infty}^{\infty} \frac{\beta(d_k)}{\mu(d_{k+1})} < \infty.$$

Then, the game (1.1) with terminal set M is terminable.

The proofs of the above two propositions are obvious.

To close this section, let us make the following remarks. In the proofs of Corollaries 3.4, 3.5 and 3.10, we have given the prototypes of the Lyapunov functions used in

Theorems 3.1, 3.3 and 3.6, respectively, namely, $\Phi(M, z) = d(M, z)$ and $\Psi(M, z) = d(M, z)^2$. Each of them has different advantages. We can see this from Corollaries 3.4, 3.5 and 3.10. It seems possible that we can give results as in Theorems 3.1 and 3.3 by using Lyapunov functions of type Ψ used in § 3.2 instead of type Φ used in § 3.1. But the proofs will be different since $\partial\Psi/\partial z = 0$ for $z \in M$. Also, it seems to us that if we used Ψ instead of Φ , then the corresponding result of Corollary 3.4 would not be as neat as the present Corollary 3.4, especially for (3.21).

Finally, we see that in the proofs of our above results, the basic idea is Lyapunov's direct method. Since in our case, we have an evader control $v(\cdot)$ in the system and our pursuit control $u(\cdot)$ is restricted to be in $\mathcal{U}[0, \infty)$, the proofs become a little bit difficult. The same idea was also used in a previous paper on differential evasion games [17].

4. Linear cases. In this section, we discuss the case that $f(z, u, v)$ is of form $Az + g(u, v)$.

4.1. The case $M = \{0\}$. Let us consider the following system:

$$(4.1) \quad \begin{aligned} \dot{z} &= Az + g(u, v), \\ z(0) &= z_0, \end{aligned}$$

where A is an $(n \times n)$ -matrix, $g: U \times V \rightarrow \mathbb{R}^n$ is a continuous function, $U \subseteq \mathbb{R}^p$ and $V \subseteq \mathbb{R}^q$ are compact sets, and the terminal set $M = \{0\}$. It is clear that in this case, (A1)–(A3) hold, and

$$(4.2) \quad N(0) = S^n \equiv \{\eta \in \mathbb{R}^n \mid |\eta| = 1\}.$$

We denote by $\sigma(A)$ the spectrum of A , A^* the conjugate of A , and $\operatorname{Re} \sigma(A) \triangleq \{\operatorname{Re} \lambda \mid \lambda \in \sigma(A)\}$. By $\operatorname{Re} \sigma(A) \leq a$, we mean that $\sigma(A) \subseteq \{\lambda \in \mathbb{C} \mid \operatorname{Re} \lambda \leq a\}$. The meaning of $\operatorname{Re} \sigma(A) < a$ is similar. The following result is concerned with some necessary conditions for certain terminabilities.

THEOREM 4.1. *Suppose $\delta > 0$, such that*

$$(4.3) \quad G_0^\delta(M) \subset C_a(M),$$

namely, the game is s -locally asymptotically terminable. Then,

$$(4.4) \quad \operatorname{Re} \sigma(A) \leq \frac{1}{\delta} \inf_{v \in V} \sup_{u \in U} |g(u, v)|.$$

In particular, if the game is (globally) asymptotically terminable, then,

$$(4.5) \quad \operatorname{Re} \sigma(A) \leq 0.$$

Proof. Since $g(\cdot, \cdot)$ is continuous and U and V are compact, we have $v_0 \in V$, such that

$$(4.6) \quad \gamma \triangleq \inf_{v \in V} \sup_{u \in U} |g(u, v)| = \sup_{u \in U} |g(u, v_0)|.$$

We let $\lambda = \max \operatorname{Re} \sigma(A)$. Suppose (4.4) is not true, i.e.,

$$(4.7) \quad \lambda > \frac{\gamma}{\delta} \geq 0.$$

Case 1. $\lambda \in \sigma(A)$. Then there exists an $\eta \in S^n$, such that

$$A^* \eta = \lambda \eta.$$

Now, let the evader take control $v(t) \equiv v_0$, and let the initial point satisfy

$$(4.8) \quad z_0 = \alpha \eta, \quad \alpha \in \left(\frac{\gamma}{\lambda}, \delta \right).$$

Then, $z_0 \in G_0^\delta(M) \subset C_a(M)$. But, for any $u(\cdot) \in \mathcal{U}[0, \infty)$, we have

$$\begin{aligned} d(M, z(t)) &= |z(t)| \geq |(\eta, z(t))| \\ &= \left| \left(\eta, e^{At} z_0 + \int_0^t e^{A(t-\tau)} g(u(\tau), v_0) d\tau \right) \right| \\ &\geq \alpha e^{\lambda t} - \gamma \int_0^t e^{\lambda(t-\tau)} d\tau \\ &\geq \left(\alpha - \frac{\gamma}{\lambda} \right) e^{\lambda t} \geq \alpha - \frac{\gamma}{\lambda} > 0 \quad \forall t \geq 0. \end{aligned}$$

This means that $z_0 \notin C_a(M)$, which is a contradiction.

Case 2. $\lambda \pm \mu i \in \sigma(A)$, $\mu > 0$. Then we can find $\xi, \zeta \in \mathbb{R}^n$, such that they are linearly independent and

$$(4.9) \quad A^*(\xi, \zeta) = (\xi, \zeta) \begin{pmatrix} \lambda & -\mu \\ \mu & \lambda \end{pmatrix}.$$

Set $L = \text{span} \{ \xi, \zeta \}$. By choosing a suitable basis for $L \oplus L^\perp \equiv \mathbb{R}^n$, we have that

$$A = \begin{pmatrix} \hat{A} & 0 \\ * & * \end{pmatrix}, \quad g = \begin{pmatrix} \hat{g} \\ * \end{pmatrix}, \quad z = \begin{pmatrix} \hat{z} \\ \tilde{z} \end{pmatrix},$$

where $*$ stands for the entries which will not concern us, and

$$\hat{A} = \lambda I + \mu J, \quad I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

By applying the orthogonal projection P_L to (4.1), we get

$$(4.10) \quad \begin{aligned} \hat{z} &= \hat{A} \hat{z} + \hat{g}(u, v), \\ \hat{z}(0) &= \hat{z}_0. \end{aligned}$$

Again, let the evader take $v(t) \equiv v_0$. Take the initial point to be

$$(4.11) \quad z_0 = \begin{pmatrix} \hat{z}_0 \\ 0 \end{pmatrix}, \quad \frac{\gamma}{\lambda} < |\hat{z}_0| < \delta.$$

Then, $z_0 \in G_0^\delta(M) \subset C_a(M)$. But for any $u(\cdot) \in \mathcal{U}[0, \infty)$, we have

$$\begin{aligned} d(M, z(t)) &= |z(t)| \geq |\hat{z}(t)| \\ &= \left| e^{\hat{A}t} \hat{z}_0 + \int_0^t e^{\hat{A}(t-\tau)} \hat{g}(u(\tau), v_0) d\tau \right| \\ &\geq |e^{\lambda t} e^{\mu J t} \hat{z}_0| - \int_0^t e^{\lambda(t-\tau)} |e^{bJ(t-\tau)} \hat{g}(u(\tau), v_0)| d\tau \\ &\geq e^{\lambda t} |\hat{z}_0| - \gamma \int_0^t e^{\lambda(t-\tau)} d\tau \geq |\hat{z}_0| - \frac{\gamma}{\lambda} > 0 \quad \forall t \geq 0, \end{aligned}$$

which contradicts $z_0 \in C_a(M)$. Hence, we must have (4.4). It is clear that if (4.4) holds for all $\delta > 0$, then (4.5) holds. Thus, the last assertion of the theorem holds. \square

By (1.15), we see that Theorem 4.1 holds if we delete the word "asymptotically" in the theorem. Next, we consider some sufficient conditions.

THEOREM 4.2. *Suppose there exist $\gamma > 0$ and an $(n \times r)$ -matrix B , such that $[A, B]$ is completely controllable and*

$$(4.12) \quad \gamma B(B'_1(0)) \subset \bigcap_{v \in V} g(U, v),$$

where $B'_1(0) \triangleq \{z \in \mathbb{R}^r \mid \|z\| \leq 1\}$. Then, the game is s -locally terminable. Furthermore, if (4.5) holds, then the game is terminable.

To prove this theorem, let us first give the following lemma.

LEMMA 4.3. *Suppose the system $[A, B]$ is completely controllable. Then for any $\gamma > 0$,*

$$(4.13) \quad 0 \in \text{Int} \bigcup_{t \geq 0} \left\{ \int_0^t e^{-A\tau} B w(\tau) d\tau \mid |w(\tau)| \leq \gamma \right\}.$$

Moreover, if (4.5) holds, then

$$(4.14) \quad \bigcup_{t \geq 0} \left\{ \int_0^t e^{-A\tau} B w(\tau) d\tau \mid |w(\tau)| \leq \gamma \right\} = \mathbb{R}^n.$$

For the proof see Chapter 6 of [9] and Chapter 4 of [15].

Proof of Theorem 4.2. Since $[A, B]$ is completely controllable, by (4.12) and (4.13), we get

$$(4.15) \quad M = \{0\} \subseteq \text{Int } C(M).$$

Thus, the game is locally terminable. Then, by Remark 1.4, the game is also s -locally terminable. Now, if (4.5) holds, then by using (4.12) and (4.14), we can get that the game is terminable. \square

Actually, the result of Theorem 4.2 is not new. It is just a version of Lemma 4.3 for differential pursuit games.

THEOREM 4.4. *Suppose that*

$$(4.16) \quad \text{Re } \sigma(A) < 0,$$

$$(4.17) \quad 0 \in \bigcap_{v \in V} \text{Co } g(U, v),$$

where $\text{Co } g(U, v)$ is the convex hull of $g(U, v)$. Then, the game is asymptotically terminable.

To prove this theorem, we need the following lemma.

LEMMA 4.5. *Let $p_1(t), \dots, p_m(t)$ be measurable, bounded functions defined on $[0, T]$, with values in \mathbb{R}^r . Let $\mu_1(t), \dots, \mu_m(t)$ be measurable nonnegative scalar functions defined on $[0, T]$, satisfying $\sum_{i=1}^m \mu_i(t) = 1$. Then for any $\varepsilon > 0$, there exists a measurable function $p(\cdot)$ with values $p(t)$ in the set $\{p_1(t), \dots, p_m(t)\}$ at any $t \in [0, T]$, and the value $p(t)$ of $p(\cdot)$ at time t only depends on $\{\mu_i(s), p_i(s) \mid s \leq t, 1 \leq i \leq m\}$, such that*

$$(4.18) \quad \left| \int_0^t \left[\sum_{i=1}^m \mu_i(\tau) p_i(\tau) - p(\tau) \right] d\tau \right| \leq \varepsilon \quad \forall t \in [0, T].$$

For the proof, see [8].

Proof of Theorem 4.4. For any $z_0 \in \mathbb{R}^n \setminus M \equiv \mathbb{R}^n \setminus \{0\}$ and any $v(\cdot) \in \mathcal{V}[0, \infty)$, by (4.17), we have $u_1(\cdot), \dots, u_{n+1}(\cdot) \in \mathcal{U}[0, \infty)$, such that

$$0 = \sum_{i=1}^{n+1} \mu_i(\tau) e^{-A\tau} g(u_i(\tau), v(\tau)), \quad \tau \in [0, \infty),$$

where $\mu_i(\tau) \geq 0$, $\sum_{i=1}^{n+1} \mu_i(\tau) = 1$. Then, by Lemma 4.5, we see that there exists a $\hat{u}(\cdot) \in \mathcal{U}[0, \infty)$, such that for $k \geq 0$,

$$\begin{aligned} & \left| \int_k^{k+1} e^{-A\tau} g(\hat{u}(\tau), v(\tau)) d\tau \right| \\ &= \left| \int_k^{k+1} \left[\sum_{i=1}^{n+1} \mu_i(\tau) e^{-A\tau} g(u_i(\tau), v(\tau)) - e^{-A\tau} g(\hat{u}(\tau), v(\tau)) \right] d\tau \right| < \frac{1}{(k+1)^2}. \end{aligned}$$

Thus, by this $\hat{u}(\cdot)$, we have (note (4.16))

$$\begin{aligned} |z(t)| &= \left| e^{At} z_0 + \int_0^t e^{A(t-\tau)} g(\hat{u}(\tau), v(\tau)) d\tau \right| \\ &\leq |e^{At} z_0| + |e^{At}| \sum_{k=1}^{\infty} \frac{1}{k^2} \rightarrow 0 \quad \text{as } t \rightarrow \infty. \end{aligned}$$

Hence, the game is asymptotically terminable. \square

4.2. Some generalizations. Now we consider the game (4.1) with terminal set M being a subspace of \mathbb{R}^n , with $\dim M < n$. We let M^\perp be the orthogonal complement of M and $\Pi: \mathbb{R}^n \rightarrow M^\perp$ be the orthogonal projection onto M^\perp . It is clear that for any $x \in \partial M \equiv M$, we have

$$(4.19) \quad N(x) = M^\perp \cap S^n \equiv \{\eta \in M^\perp \mid |\eta| = 1\},$$

$$(4.20) \quad d(M, z) = |\Pi z|.$$

Since in the present case, the terminal set M is convex and closed, Corollaries 3.4 and 3.5 hold. Let us observe the implication of (3.22) (or (3.19)). Explicitly, (3.22) is equivalent to the following:

$$\begin{aligned} 0 &> \sup_{\eta \in M^\perp \cap S^n} \sup_{v \in V} \inf_{u \in U} (\eta, Ax + g(u, v)) \\ &\geq \sup_{\eta \in M^\perp \cap S^n} (\eta, Ax) - \sup_{(u, v) \in U \times V} |g(u, v)| \quad \forall x \in M. \end{aligned}$$

This implies that

$$(\eta, Ax) = 0 \quad \forall x \in M, \quad \eta \in M^\perp.$$

That means $AM \subset (M^\perp)^\perp = M$, i.e., M is an invariant subspace of A . Let us assume this is in the rest of this section. By choosing a suitable orthonormal basis for $M^\perp \oplus M \equiv \mathbb{R}^n$, we have

$$A = \begin{pmatrix} A_1 & 0 \\ A_3 & A_4 \end{pmatrix}, \quad g = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}, \quad z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \quad z_0 = \begin{pmatrix} z_{10} \\ z_{20} \end{pmatrix}.$$

Then applying Π to (4.1), we get

$$\begin{aligned} (4.21) \quad & \dot{z}_1 = A_1 z_1 + g_1(u, v), \\ & z_1(0) = z_{10}. \end{aligned}$$

Also, by (4.20), we see that $z \in M$ if and only if $z_1 = 0$. Hence, we have reduced the game (4.1) with terminal set M to the game (4.21) with terminal set $\{0\}$. Thus, we have results similar to those in § 4.1. We state these corresponding results below.

THEOREM 4.1'. Suppose $\delta > 0$, such that

$$(4.3') \quad G_0^\delta(M) \subset C_a(M).$$

Then,

$$(4.4') \quad \operatorname{Re} \sigma(A^*|_{M^\perp}) \leq \frac{1}{\delta} \inf_{v \in V} \sup_{u \in U} |\Pi g(u, v)|.$$

In particular, if the game is globally asymptotically terminable, then,

$$(4.5') \quad \operatorname{Re} \sigma(A^*|_{M^\perp}) \leq 0.$$

THEOREM 4.2'. Suppose there exist $\gamma > 0$ and an $(m \times r)$ -matrix B , with $m = \dim M^\perp$, such that $[(A^*|_{M^\perp})^*, B]$ is completely controllable and

$$(4.12') \quad \gamma B(B_1'(0)) \subset \bigcap_{v \in V} \Pi g(U, v).$$

Then, the game is s -locally terminable. Furthermore, if (4.5') holds, then the game is terminable.

THEOREM 4.4'. Suppose that

$$(4.16') \quad \operatorname{Re} \sigma(A^*|_{M^\perp}) < 0,$$

$$(4.17') \quad 0 \in \bigcap_{v \in V} \operatorname{Co} \Pi g(U, v).$$

Then, the game is asymptotically terminable.

Remark 4.6. For the game (4.1) with terminal set M being an invariant subspace of A , (2.1) holds with $K = |A|$. (We assume $A \neq 0$.) Then under (4.12') with B being the identity on M^\perp , by Corollary 3.4, we only have

$$(4.22) \quad G_0^{\gamma/|A|}(M) \subset C(M).$$

Then, by Theorems 4.1' and 4.2', we see that in this case, the game is globally terminable if and only if (4.5') holds.

5. Examples. Our first example shows that in general, local terminability does not imply the s -local terminability (recall Remark 1.4).

Example 5.1. For simplicity, take U, V to be singletons. The system is given by (on \mathbb{R}^2)

$$(5.1) \quad \begin{aligned} \dot{x} &= -e^{-y}, \\ \dot{y} &= 1. \end{aligned}$$

The terminal set is

$$(5.2) \quad M = \{(x, y) \in \mathbb{R}^2 | x \leq 0\}.$$

We take a Lyapunov function

$$(5.3) \quad \Phi(x, y) = d(M, (x, y)) = x^+ \equiv \max\{x, 0\} \quad \forall (x, y) \in \mathbb{R}^2.$$

Then, we see that (recall (3.16))

$$(5.4) \quad \sigma(R) = -e^{-R} \quad \forall R > 0.$$

Thus, (3.5'') is satisfied but (3.5) is not satisfied. Actually, we can claim that this game is locally terminable (by Theorem 3.3) but not s -locally terminable. In fact, by (5.1), for any $(x_0, y_0) \in \mathbb{R}^n \setminus M$,

$$\begin{aligned} x(t) &= x_0 - \int_0^t e^{-y_0 - \tau} d\tau \\ &= x_0 - e^{-y_0}(1 - e^{-t}). \end{aligned}$$

Hence, $(x_0, y_0) \in C(M)$, if and only if $x_0 < e^{-y_0}$, and thus

$$(5.5) \quad C(M) = \{(x, y) | x < e^{-y}\}.$$

We see that no $\delta > 0$ exists such that

$$G_0^\delta \subset C(M).$$

Similarly, we see that

$$(5.6) \quad C_a(M) = \{(x, y) | x \leq e^{-y}\}.$$

Hence, this example also shows that the local asymptotic terminability does not necessarily imply s -local asymptotic terminability.

Example 5.2. Let us consider two objects in \mathbb{R}^3 , one pursuing the other. Suppose the pursuer is x and the evader is y , and they are subject to the following systems:

$$(5.7) \quad \begin{aligned} \dot{x} &= 0, \\ \dot{p} &= A_0 x + B_0 p + C_1 u, \end{aligned}$$

$$(5.8) \quad \begin{aligned} \dot{y} &= q, \\ \dot{q} &= A_0 y + B_0 q + C_2 v. \end{aligned}$$

The pursuer wants to “softly” catch the evader, namely, the terminal set M is

$$(5.9) \quad M = \{x = y, p = q\}.$$

Let us set

$$\begin{aligned} z_1 &= x - y, & z_2 &= p - q, \\ z_3 &= x + y, & z_4 &= p + q. \end{aligned}$$

Then, (5.7) and (5.8) become

$$(5.10) \quad \dot{z} = \begin{pmatrix} 0 & I & 0 & 0 \\ A_0 & B_0 & 0 & 0 \\ 0 & 0 & 0 & I \\ 0 & 0 & A_0 & B_0 \end{pmatrix} z + \begin{pmatrix} 0 \\ C_1 \\ 0 \\ C_1 \end{pmatrix} u + \begin{pmatrix} 0 \\ -C_2 \\ 0 \\ C_2 \end{pmatrix} v \equiv Az + Bu + Cv.$$

The terminal set M becomes

$$(5.11) \quad M = \{(0, 0, z_3, z_4) | z_3, z_4 \in \mathbb{R}^3\}.$$

It is clear that $AM \subseteq M$. Thus, we have the following proposition.

PROPOSITION 5.3. *Suppose*

$$\left[\begin{pmatrix} 0 & I \\ A_0 & B_0 \end{pmatrix}, \begin{pmatrix} 0 \\ C_1 \end{pmatrix} \right]$$

is completely controllable. Let there exist a $\gamma > 0$, such that

$$(5.12) \quad \gamma C_1(B_1^3(0)) \subset \bigcap_{v \in V} (C_1 U - C_2 v).$$

Then the “soft” capture is possible if initially the positions and the velocities of these two objects are close enough. Furthermore, if we have

$$(5.13) \quad \operatorname{Re} \sigma \begin{pmatrix} 0 & I \\ A_0 & B_0 \end{pmatrix} \leq 0,$$

then, for any initial state, the “soft” capture is possible.

The proof is immediate.

Finally, we present an example in which the Lyapunov function is not a function of $d(M, z)$. This example is a modification of that given in [3].

Example 5.4. Let the game be governed by the following system:

$$(5.14) \quad \begin{aligned} \dot{z}_1 &= z_2, \\ \dot{z}_2 &= -\sin z_1 + u - v \cos z_1, \end{aligned}$$

where $z_1, z_2 \in \mathbb{R}$, $u \in U \equiv [-a, a]$, $v \in V \equiv [-b, b]$, $a > b \geq 0$. The terminal set is $M = \{0\}$. We define

$$(5.15) \quad \Psi(z) = z_1^2 + 2z_1z_2 + z_2^2 + 2(1 - \cos z_1).$$

Then, we have

$$(5.16) \quad \begin{aligned} \sup_{c \in V} \inf_{u \in U} \left(\frac{\partial \Psi}{\partial z}(z), f(z, u, v) \right) &= 2 \sup_{v \in V} \inf_{u \in U} \{ z_1z_2 + z_2^2 + (z_1 + z_2)(u - v \cos z_1) - z_1 \sin z_1 \} \\ &\geq 2(-a + b + |z_2|)|z_1 + z_2| - 2z_1 \sin z_1. \end{aligned}$$

Hence (note $a > b$), we can find constants $c > 0$, $\varepsilon > 0$, such that

$$(5.17) \quad \sup_{v \in V} \inf_{u \in U} \left(\frac{\partial \Psi}{\partial z}(z), f(z, u, v) \right) \leq -c(|z_1|^2 + |z_2|^2) \equiv -cd(M, z)^2 \quad \forall |z_1|, |z_2| \leq \varepsilon.$$

Thus, by Theorem 3.6, the game is s -locally asymptotically terminable.

Acknowledgments. This paper is an improved version of one chapter of the author's Ph.D. thesis accomplished at Purdue University under the guidance of Professor L. D. Berkovitz. The author would like to take this opportunity to thank Professor Berkovitz for posing the problem, for suggestive discussions and for the encouragement he gave. Also, the author thanks the referee for the suggestions and criticism which led to this improved version.

REFERENCES

- [1] L. D. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, New York, 1974.
- [2] N. P. BHATIA AND G. P. SZEGÖ, *Stability Theory of Dynamical Systems*, Springer-Verlag, New York, 1970.
- [3] M. J. CORLESS AND G. LEITMANN, *Continuous state feedback guaranteeing uniform ultimate boundedness for uncertain dynamic systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 1139-1144.
- [4] R. GABASOV AND F. KIRILLOVA, *The Qualitative Theory of Optimal Processes*, Marcel Dekker, New York-Basel, 1976.
- [5] P. B. GUSYANTNIKOV, *A formulation of the linear pursuit problem*, Differentsial' nye Uravneniya, 8 (1972), pp. 1363-1371. (In Russian.) Differential Equations, 8 (1972), pp. 1047-1053. (In English.)
- [6] S. GUTMAN AND G. LEITMANN, *Stabilizing feedback control for dynamical systems with bounded uncertainty*, Proc. IEEE Conference on Decision and Control, 1976.
- [7] R. B. HOLMES, *A Course on Optimization and Best Approximation*, Lecture Notes in Mathematics 257, Springer-Verlag, New York, 1972.
- [8] B. KASKOSZ, *On a nonlinear evasion problem*, this Journal, 15 (1977), pp. 661-673.
- [9] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [10] S. LEFSCHETZ, *Stability of Nonlinear Control Systems*, Academic Press, New York, 1965.
- [11] E. F. MISHCHENKO AND L. S. PONTRYAGIN, *Linear differential games*, Dokl. Akad. Nauk. SSSR, 174 (1967), pp. 27-29. (In Russian.) Soviet Math. Dokl., 8 (1967), pp. 585-588. (In English.)
- [12] L. S. PONTRYAGIN, *Linear differential games of pursuit*, Math. Sb., 112 (1980), pp. 307-330. (In Russian.) Math. USSR Sb., 40 (1981), pp. 285-303. (In English.)
- [13] B. N. PSHENICHNYI, *Linear differential games*, Avtomat. i Telemekh., (1968) pp. 65-78. (In Russian.) Automat. Remote Control, 29 (1968), pp. 55-67. (In English.)
- [14] N. SATIMOV, *The pursuit problem in linear differential games*, Differentsial' nye Uravneniya, 9 (1973), pp. 2000-2009. (In Russian.) Differential Equations, 9 (1973), pp. 1535-1542. (In English.)
- [15] A. STRAUSS, *An Introduction to Optimal Control Theory*, Lecture Notes in Oper. Res. and Math. Econ., 3, Springer-Verlag, Berlin-New York, 1968.
- [16] J. YONG, *Feedback stabilization of uncertain dynamic systems*, J. Math. Anal. Appl., to appear.
- [17] ———, *On differential evasion games*, this Journal, 26 (1988), pp. 1-22.
- [18] ———, *On differential games of evasion and pursuit*, Ph.D. thesis, Purdue University, West Lafayette, IN, 1986.

A PRODUCT FORMULA APPROACH TO NONLINEAR OPTIMAL CONTROL PROBLEMS*

VIOREL BARBU†

Abstract. In this paper we prove the convergence of an iterative scheme of Lie-Trotter type for optimal control problems governed by nonlinear systems. The numerical implementation of the resulting algorithm is also discussed.

AMS(MOS) subject classification. 35

Key words. Lie-Trotter product formula, maximal monotone operators, nonlinear semigroups, distributed control systems, free boundary problem, optimal control, Stefan problem

0. Introduction. This work is concerned with an iterative scheme of the Lie-Trotter product formula type for optimal control problems of the following form:

$$(P) \quad \text{Minimize} \quad \int_0^T g(y(t)) \, dt + h(u) + \varphi_0(y(T))$$

$$\text{on all } (y, u) \in C([0, T]; H) \times L^2(0, T; U)$$

subject to

$$(0.1) \quad y'(t) + Ay(t) + Fy(t) \ni (Bu)(t) + f(t), \quad t \in [0, T],$$

$$y(0) = y_0.$$

Here H and U are real Hilbert spaces, $f \in L^2(0, T; H)$, $y_0 \in H$, $g: H \rightarrow \mathbb{R}$, $\varphi_0: H \rightarrow \mathbb{R}$ and $h: U \rightarrow \bar{\mathbb{R}} = (-\infty, +\infty]$ are given functions, A and F are maximal monotone operators in $H \times H$ and $B: U \rightarrow L^2(0, T; H)$ is a linear operator. (The exact hypotheses will be given later.)

In few words, the main result of this paper amounts to saying that in certain situations problem (P) can be approximated for $\varepsilon \rightarrow 0$ by the following sequence of optimal control problems

$$(P_\varepsilon) \quad \text{Minimize} \quad \int_0^T g(y(t)) \, dt + h(u) + \varphi_0(y(T))$$

$$\text{on all } y: [0, T] \rightarrow H, u \in L^2(0, T; U)$$

subject to

$$(0.2) \quad y'(t) + Ay(t) = (Bu)(t) + f(t) \quad \text{for } t \in (i\varepsilon, (i+1)\varepsilon),$$

$$y^+(i\varepsilon) = w_i(\varepsilon) \quad \text{for } i = 1, \dots, N-1, \quad y^+(0) = y_0,$$

$$(0.3) \quad w'_i + Fw_i = 0 \quad \text{in } (0, \varepsilon), \quad \varepsilon = T/N,$$

$$w_i(0) = Py^-(i\varepsilon) \quad \text{for } i = 1, 2, \dots, N-1.$$

Here P is the projection operator of H into the closed convex subset $K = \overline{D(A) \cap D(F)}$.

As we will see later the iterative scheme (0.2), (0.3) is related to the Lie-Trotter formula for nonlinear semigroups and it provides a theoretic algorithm for obtaining the solutions to problem (P). The convergence of this scheme is proved in the cases

* Received by the editors February 18, 1986; accepted for publication May 6, 1987.

† University of Iasi, 6600 Iasi, Romania.

described in Theorems 1 and 2 below which include a large class of distributed and boundary optimal control problems governed by nonlinear parabolic equations and free boundary problems (see Examples 1, 2 below).

The semigroup approximation theory has already been used by H. T. Banks (see, for instance, [1]) to derive numerical algorithms for infinite-dimensional optimal control problems. However, these results differ in spirit from our own.

The numerical implementation of this algorithm, which was written by Dr. V. Arnăutu from the University of Iasi, is discussed in § 5.

1. Preliminaries and the main result. Notation and hypotheses. (1) H and U are real Hilbert spaces with the norms $|\cdot|$ and $|\cdot|_U$, respectively. B is a linear continuous operator from U to $L^2(0, T; H)$.

(2) V is a real Hilbert space continuously and densely imbedded in H . Identifying H with its own dual and denoting by V' the dual of V we have

$$V \subset H \subset V'$$

algebraically and topologically. Assume that the injection of V into H is compact and denote by $\|\cdot\|$ the norm of V . We will denote by the same symbol (\cdot, \cdot) the scalar product of H and the pairing between V and V' .

(3) $A: V \rightarrow V'$ is a linear, continuous, symmetric and coercive operator, i.e.,

$$(1.1) \quad (Ay, y) \geq \omega \|y\|^2 \quad \forall y \in V$$

where $\omega > 0$.

(4) $F: D(F) \subset H \rightarrow 2^H$ is a maximal monotone operator of subgradient type, i.e., $F = \partial\varphi$, where $\varphi: H \rightarrow \bar{R} = (-\infty, +\infty]$ is a lower semicontinuous, convex function. (We have denoted by $\partial\varphi$ the subdifferential of φ .) We assume that for every $\lambda > 0$ and all $y \in D(A) = \{y \in V; Ay \in H\}$,

$$(1.2) \quad (Ay, F_\lambda y) \geq 0$$

where

$$F_\lambda = \lambda^{-1}(I - (I + \lambda F)^{-1}) \quad (I \text{ is the identity operator on } H).$$

(5) P is the projection operator of H into $K = \overline{D(F)}$. We shall assume that P maps V into itself and

$$(1.3) \quad (APy, Py) \leq (Ay, y) \quad \forall y \in V.$$

(6) The functions $g: H \rightarrow R$ and $\varphi_0: H \rightarrow R$ are Lipschitzian on bounded subsets, Fréchet differentiable and bounded from below by affine functions. The function $h: H \rightarrow \bar{R}$ is convex, lower semicontinuous and coercive, i.e.,

$$(1.4) \quad \lim_{|u|_U \rightarrow \infty} h(u)/|u|_U = +\infty.$$

We will assume, finally, that

$$(1.5) \quad y_0 \in D(\varphi) \cap V \quad \text{and} \quad f \in L^2(0, T; H).$$

We note that assumption (1.2) implies that the operator $A + F$ is maximal monotone in $H \times H$ and $\overline{D(A) \cap D(F)} = \overline{D(F)}$ (see for instance [6] or [3, p. 19]). Moreover, $A + F = \partial\psi$, where

$$\psi(y) = \frac{1}{2}(Ay, y) + \varphi(y).$$

Then according to standard existence results for nonlinear evolution equations of subgradient type (see [6] and [2, p. 188]), (0.1) admits a unique solution $y \in W^{1,2}([0, T]; H) \cap L^2(0, T; D(A))$. (We have denoted by $W^{1,2}([0, T]; H)$ the space of all absolutely continuous functions $y: [0, T] \rightarrow H$ such that $y' \in L^2(0, T; H)$). Moreover, since by our assumptions on V , A and F the map $u \rightarrow y^u$ is compact from U to $C([0, T]; H)$ we infer by standard device that problem (P) admits at least one solution (y^*, u^*) . (We have denoted by y^u the solution to state system (0.1).)

Now we come back to iterative scheme (0.2), (0.3) and note that for y_0 and f satisfying (1.4) this system has a unique solution $y: [0, T] \rightarrow H$ which is piecewise continuous and belongs to $W^{1,2}([i\varepsilon, (i+1)\varepsilon]; H) \cap L^\infty(0, T; V)$ for all $i = 0, 1, \dots, N-1$. Indeed, by assumption (1.2), we know that (see for instance [2, p. 183])

$$(1.6) \quad e^{-Ft}V \subset V \quad \text{for all } t \geq 0$$

and so $y^+(i\varepsilon) \in V$ for all i .

As a matter of fact we may rewrite system (0.2), (0.3) as the impulse differential equation

$$y' + Ay = Bu + f + \sum_{i=1}^N (e^{-F\varepsilon}Py^-(i\varepsilon) - y^-(i\varepsilon)) \otimes \delta(i\varepsilon),$$

where e^{-Ft} is the nonlinear semigroup generated by $-F$ and δ is the Dirac measure.

It is readily seen that for each $\varepsilon > 0$, problem (P_ε) has at least one solution $(y_\varepsilon, u_\varepsilon)$.

We set

$$(1.7) \quad \phi(u) = \int_0^T g(y^u(t)) dt + \varphi_0(y^u(T)) + h(u)$$

and

$$(1.8) \quad \phi_\varepsilon(u) = \int_0^T g(y_\varepsilon^u(t)) dt + \varphi_0(y_\varepsilon^u(T)) + h(u)$$

where y_ε^u is the solution to approximating system (0.2), (0.3).

Then we may rewrite (P) and (P_ε) as

$$(P) \quad \min \{ \phi(u); u \in U \},$$

$$(P_\varepsilon) \quad \min \{ \phi_\varepsilon(u); u \in U \}.$$

The main result is the following theorem.

THEOREM 1. Assume that beside (1) ~ (6) at least one of the following hypotheses holds:

(i) The set $\{w = Bu; h(u) \leq \lambda\}$ is compact in $L^1(0, T; H)$ for every $\lambda \in \mathbb{R}$.

(ii) $F = \partial I_C$, where C is a closed convex subset of H .

Then $\lim_{\varepsilon \downarrow 0} \inf \phi_\varepsilon = \inf \phi$, and if $\{u_\varepsilon^*\}$ is a sequence of optimal controls for problems (P_ε) then

$$(1.9) \quad \lim_{\varepsilon \downarrow 0} \phi_\varepsilon(u_\varepsilon^*) = \inf \phi.$$

Moreover, every weak limit point of $\{u_\varepsilon^*\}$ is an optimal control for problem (P).

In (ii) we have denoted by ∂I_C the subdifferential of the indicator function I_C of C , i.e.,

$$(1.10) \quad \partial I_C(y) = \{p \in H; (p, y - w) \geq 0 \forall w \in C\}.$$

Conclusion (1.9) of the theorem amounts to saying that $\{u_\varepsilon^*\}$ is a sequence of suboptimal controllers for problem (P).

Now we shall present some typical situations where Theorem 1 applies.

Example 1. Consider the distributed control system defined on an open domain Ω of R^n with a sufficiently smooth boundary:

$$\begin{aligned}
 & \frac{\partial y}{\partial t}(t, x) - \Delta y(t, x) + \gamma(y(t, x)) \\
 & = f(t, x) + \sum_{i=1}^m v_i(t) \alpha_i(x), \quad (t, x) \in Q = (0, T) \times \Omega, \\
 & y(0, x) = y_0(x), \quad x \in \Omega, \\
 & y + \alpha \frac{\partial y}{\partial \nu} = 0 \quad \text{in } \Sigma = (0, T) \times \Gamma, \\
 & \frac{dv}{dt} + D(v) = B_0 u \quad \text{a.e. } t \in [0, T], \\
 & v(0) = v_0, \quad v = (v_1, v_2, \dots, v_m).
 \end{aligned}
 \tag{1.11}$$

Here, $\alpha_i \in L^2(\Omega)$, $\alpha \geq 0$, γ is a maximal monotone graph in $R \times R$ (eventually multivalued) such that $\overline{D(\gamma)} = R$, $0 \in \gamma(0)$, D is a Lipschitzian mapping from R^m to itself and B_0 is a linear continuous operator from a control space U to $L^1(0, T; R^m)$.

Theorem 1 is applicable if we take $V = H^1(\Omega)$, $H = L^2(\Omega)$ and $A: V \rightarrow V'$, $F: H \rightarrow H$, $B: U \rightarrow L^2(Q)$ defined by

$$(Ay, z) = \int_{\Omega} \nabla y \cdot \nabla z \, dx + \alpha^{-1} \int_{\Gamma} yz \, d\sigma \quad \forall z \in H^1(\Omega),
 \tag{1.13}$$

$$(Fy)(x) = \gamma(y(x)) \quad \text{a.e. } x \in \Omega,
 \tag{1.14}$$

$$(Bu)(t, x) = \sum_{i=1}^m \alpha_i(x) v_i(t), \quad (t, x) \in Q.$$

Assumption (i) is obviously satisfied as an immediate consequence of the Arzelà theorem.

Example 2. Consider the free boundary problem

$$\begin{aligned}
 & \frac{\partial y}{\partial t} - \Delta y \geq f + Bu, \quad y \geq 0 \quad \text{a.e. in } Q, \\
 & \left(\frac{\partial y}{\partial t} - \Delta y - Bu - f \right) y = 0 \quad \text{a.e. in } Q, \\
 & y(0, x) = y_0(x) \quad \text{a.e. } x \in \Omega, \\
 & y = 0 \quad \text{in } \Sigma
 \end{aligned}
 \tag{1.15}$$

where $y_0 \in H_0^1(\Omega)$ and $y_0(x) \geq 0$ almost everywhere $x \in \Omega$.

The control system is of the form (0.1) where $V = H_0^1(\Omega)$, $H = L^2(\Omega)$, $A = -\Delta$, $F = \partial I_C$ where $C = \{y \in H_0^1(\Omega); y(x) \geq 0 \text{ a.e. } x \in \Omega\}$.

Note that $(Py)(x) = \max\{y(x), 0\}$ almost everywhere $x \in \Omega$ and it is well known that assumption (1.3) is satisfied, i.e.,

$$\|\nabla Py\|_{L^2(\Omega)} \leq \|\nabla y\|_{L^2(\Omega)} \quad \forall y \in H_0^1(\Omega).$$

Optimal control problems governed by the free boundary problems of the form (1.15) have been studied in [3], [11], [14], [15].

2. Proof of the main result. We begin by establishing a Lie–Trotter product formula for the nonhomogeneous evolution equation

$$(2.1) \quad \begin{aligned} y' + Ay + Fy &= \chi, & t \geq 0, \\ y(0) &= x \end{aligned}$$

where $\chi \in L^1(R^+; H)$, $x \in \overline{D(A) \cap D(F)} = \overline{D(F)} = K$ and A, F satisfy assumptions (1)–(5) of § 1.

We know that (2.1) has a unique integral solution $y \in C(R^+; H)$ in the sense of B  nilan (see [2], [6]). If $\chi \in L^2_{\text{loc}}(R^+; H)$ then $y \in W^{1,2}_{\text{loc}}([\delta, T]; H)$ on every interval $[\delta, T] \subset (0, +\infty)$.

Following an idea due to Dafermos and Slemrod [10] we may write (2.1) as an autonomous differential equation in the Banach space $X = H \times L^1(R^+; H)$ with the norm $\|(x, \chi)\|_X = |x| + \int_0^\infty |\chi(s)| ds$.

Let $\mathcal{A}: X \rightarrow X$ be the operator defined by

$$(2.2) \quad \begin{aligned} \mathcal{A}(x, \chi) &= \{(A + F)x - \chi(0), -\chi'\}, & (x, \chi) \in D(\mathcal{A}), \\ D(\mathcal{A}) &= \{(x, \chi) \in X; x \in D(A) \cap D(F)\}, & \chi \in W^{1,1}(R^+; H) \end{aligned}$$

and

$$(2.3) \quad S(t)(x, \chi) = (y(t), \chi_t), \quad \chi_t(s) = \chi(t + s), \quad t, s \geq 0;$$

then we may write (2.1) as

$$(2.4) \quad \frac{d}{dt} S(t)(x, \chi) + \mathcal{A}S(t)(x, \chi) \ni 0, \quad t \geq 0.$$

In other words, $S(t): K \times L^1(R^+; H) \rightarrow K \times L^1(R^+; H)$ is the semigroup of nonlinear contractions generated by the m -accretive operator \mathcal{A} via the Crandall–Liggett generation theorem [9].

On $\mathcal{H} = K \times L^1(R^+; H)$ consider the nonlinear semigroups $S_1(t)$ and $S_2(t)$ defined by

$$(2.5) \quad S_1(t)(x, \chi) = (\tilde{y}(t), \chi_t), \quad t \geq 0$$

and

$$(2.6) \quad S_2(t)(x, \chi) = (e^{-Ft}x, \chi), \quad t \geq 0$$

where

$$\tilde{y}' + A\tilde{y} = \chi \quad \text{in } R^+, \quad \tilde{y}(0) = x$$

and $e^{-Ft}x = w(t)$ is the solution to the evolution equation

$$w' + Fw = 0 \quad \text{in } R^+, \quad w(0) = x.$$

We see that $S_1(t)$ and $S_2(t)$ are generated by the operators \mathcal{A}_1 and \mathcal{A}_2 defined below:

$$\mathcal{A}_1(x, \chi) = \{Ax - \chi(0), \chi'\}, \quad \mathcal{A}_2(x, \chi) = \{Fx, 0\}.$$

Let $P: H \rightarrow K$ be the projection operator on $K = \overline{D(F)}$ and let $\Pi: X \rightarrow \mathcal{H} = K \times L^1(R^+; H)$ be defined by

$$(2.7) \quad \Pi(x, \chi) = (Px, \chi), \quad x \in K, \quad \chi \in L^1(R^+; H).$$

PROPOSITION 1. For all $x \in K$ and $\chi \in L^1(R^+; H)$ we have

$$(2.8) \quad \lim_{n \rightarrow \infty} \left(\Pi S_1 \left(\frac{t}{n} \right) S_2 \left(\frac{t}{n} \right) \right)^n (x, \chi) = S(t)(x, \chi) \quad \forall t \geq 0$$

in X , and the convergence is uniform in t on compact intervals of $R^+ = [0, \infty)$.

Proof. We will employ the nonlinear version of Chernoff theorem due to Brézis and Pazy [7].

Let $\{\Gamma(h)\}_{h>0}$ be the family of nonexpansive operators on \mathcal{H} defined by

$$\Gamma(h) = \Pi S_1(h) S_2(h)$$

and set

$$(2.9) \quad X_h = (I + \lambda h^{-1}(I - \Gamma(h)))^{-1}(x, \chi), \quad \lambda > 0$$

where I is the identity operator in X .

According to Theorem 3.2 in [7] to prove (2.8) it suffices to show that for $h \rightarrow 0$

$$(2.10) \quad X_h \rightarrow (I + \lambda \mathcal{A})^{-1}(x, \chi) \quad \text{for all } \lambda > 0.$$

Taking into account (2.5)–(2.7) we may rewrite (2.9) as

$$(2.11) \quad (h + \lambda)x^h - \lambda P \left(e^{-Ah} e^{-Fh} x^h + \int_0^h e^{-A(h-s)} y^h(s) ds \right) = hx,$$

$$(2.12) \quad (h + \lambda)y^h(s) - \lambda y^h(s + h) = h\chi(s), \quad s \geq 0$$

where $(x^h, y^h) = X_h \in X$ and e^{-At} is the C_0 -semigroup generated on H by the operator $-A$.

Inasmuch as the operators $(I + \lambda \mathcal{A})^{-1}$ and $(I + \lambda h^{-1}(I - \Gamma(h)))^{-1}$ are nonexpansive on X , without any loss of generality we may assume that $x \in D(A) \cap D(F)$ and

$$\chi, \chi' \in L^2(R^+; V) \cap L^1(R^+; V).$$

Then by (2.12) we see that $y^h \in L^1(R^+; V) \cap L^2(R^+; V)$, y^h is absolutely continuous on compact intervals in the V -norm and

$$\|y^h\|_{L^i(R^+; V)} \leq \|\chi\|_{L^i(R^+; V)}, \quad i = 1, 2,$$

$$\left\| \frac{d}{ds} y^h \right\|_{L^i(R^+; V)} \leq \left\| \frac{d\chi}{ds} \right\|_{L^i(R^+; V)}, \quad i = 1, 2.$$

Hence $\{y^h\}_{h>0}$ is compact in $L^1(R^+; H) \cap C(R^+; H)$ and since the limit is unique we may infer that for $h \rightarrow 0$

$$(2.13) \quad y^h \rightarrow w \quad \text{strongly in } H, \quad \text{uniformly in } s \text{ on compacts.}$$

Since by (2.12)

$$\int_{\mu}^{\infty} |y^h(s)| ds \leq \int_{\mu}^{\infty} |\chi(s)| ds \quad \forall \mu > 0$$

we infer by (2.13) that $\{y^h\}$ is compact in $L^1(R^+; H)$, and therefore

$$(2.14) \quad y^h \rightarrow w \quad \text{strongly in } L^1(R^+; H).$$

On the other hand, we see that for each $\psi \in W^{1,2}(R^+; H)$, $\psi(0) = 0$, we have

$$\lim_{h \rightarrow 0} \int_0^\infty \left(\frac{y^h(s+h) - y^h(s)}{h}, \psi(s) \right) ds = - \int_0^\infty (w(s), \psi'(s)) ds.$$

Hence for $h \rightarrow 0$

$$\frac{1}{h} (y^h(s+h) - y^h(s)) \rightarrow w' \quad \text{weakly in } L^2(R^+; H)$$

and letting h tend to zero in (2.12) we see that

$$(2.15) \quad w - \lambda w' = \chi \quad \text{a.e. in } R^+.$$

Next by (2.13) we see that for $h \rightarrow 0$

$$(2.16) \quad h^{-1} \int_0^h e^{-A(h-s)} y^h(s) ds \rightarrow w(0) \quad \text{strongly in } H.$$

To complete the proof of (2.10) it remains to be shown that for $h \rightarrow 0$

$$(2.17) \quad x^h \rightarrow x_\lambda^0 \quad \text{strongly in } H$$

where $x_\lambda^0 \in D(A) \cap D(F)$ is the solution to the equation

$$(2.18) \quad x_\lambda^0 + \lambda (Ax_\lambda^0 + Fx_\lambda^0) = x + \lambda w(0).$$

To this purpose we set $q^h = h^{-1} \int_0^h e^{-A(h-s)} y^h(s) ds$ and noting that $P = (1 + \partial I_K)^{-1}$ (1 is the identity operator in H) we may put (2.11) under the following form:

$$(2.19) \quad \begin{aligned} & h^{-1} (x^h - e^{-Ah} x^h) + \lambda h^{-1} e^{-Ah} (x^h - e^{-Fh} x^h) \\ & + \lambda h^{-1} \partial I_K (\lambda^{-1} (h + \lambda) x^h - h \lambda^{-1} x) \ni x + \lambda q^h - x^h. \end{aligned}$$

Let $z(s) = e^{-As} x^h$, i.e.,

$$z' + Az = 0 \quad \text{in } [0, h], \quad z(0) = x^h$$

and let u be arbitrary but fixed in V . We multiply the latter by $z(s) - u$ and integrate on $[0, h]$. After some calculation we get

$$(2.20) \quad (e^{-Ah} x^h - x^h, x^h - u) + \frac{1}{2} |e^{-Ah} x^h - x^h|^2 + \int_0^h (\varphi_1(z(s)) - \varphi_1(u)) ds \leq 0$$

where $\varphi_1(u) = \frac{1}{2} (Au, u)$.

Similarly,

$$(2.21) \quad (e^{-Fh} x^h - x^h, x^h - u) + \frac{1}{2} |e^{-Fh} x^h - x^h|^2 + \int_0^h (\varphi(e^{-Fs} x^h) - \varphi(u)) ds \leq 0.$$

Hence

$$(2.22) \quad \begin{aligned} & h^{-1} (e^{-Fh} x^h - x^h, e^{-Ah} (x^h - u)) + (2h)^{-1} |e^{-Fh} x^h \\ & - x^h|^2 + h^{-1} \int_0^h (\varphi(e^{-Fs} x^h) - \varphi(u)) ds \\ & \leq h^{-1} (e^{-Fh} x^h - x^h, e^{-Ah} x^h - x^h) + h^{-1} |e^{-Fh} x^h - x^h| |e^{-Ah} u - u|. \end{aligned}$$

Now we multiply (2.19) (scalarly in H) by $x^h - e^{-Fh}x^h + h\lambda^{-1}(x^h - x)$ and use the accretivity of e^{-Ah} along with the definition of ∂I_k (see (1.10)) to find

$$\begin{aligned} h^{-1}(x^h - e^{-Ah}x^h, x^h - e^{-Fh}x^h) &\leq -(x^h - e^{-Ah}x^h, x^h - x) - (x^h - e^{-Fh}x^h, e^{-Ah}(x^h - x)) \\ &\quad + (x + \lambda q^h - x^h, x^h - e^{-Fh}x^h) \\ &\quad + h\lambda^{-1}(x + \lambda q^h - x^h, x^h - x). \end{aligned}$$

Then combining the latter with (2.20) and (2.22) yields

$$\begin{aligned} &\lambda(h^{-1}(e^{-Ah}x^h - x^h), x^h - u) + \lambda(h^{-1}e^{-Ah}(e^{-Fh}x^h - x^h), x^h - u) \\ &\quad + \lambda h^{-1} \int_0^h (\phi(e^{-Fs}x^h) + \phi_1(e^{-As}x^h)) ds \\ (2.23) \quad &\leq \lambda(\phi(u) + \phi_1(u)) + (2h)^{-1}\lambda|e^{-Ah}u - u|^2 \\ &\quad + |x^h - e^{-Fh}x^h|(2|x^h - x| + |x + \lambda q^h - x^h|) \\ &\quad + h\lambda^{-1}|x + \lambda q^h - x^h||x^h - x|. \end{aligned}$$

Now by (2.11)

$$(2.24) \quad x^h = h(\lambda + h)^{-1}x + \lambda(h + \lambda)^{-1}P(e^{-Ah}e^{-Fh}x^h + hq^h).$$

Since assumption (1.2) implies that (see [2, p. 183])

$$\phi_1(e^{-Fh}x) \leq \phi_1(x) \quad \forall h \geq 0, \quad x \in V$$

it follows by (2.24) that

$$\|x^h\| \leq \|x\| + \lambda\|q^h\| \leq C \quad \text{for all } h > 0$$

because, as we have seen above, $\{q^h\}$ is bounded in $L^\infty(0, 1; V)$.

We may conclude therefore that $x_{h>0}^h$ is a compact subset of H . Hence on a subsequence again denoted h , we have

$$(2.25) \quad x^h \rightarrow x_\lambda^0 \quad \text{strongly in } H.$$

Since the functions ϕ and ϕ_1 are lower semicontinuous it follows by the Fatou lemma that

$$(2.26) \quad \liminf_{h \rightarrow 0} h^{-1} \int_0^h (\phi(e^{-Fs}x^h) + \phi_1(e^{-As}x^h)) ds \geq \phi(x^0) + \phi_1(x^0).$$

Next from (2.25) and the strong continuity of e^{-Ft} and e^{-At} we infer that

$$(2.27) \quad |e^{-Ah}x^h - x^h| + |e^{-Fh}x^h - x^h| \rightarrow 0 \quad \text{as } h \rightarrow 0.$$

Now once again, using (2.19) and the definition of ∂I_k , we get

$$\begin{aligned} &\lambda h^{-1}(\partial I_k(\lambda^{-1}(h + \lambda)x^h - h\lambda^{-1}x), x^h - u) \\ &\quad \geq -(\partial I_k(\lambda^{-1}(h + \lambda)x^h - h\lambda^{-1}x), x^h - x) \\ &\quad = -(e^{-Ah}x^h - x^h, x^h - x) - (e^{-Fh}x^h - x^h, e^{-Ah}(x^h - x)) - \lambda^{-1}h(x + \lambda q^h - x^h, x^h - x) \end{aligned}$$

and by (2.27) we conclude that

$$(2.28) \quad \liminf_{h \rightarrow 0} h^{-1}(\partial I_k(\lambda^{-1}(h + \lambda)x^h - h\lambda^{-1}x), x^h - u) \geq 0.$$

Along with (2.19), (2.23) and (2.26) the latter yields

$$-\lim_{h \rightarrow 0} (x + \lambda q^h - x^h, x^h - u) \leq \lambda (\zeta(u) - \zeta(x_\lambda^0))$$

where $\zeta = \varphi + \varphi_1$. Then by (2.16) and (2.17) we have

$$(x_\lambda^0 - \lambda w(0) - x, x_\lambda^0 - u) \leq \lambda (\zeta(u) - \zeta(x_\lambda^0)).$$

Since u is arbitrary in H and $\partial\zeta = A + F$ we conclude that

$$x + \lambda w(0) - x_\lambda^0 \in \lambda (A + F)(x_\lambda^0).$$

Hence x_λ^0 is the unique solution to (2.18) and the proof of Proposition 1 is complete.

In the following we shall denote by y^u the solution to control system (0.1) and by y_ε^u the solution to system (0.2), (0.3).

PROPOSITION 2. *For each $u \in U$ we have*

$$(2.29) \quad y_\varepsilon^u(t) \rightarrow y^u(t) \quad \text{strongly in } H \text{ for all } t \in [0, T].$$

If $Bu_{\varepsilon_n} \rightarrow Bu$ strongly in $L^1(0, T; U)$ for $\varepsilon_n \rightarrow 0$ then

$$(2.30) \quad y_{\varepsilon_n}^u(t) \rightarrow y^u(t) \quad \text{strongly in } H \text{ for } t \in [0, T].$$

Proof. We extend $(Bu)(t)$ with 0 on $[T, +\infty)$ and apply Proposition 1. Then the sequence $\{\tilde{y}_\varepsilon\}_{\varepsilon > 0}$ defined by

$$(2.31) \quad (\tilde{y}_\varepsilon(t), \tilde{u}_\varepsilon(t)) = (\Pi S_1(\varepsilon) S_2(\varepsilon))^i (y_0, Bu) \quad \text{for } t \in [i\varepsilon, (i+1)\varepsilon]$$

or equivalently

$$(2.31)' \quad \tilde{y}_\varepsilon = P(y_\varepsilon^u)^-(i\varepsilon) \quad \text{for } t \in [i\varepsilon, (i+1)\varepsilon]$$

is strongly convergent in H to $y^u(t)$ for all $t \in [0, T]$. On the other hand, it follows by (0.2), (0.3) that

$$(2.32) \quad |(y_\varepsilon^u)^-(i\varepsilon) - (y_\varepsilon^u)^+(i\varepsilon)| = |(y_\varepsilon^u)^-(i\varepsilon) - e^{-F\varepsilon} P(y_\varepsilon^u)^-(i\varepsilon)|$$

and by (0.2)

$$(2.33) \quad y_\varepsilon^-(i\varepsilon) = e^{-A\varepsilon} e^{-F\varepsilon} P y_\varepsilon^-((i-1)\varepsilon) + \int_{(i-1)\varepsilon}^{i\varepsilon} e^{-A(i\varepsilon-s)} (Bu + f) ds.$$

Note also that

$$\begin{aligned} & \int_{i\varepsilon}^t |(y_\varepsilon^u)'|^2 ds + (A y_\varepsilon^u(t), y_\varepsilon^u(t)) \\ & \leq (A(y_\varepsilon^u)^+(i\varepsilon), (y_\varepsilon^u)^+(i\varepsilon)) + \int_{i\varepsilon}^t (|Bu|^2 + |f|^2) ds \quad \text{for } t \in [i\varepsilon, (i+1)\varepsilon]. \end{aligned}$$

Since as a consequence of assumptions (1.2), (1.3)

$$\begin{aligned} (A(y_\varepsilon^u)^+(i\varepsilon), (y_\varepsilon^u)^+(i\varepsilon)) &= (A e^{-F\varepsilon} P(y_\varepsilon^u)^-(i\varepsilon), e^{-F\varepsilon} P(y_\varepsilon^u)^-(i\varepsilon)) \\ &\leq (A P(y_\varepsilon^u)^-(i\varepsilon), P(y_\varepsilon^u)^-(i\varepsilon)) \\ &\leq (A(y_\varepsilon^u)^-(i\varepsilon), (y_\varepsilon^u)^-(i\varepsilon)) \end{aligned}$$

the latter yields

$$(2.34) \quad \sum_{i=0}^{N-1} \int_{i\varepsilon}^{(i+1)\varepsilon} |(y_\varepsilon^u)'|^2 ds + (A y_\varepsilon^u(t), y_\varepsilon^u(t)) \leq (A y_0, y_0) + \int_0^T (|Bu|^2 + |f|^2) ds.$$

In particular, it follows that $\{y_\varepsilon^u\}$ is bounded in $L^\infty(0, T; V)$ and $\{(y_\varepsilon^u)^-(i\varepsilon)\}$ is compact in H .

Now since $e^{-A\varepsilon}D(F) \subset D(F)$ for all $\varepsilon > 0$ we see by (2.33) that

$$|y_\varepsilon^-(i\varepsilon) - Py_\varepsilon^-(i\varepsilon)| \leq \int_{(i-1)\varepsilon}^{i\varepsilon} |Bu + f| \, ds \leq C\varepsilon^{1/2} \quad \text{for all } i.$$

Along with (2.32) the latter yields

$$(2.35) \quad \lim_{\varepsilon \rightarrow 0} |(y_\varepsilon^u)^-(i\varepsilon) - (y_\varepsilon^u)^+(i\varepsilon)| = 0$$

because $\{y_\varepsilon^-(i\varepsilon)\}$ is compact and e^{-Ft} is continuous in t .

Now again by (0.3) we have for $t \in [i\varepsilon, (i+1)\varepsilon]$

$$|y_\varepsilon^u(t) - (y_\varepsilon^u)^+(i\varepsilon)|^2 \leq C(\varepsilon \|(y_\varepsilon^u)^+(i\varepsilon)\|^2 + \int_{i\varepsilon}^t (|Bu|^2 + |f|^2) \, ds)$$

and so by (2.31)', (2.35) we infer that

$$\lim_{\varepsilon \rightarrow 0} y_\varepsilon^u(t) = \lim_{\varepsilon \rightarrow 0} \tilde{y}_\varepsilon(t) = y^u(t) \quad \text{strongly in } H \text{ for } t \in [0, T]$$

as claimed.

The second part of Proposition 2 follows by (2.31) and the previous discussion if we take into account the fact that for each $\varepsilon > 0$ the operator $\Pi S_1(\varepsilon)S_2(\varepsilon)$ is nonexpansive on X . An immediate consequence of Proposition 2 is that $\lim_{\varepsilon \rightarrow 0} \phi_\varepsilon(u) = \phi(u)$ for all $u \in U$.

Proof of Theorem 1. Let u_ε^* be an optimal controller for problem (P_ε) and let y_ε^* be the corresponding solution to system (0.2), (0.3). By assumption (1.4), $\{u_\varepsilon^*\}$ is bounded in U and so on a subsequence $\varepsilon_n \rightarrow 0$ we have

$$(2.36) \quad u_{\varepsilon_n} \rightarrow u^* \quad \text{weakly in } U$$

while by Proposition 2

$$(2.37) \quad y_{\varepsilon_n}^*(t) \rightarrow y^{u^*}(t) \quad \text{strongly in } H \text{ for all } t \in [0, T].$$

(We set $y_{\varepsilon_n}^* = y_{\varepsilon_n \varepsilon_n}^*$.) Since by estimate (2.34), $\{y_{\varepsilon_n}^*\}$ is bounded in $L^\infty(0, T; H)$ we see by (2.37) that $g(y_{\varepsilon_n}^*) \rightarrow g(y^{u^*})$ in $L^1(0, T)$. Finally, since h is weakly lower semicontinuous, it follows by (2.36) that

$$\liminf_{n \rightarrow \infty} h(u_{\varepsilon_n}^*) \geq h(u^*).$$

Then if we let ε_n tend to zero in the obvious inequality (\tilde{u}^* is an optimal control for problem (P))

$$\phi_{\varepsilon_n}(u_{\varepsilon_n}) \leq \phi_{\varepsilon_n}(\tilde{u}^*)$$

it follows by (2.29) that

$$(2.38) \quad \phi(u^*) \leq \liminf_{n \rightarrow \infty} \phi_{\varepsilon_n}(u_{\varepsilon_n}) \leq \phi(\tilde{u}^*).$$

Hence

$$\lim_{\varepsilon_n \rightarrow 0} \phi_{\varepsilon_n}(u_{\varepsilon_n}) = \phi(\tilde{u}^*) = \inf \{\phi(u); u \in U\}$$

and u^* is an optimal controller as claimed. To prove (1.9) we set $\tilde{y}_\varepsilon = y^{u^*}$ and note that by (2.36) and the Arzelà theorem, we have

$$\tilde{y}_{\varepsilon_n} \rightarrow y^* = y^{u^*} \quad \text{in } C([0, T]; H).$$

Hence

$$\int_0^T g(y_{\varepsilon_n}(t)) dt + \varphi_0(\tilde{y}_{\varepsilon_n}(T)) \rightarrow \int_0^T g(y^*(t)) dt + \varphi_0(y^*(T))$$

and by (2.38) we see that $h(u_{\varepsilon_n}) \rightarrow h(u^*)$. Hence $\lim_{n \rightarrow \infty} \phi(u_{\varepsilon_n}^*) = \phi(u^*)$ and since $\{\varepsilon_n\}$ is arbitrary (1.9) follows.

Assume now that Hypothesis (ii) holds, i.e., $F = \partial I_C$. Then $e^{-\varepsilon F} P = P = (1 + \varepsilon \partial I_C)^{-1}$ for all $\varepsilon > 0$, $K = C$ and system (0.3), (0.4) becomes

$$(2.39) \quad \begin{aligned} y'(t) + Ay(t) &= (Bu)(t) + f(t) \quad \text{a.e. } t \in [i\varepsilon, (i+1)\varepsilon], \\ y^+(i\varepsilon) &= Py^-(i\varepsilon). \end{aligned}$$

Let $\{u_{\varepsilon_n}^*\}$ be a weakly convergent subsequence of $\{u_\varepsilon^*\}$ and let u^* be its weak limit (see (2.36)). We see $u_n = u_{\varepsilon_n}^*$ and $y_n = y_{\varepsilon_n}^*$. For simplicity we set $\varepsilon_n = \varepsilon$ and $N_n = N$.

By estimate (2.34) we have

$$(2.40) \quad \sum_{i=0}^{N-1} \int_{i\varepsilon}^{(i+1)\varepsilon} |y_n'(t)|^2 dt + \|y_n(t)\|^2 \leq C \left(\|y_0\|^2 + \int_0^T (|u_n|_U^2 + |f|^2) dt \right)$$

and therefore

$$(2.40)' \quad \int_0^T |Ay_n(t)|^2 dt \leq C.$$

On the other hand, we have

$$(2.41) \quad \begin{aligned} & \sum_{i=0}^N |y_n^+(i\varepsilon) - y_n^-(i\varepsilon)| \\ & \leq \sum_{i=0}^N |Py_n^-(i\varepsilon) - e^{-A\varepsilon} Py_n^-((i-1)\varepsilon)| + \int_0^T (|Bu_n(s)| + |f(s)|) ds. \end{aligned}$$

Since by virtue of assumption (1.2), $e^{-At}C \subset C$ for all $t \geq 0$ (see [6] or [2, p. 183]) we have

$$\begin{aligned} |Py_n^-(i\varepsilon) - e^{-A\varepsilon} Py_n^-((i-1)\varepsilon)| &= |Py_n^-(i\varepsilon) - P e^{-A\varepsilon} Py_n^-((i-1)\varepsilon)| \\ &\leq |y_n^-(i\varepsilon) - e^{-A\varepsilon} Py_n^-((i-1)\varepsilon)| \\ &\leq \int_{(i-1)\varepsilon}^{i\varepsilon} (|Bu_n| + |f|) ds \end{aligned}$$

and by (2.41) we conclude that

$$(2.42) \quad \sum_{i=0}^N |y_n^+(i\varepsilon) - y_n^-(i\varepsilon)| \leq 2 \int_0^T (|Bu_n| + |f|) ds.$$

The functions y_n are of bounded variation on $[0, T]$ and by estimates (2.40), we see that

$$\bigvee_0^T y_n + \|y_n(t)\| \leq C \quad \text{for all } n \text{ and } t \in [0, T]$$

where $\bigvee_0^T y_n$ stands for the variation of $y_n : [0, T] \rightarrow H$. Since the injection of V into

H is compact we conclude, by virtue of the infinite-dimensional Helly theorem, that on a subsequence, again denoted y_n , we have

$$(2.43) \quad y_n(t) \rightarrow y^*(t) \text{ strongly in } H \text{ for all } t \in [0, T] \text{ and weak* in } L^\infty(0, T; V).$$

By estimate (2.40)' we see that $Ay^* \in L^2(0, T; H)$.

Now let $y \in C$ be arbitrary but fixed and let $t \in [k\varepsilon, (k+1)\varepsilon]$, $s \in [i\varepsilon, (i+1)\varepsilon]$, $i < k$, two points on the interval $[0, T]$. By (2.39) we have

$$\begin{aligned} & \frac{1}{2} (|y_n(t) - y|^2 - |y_n^+(k\varepsilon) - y|^2) + \frac{1}{2} \sum_{j=i}^k (|y_n^-(j\varepsilon) - y|^2 - |y_n^+((j-1)\varepsilon) - y|^2) \\ & \quad + \frac{1}{2} (|y_n^-((i+1)\varepsilon) - y|^2 - |y^n(s) - y|^2) \\ & = \int_s^t (Bu_n + f - Ay_n, y_n - y) d\tau. \end{aligned}$$

Hence

$$\begin{aligned} \frac{1}{2} |y_n(t) - y|^2 &= \int_s^t (Bu_n + f - Ay_n, y_n - y) d\tau + \frac{1}{2} |y_n(s) - y|^2 \\ & \quad + \frac{1}{2} \sum_{j=i}^{k+1} (|y_n^+(j\varepsilon) - y|^2 - |y_n^-(j\varepsilon) - y|^2). \end{aligned}$$

On the other hand, we have

$$\begin{aligned} \frac{1}{2} (|y_n^+(j\varepsilon) - y|^2 - |y_n^-(j\varepsilon) - y|^2) &\leq (y_n^+(j\varepsilon) - y_n^-(j\varepsilon), y_n^+(j\varepsilon) - y) \\ &= (Py_n^-(j\varepsilon) - y_n^-(j\varepsilon), Py_n^-(j\varepsilon) - y) \leq 0 \end{aligned}$$

because $1 - P = \lambda(\partial I_C)_\lambda \in \lambda \partial I_C P$ for all $\lambda > 0$. Hence

$$\frac{1}{2} (|y_n(t) - y|^2 - |y_n(s) - y|^2) \leq \int_s^t (Bu_n + f - Ay_n, y_n - y) d\tau.$$

Letting n tend to ∞ in the previous inequality we get

$$(2.44) \quad \frac{1}{2} (|y^*(t) - y|^2 - |y^*(s) - y|^2) \leq \int_s^t (Bu^* + f - Ay^*, y^* - y) d\tau$$

for $0 \leq s \leq t \leq T$.

Before proceeding further let us observe that $y^*(t) \in C$ almost everywhere $t \in [0, T]$. Indeed by (2.39) we have

$$y_n(t) = e^{-A(t-i\varepsilon)} Py_n^-(i\varepsilon) + \int_{i\varepsilon}^t e^{-A(t-s)} (Bu_n + f) ds \quad \text{for } t \in [i\varepsilon, (i+1)\varepsilon]$$

and therefore

$$|y_n(t) - Py_n(t)| \leq \left| \int_{i\varepsilon}^t e^{-A(t-s)} (Bu_n + f) ds \right| \leq C\varepsilon_n^{1/2} \quad \text{for } t \in [i\varepsilon_n, (i+1)\varepsilon_n]$$

because $e^{-At}C \subset C$. Hence $y^*(t) = Py^*(t)$ as claimed.

Now in (2.44) take $y = y^*(s)$. By Gronwall's lemma we obtain

$$|y^*(t) - y^*(s)| \leq \int_s^t |Bu^* + f + Ay^*| d\tau \quad \text{for } 0 \leq s \leq t \leq T$$

and therefore $y^*: [0, T] \rightarrow H$ is absolutely continuous and almost everywhere differentiable. Now (2.44) yields

$$(2.45) \quad (y^*(t) - y^*(s), y^*(s) - y) \leq \int_s^t (Bu^*(\tau) + f(\tau) - Ay^*(\tau), y^*(\tau) - y) d\tau.$$

Then dividing by $t-s$ and letting s tend to zero we see that

$$(y^{*'}(t) + Ay^*(t) - Bu^*(t) - f(t), y^*(t) - y) \leq 0 \quad \text{a.e. } t \in [0, T]$$

for all $y \in C$. Hence

$$y^{*'} + Ay^* + \partial I_C(y^*) \ni f + Bu^* \quad \text{a.e. } t \in [0, T].$$

In other words, we have proved that $y^* = y^{u^*}$. From now on the proof is identical with that of first case.

3. Problem P_ε in a particular case. We will consider here the special case of free boundary problem (1.15), where for the sake of simplicity we take $\varphi_0 \equiv 0$. In this case system (0.2), (0.3) becomes

$$\begin{aligned} (3.1) \quad & y_t - \Delta y = Bu \quad \text{in } Q_\varepsilon^i = (i\varepsilon, (i+1)\varepsilon) \times \Omega, \\ & y = 0 \quad \text{in } \Sigma_\varepsilon^i = (i\varepsilon, (i+1)\varepsilon) \times \Gamma, \\ & y(0, x) = y_0(x), \quad x \in \Omega, \\ & y^+(i\varepsilon, x) = Py^-(i\varepsilon, x), \quad x \in \Omega \end{aligned}$$

where $Pz = \max(z, 0)$ for all $z \in L^2(\Omega)$ and $y_0 \in H_0^1(\Omega)$, $y_0 \geq 0$ a.e. in Ω .

We may approximate system (3.1) by the following the family of penalized smooth control systems

$$\begin{aligned} (3.2) \quad & y_t - \Delta y = Bu \quad \text{in } Q_\varepsilon^i, \\ & y = 0 \quad \text{in } \Sigma_\varepsilon^i, \\ & y^+(i\varepsilon, x) = (1 + \beta_\lambda)^{-1} y^-(i\varepsilon, x), \quad y^+(0, x) = y_0(x) \quad \text{in } \Omega \end{aligned}$$

where $\lambda > 0$ and

$$(3.3) \quad \beta_\lambda(r) = \begin{cases} \lambda^{-1}r + 2^{-1} & \text{for } r \leq -\lambda, \\ -(2\lambda^2)^{-1}r^2 & \text{for } -\lambda \leq r \leq 0, \\ 0 & \text{for } r > 0. \end{cases}$$

We may obtain first-order necessary conditions of optimality for problem (P_ε) by using the method developed in [3]. To this aim consider an optimal pair $(y_\varepsilon^*, u_\varepsilon^*)$ to problem (P_ε) .

Let $(y_\varepsilon^\lambda, u_\varepsilon^\lambda)$ be an optimal pair for control problem with state equation (3.2) and cost functional

$$\int_0^T g(y(t)) dt + h(u) + \frac{1}{2} \|u - u_\varepsilon^*\|_U^2.$$

It follows that for $\lambda \rightarrow 0$

$$u_\varepsilon^\lambda \rightarrow u_\varepsilon^* \quad \text{strongly in } U,$$

$$y_\varepsilon^\lambda \rightarrow y_\varepsilon^* \quad \text{strongly in every } C([i\varepsilon, (i+1)\varepsilon]; H) \text{ and} \\ \text{weakly in } W^{1,2}([i, (i+1)\varepsilon]; H) \cap L^2((i\varepsilon, (i+1)\varepsilon); H_0^1(\Omega)).$$

The optimality conditions for this problem can be written as

$$\begin{aligned} (3.4) \quad & (p_\varepsilon^\lambda)_t + \Delta p_\varepsilon^\lambda = \nabla g(y_\varepsilon^\lambda) \quad \text{in } Q_\varepsilon^i, \\ & (p_\varepsilon^\lambda)^-((i+1)\varepsilon) = (1 + \beta_\lambda'((y_\varepsilon^\lambda)^+((i+1)\varepsilon)))^{-1} (p_\varepsilon^\lambda)^+((i+1)\varepsilon) \quad \text{in } \Omega, \\ & (p_\varepsilon^\lambda)^-(T) = 0, \end{aligned}$$

$$(3.5) \quad B^* p_\varepsilon^\lambda \in \partial h(u_\varepsilon^\lambda) + u_\varepsilon^\lambda - u_\varepsilon^*.$$

Multiplying (3.4) by p_ε^λ and integrating by parts, we obtain after some calculations the following estimates:

$$\int_0^T \|p_\varepsilon^\lambda(t)\|^2 dt + |p_\varepsilon^\lambda(t)| \leq C \quad \text{for all } \lambda > 0,$$

$$\sum_{i=0}^{N-1} \int_{i\varepsilon}^{(i+1)\varepsilon} \|(p_\varepsilon^\lambda)_i\|_*^2 dt \leq C \quad \text{for all } \lambda > 0.$$

(We have denoted as usual by $\|\cdot\|$, $|\cdot|$ and $\|\cdot\|_*$ the norms in $V = H_0^1(\Omega)$, $H = L^2(\Omega)$ and $V' = H^{-1}(\Omega)$.)

Then by the Helly theorem we infer that on a subsequence again denoted λ , we have

$$p_\varepsilon^\lambda(t) \rightarrow p(t) \quad \text{strongly in } H^{-1}(\Omega) \text{ for every } t \in [0, T],$$

$$p_\varepsilon^\lambda \rightarrow p \quad \text{weakly in } L^2(0, T; H_0^1(\Omega)), \text{ weak-star in } L^\infty(0, T; L^2(\Omega)).$$

Recalling the inequality (see [12])

$$\|p_\varepsilon^\lambda - p\|_{L^2(\Omega)} \leq \delta \|p_\varepsilon^\lambda - p\|_{H_0^1(\Omega)} + \eta(\delta) \|p_\varepsilon^\lambda - p\|_{H^{-1}(\Omega)}$$

where $\eta(\delta) \rightarrow 0$ as $\delta \rightarrow 0$, we conclude that

$$p_\varepsilon^\lambda \rightarrow p \quad \text{strongly in } L^2(Q).$$

Then letting λ tend to zero in (3.4), (3.5) we see that

$$p_t + \Delta p = \nabla g(y_\varepsilon^*) \quad \text{in } Q_\varepsilon^i,$$

$$p^-((i+1)\varepsilon, x) = \mu_i p^+((i+1)\varepsilon, x), \quad x \in \Omega; \quad i = 0, 1, \dots, N-2,$$

$$p^-(T, x) = 0$$

where

$$(3.6) \quad \mu_i = w - \lim_{\lambda \rightarrow 0} (1 + \beta_\lambda'((y_\varepsilon^\lambda)^+((i+1)\varepsilon)))^{-1} \quad \text{in } L^2(\Omega).$$

We set $E_\lambda^1 = \{x \in \Omega; (y_\varepsilon^\lambda)^+((i+1)\varepsilon, x) \leq -\lambda\}$ and

$$E_\lambda^2 = \{x \in \Omega; -\lambda < (y_\varepsilon^\lambda)^+((i+1)\varepsilon, x) \leq 0\}.$$

We have

$$(3.7) \quad \beta_\lambda((y_\varepsilon^\lambda)^+((i+1)\varepsilon)) = \beta_\lambda'((y_\varepsilon^\lambda)^+((i+1)\varepsilon))(y_\varepsilon^\lambda)^+((i+1)\varepsilon) \\ + 2^{-1}\chi_\lambda^1(x) + (2\chi_\lambda^2)^{-1}((y_\varepsilon^\lambda)^+((i+1)\varepsilon)^2)\chi_\lambda^2(x)$$

where χ_λ^1 , χ_λ^2 are the characteristic functions of E_λ^1 and E_λ^2 , respectively.

Since $\{\beta_\lambda((y_\varepsilon^\lambda)^+((i+1)\varepsilon))\}$ is bounded in $L^2(\Omega)$ and weakly convergent to some $\nu \in \partial I_C((y_\varepsilon^*)^+((i+1)\varepsilon))$, i.e., $\nu(x) = 0$ in $\{x; (y_\varepsilon^*)^+((i+1)\varepsilon) > 0\}$, $\nu(x) \leq 0$ in Ω , we see by (3.6) and (3.7) that

$$\mu_i(x) = 1 \quad \text{in } [(y_\varepsilon^*)^+((i+1)\varepsilon) > 0],$$

$$\mu_i(x) = 0 \quad \text{in } [(y_\varepsilon^*)^+((i+1)\varepsilon) = 0].$$

PROPOSITION 3. *Let $(y_\varepsilon^*, u_\varepsilon^*)$ be any optimal pair in problem (P_ε) with state equation (3.1). Then we have $p \in BV([0, T]; H^{-1}(\Omega)) \cap L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$ such*

that $p_i \in W^{1,2}([\varepsilon, (i+1)\varepsilon]; H^{-1}(\Omega))$ for $i = 0, 1, \dots, N-1$ and

$$(3.8) \quad \begin{aligned} p_i + \Delta p &= \nabla g(y_\varepsilon^*) \quad \text{in } Q_\varepsilon^i \text{ for all } i, \\ p^-((i+1)\varepsilon, x) &= p^+((i+1)\varepsilon, x) \quad \text{in } [x; (y_\varepsilon^*)^+((i+1)\varepsilon, x) > 0], \\ p^-((i+1)\varepsilon, x) &= 0 \quad \text{in } [x; (y_\varepsilon^*)^+((i+1)\varepsilon, x) = 0], \\ p^-(T, x) &= 0 \quad \text{in } \Omega, \end{aligned}$$

$$(3.9) \quad B^*p \in \partial h(u_\varepsilon^*).$$

To solve problem (P_ε) numerically we shall use the following gradient type algorithm (δ is a prescribed precision, MAXITER is the prescribed maximum number of iterations and $\varepsilon = T/N$):

Step 0. Choose $u^{(0)} \in L^2(0, T; U)$;

$$n := 0, \quad \text{iter} := 1.$$

Step 1. Compute y_n solving

$$\begin{aligned} \frac{\partial y_n}{\partial t} - \Delta y_n &= Bu^{(n)} \quad \text{on } Q_\varepsilon^i, \\ y_n &= 0 \quad \text{on } \Sigma_\varepsilon^i, \quad i = 0, 1, \dots, N-1, \\ y_n^+(i\varepsilon, x) &= \max(y_n^+(i\varepsilon, x), 0) \quad \text{for } x \in \Omega, \quad i = 1, 2, \dots, N-1, \\ y_n^+(0, x) &= y_0(x) \quad \text{for } x \in \Omega. \end{aligned}$$

Step 2. Compute p_n solving

$$\begin{aligned} \frac{\partial p_n}{\partial t} + \Delta p_n &= \nabla g(y_n) \quad \text{on } Q_\varepsilon^i, \\ p_n^{(n)} &= 0 \quad \text{on } \Sigma_\varepsilon^i \quad \text{for } i = 0, 1, \dots, N-1, \\ p_n^-((i+1)\varepsilon, x) &= \begin{cases} p_n^+((i+1)\varepsilon, x) & \text{if } y_n^+((i+1)\varepsilon, x) > 0, \\ 0 & \text{if } y_n^+((i+1)\varepsilon, x) = 0, \end{cases} \\ p_n^-(T, x) &= 0 \quad \text{for } x \in \Omega. \end{aligned}$$

Step 3. Compute $w^{(n)} \in \partial h(u^{(n)}) - B^*p_n$;
compute ρ_n —the steplength of the gradient method.

Step 4. Test: $\|\rho_n w^{(n)}\| \leq \delta$?

-YES \rightarrow STOP (the algorithm is convergent).

-NO \rightarrow iter := iter + 1;

Test: iter \leq MAXITER?

-YES $\rightarrow u^{(n+1)} := u^{(n)} - \rho_n w^{(n)}$;

$n := n + 1$;

GO TO Step 1.

-NO \rightarrow STOP (the algorithm is not convergent).

Remark 3.1. The convergence test may also deal with the difference $\phi_\varepsilon(u^{(n+1)}) - \phi_\varepsilon(u^{(n)})$.

This gradient algorithm yields a sequence $\{\phi_\varepsilon(u^{(n)})\}$ which is convergent to a value corresponding to a stationary point of ϕ_ε [8, p. 70]. Moreover, if the sequence above is convergent to $\inf \phi_\varepsilon$ then $\{u^{(n)}\}$ is a minimizing sequence for problem (P), that is

$$\lim_{\substack{\varepsilon \rightarrow 0 \\ n \rightarrow \infty}} \phi_\varepsilon(u^{(n)}) = \inf \phi.$$

A numerical example will be presented in § 5.

4. Product formula for a boundary optimal control problem. We shall consider here the following problem: Minimize

$$(4.1) \quad \int_0^T g(y(t)) dt + h(u) + \varphi_0(y(T))$$

on all $y \in L^2(Q)$ and $u \in L^2(\Sigma_1)$ subject to

$$(4.2) \quad \begin{aligned} y_t - \Delta y &= f_0 \quad \text{in } [(t, x) \in Q; y(t, x) > 0], \\ y_t - \Delta y &\geq f_0, \quad y \geq 0 \quad \text{in } Q = (0, T) \times \Omega, \\ y(0, x) &= y_0(x), \quad x \in \Omega, \\ \frac{\partial y}{\partial \nu} + \alpha y &= \int_0^t (Bu)(s) ds \quad \text{for } (t, x) \in \Sigma_1 = (0, T) \times \Gamma_1, \\ y &= 0 \quad \text{in } \Sigma_2 = (0, T) \times \Gamma_2. \end{aligned}$$

Here Ω is a bounded and open subset of R^n with a sufficiently smooth boundary $\Gamma = \Gamma_1 \cup \Gamma_2$, where $\Gamma_1 \cap \Gamma_2 = \emptyset$, B is a linear continuous operator from a Hilbert space of controllers U to $L^2(\Sigma_1)$, $\alpha \geq 0$, and y_0, f_0 are given functions such that

$$(4.3) \quad f_0 \in W^{1,2}([0, T]; L^2(\Omega)),$$

$$(4.4) \quad y_0 \in H^1(\Omega), \quad y_0 \geq 0 \quad \text{in } \Omega, \quad y_0 = 0 \quad \text{in } \Gamma_2.$$

In regard to the functions $g: L^2(\Omega) \rightarrow R$, $\varphi_0: L^2(\Omega) \rightarrow R$ and $h: U \rightarrow \bar{R}$ we will assume that hypothesis (6) in § 1 holds.

It is well known (see, for instance, [3], [4]) that the one-phase Stefan problem

$$\begin{aligned} \theta_t - \Delta \theta &= 0 \quad \text{in } \{(t, x) \in Q; l(x) < t < T\}, \\ \theta &= 0 \quad \text{in } \{(t, x) \in Q; l(x) \geq t\}, \\ \nabla x \theta \cdot \nabla l(x) &= -\rho \quad \text{in } \{(t, x) \in Q; t = l(x)\}, \\ \frac{\partial \theta}{\partial \nu} + \alpha(\theta - w) &= 0 \quad \text{in } \Sigma_1, \quad \theta = 0 \quad \text{in } \Sigma_2, \\ \theta(x, 0) &= \theta_0(x) \quad \text{in } \Omega_0; \quad \theta(0, x) = 0 \quad \text{in } \Omega \setminus \Omega_0 \end{aligned}$$

(here $t = l(x)$ is the equation of liquid-solid interface) can be put into the form (4.2), where

$$y(t, x) = \int_0^t \theta(s, x) \chi(s, x) ds$$

(χ is the characteristic function of the set $\{(t, x); l(x) \leq t\}$) and

$$(4.5) \quad (Bu)(s, \sigma) = w(s, \sigma) \quad \text{a.e. } (s, \sigma) \in \Sigma_1,$$

$$(4.6) \quad f_0(x) = \theta_0(x) \quad \text{for } x \in \Omega_0, \quad f_0(x) = -\rho \quad \text{for } x \in \Omega \setminus \Omega_0.$$

Under assumptions (4.3), (4.4) the free boundary problem admits a unique solution $y \in W^{1,2}([0, T]; L^2(\Omega)) \cap L^\infty(0, T; V)$ (see [3, p. 157]). Here $V = \{y \in H^1(\Omega); y = 0 \text{ in } \Gamma_2\}$.

First-order necessary conditions for problems of this form have been established in [3], [4], [11].

We associate to problem (4.2) the approximation scheme

$$\begin{aligned}
 (4.7) \quad & y_t - \Delta y = f_0 \quad \text{in } Q_\varepsilon^i = (i\varepsilon, (i+1)\varepsilon) \times \Omega, \\
 & \frac{\partial y}{\partial \nu} + \alpha y = v \quad \text{in } \Sigma_\varepsilon^{1,i} = (i\varepsilon, (i+1)\varepsilon) \times \Gamma_1, \\
 & y = 0 \quad \text{in } \Sigma_\varepsilon^{2,i} = (i\varepsilon, (i+1)\varepsilon) \times \Gamma_2, \\
 & y^+(i\varepsilon, x) = Py^-(i\varepsilon, x), \quad x \in \Omega, \quad i = 1, \dots, N-1 \\
 & y(0, x) = y_0(x), \quad x \in \Omega
 \end{aligned}$$

where

$$\forall(t) = \int_0^t (Bu)(s) \, ds \quad \text{for } t \in [0, T], \quad Py = \max(y, 0).$$

Note that for every $\varepsilon > 0$, problem (4.7) has a unique solution $y \in L^\infty(0, T; V) \cap W^{1,2}([i\varepsilon, (i+1)\varepsilon]; L^2(\Omega))$ on every interval $[i\varepsilon, (i+1)\varepsilon]$. We set

$$(4.8) \quad \psi(u) = \int_0^T g(y''(t)) \, dt + h(u) + \varphi_0(y''(T))$$

and

$$(4.9) \quad \psi_\varepsilon(u) = \int_0^T g(y_\varepsilon''(t)) \, dt + h(u) + \varphi_0(y_\varepsilon''(T))$$

where y'' and y_ε'' are the solutions to (4.2) and (4.7), respectively.

Then we consider the optimal control problem with cost functional (4.1) and state equation (4.7), i.e.,

$$(4.10) \quad \min \{\psi_\varepsilon(u); u \in U\}.$$

We have for the approximating problem (4.10) a convergence of the type shown in Theorem 1.

THEOREM 2. *Let $\{u_\varepsilon^*\}$ be a sequence of optimal controllers for problems (4.10). Then*

$$(4.11) \quad \lim_{\varepsilon \rightarrow 0} \psi(u_\varepsilon^*) = \inf \{\psi(u); u \in U\}$$

and

$$(4.12) \quad \liminf_{\varepsilon \rightarrow 0} \psi_\varepsilon = \inf \psi.$$

Moreover, every weak limit point of $\{u_\varepsilon^*\}$ for $\varepsilon \rightarrow 0$ is an optimal controller for problem (4.1).

Proof. Let $(y_\varepsilon^*, u_\varepsilon^*)$ be an optimal pair for problem (4.10). Then as seen earlier in the proof of Theorem 1 there is a subsequence $\varepsilon_n \rightarrow 0$ such that

$$(4.13) \quad u_{\varepsilon_n}^* \rightarrow u^* \quad \text{weakly in } U.$$

We postpone for the time being the verification of Lemma 1 below.

LEMMA 1. By (4.13) it follows that

$$(4.14) \quad y_{\varepsilon_n}(t) \rightarrow y^*(t) \quad \text{strongly in } H \text{ for all } t \in [0, T] \text{ and weak star in } L^\infty(0, T; V)$$

where $y^* = y^{u^*}$ is the solution to system (4.61), where $u = u^*$.

Then by (4.13), (4.14) we infer that

$$\liminf_{n \rightarrow \infty} h(u_{\varepsilon_n}) \geq h(u^*), \quad \lim_{n \rightarrow \infty} g(y_{\varepsilon_n}) = g(y^*) \quad \text{in } L^1(0, T)$$

and this implies that

$$(4.15) \quad \lim_{n \rightarrow \infty} \psi_{\varepsilon_n}(u_{\varepsilon_n}) = \psi(u^*) = \inf \psi.$$

Now if $z_\varepsilon = y^{u_\varepsilon^*}$ then we have the estimate (see, for instance, [3, p. 157])

$$\|z_\varepsilon(t)\|_V + \|z'_\varepsilon(t)\|_{L^2(\Omega)}^2 dt \leq C(1 + \|Bu_\varepsilon\|_{L^2(\Sigma_1)}^2)$$

and this implies via the Arzelá theorem that $\{z_\varepsilon\}$ is compact in $C([0, T]; L^2(\Omega))$. Hence without loss of generality we may assume that

$$z_{\varepsilon_n} \rightarrow y^* \quad \text{in } C([0, T]; L^2(\Omega)).$$

Since $h(u_{\varepsilon_n}) \rightarrow h(u^*)$ as easily follows by (4.15), we infer that $\psi(u_{\varepsilon_n}) \rightarrow \psi(u^*)$ as claimed.

Proof of Lemma 1. The proof is similar to part (ii) of Theorem 2 with some modifications. We set $u_{\varepsilon_n} = u_n$, $y_{\varepsilon_n} = y_n$ and $v_n(t) = \int_0^t (Bu_n) ds$. By (4.7) we have (setting $\varepsilon_n = \varepsilon$)

$$\begin{aligned} & |y_n^-((i+1)\varepsilon)|_2^2 + \int_{Q_\varepsilon^i} |\nabla y_n|^2 dx dt + \alpha \int_{\Sigma_\varepsilon^{1,i}} |y_n|^2 dx dt \\ & \leq \int_{Q_\varepsilon^i} |f_0|^2 dx dt + \int_{\Sigma_\varepsilon^{1,i}} |v_n|^2 dx dt + |y_n^+(i\varepsilon)|_2^2 \end{aligned}$$

and

$$\begin{aligned} & \int_{i\varepsilon}^{(i+1)\varepsilon} |(y_n)_t|_2^2 dt + |\nabla y_n^-((i+1)\varepsilon)|_2^2 + \alpha \int_{\Gamma_1} |y_n^-((i+1)\varepsilon, x)|^2 dx \\ & \leq |\nabla y_n^+(i\varepsilon)|_2^2 + \alpha \int_{\Gamma_1} |y_n^+(i\varepsilon, x)|^2 dx + \int_{Q_\varepsilon^i} |f_0|^2 dx dt + \int_{\Sigma_\varepsilon^{1,i}} |v_n'|^2 dx dt \end{aligned}$$

where $|\cdot|_2$ stands for $L^2(\Omega)$ -norm. Combining these two inequalities we obtain

$$\begin{aligned} (4.16) \quad & \sum_{i=0}^{N-1} \int_{i\varepsilon}^{(i+1)\varepsilon} |(y_n)_t|_2^2 dt + \|y_n(t)\|^2 \\ & \leq \|y_0\|^2 + \int_Q |f_0|^2 dx dt + \int_{\Sigma_1} |Bu_n|^2 dx dt \end{aligned}$$

where $\|\cdot\|$ is the $H^1(\Omega)$ -norm.

In order to estimate

$$(4.17) \quad \sum_{i=0}^{N-1} |y_n^+(i\varepsilon) - y_n^-(i\varepsilon)|_2 = \sum_{i=0}^{N-1} |Py_n^-(i\varepsilon) - y_n^-(i\varepsilon)|_2$$

we observe that $y_n^-(i\varepsilon) = z_n(i\varepsilon) + \xi_n(i\varepsilon)$, where

$$\begin{aligned} (4.18) \quad & (z_n)_t - \Delta z_n = f_0 \quad \text{in } Q_\varepsilon^i, \\ & \frac{\partial z_n}{\partial \nu} + \alpha z_n = v_n \quad \text{in } \Sigma_\varepsilon^{1,i}, \quad z_n = 0 \quad \text{in } \Sigma_\varepsilon^{2,i}, \\ & z_n((i-1)\varepsilon) = 0 \quad \text{in } \Omega \end{aligned}$$

and

$$(4.19) \quad \begin{aligned} (\xi_n)_t - \Delta \xi_n &= 0 \quad \text{in } Q_\varepsilon^i, \\ \frac{\partial \xi_n}{\partial \nu} + \alpha \xi_n &= 0 \quad \text{in } \Sigma_\varepsilon^{1,i}, \quad \xi_n = 0 \quad \text{in } \Sigma_\varepsilon^{2,1}, \\ \xi_n((i-1)\varepsilon) &= Py_n^-((i-1)\varepsilon) \quad \text{in } \Omega. \end{aligned}$$

We have

$$\begin{aligned} |Py_n^-(i\varepsilon) - y_n^-(i\varepsilon)|_2 &\leq |Py_n^-(i\varepsilon) - \xi_n(i\varepsilon)|_2 + |z_n(i\varepsilon)|_2 \\ &\leq |Py_n^-(i\varepsilon) - Pe^{-\bar{\Lambda}\varepsilon} Py_n^-((i-1)\varepsilon)|_2 + \left(\int_{Q_\varepsilon^i} |f_0|^2 dx dt \right)^{1/2} \\ &\quad + \left(\int_{\Sigma_\varepsilon^{1,i}} |v_n|^2 dx dt \right)^{1/2} \\ &\leq 2(\|f_0\|_{L^2(Q_\varepsilon^i)} + \|v_n\|_{L^2(\Sigma_\varepsilon^{1,i})}) \end{aligned}$$

where $A\xi = \Delta \xi$ for $\xi \in D(A) = \{\xi \in H^1(\Omega); \partial \xi / \partial \nu + \alpha \xi = 0 \text{ in } \Gamma_1; \xi = 0 \text{ in } \Gamma_2\}$.

Putting the latter estimate in (4.17) we get

$$(4.20) \quad \sum_{i=0}^{N-1} |y_n^+(i\varepsilon) - y_n^-(i\varepsilon)|_2 \leq 2(\|f_0\|_{L^2(Q)} + \|v_n\|_{L^2(\Sigma_1)}) \leq C.$$

Hence the total variation of y_n on $[0, T]$ is uniformly bounded. We may therefore apply the Helly theorem to sequence $\{y_n\}$ to conclude that on a subsequence, again denoted $\{y_n\}$, we have

$$(4.21) \quad y_n(t) \rightarrow y^*(t) \quad \text{strongly in } H \text{ for all } t \in [0, T] \text{ and weak star in } L^\infty(0, T; V).$$

Now let $K = \{y \in V; y(x) \geq 0 \text{ almost everywhere } x \in \Omega\}$ and let y be any element of K . Let $s \leq t$ be two arbitrary points of $[0, T]$. If $s \in [i\varepsilon, (i+1)\varepsilon]$ and $t \in [k\varepsilon, (k+1)\varepsilon]$, where $i \leq k$, by (4.7) we obtain

$$\begin{aligned} &\frac{1}{2} (|y_n(t) - y|_2^2 - |y_n^+(k\varepsilon) - y|_2^2) + \frac{1}{2} \sum_{j=i}^k (|y_n^-(j\varepsilon) - y|_2^2 - |y_n^+((j-1)\varepsilon) - y|_2^2) \\ &\quad + \frac{1}{2} (|y_n^-((i+1)\varepsilon) - y|_2^2 - |y_n(s) - y|_2^2) \\ &= - \int_s^t d\tau \left(\int_\Omega \nabla y_n(\tau, x) \cdot \nabla (y_n(\tau, x) - y(x)) dx \right) \\ &\quad - \int_{\Gamma_1} (v_n(\tau, x) - \alpha y_n(\tau, x)) \cdot (y_n(\tau, x) - y(x)) dx \end{aligned}$$

and arguing as in the proof of Theorem 1 we obtain

$$\begin{aligned} &\frac{1}{2} (|y_n(t) - y|_2^2 - |y_n(s) - y|_2^2) \\ &\leq - \int_s^t d\tau \left(\int_\Omega (\nabla y_n \cdot \nabla (y_n - y) - f_0(y_n - y)) dx + \int_{\Gamma_1} (v_n - \alpha y_n)(y_n - y) dx \right) \end{aligned}$$

and if we let n tend to $+\infty$ it follows by (4.21) that

$$(4.22) \quad \begin{aligned} &\frac{1}{2} (|y^*(t) - y|_2^2 - |y^*(s) - y|_2^2) + \int_s^t d\tau \left(\int_\Omega (\nabla y^* \cdot \nabla (y^* - y) - f_0(y^* - y)) dx \right) \\ &\quad + \alpha \int_s^t d\tau \int_\Omega y^*(y^* - y) dx - \int_s^t d\tau \int_{\Gamma_1} v^*(y^* - y) dx \leq 0 \end{aligned}$$

where $v^*(t) = \int_0^t Bu^* d\tau$.

By (4.1) and (4.19) we see that for $t \in [(i-1)\varepsilon, i\varepsilon]$

$$y_n(t) = \xi_n(t) + z_n(t) = e^{-A(t-(i-1)\varepsilon)} P y_n^-((i-1)\varepsilon) + z_n(t).$$

Since e^{-At} maps K into itself we have

$$(4.23) \quad |y_n(t) - P y_n(t)|_2 \leq 2|z_n(t)|_2 \quad \text{for } t \in [(i-1)\varepsilon, i\varepsilon].$$

On the other hand, by (4.18) we see that

$$|z_n(t)|_2 \leq \int_{(i-1)\varepsilon}^{i\varepsilon} (|f_0|_2 + \|v_n\|_{L^2(\Gamma_1)}) ds \leq C\varepsilon^{1/2} (\|f_0\|_{L^2(Q)} + \|v_n\|_{L^2(\Gamma_1)}).$$

Hence $y^*(t) \in K$ almost everywhere $t \in [0, T]$.

Now if we take $y = y^*(s)$ in (4.22) and use Gronwall's lemma we find after some computation that

$$|y^*(t) - y^*(s)|_2^2 \leq C \left(\int_s^t (\|v^*(\tau)\|_{L^2(\Gamma_1)}^2 + |f_0(\tau)|_2^2) d\tau \right) \quad \text{for all } s \leq t \leq T.$$

Hence $y^*: [0, T] \rightarrow L^2(\Omega)$ is absolutely continuous and $y_i^* \in L^2(Q)$, i.e., $y^* \in W^{1,2}([0, T]; L^2(\Omega))$.

Now by (4.22) we have

$$\begin{aligned} (y^*(t) - y^*(s), y^*(s) - y)_2 + \int_s^t d\tau \int_{\Omega} (\nabla y^* \cdot \nabla (y^* - y) - f_0(y^* - y)) dx \\ + \int_s^t d\tau \int_{\Gamma_1} (\alpha y^* - v^*)(y^* - y) dx \leq 0 \end{aligned}$$

for all $y \in K$ and all $s \leq t \leq T$. (Here $(\cdot, \cdot)_2$ is the scalar product of $L^2(\Omega)$.)

Then dividing by $t - s$ and letting s tend to t we obtain

$$\begin{aligned} (4.24) \quad (y_i^*(t), y^*(t) - y)_2 + \int_{\Omega} \nabla y^*(t, x) \cdot \nabla (y^*(t, x) - y(x)) dx \\ + \int_{\Gamma_1} (\alpha y^*(t, x) - v^*(t, x))(y^*(t, x) - y(x)) dx \\ \leq \int_{\Omega} f_0(t, x)(y^*(t, x) - y(x)) dx \quad \text{a.e. } t \in [0, T] \quad \text{for all } y \in K. \end{aligned}$$

Inequality (4.24) shows that y^* is the solution to (4.1) thereby completing the proof of Lemma 1 and of Theorem 2.

5. Numerical results. In this section, we present numerical results for the algorithm described in § 3.

The algorithm was tested on the following example: $\Omega = (0, 1)$, $T = 1$. Let us consider a little more generally that $\Omega = (a, b)$. The cost functional is

$$\phi(u, t) = \frac{1}{2} \int_0^T \int_a^b (y(t, x) - y_0(x))^2 dx dt + \frac{1}{2} \int_0^T \int_a^b u^2(t, x) dx dt$$

and B is the identity operator.

The problem was discretized by use of finite differences. Consider the grid defined by the following equidistant points:

$$0 = t_1 < t_2 < \dots < t_{N+1} = T$$

with the steplength $\varepsilon = T/N$ and

$$a = x_1 < x_2 < \cdots < x_{M+1} = b$$

with the steplength $h = (b - a)/M$. The state y is approximated by two matrices, namely YL and YR , whose components are defined by

$$YL(i, j) = y(x_i, t_j - 0), \quad YR(i, j) = y(x_i, t_j + 0).$$

The adjoint state p is approximated the same way by the matrices PL and PR . The control u is represented by the matrix U with components

$$U(i, j) = u(x_i, \tau_j)$$

where $\tau_j = (t_j + t_{j+1})/2$. w defined in Step 3 is also represented by only one matrix W . In what follows we shall omit the iteration index (n).

The state system in Step 1 was solved ascending with respect to time levels, level by level. We have from the initial condition

$$YR(i, 1) = y_0(x_i) \quad \text{for } i = 1, 2, \dots, M+1.$$

Assume now that we know the values corresponding to level j . In order to reach the level $j+1$ we proceed as follows:

(i) We use the implicit method to discretize the differential equation. We obtain

$$(5.1) \quad \begin{aligned} (YL(i, j+1) - YR(i, j))/\varepsilon &= (YL(i+1, j+1) - 2^* YL(i, j+1) \\ &\quad + YL(i-1, j+1))/h^2 + U(i, j) \end{aligned}$$

for $i = 1, 2, \dots, M$.

From the boundary conditions we know that

$$YL(1, j+1) = YL(M+1, j+1) = 0.$$

Hence we obtain an algebraic linear system with respect to the unknowns $YL(i, j+1)$, $i = 1, 2, \dots, M$. Its matrix is a band one having the structure

$$(5.2) \quad \begin{bmatrix} 1+2c & -c & 0 & \cdots & 0 & 0 & 0 \\ -c & 1+2c & -c & \cdots & 0 & 0 & 0 \\ \vdots & & & & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -c & 1+2c & -c \\ 0 & 0 & 0 & \cdots & 0 & -c & 1+2c \end{bmatrix}$$

where $c = \varepsilon/h^2$.

(ii) Solving the system above we may compute

$$YR(i, j+1) = \max(YL(i, j+1), 0) \quad \text{for } i = 1, 2, \dots, M+1.$$

The adjoint state system in Step 2 was solved descending with respect to time levels, level by level. We have from the final condition

$$PL(i, N+1) = 0 \quad \text{for } i = 1, 2, \dots, M+1.$$

Assume that all values on level $j+1$ are known. In order to reach the level j we proceed as follows:

(i) We use the implicit method to discretize the differential equation, thus obtaining

$$(5.3) \quad \begin{aligned} & (PL(i, j+1) - PR(i, j))/\varepsilon + (PR(i+1, j) - 2*PR(i, j) + PR(i-1, j))/h^2 \\ & = g'(YR(i, j)) \quad \text{for } i=2, \dots, M. \end{aligned}$$

We have from the boundary conditions

$$PR(1, j) = PR(M+1, j) = 0$$

and we obtain an algebraic linear system with respect to the unknowns $PR(i, j)$, $i=2, \dots, M$. If all equations (5.3) are multiplied by (-1) then the matrix of this system is again (5.2).

(ii) We may now compute

$$PL(i, j) = \begin{cases} PR(i, j) & \text{if } YR(i, j) > 0, \\ 0 & \text{if } YR(i, j) = 0 \end{cases}$$

for $i=1, 2, \dots, M+1$. Of course the test dealing with $YR(i, j)$ and 0 must be carefully implemented with respect to the floating point arithmetic.

It is very important to point out now that all systems to be solved in Steps 1 and 2 have the same matrix (constant with respect to time levels). Therefore this matrix, namely (5.2), is inverted once and all linear systems are solved by multiplication of the inverse matrix times the corresponding right-hand side vector. Hence a large amount of time is saved concerning computations. This is a direct result of decoupling the operators A and F . In the presence of both operators the resulting nonlinear system cannot be solved in such a simple way. Moreover, in our case even the structure of the computer program is more simple and therefore the programming task is easier.

Concerning the computations of w in Step 3, a general idea is to replace the function h by a smooth approximation. Its gradient will be taken instead of the subdifferential ∂h . No regularization was necessary in our particular case since the function h is quadratic. $p(x_i, \tau_j)$ was interpolated by $(PR(i, j) + P(i, j+1))/2$.

Setting the steplength ρ of the gradient method is known to be a delicate problem. Moreover, in our case as in the general case of control problems it is necessary to solve (numerically) the state system in order to evaluate the cost functional for a given control. We have used an interpolation method for the one-dimensional minimization to set ρ at each iteration (see also [12, p. 14]).

We now give more information about the computational complexity of the algorithm. We omit the inversion of the band matrix (5.2) since it is done only once. It enables us to solve an $n \times n$ system of linear equations using at most $2n^2$ arithmetical operations. Let us recall that by $O(n)$ we mean such a quantity that $\lim_{n \rightarrow \infty} O(n)/n = l$, with $0 < l < \infty$ (a polynomial of first degree in n). The computational complexity of the algorithm is given by:

- Solving the system in Step 1— $O(M^2N)$;
- Solving the system in Step 2— $O(M^2N)$;
- Computing w in Step 3— $O(MN)$;
- Computing ρ in Step 3— $kO(M^2N)$, where k is the number of evaluations of the cost functional (a double integral requires $O(MN)$ operations). It follows that a loop in our algorithm involves $(k+2)O(M^2N)$.

The numerical tests were made for $\varepsilon, h < 10^{-1}$. The number of iterations (loops) necessary to satisfy the convergence criterion was only 2. The precision obtained for the test in Step 4 was 10^{-3} and for the one in Remark 3.1 was 10^{-7} . The CPU time spent by an IRIS-50 computer (using 6 hexadecimal digits per number) for the two loops was 8.64 seconds.

The computed (sub)optimal control is symmetric with respect to x , that is $U(i, j) = U(k, j)$ for $i + k = M + 2$ and for every j . In other words $u(x_i, \tau_j) = u(x_k, \tau_j)$ for $x_i + x_k = a + b$ and for every j .

Some values $u(x_i, \tau_j)$ are given in Table 1.

Here $yE - p$ means y times 10^{-p} . Values for $x_i \geq 0.5$ may be obtained from the symmetry property given above.

The behaviour of the functions $t \rightarrow u(x, t)$ for different fixed values of x is given in Fig. 1. The matrix U has been scaled in the following sense: if $u_{\min} = \min \{U(i, j); i = 1, \dots, M+1, j = 1, \dots, N\}$ and $u_{\max} = \max \{U(i, j); i = 1, \dots, M+1, j = 1, \dots, N\}$ then the interval $[u_{\min}, u_{\max}]$ is divided into 50 subintervals of equal length numbered from 1 to 50. Every value $U(i, j)$ belongs to one of these subintervals.

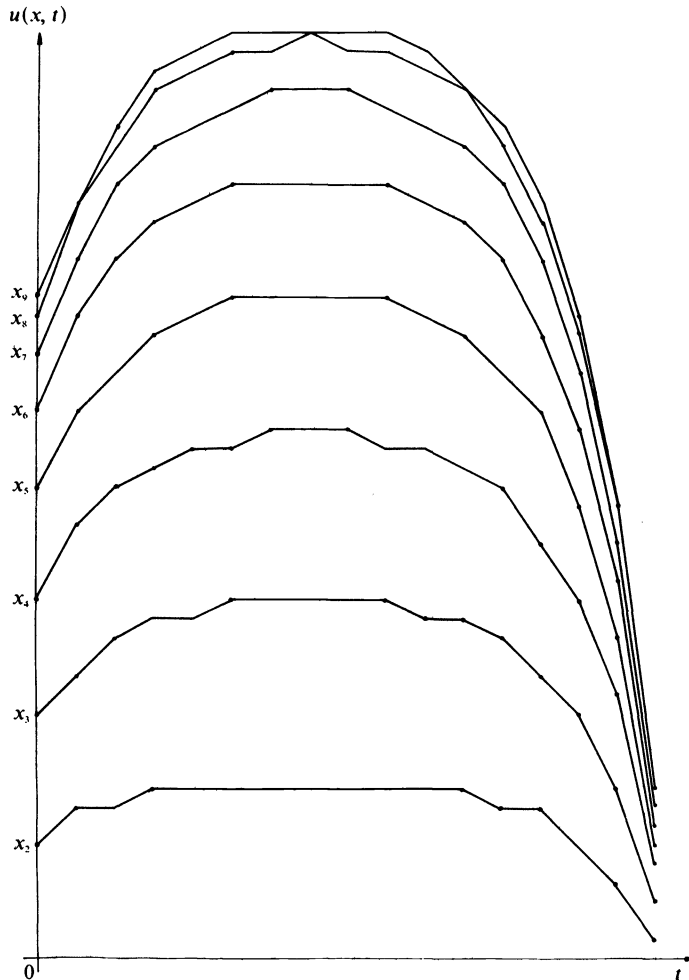


FIG. 1

TABLE 1

$\tau_j \setminus x_i$	0.0625	0.125	0.25	0.375
0.0294	0.347372E-2	0.678692E-2	0.124198E-1	0.161045E-1
0.1470	0.442166E-2	0.864465E-2	0.158430E-1	0.205661E-1
0.2647	0.479155E-2	0.937017E-2	0.171834E-1	0.223170E-1
0.3824	0.491861E-2	0.961941E-2	0.176439E-1	0.229187E-1
0.5000	0.491650E-2	0.961527E-2	0.176363E-1	0.229087E-1
0.6176	0.478334E-2	0.935406E-2	0.171536E-1	0.222781E-1
0.7353	0.440011E-2	0.860234E-2	0.157646E-1	0.204634E-1
0.8569	0.342352E-2	0.668698E-2	0.122272E-1	0.158438E-1
0.9706	0.944482E-3	0.184004E-2	0.334439E-2	0.431399E-2

The corresponding number is assigned to it. The graphical representations are given for the scaled values. Here $x_2 = 0.0625$, $x_3 = 0.125$, $x_4 = 0.1875$, $x_5 = 0.25$, $x_6 = 0.3125$, $x_7 = 0.375$, $x_8 = 0.4375$, $x_9 = 0.5$. For $x > 0.5$ the graphics are the same because of the symmetry property explained above.

REFERENCES

- [1] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: numerical methods based on averaging approximations*, this Journal, 16 (1978), pp. 169–208.
- [2] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, the Netherlands, 1976.
- [3] ———, *Optimal Control of Variational Inequalities*, Research Notes in Mathematics 100, Pitman, London–Boston–Melbourne, 1984.
- [4] ———, *Optimal control for free boundary problems*. Conferenze Seminario Matematico di Bari, Bari, Italy, 1985.
- [5] V. BARBU AND T. PRECUPANU, *Convexity and Optimization in Banach Spaces*, Reidel, Dordrecht–Boston, 1986.
- [6] H. BREZIS, *Opérateurs maximaux monotones et semigroupes de contractions dans les espaces de Hilbert*, North–Holland, Amsterdam, 1973.
- [7] H. BREZIS AND A. PAZY, *Convergence and approximation of semigroups of nonlinear operators in Banach spaces*, J. Funct. Anal., 9 (1971), pp. 63–74.
- [8] J. CEA, *Lectures on Optimization Theory and Algorithms*, Tata Institute of Fundamental Research, Bombay; Springer–Verlag, Berlin–New York, 1978.
- [9] M. G. CRANDALL AND T. LIGGETT, *Generation of semigroups of non-linear transformations on general Banach spaces*, Amer. J. Math., 93 (1971), pp. 265–298.
- [10] C. DAFERMOS AND M. SLEMROD, *Asymptotic behaviour of nonlinear contraction semigroups*, J. Funct. Anal., 13 (1973), pp. 97–106.
- [11] A. FRIEDMAN, *Optimal control for parabolic variational inequalities*, this Journal, 24 (1986), pp. 439–451.
- [12] J. LEGRAS, *Algorithmes et programmes d'optimisation non linéaire avec contraintes*, Masson, Paris, 1980.
- [13] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Gauthier–Villars, Paris, 1969.
- [14] F. MIGNOT AND J. P. PUEL, *Contrôle optimal d'un système gouverné par une inéquation variationnelle parabolique*, C.R. Acad. Sci. Paris, 298 (1984), pp. 277–280.
- [15] C. SAGUEZ, *Contrôle optimal de système à frontière libre*, Thèse de l'Université de Technologie de Compiègne, 1980.

STABILITY AND SENSITIVITY ANALYSIS IN CONVEX VECTOR OPTIMIZATION*

TETSUZO TANINO†

Abstract. In this paper stability and sensitivity of the efficient set in convex vector optimization are considered. The perturbation map is defined as a set-valued map. It associates with each parameter vector the set of all minimal points of the parametrized feasible set with respect to an ordering cone in the objective space. Sufficient conditions for the upper and lower semicontinuity of the perturbation map are obtained. Because of the convexity assumptions, the conditions obtained are fairly simple if compared to those in the general case. Moreover, a complete characterization of the contingent derivative of the perturbation map is obtained under some assumptions. It provides quantitative information on the behavior of the perturbation map.

Key words. stability, convex vector optimization, upper and lower semicontinuity, perturbation map, contingent derivative

AMS(MOS) subject classifications. 49B50, 54C60, 90C25

1. Introduction. In this paper we consider a family of parametrized vector optimization problems:

$$(1.1) \quad \begin{aligned} &P\text{-minimize} \quad f(x, u) = (f_1(x, u), \dots, f_p(x, u)) \\ &\text{subject to} \quad x \in X(u) \subset \mathbb{R}^n. \end{aligned}$$

Here x is an n -dimensional decision variable, u is an m -dimensional parameter vector, f_i ($i = 1, \dots, p$) is a real-valued objective function on $\mathbb{R}^n \times \mathbb{R}^m$, X is a set-valued map from \mathbb{R}^m to \mathbb{R}^n , which specifies a feasible decision set, and P is a nonempty pointed closed convex ordering cone in \mathbb{R}^p . We can define another set-valued map Y from \mathbb{R}^m to \mathbb{R}^p by

$$(1.2) \quad Y(u) := \{y \in \mathbb{R}^p \mid y = f(x, u) \text{ for some } x \in X(u)\}.$$

$Y(u)$ is the parametrized feasible set in the objective space. The cone P induces a partial order \leq_P on \mathbb{R}^p , that is, we define the relation \leq_P by

$$(1.3) \quad y \leq_P y' \Leftrightarrow y' - y \in P \quad \text{for } y, y' \in \mathbb{R}^p.$$

This relation \leq_P is reflexive, antisymmetric and transitive. In the problem (1.1), we aim to obtain all the minimal points of the feasible set $Y(u)$ with respect to the order \leq_P . In other words, the solution set in the objective space to the problem (1.1) is given by

$$(1.4) \quad \begin{aligned} \text{Min}_P Y(u) &= \{\hat{y} \in Y(u) \mid \text{there exists no } y \neq \hat{y} \text{ such that } y \leq_P \hat{y}\} \\ &= \{\hat{y} \in Y(u) \mid (Y(u) - \hat{y}) \cap (-P) = \{0\}\}. \end{aligned}$$

Therefore, we can define another set-valued map W from the parameter space \mathbb{R}^m to the objective space \mathbb{R}^p by

$$(1.5) \quad W(u) := \text{Min}_P Y(u).$$

W is often called the perturbation map for (1.1).

* Received by the editors March 24, 1986; accepted for publication (in revised form) June 26, 1987. This research was conducted when the author stayed at the International Institute for Applied Systems Analysis, Laxenburg, Austria.

† Department of Mechanical Engineering II, Tohoku University, Sendai 980, Japan.

In usual scalar optimization where $p = 1$ and $P = \mathbb{R}_+$ (=the set of nonnegative real numbers) W is at most single valued and so it can be identified with the function

$$(1.6) \quad w(u) := \min \{f(x, u) \mid x \in X(u)\}.$$

And the stability and sensitivity analysis in scalar optimization is mainly a study of continuity properties and derivatives of the function w . In case of vector optimization, we investigate the behavior of the set-valued map W .

Some results for general vector optimization problems from this point of view can be seen, for example, in [2], [7] for stability and in [6] for sensitivity. In this paper we consider the case in which convexity is assumed. It is shown that the convexity assumption considerably simplifies the sufficient conditions for the semicontinuity of the perturbation map W and also makes it possible to characterize the contingent derivative of W completely.

2. Convexity assumption and preliminary results. Throughout this paper we assume the following convexity on the feasible decision set map X and the objective function f .

Convexity assumption (CA).

(1) The set-valued map X is convex, i.e., the graph of X which is defined by

$$(2.1) \quad \text{graph } X = \{(u, x) \mid x \in X(u)\}$$

is a convex set in $\mathbb{R}^m \times \mathbb{R}^n$. In other words, for any $u^1, u^2 \in \mathbb{R}^m$ and any α , $0 \leq \alpha \leq 1$,

$$(2.2) \quad \alpha X(u^1) + (1 - \alpha)X(u^2) \subset X(\alpha u^1 + (1 - \alpha)u^2).$$

(2) The function f is P -convex, i.e., for any $(x^1, u^1), (x^2, u^2) \in \mathbb{R}^n \times \mathbb{R}^m$ and any α , $0 \leq \alpha \leq 1$,

$$\alpha f(x^1, u^1) + (1 - \alpha)f(x^2, u^2) \in f(\alpha x^1 + (1 - \alpha)x^2, \alpha u^1 + (1 - \alpha)u^2) + P.$$

LEMMA 2.1. *If P is a pointed closed convex cone and f is P -convex, then f is continuous.*

Proof. Since P is a pointed closed convex cone, the interior of the negative polar cone P° of P is not empty¹. It is easy to prove that $-\langle \mu, f(x, u) \rangle$ is convex as a function of (x, u) for $\mu \in P^\circ$. Hence $\langle \mu, f(\cdot, \cdot) \rangle$ is continuous [3, Cor. 10.1.1]. Take $\bar{\mu} \in \text{int } P^\circ$ and $\bar{\mu} + \delta e^i \in P^\circ$ for sufficiently small $\delta > 0$, where e^i is the i th unit vector in \mathbb{R}^p . Then both $\langle \bar{\mu}, f(\cdot, \cdot) \rangle$ and $\langle \bar{\mu} + \delta e^i, f(\cdot, \cdot) \rangle$ are continuous and hence $f_i(\cdot, \cdot)$ is continuous ($i = 1, \dots, p$). Namely f is continuous. \square

PROPOSITION 2.1. *Under the convexity assumption (CA), the set-valued map Y defined by (1.2) is P -convex, i.e., for any $u^1, u^2 \in \mathbb{R}^m$ and α , $0 \leq \alpha \leq 1$,*

$$(2.3) \quad \alpha Y(u^1) + (1 - \alpha)Y(u^2) \subset Y(\alpha u^1 + (1 - \alpha)u^2) + P.$$

In other words, the graph of the set-valued map $Y + P$ is convex. Here $Y + P$ is defined by

$$(2.4) \quad (Y + P)(u) := Y(u) + P \quad \text{for each } u \in \mathbb{R}^m.$$

Proof. This proposition can be easily proved. \square

Now we introduce concepts of semicontinuity of set-valued maps. Let F be a set-valued map from \mathbb{R}^m to \mathbb{R}^p hereafter in this section. We denote it by $F: \mathbb{R}^m \rightrightarrows \mathbb{R}^p$.

DEFINITION 2.1. (1) F is said to be upper semicontinuous at $\hat{u} \in \mathbb{R}^m$ if $u^k \rightarrow \hat{u}$, $y^k \in F(u^k)$ and $y^k \rightarrow \hat{y}$ all imply that $\hat{y} \in F(\hat{u})$.

(2) F is said to be lower semicontinuous at $\hat{u} \in \mathbb{R}^m$ if $u^k \rightarrow \hat{u}$ and $\hat{y} \in F(\hat{u})$ imply the existence of an integer K and a sequence $\{y^k\} \subset \mathbb{R}^p$ such that $y^k \in F(u^k)$ for $k \geq K$ and $y^k \rightarrow \hat{y}$.

¹ $P^\circ = \{\mu \in \mathbb{R}^p \mid \langle \mu, d \rangle \leq 0 \text{ for all } d \in P\}$, where $\langle \cdot, \cdot \rangle$ denotes the inner product.

(3) F is said to be continuous at $\hat{u} \in \mathbb{R}^m$ if it is both upper and lower semicontinuous at \hat{u} .

Remark 2.1. F is upper semicontinuous on \mathbb{R}^m if and only if graph F is a closed set in $\mathbb{R}^m \times \mathbb{R}^p$.

We shall provide lemmas concerning the semicontinuity of convex set-valued maps. Given F and $\hat{y} \in \mathbb{R}^p$, we define the function ρ from \mathbb{R}^m to $\mathbb{R} \cup \{+\infty\}$ by

$$(2.5) \quad \rho(u) = \text{dist}(\hat{y}, F(u)) := \inf \{\|y - \hat{y}\| \mid y \in F(u)\}.$$

If $F(u) = \emptyset$, let $\rho(u) = +\infty$. The domain of the set-valued map F is defined and denoted by

$$(2.6) \quad \text{dom } F := \{u \in \mathbb{R}^m \mid F(u) \neq \emptyset\}.$$

Clearly $\text{dom } \rho = \{u \in \mathbb{R}^m \mid \rho(u) < +\infty\} = \text{dom } F$.

LEMMA 2.2. *If F is convex, then the function ρ defined by (2.5) is a convex function.*

Proof. Let $u^1, u^2 \in \text{dom } \rho$, which is a convex set, and $0 \leq \alpha \leq 1$. Since F is convex,

$$\alpha F(u^1) + (1 - \alpha)F(u^2) \subset F(\alpha u^1 + (1 - \alpha)u^2)$$

and hence

$$\begin{aligned} \rho(\alpha u^1 + (1 - \alpha)u^2) &= \inf \{\|y - \hat{y}\| \mid y \in F(\alpha u^1 + (1 - \alpha)u^2)\} \\ &\leq \inf \{\|y - \hat{y}\| \mid y \in \alpha F(u^1) + (1 - \alpha)F(u^2)\} \\ &= \inf \{\|\alpha y^1 + (1 - \alpha)y^2 - \hat{y}\| \mid y^1 \in F(u^1), y^2 \in F(u^2)\} \\ &\leq \inf \{\alpha \|y^1 - \hat{y}\| + (1 - \alpha)\|y^2 - \hat{y}\| \mid y^1 \in F(u^1), y^2 \in F(u^2)\} \\ &= \alpha \inf \{\|y^1 - \hat{y}\| \mid y^1 \in F(u^1)\} + (1 - \alpha) \inf \{\|y^2 - \hat{y}\| \mid y^2 \in F(u^2)\} \\ &= \alpha \rho(u^1) + (1 - \alpha)\rho(u^2). \end{aligned}$$

LEMMA 2.3. *If F is convex and $\hat{u} \in \text{int}(\text{dom } F)$, then F is lower semicontinuous at \hat{u} .*

Proof. Let $u^k \rightarrow \hat{u}$ and $\hat{y} \in F(\hat{u})$. Define the function ρ by (2.5). Then, from Lemma 2.2, ρ is a convex function and $\text{dom } \rho = \text{dom } F$. Since $\hat{u} \in \text{int}(\text{dom } \rho)$ and $u^k \rightarrow \hat{u}$, there exists a number K such that $u^k \in \text{dom } \rho$ for any $k \geq K$. For each u^k ($k \geq K$), from the definition of $\rho(u^k)$, there exists $y^k \in F(u^k)$ such that

$$\|y^k - \hat{y}\| < \rho(u^k) + \frac{1}{k}.$$

Since the convex function ρ is continuous at $\hat{u} \in \text{int}(\text{dom } \rho)$ and $\rho(\hat{u}) = 0$, by taking the limit of the above inequality, $\|y^k - \hat{y}\| \rightarrow 0$ as $k \rightarrow \infty$, namely, $y^k \rightarrow \hat{y}$. Therefore F is lower semicontinuous at \hat{u} . \square

Remark 2.2. Since the spaces considered here are all finite dimensional, the assumption in Lemma 2.3 is weaker than in the result of Aubin and Ekeland [1, p. 131], where F is assumed to be not only convex but also upper semicontinuous.

Remark 2.3. The following example illustrates that the condition $\hat{u} \in \text{int}(\text{dom } F)$ is essential in Lemma 2.3. Let $F: \mathbb{R}^2 \rightrightarrows \mathbb{R}$ be defined by

$$F(u) = \begin{cases} \{y \in \mathbb{R} \mid y \geq \alpha\} & \text{if } (u_1 - \alpha)^2 + (u_2)^2 = \alpha^2 \text{ for } \alpha > 0, \quad u \neq (0, 0), \\ \{y \in \mathbb{R} \mid y \geq 0\} & \text{if } u = (0, 0), \\ \emptyset & \text{otherwise.} \end{cases}$$

Then, for $u^k = (1 - \cos(\pi/k), \sin(\pi/k))$, $F(u^k) = \{y \mid y \geq 1\}$ for all $k = 1, 2, \dots$. Clearly $u^k \rightarrow (0, 0)$. However, by taking $0 \in F((0, 0))$, we can easily see that F is not lower semicontinuous at $\hat{u} = (0, 0)$.

LEMMA 2.4. *If F is convex, $\hat{u} \in \text{int}(\text{dom } F)$ and $F(\hat{u})$ is a closed set, then F is upper semicontinuous (and therefore continuous in view of Lemma 2.3) at \hat{u} .*

Proof. Let $u^k \rightarrow \hat{u}$, $y^k \in F(u^k)$ and $y^k \rightarrow \hat{y}$. Define ρ as in (2.5). Then ρ is a convex function from Lemma 2.2. Hence ρ is continuous at $\hat{u} \in \text{int}(\text{dom } F) = \text{int}(\text{dom } \rho)$. On the other hand, taking the limit of the inequality

$$0 \leq \rho(u^k) \leq \|y^k - \hat{y}\|,$$

as $k \rightarrow \infty$, we can prove that $\rho(\hat{u}) = 0$. Since $F(\hat{u})$ is a closed set, this implies $\hat{y} \in F(\hat{u})$. Hence F is upper semicontinuous at \hat{u} . \square

Remark 2.4. It is easily understood that the closedness of $F(\hat{u})$ is very important in the above lemma. The following example illustrates the inevitability of the condition $\hat{u} \in \text{int}(\text{dom } F)$. Let $F: \mathbb{R} \rightrightarrows \mathbb{R}$ be defined by

$$F(u) = \begin{cases} \{y \mid y \geq 0\} & \text{if } u > 0, \\ \{y \mid y \geq 1\} & \text{if } u = 0, \\ \emptyset & \text{if } u < 0. \end{cases}$$

Then, for $u^k = 1/k$, $y^k = 0 \in F(u^k)$ ($k = 1, 2, \dots$). However, the limit 0 of $\{y^k\}$ is not contained in $F(0)$.

3. Upper semicontinuity of the perturbation map. In this section we shall consider sufficient conditions for the upper semicontinuity of the perturbation map W . First we provide sufficient conditions in terms of the feasible set map Y .

THEOREM 3.1. *If the following three conditions are satisfied, then the perturbation map W is upper semicontinuous at $\hat{u} \in \mathbb{R}^m$:*

- (1) $\hat{u} \in \text{int}(\text{dom } Y)$;
- (2) Y is upper semicontinuous at \hat{u} ;
- (3) $W(\hat{u}) = w - \text{Min}_P Y(\hat{u})$, where $w - \text{Min}_P Y(\hat{u})$ is the set of all weakly P -minimal points of $Y(\hat{u})$, i.e.,

$$(3.1) \quad w - \text{Min}_P Y(\hat{u}) := \{y \in Y(\hat{u}) \mid (Y(\hat{u}) - y) \cap (-\text{int } P) = \emptyset\}.$$

Proof. Let $u^k \rightarrow \hat{u}$, $y^k \in W(u^k)$ and $y^k \rightarrow \hat{y}$. Since Y is upper semicontinuous at \hat{u} , $\hat{y} \in Y(\hat{u})$. Hence, if we suppose that $\hat{y} \notin W(\hat{u}) = w - \text{Min}_P Y(\hat{u})$, then there exists $\bar{y} \in Y(\hat{u})$ such that $\hat{y} - \bar{y} \in \text{int } P$. Since $\hat{u} \in \text{int}(\text{dom } Y) = \text{int}(\text{dom } (Y + P))$ and $Y + P$ is convex, $Y + P$ is lower semicontinuous at \hat{u} from Lemma 2.3. Namely there exist a sequence $\{\bar{y}^k\} \subset \mathbb{R}^p$ and a number K such that

$$\bar{y}^k \rightarrow \bar{y} \quad \text{and} \quad \bar{y}^k \in Y(u^k) + P \quad \text{for } k \geq K$$

since $y^k - \bar{y}^k \rightarrow \hat{y} - \bar{y} \in \text{int } P$, $y^k - \bar{y}^k \in \text{int } P$ for all k sufficiently large. However, this contradicts that $y^k \in W(u^k) = \text{Min}_P Y(u^k) = \text{Min}_P (Y(u^k) + P)$ (see [5, Prop. 3.1.2]). Therefore $\hat{y} \in W(\hat{u})$, as was to be proved. \square

Remark 3.1. We can guarantee the upper semicontinuity of W under the following conditions without the convexity assumption (CA) [7]:

- (i) Y is continuous at \hat{u} ;
- (ii) $W(\hat{u}) = w - \text{Min}_P Y(\hat{u})$.

If we compare these conditions with Theorem 3.1, the following can be observed: we can replace the lower semicontinuity condition of Y by the weaker condition $\hat{u} \in \text{int}(\text{dom } Y)$ under the convexity assumption.

Now we shall derive sufficient conditions for the upper semicontinuity of W , which are described in terms of the feasible decision set map X and the objective

function f . For the purpose we shall introduce a set-valued map \tilde{X} from $\mathbb{R}^m \times \mathbb{R}^p$ to \mathbb{R}^n as follows:

$$(3.2) \quad \tilde{X}(u, y) := \{x \in X(u) \mid f(x, u) = y\}.$$

The following proposition provides sufficient conditions for the upper semicontinuity of Y at \hat{u} .

PROPOSITION 3.1. *If $\hat{u} \in \text{int}(\text{dom } X)$, if $X(\hat{u})$ is a closed set and if the map \tilde{X} is locally bounded at $(\hat{u}, \hat{y})^2$ for any $\hat{y} \in \{y \mid (\hat{u}, y) \in \text{cl}(\text{graph } Y)\}$, then Y is upper semicontinuous at \hat{u} .*

Proof. Let $u^k \rightarrow \hat{u}$, $y^k \in Y(u^k)$ and $y^k \rightarrow \hat{y}$. Then there exists a sequence $\{x^k\} \subset \mathbb{R}^n$ such that $x^k \in \tilde{X}(u^k, y^k)$ for all $k = 1, 2, \dots$. Since \tilde{X} is locally bounded at (\hat{u}, \hat{y}) , $\{x^k\}$ has a convergent subsequence. By taking the subsequence if necessary, we may assume that $\{x^k\}$ converges to some \hat{x} . From Lemma 2.4, X is upper semicontinuous at \hat{u} and so $\hat{x} \in X(\hat{u})$. On the other hand, since f is continuous from Lemma 2.1, $f(\hat{x}, \hat{u}) = \hat{y}$. Therefore $\hat{y} \in Y(\hat{u})$ and Y is upper semicontinuous at \hat{u} . \square

Remark 3.2. If X is locally bounded at \hat{u} , then \tilde{X} is clearly locally bounded at (\hat{u}, y) for any $y \in \mathbb{R}^p$.

Now we can prove the following theorem.

THEOREM 3.2. *If the following four conditions are satisfied, then the set-valued map W is upper semicontinuous at \hat{u} :*

- (1) $\hat{u} \in \text{int}(\text{dom } X)$.
- (2) $X(\hat{u})$ is a closed set.
- (3) \tilde{X} is locally bounded at (\hat{u}, \hat{y}) for any $\hat{y} \in \{y \mid (\hat{u}, y) \in \text{cl}(\text{graph } Y)\}$.
- (4) $W(\hat{u}) = w - \text{Min}_P Y(\hat{u})$.

Proof. From (1), $\hat{u} \in \text{int}(\text{dom } Y)$. From (1)–(3), in view of Proposition 3.1, Y is upper semicontinuous at \hat{u} . Hence W is upper semicontinuous at \hat{u} by Theorem 3.1. \square

Remark 3.3. The following examples illustrate that each condition in the above theorem is essential.

(1) Take F in Remark 2.3 as X and let $f(x, u) = x$ and $P = \mathbb{R}_+$. Then $\hat{u} = (0, 0) \notin \text{int}(\text{dom } X)$ and

$$W(u) = \begin{cases} \{\alpha\} & \text{if } (u_1 - \alpha)^2 + (u_2)^2 = \alpha^2 \text{ for } \alpha > 0, \quad u \neq (0, 0), \\ \{0\} & \text{if } u = (0, 0), \\ \emptyset & \text{otherwise,} \end{cases}$$

which is not upper semicontinuous at $(0, 0)$.

(2) Let $m = p = n = 1$, $P = \mathbb{R}_+$, $f(x, u) = x$ and

$$X(u) = \begin{cases} \{x \mid x \geq 0\} & \text{if } u \neq 0, \\ \{x \mid x > 0\} & \text{if } u = 0. \end{cases}$$

Then $W(u) = \{0\}$ if $u \neq 0$ and $W(0) = \emptyset$. Hence W is not upper semicontinuous at 0.

(3) Let $m = n = p = 1$, $P = \mathbb{R}_+$ and $X(u) = \mathbb{R}$ for any $u \in \mathbb{R}$. Let C be a convex set in $\mathbb{R} \times \mathbb{R}$ defined by

$$C = \{(u, x) \mid ux \geq 1, u > 0\}$$

and f be defined by

$$f(x, u) = d((u, x), C) = \inf \{\|(u, x) - (u', x')\| \mid (u', x') \in C\}.$$

² A set-valued map F is said to be locally bounded at \hat{u} if there exists a neighborhood N of \hat{u} such that $\bigcup_{u \in N} F(u)$ is bounded.

Then f is P -convex and

$$Y(u) = \begin{cases} \{y \in \mathbb{R} \mid y \geq 0\} & \text{if } u > 0, \\ \{y \in \mathbb{R} \mid y > -u\} & \text{if } u \leq 0. \end{cases}$$

Hence

$$W(u) = \begin{cases} \{0\} & \text{if } u > 0, \\ \emptyset & \text{if } u \leq 0, \end{cases}$$

which is not upper semicontinuous at 0. In this example, \tilde{X} is not locally bounded at $(0, 0)$.

4. Lower semicontinuity of the perturbation map. In this section we consider sufficient conditions for the lower semicontinuity of the map W . First we should introduce several concepts.

DEFINITION 4.1. A set S in \mathbb{R}^p is said to be P -minicomplete if

$$(4.1) \quad S \subset \text{Min}_P S + P.$$

Remark 4.1. Since $\text{Min}_P S \subset S$, if S is P -minicomplete,

$$(4.2) \quad S + P = \text{Min}_P S + P.$$

DEFINITION 4.2. For a nonempty set S in \mathbb{R}^p , its recession cone S^+ is defined by

$$(4.3) \quad S^+ = \{y \in \mathbb{R}^p \mid \text{there exist sequences } \{\lambda_k\} \subset \mathbb{R} \text{ and } \{y^k\} \subset \mathbb{R}^p \text{ such that } \lambda_k > 0, \lambda_k \rightarrow 0, \lambda_k y^k \rightarrow y \text{ and } y^k \in S \text{ for all } k\}.$$

Remark 4.2. S^+ is a closed cone which contains the origin. Moreover, if S is a nonempty closed convex set, S^+ coincides with the set 0^+S which is defined by

$$(4.4) \quad \begin{aligned} 0^+S &= \{y \in \mathbb{R}^p \mid \bar{y} + \lambda y \in S \forall \lambda \geq 0, \forall \bar{y} \in S\} \\ &= \{y \in \mathbb{R}^p \mid S + y \subset S\} \end{aligned}$$

and therefore it is a closed convex cone [3, Thm. 8.2].

LEMMA 4.1 (Sawaragi et al. [5, Lemma 3.2.1]). A nonempty set S is bounded if and only if $S^+ = \{0\}$.

LEMMA 4.2 (Sawaragi et al. [5, Lemma 3.2.3]). Let S_1 and S_2 be nonempty closed sets. If $S_1^+ \cap (-S_2^+) = \{0\}$, then $S_1 + S_2$ is also a nonempty closed set.

In view of the above two lemmas, the following concept plays an important role in this section.

DEFINITION 4.3. A nonempty set S in \mathbb{R}^p is said to be P -bounded if

$$(4.5) \quad S^+ \cap (-P) = \{0\}.$$

LEMMA 4.3 (Sawaragi et al. [5, Thm. 3.2.12]). If $S \subset \mathbb{R}^p$ is a nonempty closed convex set, the following statements are equivalent:

- (1) S is P -bounded.
- (2) $\text{Min}_P S \neq \emptyset$.
- (3) S is P -minicomplete.

LEMMA 4.4. Suppose that F is P -convex, $\hat{u} \in \text{int}(\text{dom } F)$, and $F(\hat{u})$ is P -bounded. Then there exists a neighborhood N of \hat{u} such that $F(u)$ is P -bounded for all $u \in N$.

Proof. If the conclusion of the lemma were not true, there would exist sequences $\{u^k\} \in \mathbb{R}^m$ and $\{d^k\} \subset \mathbb{R}^p$ such that $u^k \rightarrow \hat{u}$, $d^k \neq 0$ and

$$-d^k \in [F(u^k)]^+ \cap (-P).$$

Since $[F(u^k)]^+ \cap (-P)$ is a cone, we may assume that $\|d^k\| = 1$ for all k . By taking a subsequence if necessary, we may assume that $\{d^k\}$ converges to some d . Since P is closed, $d \in P$. Moreover, $\|d\| = 1$ and so $d \neq 0$. Since $-d^k \in [F(u^k)]^+$, there exist sequences $\{\lambda_{kl}\} \subset \mathbb{R}$, $\{d^{kl}\} \subset -F(u^k)$ such that $\lambda_{kl} > 0$,

$$\lambda_{kl} \rightarrow 0 \quad \text{and} \quad \lambda_{kl} d^{kl} \rightarrow d^k \quad \text{as } l \rightarrow \infty.$$

If we take l sufficiently large,

$$\lambda_{kl} < \frac{1}{k} \quad \text{and} \quad \|\lambda_{kl} d^{kl} - d^k\| < \frac{1}{k}.$$

By choosing those λ_{kl} and d^{kl} as $\bar{\lambda}_k$ and \bar{d}^k , respectively, we can construct sequences $\{\bar{\lambda}_k\}$ and $\{\bar{d}^k\}$ satisfying

$$-\bar{d}^k \in F(u^k), \quad 0 < \bar{\lambda}_k < \frac{1}{k}, \quad \|\bar{\lambda}_k \bar{d}^k - d^k\| < \frac{1}{k}.$$

When $k \rightarrow \infty$, $\bar{\lambda}_k \rightarrow 0$ and $\bar{\lambda}_k \bar{d}^k \rightarrow d$. Now take an arbitrary $\tilde{y} \in F(\hat{u})$. Since $2\hat{u} - u^k \rightarrow \hat{u}$ and $F + P$ is lower semicontinuous at \hat{u} by Lemma 2.3, there exist a sequence $\{\tilde{y}^k\}$ and a number K such that

$$\tilde{y}^k \rightarrow \tilde{y} \quad \text{and} \quad \tilde{y}^k \in F(2\hat{u} - u^k) + P \quad \text{for } k \geq K.$$

Since F is P -convex,

$$\frac{1}{2}(\tilde{y}^k - \bar{d}^k) \in F(\hat{u}) + P \quad \text{for } k \geq K.$$

Moreover, $2\bar{\lambda}_k \cdot \frac{1}{2}(\tilde{y}^k - \bar{d}^k) \rightarrow -d$. This implies that $-d \in [F(\hat{u}) + P]^+$ and hence $[F(\hat{u}) + P]^+ \cap (-P) \neq \{0\}$. In view of Lemma 3.2.4 of [5], this means that $F(\hat{u})$ is not P -bounded, which is a contradiction. Hence $F(u)$ is P -bounded for all u in a certain neighborhood of \hat{u} . \square

Now we can obtain sufficient conditions for the lower semicontinuity of W .

THEOREM 4.1. *If the following conditions are satisfied, then the perturbation map W is lower semicontinuous at \hat{u} :*

- (1) $\hat{u} \in \text{int}(\text{dom } Y)$.
- (2) $Y + P$ is upper semicontinuous in a neighborhood of \hat{u} .

Proof. If $W(\hat{u}) = \emptyset$, the theorem is trivial. Hence we suppose that $W(\hat{u}) \neq \emptyset$. Let $u^k \rightarrow \hat{u}$ and $\hat{y} \in W(\hat{u})$. From Lemma 2.3, $Y + P$ is lower semicontinuous at \hat{u} and hence there exist a sequence $\{y^k\}$ and a number K_1 such that

$$y^k \rightarrow \hat{y} \quad \text{and} \quad y^k \in (u^k) + P \quad \text{for all } k \geq K_1.$$

Since $Y(\hat{u}) + P$ is a nonempty closed convex set and $\text{Min}_P(Y(\hat{u}) + P) = W(\hat{u}) \neq \emptyset$, $Y(\hat{u}) + P$ is P -bounded from Lemma 4.3. Therefore, in view of Lemma 4.4, $Y(u) + P$ is P -bounded for all u in a certain neighborhood N of \hat{u} . (Note that $\hat{u} \in \text{int}(\text{dom } Y)$.) From Lemma 4.3 and Remark 4.1, this implies that

$$W(u) + P = (Y(u) + P) + P = Y(u) + P$$

in a neighborhood of \hat{u} . Hence there exist a sequence $\{\hat{y}^k\}$ and a number $K_2 \geq K_1$ such that

$$y^k - \hat{y}^k \in P \quad \text{and} \quad \hat{y}^k \in W(u^k) \quad \text{for } k \geq K_2.$$

First we will show that $\{\hat{y}^k\}$ is bounded. If this were not the case, from Lemma 4.1, we can take a subsequence of $\{\hat{y}^k\}$, for which there exist a sequence $\{\lambda_k\}$ of positive

numbers and a nonzero vector \tilde{y} such that $\lambda_k \rightarrow 0$ and $\lambda_k \hat{y}^k \rightarrow \tilde{y}$. Since $\lambda_k(y^k - \hat{y}^k) \in P$ and $y^k \rightarrow \hat{y}$, the limit $-\tilde{y}$ of $\{\lambda_k(y^k - \hat{y}^k)\}$ is contained in P . Take an arbitrary $\bar{y} \in Y(\hat{u}) + P$. Then there exist a sequence $\{\bar{y}^k\}$ and a number $K_3 \geq K_2$ such that

$$\bar{y}^k \rightarrow \bar{y} \quad \text{and} \quad \bar{y}^k \in Y(2\hat{u} - u^k) + P \quad \text{for } k \geq K_3,$$

since $Y + P$ is lower semicontinuous at \hat{u} . Then, from the convexity of $Y + P$,

$$\frac{1}{2}(\hat{y}^k + \bar{y}^k) \in Y(\hat{u}) + P \quad \text{for } k \geq K_3.$$

Moreover, $\lambda_k(\hat{y}^k + \bar{y}^k) \rightarrow \tilde{y}$. This implies that $\tilde{y} \in [Y(\hat{u}) + P]^+$ and hence leads to a contradiction to the P -boundedness of $Y(\hat{u}) + P$. Therefore $\{\hat{y}^k\}$ must be bounded. Hence $\{\hat{y}^k\}$ has a cluster point, which is denoted by y' . Since $y^k - \hat{y}^k \in P$ and $y^k \rightarrow \hat{y}$, $\hat{y} - y' \in P$. Since $Y + P$ is upper semicontinuous at \hat{u} , $y' \in Y(\hat{u}) + P$. Recalling that $\hat{y} \in W(\hat{u})$, we can conclude that $y' = \hat{y}$. In other words, \hat{y} is the unique cluster point for the bounded sequence $\{\hat{y}^k\} \rightarrow \hat{y}$. Therefore $\hat{y}^k \rightarrow \hat{y}$, which indicates that W is lower semicontinuous at \hat{u} . \square

Remark 4.3. We can generate the lower semicontinuity of W under the following conditions without the convexity assumption (CA) [7]:

- (i) Y is continuous at \hat{u} ;
- (ii) Y is locally bounded at \hat{u} ;
- (iii) $Y(u)$ is P -minicomplete for every u near \hat{u} .

Theorem 4.1 considerably simplifies the above result.

The following proposition shows that $Y + P$ is often upper semicontinuous when $W(\hat{u})$ is not empty.

PROPOSITION 4.1. *If $X(u)$ is a nonempty closed set for every u near \hat{u} , $W(\hat{u}) \neq \emptyset$ and $\tilde{X}(\hat{u}, \hat{y})$ is bounded for some $\hat{y} \in W(\hat{u})$, then $Y(u)$ is a P -bounded closed set in a neighborhood of \hat{u} . In this case $Y(u) + P$ is also a closed set by Lemma 4.2 and therefore the set-valued map $Y + P$ is upper semicontinuous in a neighborhood of \hat{u} .*

Proof. (a) First we shall prove that $Y(u)$ is a closed set in some neighborhood of \hat{u} . If this were not true, we can consider sequences $\{u^k\}$ and $\{y^k\}$ such that

$$u^k \rightarrow \hat{u} \quad \text{and} \quad y^k \in \text{cl } Y(u^k) \setminus Y(u^k).$$

Corresponding to each y^k , there exists a sequence $\{x^{kl}\} \subset X(u^k)$ such that $f(x^{kl}, u^k) \rightarrow y^k$ as $l \rightarrow \infty$. Take k sufficiently large so that $X(u^k)$ is closed. If $\{x^{kl}\}_{l=1,2,\dots}$ has a convergent subsequence, the limit x^k of it is contained in $X(u^k)$. Since f is continuous, $f(x^k, u^k) = y^k$, which contradicts that $y^k \notin Y(u^k)$. Hence, if k is sufficiently large, $\{x^{kl}\}_{l=1,2,\dots}$ has not a convergent subsequence and so $\|x^{kl}\| \rightarrow +\infty$ as $l \rightarrow \infty$. We may assume that the sequence $\{x^{kl}/\|x^{kl}\|\}_{l=1,2,\dots}$ converges to some \bar{x}^k as $l \rightarrow \infty$. Furthermore, since $\|\bar{x}^k\| = 1$ for all k , we may also assume without loss of generality that $\{\bar{x}^k\}$ converges to a vector \bar{x} . In this case $\|\bar{x}\| = 1$, i.e., $\bar{x} \neq 0$. From the assumptions, we can take $\hat{y} \in W(\hat{u})$ for which $\tilde{X}(\hat{u}, \hat{y})$ is bounded. Let $\hat{x} \in \tilde{X}(\hat{u}, \hat{y})$. Since X is lower semicontinuous at \hat{u} from Lemma 2.3, there exist a sequence $\{\hat{x}^k\}$ and a number K such that

$$\hat{x}^k \rightarrow \hat{x} \quad \text{and} \quad \hat{x}^k \in X(u^k) \quad \text{for } k \geq K.$$

Let $k \geq K$. For an arbitrary $\alpha \geq 0$, $0 \leq \alpha/\|x^{kl}\| \leq 1$ for all k sufficiently large, for $\|x^{kl}\| \rightarrow +\infty$ as $l \rightarrow \infty$. Since X is convex,

$$\left(1 - \frac{\alpha}{\|x^{kl}\|}\right)\hat{x}^k + \frac{\alpha}{\|x^{kl}\|}x^{kl} \in X(u^k).$$

Taking the limit when $l \rightarrow \infty$, we obtain from the closedness of $X(u^k)$,

$$(4.6) \quad \hat{x}^k + \alpha\bar{x}^k \in X(u^k) \quad \text{for all } k \text{ sufficiently large.}$$

Since f is P -convex,

$$f\left(\left(1 - \frac{\alpha}{\|x^{kl}\|}\right)\hat{x}^k + \frac{\alpha}{\|x^{kl}\|}x^{kl}, u^k\right) \leq_P \left(1 - \frac{\alpha}{\|x^{kl}\|}\right)f(\hat{x}^k, u^k) + \frac{\alpha}{\|x^{kl}\|}f(x^{kl}, u^k).$$

Let $l \rightarrow \infty$. Then, since $f(x^{kl}, u^k) \rightarrow y^k$,

$$(4.7) \quad f(\hat{x}^k + \alpha \bar{x}^k, u^k) \leq_P f(\hat{x}^k, u^k) \quad \text{for } k \text{ sufficiently large.}$$

Take the limit of (4.6) and (4.7) as $k \rightarrow \infty$. Then, since X is upper semicontinuous at \hat{u} from Lemma 2.4 and f is continuous, $\hat{x} + \alpha \bar{x} \in X(\hat{u})$ and

$$f(\hat{x} + \alpha \bar{x}, \hat{u}) \leq_P f(\hat{x}, \hat{u}) = \hat{y}.$$

Since $\hat{y} \in W(\hat{u})$, these imply that $f(\hat{x} + \alpha \bar{x}, \hat{u}) = \hat{y}$, i.e., $\hat{x} + \alpha \bar{x} \in \tilde{X}(\hat{u}, \hat{y})$ for all $\alpha \geq 0$. However, this contradicts the boundedness of $\tilde{X}(\hat{u}, \hat{y})$. Hence $Y(u)$ must be a closed set for every u in a certain neighborhood of \hat{u} .

(b) Next, we shall prove that $Y(\hat{u})$ is P -bounded. Let $y \in [Y(\hat{u})]^+ \cap (-P)$. There exist sequences $\{\lambda_k\} \subset R$ and $\{x^k\} \subset X(\hat{u})$ such that $\lambda_k > 0$, $\lambda_k \rightarrow 0$ and $\lambda_k f(x^k, \hat{u}) \rightarrow y$. Then, for all k sufficiently large, $\lambda_k x^k + (1 - \lambda_k)\hat{x} \in X(\hat{u})$ and

$$(4.8) \quad f(\lambda_k x^k + (1 - \lambda_k)\hat{x}, \hat{u}) \leq_P \lambda_k f(x^k, \hat{u}) + (1 - \lambda_k)f(\hat{x}, \hat{u})$$

due to the P -convexity of f . The right-hand side of (4.8) converges to $y + \hat{y}$. First we assume that $\{\lambda_k x^k\}$ has no convergent subsequence. Then $\lambda_k \|x^k\| \rightarrow +\infty$. We may assume without loss of generality that $\{x^k / \|x^k\|\}$ converges to a vector \tilde{x} with $\|\tilde{x}\| = 1$. For any $\alpha \geq 0$, $0 \leq \alpha / \|x^k\| \leq 1$ for all k sufficiently large and so

$$\frac{\alpha}{\|x^k\|}x^k + \left(1 - \frac{\alpha}{\|x^k\|}\right)\hat{x} \in X(\hat{u})$$

from the convexity of $X(\hat{u})$. Since $X(\hat{u})$ is a closed set, the limit of the above relation implies that $\hat{x} + \alpha \tilde{x} \in X(\hat{u})$. Moreover, since f is P -convex,

$$f\left(\frac{\alpha}{\|x^k\|}x^k + \left(1 - \frac{\alpha}{\|x^k\|}\right)\hat{x}, \hat{u}\right) \leq_P \frac{\alpha}{\lambda_k \|x^k\|}\lambda_k f(x^k, \hat{u}) + \left(1 - \frac{\alpha}{\|x^k\|}\right)f(\hat{x}, \hat{u})$$

for all k sufficiently large. Thus, as the limit of the above inequality, we have

$$f(\hat{x} + \alpha \tilde{x}, \hat{u}) \leq_P \hat{y}.$$

Since $\hat{y} \in W(\hat{u})$, $f(\hat{x} + \alpha \tilde{x}, \hat{u}) = \hat{y}$. This implies that $\hat{x} + \alpha \tilde{x} \in \tilde{X}(\hat{u}, \hat{y})$ for all $\alpha \geq 0$, which contradicts the boundedness of $\tilde{X}(\hat{u}, \hat{y})$. Hence $\{\lambda_k x^k\}$ necessarily has a convergent subsequence whose limit is denoted by x . We may assume that $\lambda_k x^k \rightarrow x$ from the first. Since $X(\hat{u})$ is closed, from the limit of $\lambda_k x^k + (1 - \lambda_k)\hat{x} \in X(\hat{u})$, $x + \hat{x} \in X(\hat{u})$. Therefore the limit of the left-hand side of (4.8), which is $f(x + \hat{x}, \hat{u})$, belongs to $Y(\hat{u})$. Since (4.8) leads to

$$f(x + \hat{x}, \hat{u}) \leq_P y + \hat{y}$$

when $k \rightarrow \infty$, $y \in -P$ and $\hat{y} \in W(\hat{u})$, y must be equal to the zero vector. Thus $Y(\hat{u})$ is P -bounded.

(c) Finally, the result proved just above and Lemma 4.4 imply that $Y(u)$ is P -bounded in a neighborhood of \hat{u} . This completes the proof of the proposition. \square

Now we can immediately obtain the following result by combining Theorem 4.1 and Proposition 4.1.

THEOREM 4.2. *If the following conditions are satisfied, then the perturbation map W is lower semicontinuous at \hat{u} :*

- (1) $\hat{u} \in \text{int}(\text{dom } X)$.
- (2) $X(u)$ is a closed set for every u near \hat{u} .
- (3) When $W(\hat{u}) \neq \emptyset$, $\tilde{X}(\hat{u}, \hat{y})$ is bounded for some $\hat{y} \in W(\hat{u})$.

Remark 4.4. The following examples show that each condition in the above theorem is essential.

(1) Consider the case in Remark 3.3(1). Then we can easily understand that the condition $\hat{u} \in \text{int}(\text{dom } X)$ is essential.

(2) Let $m = n = p = 1$, $P = \mathbb{R}_+$,

$$X(u) = \begin{cases} \{x \in \mathbb{R} \mid x > u^2\} & \text{if } u \neq 0, \\ \{x \in \mathbb{R} \mid x \geq 0\} & \text{if } u = 0, \end{cases}$$

and $f(x, u) = x$. Then

$$W(u) = \begin{cases} \emptyset & \text{if } u \neq 0, \\ \{0\} & \text{if } u = 0, \end{cases}$$

which is clearly not lower semicontinuous at $\hat{u} = 0$.

(3) Let $m = n = p = 1$, $P = \mathbb{R}_+$, $X(u) = \mathbb{R}_+$ and

$$f(x, u) = \begin{cases} 0 & \text{if } x \geq 0, \quad u = 0, \\ |u| e^{-x/|u|} & \text{if } x \geq 0, \quad u \neq 0, \\ |u| - x & \text{if } x < 0. \end{cases}$$

Then

$$Y(u) = \begin{cases} \{0\} & \text{if } u = 0, \\ \{y \mid 0 < y \leq |u|\} & \text{if } u \neq 0 \end{cases}$$

and so

$$W(u) = \begin{cases} \{0\} & \text{if } u = 0, \\ \emptyset & \text{if } u \neq 0. \end{cases}$$

Now $\tilde{X}(0, 0) = \mathbb{R}_+$, which is not bounded, and W is not lower semicontinuous at $\hat{u} = 0$.

5. Contingent derivative of the perturbation map. In this section we will show some quantitative results concerning the behavior of the perturbation map by using the concept of contingent derivatives of set-valued maps. The author has already provided an “inner” approximation of the contingent derivative of the perturbation map for general multi-objective optimization problems [6]. In this paper, a complete characterization of the contingent derivative will be obtained under the convexity assumption (CA) and some additional conditions.

First we briefly review the concept of contingent derivatives for set-valued maps.

DEFINITION 5.1. Let S be a nonempty subset of \mathbb{R}^q and $\hat{v} \in \mathbb{R}^q$. The set $T_S(\hat{v})$ defined by

$$(5.1) \quad T_S(\hat{v}) := \{v \in \mathbb{R}^q \mid \text{there exist sequences } \{h_k\} \subset \mathring{\mathbb{R}}_+ \text{ and } \{v^k\} \subset \mathbb{R}^q \\ \text{such that } h_k \rightarrow 0, v^k \rightarrow v \text{ and } \hat{v} + h_k v^k \in S \text{ for all } k\}$$

is called the contingent cone to S at \hat{v} , where $\mathring{\mathbb{R}}_+$ is the set of all positive real numbers.

DEFINITION 5.2. Let F be a set-valued map from \mathbb{R}^m to \mathbb{R}^p and $\bar{y} \in F(\bar{u})$. The set-valued map $DF(\bar{u}, \bar{y})$ from \mathbb{R}^m to \mathbb{R}^p defined by the following is called the contingent derivative of F at (\bar{u}, \bar{y}) :

$$(5.2) \quad y \in DF(\bar{u}, \bar{y})(u) \quad \text{iff } (u, y) \in T_{\text{graph } F}(\bar{u}, \bar{y}).$$

In other words, $y \in DF(\bar{u}, \bar{y})(u)$ if and only if there exist sequences $\{h_k\} \subset \mathring{\mathbb{R}}_+$, $\{u^k\} \subset \mathbb{R}^m$ and $\{y^k\} \subset \mathbb{R}^p$ such that $h_k \rightarrow 0$, $u^k \rightarrow u$, $y^k \rightarrow y$ and

$$\bar{y} + h_k y^k \in F(\bar{u} + h_k u^k) \quad \forall k.$$

The purpose of this section is to provide a complete characterization of the contingent derivative of the perturbation map. Throughout this section let \hat{y} be a P -minimal point of $Y(\hat{u})$, i.e., $\hat{y} \in W(\hat{u})$. First we can simplify Theorem 3.2 in [6] under the convexity assumption (CA) as in the following theorem.

THEOREM 5.1. *If $Y(u)$ is P -minicomplete for every u near \hat{u} , then*

$$(5.3) \quad \text{Min}_P DY(\hat{u}, \hat{y})(u) \subset DW(\hat{u}, \hat{y})(u) \quad \forall u \in \mathbb{R}^m.$$

Proof. Let $y \in \text{Min}_P DY(\hat{u}, \hat{y})(u)$. Since $y \in DY(\hat{u}, \hat{y})(u)$, there exist sequences $\{h_k\} \subset \mathring{\mathbb{R}}_+$, $\{u^k\} \subset \mathbb{R}^m$ and $\{y^k\} \subset \mathbb{R}^p$ such that $h_k \rightarrow 0$, $u^k \rightarrow u$, $y^k \rightarrow y$ and

$$\hat{y} + h_k y^k \in Y(\hat{u} + h_k u^k) \quad \forall k.$$

Since $Y(u)$ is P -minicomplete for every u near \hat{u} , there exists a sequence $\{\bar{y}^k\} \subset \mathbb{R}^p$ such that

$$(5.4) \quad \hat{y} + h_k \bar{y}^k \in W(\hat{u} + h_k u^k) \quad \text{and} \quad y^k - \bar{y}^k \in P$$

for all k sufficiently large. We may assume (5.4) for all k . Suppose that $\{\bar{y}^k\}$ has no convergent subsequence. Then $\|\bar{y}^k\| \rightarrow +\infty$. There exist sequences $\{x^k\}$ and $\{\bar{x}^k\}$ in \mathbb{R}^n such that

$$\begin{aligned} \hat{x} + h_k x^k &\in X(\hat{u} + h_k u^k), \\ f(\hat{x} + h_k x^k, \hat{u} + h_k u^k) &= \hat{y} + h_k y^k, \\ \hat{x} + h_k \bar{x}^k &\in X(\hat{u} + h_k u^k), \\ f(\hat{x} + h_k \bar{x}^k, \hat{u} + h_k u^k) &= \hat{y} + h_k \bar{y}^k. \end{aligned}$$

For any α satisfying $0 \leq \alpha \leq 1$, we have

$$\hat{x} + h_k(\alpha x^k + (1 - \alpha)\bar{x}^k) \in X(\hat{u} + h_k u^k)$$

from the convexity of X . Moreover, from the P -convexity of f ,

$$\begin{aligned} \hat{y} + h_k y^k(\alpha) &:= f(\hat{x} + h_k(\alpha x^k + (1 - \alpha)\bar{x}^k), \hat{u} + h_k u^k) \\ &\leq_P \hat{y} + h_k(\alpha y^k + (1 - \alpha)\bar{y}^k) \\ &\leq_P \hat{y} + h_k y^k. \end{aligned}$$

And, since f is continuous,

$$\begin{aligned} \hat{y} + h_k y^k(\alpha) &\rightarrow \hat{y} + h_k \bar{y}^k \quad \text{as } \alpha \rightarrow 0, \\ \hat{y} + h_k y^k(\alpha) &\rightarrow \hat{y} + h_k y^k \quad \text{as } \alpha \rightarrow 1. \end{aligned}$$

Since $\|\bar{y}^k\| \rightarrow +\infty$ and $y^k \rightarrow y$, by taking α_k appropriately close to 1, we have

$$\varepsilon h_k \leq \|\hat{y} + h_k y^k - (\hat{y} + h_k y^k(\alpha_k))\| \leq h_k \quad \forall k \text{ sufficiently large}$$

where ε is a fixed number such that $0 < \varepsilon < 1$. Taking this $y^k(\alpha_k)$ as \tilde{y}^k , we see that

$$\varepsilon \leq \|y^k - \tilde{y}^k\| \leq 1 \quad \forall k \text{ sufficiently large.}$$

Since $y^k \rightarrow y$, the sequence $\{\tilde{y}^k\}$ is bounded and so we may assume without loss of generality that $\{\tilde{y}^k\}$ converges to a vector \tilde{y} . It is clear that $\tilde{y} \in DY(\hat{u}, \hat{y})(u)$. Since

$\|y^k - \tilde{y}^k\| \geq \varepsilon$ for all k sufficiently large, $\|y - \tilde{y}\| \geq \varepsilon$, that is, $y \neq \tilde{y}$. Since $y^k - \tilde{y}^k \in P$, $y - \tilde{y} \in P$. However, these contradict the assumption that $y \in \text{Min}_P DY(\hat{u}, \hat{y})(u)$. Therefore $\{\tilde{y}^k\}$ always has a convergent subsequence. Hence we may assume from the first that $\tilde{y}^k \rightarrow \bar{y}$. Then $\bar{y} \in DW(\hat{u}, \hat{y})(u) \subset DY(\hat{u}, \hat{y})(u)$ and $y^k - \tilde{y}^k \rightarrow y - \bar{y} \in P$. Since $y \in \text{Min}_P DY(\hat{u}, \hat{y})(u)$, $y = \bar{y}$. This implies that $y \in DW(\hat{u}, \hat{y})(u)$, and completes the proof of the theorem. \square

Remark 5.1. We can see from the example in Remark 4.4(3) that the P -mini-completeness condition is essential for Theorem 5.1. There, $DW(\hat{u}, \hat{y})(u) = \emptyset$ for $\hat{u} = 0$, $\hat{y} = 0$ and $u \neq 0$. However $DY(\hat{u}, \hat{y})(u) = [0, |u|]$ and $\text{Min}_P DY(\hat{u}, \hat{y})(u) = \{0\}$ for $u \neq 0$.

Next we consider sufficient conditions for the converse inclusion of (5.3).

DEFINITION 5.3. Let S be a nonempty set in \mathbb{R}^p and $\hat{v} \in \mathbb{R}^p$. The normal cone $N_S(\hat{v})$ to S at \hat{v} is the negative polar cone of the tangent cone $T_S(\hat{v})$, i.e.,

$$(5.5) \quad N_S(\hat{v}) = [T_S(\hat{v})]^\circ = \{\mu \in \mathbb{R}^p \mid \langle \mu, v \rangle \leq 0 \ \forall v \in T_S(\hat{v})\}.$$

When S is a convex set and $\hat{v} \in S$,

$$(5.6) \quad N_S(\hat{v}) = \{\mu \in \mathbb{R}^p \mid \langle \mu, \hat{v} \rangle \geq \langle \mu, v \rangle \ \forall v \in S\}.$$

DEFINITION 5.4. Let S be a nonempty P -convex set in \mathbb{R}^p . If a point $\hat{y} \in \text{Min}_P S$ satisfies the condition

$$(5.7) \quad N_{S+P}(\hat{y}) \subset \text{int } P^\circ \cup \{0\},$$

then \hat{y} is called the normally P -minimal point of S .

Remark 5.2. A point $\hat{y} \in S$ is said to be the properly P -minimal point of S if

$$(5.8) \quad T_{S+P}(\hat{y}) \cap (-P) = \{0\}^3.$$

If \hat{y} is a properly P -minimal point of a convex set, there exists a vector $\mu \in N_{S+P}(\hat{y}) \cap \text{int } P^\circ$. The relation (5.7) is a stronger requirement than the existence of such μ as long as $\hat{y} \in \text{Min}_P S$. In other words, the normal P -minimality is a stronger concept than the proper P -minimality. From the geometric viewpoint, the latter implies the existence of the supporting hyperplane to S at \hat{y} with the normal vector μ in $\text{int } P^\circ$ and, on the other hand, the former implies that all the normal vectors of the supporting hyperplanes to S at \hat{y} belong to $\text{int } P^\circ$. (The existence of such a hyperplane is guaranteed by the fact that $\hat{y} \in \text{Min}_P S$.)

Remark 5.3. It is not difficult to show that the normal P -minimality of \hat{y} to a convex set S is equivalent to the following condition:

$$(5.9) \quad \text{int } T_{S+P}(\hat{y}) \cup \{0\} \supset P.$$

THEOREM 5.2. If $\hat{u} \in \text{int}(\text{dom } Y)$ and \hat{y} is a normally P -minimal point of $Y(\hat{u})$, then

$$(5.10) \quad DW(\hat{u}, \hat{y})(u) \subset \text{Min}_P DY(\hat{u}, \hat{y})(u) \quad \forall u \in \mathbb{R}^m.$$

Proof. Let $y \in DW(\hat{u}, \hat{y})(u)$. Of course $y \in DY(\hat{u}, \hat{y})(u)$. Hence if we assume that $y \notin \text{Min}_P DY(\hat{u}, \hat{y})(u)$, there exists $\bar{y} \in DY(\hat{u}, \hat{y})(u)$ such that $y - \bar{y} \in P \setminus \{0\}$. Since $\bar{y} \in DY(\hat{u}, \hat{y})(u)$, there exist sequences $\{\bar{h}_k\} \subset \mathbb{R}_+$, $\{\bar{u}^k\} \subset \mathbb{R}^m$ and $\{\bar{y}^k\} \subset \mathbb{R}^p$ such that $\bar{h}_k \rightarrow 0$, $\bar{u}^k \rightarrow u$, $\bar{y}^k \rightarrow \bar{y}$ and

$$\hat{y} + \bar{h}_k \bar{y}^k \in Y(\hat{u} + \bar{h}_k \bar{u}^k) \quad \forall k.$$

³ There are several definitions of the proper P -minimality (see, e.g., [5]). However they coincide under the convexity assumption.

On the other hand, since $y \in DW(\hat{u}, \hat{y})(u)$, there exist sequences $\{h_k\} \in \mathring{\mathbb{R}}_+$, $\{u^k\} \subset \mathbb{R}^m$ and $\{y^k\} \subset \mathbb{R}^p$ such that $h_k \rightarrow 0$, $u^k \rightarrow u$, $y^k \rightarrow y$ and

$$\hat{y} + h_k y^k \in W(\hat{u} + h_k u^k) \quad \forall k.$$

Since $h_k \rightarrow 0$, we may assume that $h_k \leq \bar{h}_k$ by taking a subsequence if necessary. Since $\hat{y} + h_k y^k \in W(\hat{u} + h_k u^k)$, $(\hat{u} + h_k u^k, \hat{y} + h_k y^k)$ is a boundary point of the convex set $\text{graph}(Y + P)$. Hence there exist a vector $(\lambda^k, \mu^k) \in \mathbb{R}^m \times \mathbb{R}^p$ such that

$$(5.11) \quad \langle \lambda^k, \hat{u} + h_k u^k \rangle + \langle \mu^k, \hat{y} + h_k y^k \rangle \geq \langle \lambda^k, u' \rangle + \langle \mu^k, y' \rangle \quad \forall (u', y') \in \text{graph}(Y + P)$$

for each k . Since we may normalize these vectors so that $\|(\lambda^k, \mu^k)\| = 1$, we may assume that $\{(\lambda^k, \mu^k)\}$ converges to a nonzero vector $(\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}^p$. By taking the limit of (5.11) as $k \rightarrow \infty$, we see that

$$(5.12) \quad \langle \lambda, \hat{u} \rangle + \langle \mu, \hat{y} \rangle \geq \langle \lambda, u' \rangle + \langle \mu, y' \rangle \quad \forall (u', y') \in \text{graph}(Y + P).$$

Since $\hat{u} \in \text{int}(\text{dom } Y)$, $\mu \neq 0$. Take an arbitrary $\tilde{y} \in Y(\hat{u}) + P$. From Lemma 2.3, the set-valued map $Y + P$ is lower semicontinuous at \hat{u} and so there exist a sequence $\{\tilde{y}^k\} \subset \mathbb{R}^p$ and a number $K > 0$ such that $\tilde{y}^k \rightarrow \tilde{y}$ and

$$(5.13) \quad \tilde{y}^k \in Y(\hat{u} + h_k u^k) + P \quad \text{for } k \geq K.$$

From (5.11), for $k \geq K$

$$\langle \lambda^k, \hat{u} + h_k u^k \rangle + \langle \mu^k, \hat{y} + h_k y^k \rangle \geq \langle \lambda^k, \hat{u} + h_k u^k \rangle + \langle \mu^k, \tilde{y}^k \rangle.$$

Letting $k \rightarrow \infty$, we have that

$$\langle \mu, \hat{y} \rangle \geq \langle \mu, \tilde{y} \rangle.$$

This implies that $\mu \in N_{Y(\hat{u})+P}(\hat{y})$. Since \hat{y} is a normally P -minimal point of $Y(\hat{u})$, $\mu \in \text{int } P^\circ$. Since $y - \bar{y} \in P \setminus \{0\}$,

$$(5.14) \quad \langle \mu, y \rangle < \langle \mu, \bar{y} \rangle.$$

Recalling that $\hat{y} + \bar{h}_k \bar{y}^k \in Y(\hat{u} + \bar{h}_k \bar{u}^k)$, $\hat{y} \in Y(\hat{u})$ and $h_k \leq \bar{h}_k$, we obtain that

$$\hat{y} + h_k \bar{y}^k \in Y(\hat{u} + h_k \bar{u}^k) + P$$

from the P -convexity of Y . Hence, from (5.11),

$$\langle \lambda^k, \hat{u} + h_k u^k \rangle + \langle \mu^k, \hat{y} + h_k y^k \rangle \geq \langle \lambda^k, \hat{u} + h_k \bar{u}^k \rangle + \langle \mu^k, \hat{y} + h_k \bar{y}^k \rangle,$$

i.e.,

$$\langle \lambda^k, u^k \rangle + \langle \mu^k, y^k \rangle \geq \langle \lambda^k, \bar{u}^k \rangle + \langle \mu^k, \bar{y}^k \rangle.$$

By taking the limit as $k \rightarrow \infty$, we have that

$$\langle \lambda, u \rangle + \langle \mu, y \rangle \geq \langle \lambda, u \rangle + \langle \mu, \bar{y} \rangle.$$

That is,

$$\langle \mu, y \rangle \geq \langle \mu, \bar{y} \rangle,$$

which contradicts (5.14). Therefore $y \in \text{Min}_P DY(\hat{u}, \hat{y})(u)$, as was to be proved. \square

Remark 5.4. The following examples show that the conditions in Theorem 5.2 are essential.

(1) ($\hat{u} \notin \text{int}(\text{dom } Y)$). Let $m = 2$, $n = p = 1$, $P = \mathbb{R}_+$,

$$X(u) = \begin{cases} \{x \mid x \geq 0\} & \text{if } u_1 \geq 0, \quad u_2 > 0, \\ \{x \mid x \geq u_1\} & \text{if } u_1 \geq 0, \quad u_2 = 0, \\ \emptyset & \text{otherwise,} \end{cases}$$

and $f(x, u) = x$. Then $Y(u) = X(u)$ and

$$W(u) = \begin{cases} \{0\} & \text{if } u_1 \geq 0, \quad u_2 > 0, \\ \{u_1\} & \text{if } u_1 \geq 0, \quad u_2 = 0, \\ \emptyset & \text{otherwise.} \end{cases}$$

Let $\hat{u} = (0, 0) \notin \text{int}(\text{dom } Y)$, $\hat{y} = 0$ and $u = (1, 0)$. Then $DW(\hat{u}, \hat{y})(u) = \{0, 1\}$ and $DY(\hat{u}, \hat{y})(u) = \{y \mid y \geq 0\}$. Hence $DW(\hat{u}, \hat{y})(u) \not\subset \text{Min}_P DY(\hat{u}, \hat{y})(u)$.

(2) (\hat{y} is not normally P -minimal). Let $m = 1$, $n = p = 2$, $P = \mathbb{R}_+^2$,

$$X(u) = \begin{cases} \{x \in \mathbb{R}^2 \mid x_1 \geq 0, x_2 \geq |u| e^{-x_1/|u|}\} & \text{if } u \neq 0, \\ \{x \in \mathbb{R}^2 \mid x_1 \geq 0, x_2 \geq 0\} & \text{if } u = 0, \end{cases}$$

and $f(x, u) = (x_1, x_2)$. Then $Y(u) = X(u)$ and

$$W(u) = \begin{cases} \{y \in \mathbb{R}^2 \mid y_1 \geq 0, y_2 = |u| e^{-y_1/|u|}\} & \text{if } u \neq 0, \\ \{(0, 0)\} & \text{if } u = 0. \end{cases}$$

Let $\hat{u} = 0$ and $\hat{y} = (0, 0)$. Then \hat{y} is not a normally P -minimal point of $Y(\hat{u})$, though it is properly P -minimal. In this case $(0, 0) \in DW(\hat{u}, \hat{y})(0) \subset DY(\hat{u}, \hat{y})(0)$ and $(1, 0) \in DW(\hat{u}, \hat{y})(0)$. Hence $DW(\hat{u}, \hat{y})(0) \not\subset \text{Min}_P DY(\hat{u}, \hat{y})(0)$.

Now we can consider the case in which every objective function f_i is differentiable.

DEFINITION 5.5. Let F be a set-valued map from \mathbb{R}^m to \mathbb{R}^p and $\bar{y} \in F(\bar{u})$. F is said to be upper pseudo-Lipschitzian at (\bar{u}, \bar{y}) if there exist neighborhood N_1 and N_2 of \bar{u} and \bar{y} , respectively, and a positive number M such that

$$(5.15) \quad F(u) \cap N_2 \subset F(\bar{u}) + M\|u - \bar{u}\|B \quad \forall u \in N_1.$$

PROPOSITION 5.1. If $X(\hat{u})$ is a closed set and $\tilde{X}(\hat{u}, \hat{y})$ is bounded, then \tilde{X} is locally bounded at (\hat{u}, \hat{y}) .

Proof. Suppose that the conclusion of the proposition is not true. Then there exist sequences $\{u^k\} \subset \mathbb{R}^m$, $\{y^k\} \subset \mathbb{R}^p$ and $\{x^k\} \subset \mathbb{R}^n$ such that $u^k \rightarrow \hat{u}$, $y^k \rightarrow \hat{y}$, $\|x^k\| \rightarrow +\infty$ and

$$x^k \in X(u^k) \quad \text{and} \quad f(x^k, u^k) = y^k \quad \forall k.$$

We may assume without loss of generality that $\{x^k / \|x^k\|\}$ converges to a nonzero vector x . Let $\alpha > 0$. Since $\|x^k\| \rightarrow +\infty$, $0 < \alpha / \|x^k\| \leq 1$ for all k sufficiently large. Hence, from the convexity of X ,

$$(5.16) \quad \left(1 - \frac{\alpha}{\|x^k\|}\right)\hat{x} + \frac{\alpha}{\|x^k\|}x^k \in X\left(\left(1 - \frac{\alpha}{\|x^k\|}\right)\hat{u} + \frac{\alpha}{\|x^k\|}u^k\right).$$

Since X is upper semicontinuous at \hat{u} from Lemma 2.4, by taking the limit of (5.16) as $k \rightarrow \infty$, we see that

$$\hat{x} + \alpha x \in X(\hat{u}).$$

Since f is P -convex,

$$f\left(\left(1 - \frac{\alpha}{\|x^k\|}\right)(\hat{x}, \hat{u}) + \frac{\alpha}{\|x^k\|}(x^k, u^k)\right) \leq_P \left(1 - \frac{\alpha}{\|x^k\|}\right)f(\hat{x}, \hat{u}) + \frac{\alpha}{\|x^k\|}f(x^k, u^k).$$

Letting $k \rightarrow \infty$, we have

$$f(\hat{x} + \alpha x, \hat{u}) \leq_P \hat{y}.$$

Since $\hat{y} \in W(\hat{u})$, $f(\hat{x} + \alpha x, \hat{u}) = \hat{y}$. Hence $\hat{x} + \alpha x \in \tilde{X}(\hat{u}, \hat{y})$ for any $\alpha > 0$. However, this contradicts the boundedness of $\tilde{X}(\hat{u}, \hat{y})$. Therefore \tilde{X} is locally bounded at (\hat{u}, \hat{y}) .

PROPOSITION 5.2. *If $X(\hat{u})$ is a closed set, if $\tilde{X}(\hat{u}, \hat{y})$ is a singleton, i.e., $\tilde{X}(\hat{u}, \hat{y}) = \{\hat{x}\}$ and if \tilde{X} is upper pseudo-Lipschitzian at $(\hat{u}, \hat{y}, \hat{x})$, then*

$$(5.17) \quad DY(\hat{u}, \hat{y})(u) = \nabla_x f(\hat{x}, \hat{u}) \cdot DX(\hat{u}, \hat{x})(u) + \nabla_u f(\hat{x}, \hat{u}) \cdot u \quad \forall u \in \mathbb{R}^m.$$

Proof. It has been already proved that

$$DY(\hat{u}, \hat{y})(u) \supset \nabla_x f(\hat{x}, \hat{u}) \cdot DX(\hat{u}, \hat{x})(u) + \nabla_u f(\hat{x}, \hat{u}) \cdot u$$

[6, Prop. 4.1]. So we shall prove the converse inclusion here. Let $y \in DY(\hat{u}, \hat{y})(u)$. Then there exist sequences $\{h_k\} \subset \mathbb{R}_+$, $\{u^k\} \subset \mathbb{R}^m$ and $\{y^k\} \subset \mathbb{R}^p$ such that $h_k \rightarrow 0$, $u^k \rightarrow u$, $y^k \rightarrow y$ and $\hat{y} + h_k y^k \in Y(\hat{u} + h_k u^k)$ for all k . Hence there exists another sequence $\{x^k\} \subset \mathbb{R}^n$ such that

$$\hat{x} + h_k x^k \in \tilde{X}(\hat{u} + h_k u^k, \hat{y} + h_k y^k) \quad \forall k.$$

From Proposition 5.1, the sequence $\{h_k x^k\}$ is bounded and so has a convergent subsequence. We may assume from the first that $h_k x^k \rightarrow x \in \mathbb{R}^n$. Since X is upper semicontinuous at \hat{u} and f is continuous,

$$\hat{x} + x \in \tilde{X}(\hat{u}, \hat{y}).$$

Since $\tilde{X}(\hat{u}, \hat{y}) = \{\hat{x}\}$, $x = 0$. Namely, $h_k x^k \rightarrow 0$. Since \tilde{X} is upper pseudo-Lipschitzian at $(\hat{u}, \hat{y}, \hat{x})$, there exists $M > 0$ such that, for any k sufficiently large,

$$\|\hat{x} + h_k x^k - \hat{x}\| \leq M \|(\hat{u} + h_k u^k, \hat{y} + h_k y^k) - (\hat{u}, \hat{y})\|,$$

i.e.,

$$\|x^k\| \leq M \|(u^k, y^k)\|.$$

Since $u^k \rightarrow u$ and $y^k \rightarrow y$, $\{x^k\}$ is bounded. Hence we may assume that $x^k \rightarrow \bar{x}$. Then clearly $\bar{x} \in DX(\hat{u}, \hat{y})(u)$ and

$$\begin{aligned} y &= \lim_{k \rightarrow \infty} y^k = \lim_{k \rightarrow \infty} \frac{f(\hat{x} + h_k x^k, \hat{u} + h_k u^k) - f(\hat{x}, \hat{u})}{h_k} \\ &= \nabla_x f(\hat{x}, \hat{u}) \cdot \bar{x} + \nabla_u f(\hat{x}, \hat{u}) \cdot u. \end{aligned}$$

Therefore $y \in \nabla_x f(\hat{x}, \hat{u}) \cdot DX(\hat{u}, \hat{x})(u) + \nabla_u f(\hat{x}, \hat{u}) \cdot u$. This completes the proof. \square

Thus, from Theorems 5.1 and 5.2 and Proposition 5.2, we have the following theorem which provides a complete characterization of the contingent derivative of the perturbation map W .

THEOREM 5.3. *If the following conditions (1)–(5) are satisfied, then*

$$(5.18) \quad DW(\hat{u}, \hat{y})(u) = \text{Min}_P [\nabla_x f(\hat{x}, \hat{u}) \cdot DX(\hat{u}, \hat{x})(u) + \nabla_u f(\hat{x}, \hat{u}) \cdot u] \quad \forall u \in \mathbb{R}^m.$$

- (1) $\hat{u} \in \text{int}(\text{dom } Y)$,
- (2) \hat{y} is a normally P -minimal point of $Y(\hat{u})$,
- (3) $X(u)$ is a closed set for every u in a neighborhood of \hat{u} ,
- (4) $\tilde{X}(\hat{u}, \hat{y})$ is a singleton, i.e., $\tilde{X}(\hat{u}, \hat{y}) = \{\hat{x}\}$,
- (5) \tilde{X} is upper pseudo-Lipschitzian at $(\hat{u}, \hat{y}, \hat{x})$.

Finally we briefly mention sufficient conditions for the pseudo-Lipschitzian property of \tilde{X} . The following proposition can be obtained by applying Theorem 4.12 in Rockafellar [4].

PROPOSITION 5.3. *If the following two conditions are satisfied, then \tilde{X} is (upper) pseudo-Lipschitzian at $(\hat{u}, \hat{y}, \hat{x})$:*

- (1) $X(u)$ is a closed set for every u in a neighborhood of \hat{u} ,
 - (2) If $\sum_{i=1}^p \alpha_i \nabla_x f_i(\hat{x}, \hat{u}) + \nu = 0$ for some $(\lambda, \nu) \in N_{\text{graph } X}(\hat{u}, \hat{x})$, then
- $$(5.19) \quad \alpha_i = 0 \quad \text{for } i = 1, \dots, p \text{ and } \lambda = 0.$$

Remark 5.5. When $X(u)$ is specified by inequality constraints such as

$$X(u) = \{x \in \mathbb{R}^n \mid g(x) \leq u\}$$

condition (2) in Proposition 5.3 is nothing but the Mangasarian-Fromovitz constraint qualification at \hat{x} for the set

$$\tilde{X}(\hat{u}, \hat{y}) = \{x \in \mathbb{R}^n \mid f(x, \hat{u}) - \hat{y} = 0, g(x) - \hat{u} \leq 0\}.$$

In view of Proposition 5.3, we can replace the condition (5) in Theorem 5.3 by the condition (2) in Proposition 5.3.

6. Conclusion. We have obtained sufficient conditions for the upper and lower semicontinuity of the perturbation map, which provides the set of all cone minimal points depending upon the parameter vector, in convex vector optimization. It has been shown that the convexity assumption considerably simplifies the results in the general case. We have also provided a complete characterization of the contingent derivative of the perturbation map when the nominal point is normally minimal.

REFERENCES

- [1] J.-P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, John Wiley, New York, 1984.
- [2] P. H. NACCACHE, *Stability in multicriteria optimization*, J. Math. Anal. Appl., 68 (1979), pp. 441–453.
- [3] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.
- [4] ———, *Lipschitzian properties of multifunctions*, Nonlinear Anal. Theory Methods Appl., 9 (1985), pp. 867–885.
- [5] Y. SAWARAGI, H. NAKAYAMA AND T. TANINO, *Theory of Multiobjective Optimization*, Academic Press, New York, 1985.
- [6] T. TANINO, *Sensitivity analysis in multiobjective optimization*, J. Optim. Theory Appl., 56 (1988), pp. 479–499.
- [7] T. TANINO AND Y. SAWARAGI, *Stability of nondominated solutions in multicriteria decision-making*, J. Optim. Theory Appl., 30 (1980), pp. 229–253.

POSITIVE SEMIDEFINITE MATRICES: CHARACTERIZATION VIA CONICAL HULLS AND LEAST-SQUARES SOLUTION OF A MATRIX EQUATION*

J. C. ALLWRIGHT†

Abstract. Any real symmetric $n \times n$ matrix A can be described by an $n(n+1)/2$ -component vector. Positive semidefiniteness of A is characterized by the associated vector belonging to the conical hull of a suitable convex set. This characterization is used to facilitate least-squared error solution, with respect to such A , of $F = AG$, where F and G are given matrices. The solution method involves finding the point in the conical hull of a convex set which is nearest to a vector. An algorithm is given for solving that proximal point problem.

Key words. positive semidefinite matrices, least-squared error solution of matrix equations, optimization on cones

AMS (MOS) subject classifications. 15A57, 65F99, 90C25

1. Introduction. A real symmetric $n \times n$ matrix A can be identified uniquely by an $n(n+1)/2$ component vector $\overline{\text{vec}}(A)$ associated with the entries of A . In § 2 it is shown that A is positive semidefinite if and only if $\overline{\text{vec}}(A)$ lies in the conical hull of a suitable specified convex set Ω , and that A is positive definite if and only if $\overline{\text{vec}}(A)$ lies in the interior of that conical hull. That characterization of positive semidefiniteness is used in § 3 to solve the problem of least-squared error solution, with respect to symmetric positive semidefinite A , of the equation $F = AG$. The error measure used is the Frobenius norm of $F - AG$, $\|F - AG\|_F$. The solution method involves finding the closest point in a conical hull to a vector. An algorithm for solving that proximal point problem is given in § 4. Approximation issues are addressed throughout. Computational effort and numerical results for an example are the subject of § 5.

Some algorithms for optimization problems associated with symmetric positive semidefinite matrices have received attention in the recent literature, e.g., [1], [2]. Least-squared error solution of $F = AG$ with respect to such matrices A is a fundamental problem which does not seem to have been tackled before. A result of Higham [3, Lemma 2.5] reveals that for the case when $F = F'$ (i.e., when F is symmetric) and $G = I$, an approximation to the minimizer \hat{A} of $\|F - AG\|_F$ is given by the polar factor H in the polar decomposition $F = UH$, with $\|F - HG\|_F \leq 2\|F - \hat{A}G\|_F$. In fact, for general F when $G = I$, the global minimizer \hat{A} of $\|F - AG\|_F$ is actually $\frac{1}{2}[F + F']_{(0)}$, where the notation $[M]_{(0)}$ denotes the result of changing to 0 all the negative eigenvalues in the spectral form of M [4, Thm. 2.1.3]. Unfortunately, there does not seem to be a simple formula for the global minimizer \hat{A} for the case of general G —hence the analysis and algorithm of this paper. The original motivation for consideration of least-squared error solution of $F = AG$ actually arose in connexion with the development of new methods for estimating the inverse Hessian matrix in quasi-Newton algorithms.

The approach adopted here is more attractive computationally than the direct use as constraints in an optimization algorithm of the conventional tests for positive

* Received by the editors October 1, 1984; accepted for publication (in revised form) July 27, 1987.

† Department of Electrical Engineering, Imperial College of Science and Technology, London SW7 2BT, U.K.

semidefiniteness (nonnegativity of all of the eigenvalues of A or of all the principal minors), since they depend in a complicated way on the entries of A .

A fact is that a symmetric matrix A is positive semidefinite iff it can be written as $A = B^2$ for some symmetric B . The obvious way to use that fact is to consider minimization with respect to symmetric B of the Frobenius norm of $F - B^2G$. Unfortunately, the resulting optimization problem is, in general, nonconvex in B . The approach of this paper has the advantage that it involves only convex optimization problems, even though the representation of positive semidefinite A as B^2 plays an important role in one of the conical hulls, which is used later.

Proofs are given at the end of the section concerned. The following notation is used throughout.

The range of a real matrix B is written $R[B]$, its null space $N[B]$, the dimension of its null space N_B and its Moore–Penrose pseudoinverse B^\dagger . The smallest singular value of B is written $\sigma_{\min}[B]$, and when B is symmetric its most negative eigenvalue is denoted by $\lambda_{\min}[B]$. For symmetric positive semidefinite B , $[B]_{(\alpha)}$ denotes the matrix obtained by replacing all the zero eigenvalues in the spectral form of B by the real scalar α .

The Kronecker product of matrices B and C is written $B \otimes C$.

The Frobenius-norm of B is denoted $\|B\|_F$ and the 2-norm of a vector or matrix is written $\|\cdot\|$.

The sets of nonnegative and strictly positive reals are written as R_{\geq} and $R_{>}$, respectively. For $A \in R^{n \times n}$, the notation $A \geq 0$ means that A is positive semidefinite and $A > 0$ means that A is positive definite.

The convex hull of a set $X \neq \emptyset$ is denoted by $\text{conv}[X]$, the conical hull of X (i.e., $\text{conv}[\{\alpha x: \alpha \in R_{\geq}, x \in X\}]$) by $\text{cone}[X]$, and the interior by $\text{int}[X]$. The set BX is $\{Bx: x \in X\}$ and the line $\{\alpha x + (1 - \alpha)y: \alpha \in [0, 1]\}$ between two points x and y is written line $\{x, y\}$.

Minimization and minimizers will always refer to global minimization and global minimizers. A minimizer of $f: R^n \rightarrow R$ with respect to x from a set X will often be denoted \hat{x} . The unique point \hat{x} in a closed, convex set X which minimizes $\|d - x\|$ with respect to $x \in X$ is often called $\text{minpoint}[d, X]$ and $\text{mindist}[d, X]$ is the corresponding minimal distance $\|d - \hat{x}\|$. Notation regarding minimization will only be used when the minimum exists, even when the statement and proof of existence of the minimum actually appear later in the paper.

Two minimization problems will be said to be equivalent if they have the same minimal value, and a global minimizer of either may be found easily from a global minimizer of the other.

2. Conical hull characterizations of the set of symmetric positive semidefinite matrices. For $C \in R^{m \times n}$, let $\text{vec}(C)$ be the following vector containing all the entries of C :

$$\text{vec}(C) = [c_{1*} c_{2*} \cdots c_{m*}]' \in R^{mn}$$

where c_{i*} denotes row i of C .

Let

$$S^n = \{A \in R^{n \times n}: A' = A\},$$

$$S_{\geq}^n = \{A \in R^{n \times n}: A' = A \geq 0\},$$

$$S_{>}^n = \{A \in R^{n \times n}: A' = A > 0\}.$$

Consider the linear subspace $\text{vec}(S^n) = \{\text{vec}(A) : A \in S^n\} \subset R^{n^2}$ of all vectors $\text{vec}(A)$ associated with symmetric A . It has dimension $r = n(n+1)/2$. Suppose w_1, w_2, \dots, w_r is an orthonormal basis-set for $\text{vec}(S^n)$. For example, suitable w_i for $n=2$ might be $w_1 = [1 \ 0 \ 0 \ 0]'$, $w_2 = [0 \ 2^{-1/2} \ 2^{-1/2} \ 0]'$, $w_3 = [0 \ 0 \ 0 \ 1]'$. Consequently

$$(2.1) \quad W = [w_1 w_2 \dots w_r] \in R^{n^2 \times r}$$

is a basis-matrix for $\text{vec}(S^n)$ and

$$(2.2) \quad W'W = I_{r \times r}, \quad R[W'] = R', \quad R[W] = \text{vec}(S^n).$$

For symmetric A , let $\overline{\text{vec}}(A)$ denote the vector of coordinates of $\text{vec}(A)$ with respect to the basis-set w_1, w_2, \dots, w_r i.e., with respect to the columns of W . Then, in view of (2.1) and (2.2)

$$(2.3) \quad \text{vec}(A) = W \overline{\text{vec}}(A) \in R^{n^2}, \quad \overline{\text{vec}}(A) = W' \text{vec}(A) \in R^r.$$

Let

$$(2.4) \quad U = \{B \in R^{n \times n} : B' = B \text{ and } \|B\|_F = 1\},$$

$$(2.5) \quad \Omega = \text{conv}\{\Psi\}, \quad \Psi = \{\text{vec}(B^2) : B \in U\} \subset R^{n^2}.$$

The functions $\text{vec}^{-1} : R^{n^2} \rightarrow R^{n \times n}$ and $\overline{\text{vec}}^{-1} : R^r \rightarrow S^n$ will be useful later.

The characterizations of positive semidefiniteness which follow are based on the idea that any $A \in S_{\geq}^n$ can be represented either as $A = \alpha B^2$ for some $\alpha \in R_{\geq}$ and some symmetric B with $\|B\|_F = 1$, or as $A = \alpha B$ for some $\alpha \in R_{\geq}$ and some $B \in S_{\geq}^n$ with $\|B\|_F \leq 1$. In detail, the two characterizations of S_{\geq}^n are

$$S_{\geq}^n = \text{cone}[\Xi], \quad \text{where } \Xi = \{B^2 : B \in S^n, \|B\|_F = 1\},$$

$$S_{\geq}^n = \text{cone}[\tilde{\Xi}], \quad \text{where } \tilde{\Xi} = \{B : B \in S_{\geq}^n, \|B\|_F \leq 1\}.$$

For optimization purposes it is convenient to consider conical hull characterizations of $\overline{\text{vec}}(S_{\geq}^n) := \{\overline{\text{vec}}(A) : A \in S_{\geq}^n\}$ and of $\overline{\text{vec}}(S_{>}^n) := \{\overline{\text{vec}}(A) : A \in S_{>}^n\}$, which are cones of vectors associated with $n \times n$ symmetric positive semidefinite matrices and positive definite matrices, respectively. Those cones are next characterized in terms of Ω , using an approach based on the fact that $S_{\geq}^n = \text{cone}[\Xi]$. Some useful properties of Ω and Ψ , both of (2.5), are also given.

THEOREM 2.1.

- (i) $\overline{\text{vec}}(S_{\geq}^n) = \text{cone}[\Omega]$, $\overline{\text{vec}}(S_{>}^n) = \text{cone}[W'\Omega]$;
- (ii) $\overline{\text{vec}}(S_{>}^n) = \text{int cone}[W'\Omega]$;
- (iii) Ψ is a compact set; Ω is a nonempty convex compact set with $\text{mindist}[0, \Omega] = n^{-1/2}$ and $\text{cone}[\Omega]$ is a nonempty closed convex cone;
- (iv) $\text{mindist}[0, LW'\Omega] \geq \sigma_{\min}[L]n^{-1/2}$ for all $L \in R^{r \times r}$;
- (v) For $g \in R^r$ and $L \in R^{r \times r}$: $\min_{\gamma \in LW'\Omega} g'\gamma = \lambda_{\min}[Z]$

and a minimizing γ is $\hat{\gamma} = LW' \text{vec}(\nu\nu')$. Here

$$(2.6) \quad Z = \frac{[\text{vec}^{-1}(\bar{g}) + \text{vec}^{-1}(\bar{g})']}{2} \in R^{n \times n} \quad \text{for } \bar{g} = WL'g$$

and ν is a normalized eigenvector of Z corresponding to the minimal eigenvalue $\lambda_{\min}[Z]$ of Z . \square

Minimization of $\|F - AG\|_F$, to be performed later, involves the determination of the point in $\text{cone}[LW'\Omega]$ which is nearest to a vector u , for a specific $L \in R^{r \times r}$. An algorithm for doing that is given in § 4. Actually any set Ω could be used so long as

cone $[\Omega] = S_{\equiv}^n$ and properties (4.3)-(4.5) of § 4 are valid. An alternative to Ω of (2.5) is that used in an earlier version of this paper [5]. Yet another corresponds to the characterization above of S_{\equiv}^n as cone $[\tilde{\Xi}]$ and is

$$\tilde{\Omega} = \{\text{vec}(B) : B \in S_{\equiv}^n \text{ and } \|B\|_F \leq 1\}.$$

The following results guarantee that properties (4.3)-(4.5) are actually valid when Ω is replaced by $\tilde{\Omega}$.

THEOREM 2.2.

- (i) $\text{vec}(S_{\equiv}^n) = \text{cone}[\tilde{\Omega}]$; $\overline{\text{vec}(S_{\equiv}^n)} = \text{cone}[W'\tilde{\Omega}]$; $\overline{\text{vec}(S_{\equiv}^n)} = \text{int cone}(W'\tilde{\Omega})$.
- (ii) $\tilde{\Omega}$ is a nonempty, closed convex set.
- (iii) Suppose $L \in R^{r \times r}$. Then, for any $x \in LW'\tilde{\Omega}$, there exists a point y along the ray through 0 and x such that $y \in LW'\tilde{\Omega}$ and $\|y\| \geq \sigma_{\min}[L]$.
- (iv) Suppose that $L \in R^{r \times r}$ and that Z of (2.6) has the spectral form $Z = V\Lambda V'$, where $\Lambda = \text{diag}(\lambda_1[Z], \dots, \lambda_n[Z])$. Let $\hat{\lambda} = [\sum_{i \in I} \lambda_i[Z]^2]^{1/2}$, where $I = \{i: \lambda_i[Z] < 0\}$. Then for any $g \in R^r$:

$$\min_{\gamma \in LW'\tilde{\Omega}} g'\gamma = \begin{cases} 0 & \text{if } \lambda_{\min}[Z] \geq 0, \\ -\hat{\lambda} & \text{otherwise} \end{cases}$$

and a minimizing γ from $LW'\tilde{\Omega}$ is $\tilde{\gamma} = 0$ if $\lambda_{\min}[Z] \geq 0$, and is $\tilde{\gamma} = LW' \text{vec}[V\hat{\Lambda}V']$ otherwise, where $\hat{\Lambda} = \text{diag}\{\hat{\lambda}_1, \dots, \hat{\lambda}_n\}$ for $\hat{\lambda}_i = -\min\{0, \lambda_i\}/\hat{\lambda}$. \square

These results for $\tilde{\Omega}$ will not be proved here as the proofs are similar to those of the results concerning Ω in Theorem 2.1, which will be given shortly.

Considerable numerical experience will be needed before it can be decided which of the possible sets Ω is best from the computational point of view. If $\tilde{\Omega}$ were used by the algorithm of § 4 instead of Ω of (2.5), then that algorithm would require the minimization of $g'\gamma$ (for an appropriate g) with respect to $\gamma \in LW'\tilde{\Omega}$ at each iteration, instead of with respect to $\gamma \in LW'\Omega$ for Ω of (2.5). Since it is clear from Theorems 2.1 and 2.2 that such minimization is more expensive for $\tilde{\Omega}$ than for Ω , and since Ω of [5] is about as easy to use as Ω of (2.5) but is defined in a less direct way, the set Ω of (2.5) will be used in the rest of this paper.

The proof of Theorem 2.1 is facilitated by the following lemmas. The proof of the first is omitted as it is straightforward.

LEMMA 2.1. (i) For any $A, B \in R^{n \times n}$ and any $x, y \in R^n$

$$\begin{aligned} \text{trace}[AB] &= \text{trace}[BA] = \text{trace}[B'A'], \\ x'y &= \text{trace}[\text{vec}^{-1}(x) \text{vec}^{-1}(y)']. \end{aligned}$$

(ii) For any $B \in U$ and for \bar{g} and Z of (2.6)

$$\bar{g}' \text{vec}(B^2) = \text{trace}[BZB].$$

(iii) For any $C \in R^{m \times n}$, any $A \in S^n$ and any orthogonal $P, Q \in R^{n \times n}$

$$\|C\|_F = \|\text{vec}(C)\|, \quad \|\overline{\text{vec}(A)}\| = \|\text{vec}(A)\|, \quad \|A\|_F^2 = \text{trace}[A^2], \quad \|PAQ\|_F = \|A\|_F.$$

(iv) For orthogonal $V \in R^{n \times n}$

$$U = \{VBV': B \in U\}. \quad \square$$

LEMMA 2.2. For $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$

$$\min_{C \in U} \text{trace}[C\Lambda C] = \lambda_n. \quad \square$$

Proof of Lemma 2.2. For any $C \in U$

$$\text{trace}[C\Lambda C] = \sum_{i=1}^n (C^2)_{ii} \lambda_i \geq \lambda_n \text{trace}[C^2] = \lambda_n \|C\|_F^2 = \lambda_n.$$

Further, for $C = e_n e_n'$, where $e_n = (0 \ 0 \ 0 \ \cdots \ 0 \ 1)'$, $C \in U$ and $\text{trace}[C \wedge C] = \lambda_n$. The result follows. \square

Proof of Theorem 2.1. (i) Let $\bar{\Sigma}_i$ denote $\sum_{i=1}^{n^2+1}$. By the Caratheodory representation theorem [6, Thm. 17.1], since Ω [of (2.5)] is a convex subset of R^{n^2} , every $x \in \Omega$ can be represented as

$$(2.7) \quad x = \bar{\Sigma}_i \alpha_i \text{vec}[B_i^2]$$

for some $\alpha_i \in R_{\geq}$ with $\bar{\Sigma}_i \alpha_i = 1$ and for some $B_i \in U$. Hence every $x \in \text{cone}[\Omega]$ can be written as $x = \bar{\Sigma}_i (\alpha \alpha_i) \text{vec}[B_i^2]$ for some $\alpha \in R_{\geq}$. The matrix X corresponding to x is $X = \text{vec}^{-1}(x) = \bar{\Sigma}_i (\alpha \alpha_i) B_i^2$ and consequently is positive semidefinite. Hence $\text{cone}[\Omega] \subset \text{vec}[S_{\geq}^n]$.

Also, $\text{vec}[S_{\geq}^n] \subset \text{cone}[\Omega]$ since any $X \in S_{\geq}^n$ can be written as $X = \alpha B^2$ for some $\alpha \in R_{\geq}$ and for some $B \in U$ so that $\text{vec}[X] \in \text{cone}[\Omega]$.

Hence $\text{vec}[S_{\geq}^n] = \text{cone}[\Omega]$ and consequently, in view of (2.3), $\overline{\text{vec}[S_{\geq}^n]} = W' \text{cone}[\Omega] = \text{cone}[W'\Omega]$, which proves Theorem 2.1(i).

(ii) Now

$$X \in S_{>}^n \Leftrightarrow \text{there exists } \delta > 0 \text{ such that } X + Y \in S_{\geq}^n, \text{ whenever } Y \in S^n \text{ and } \|Y\|_F < \delta;$$

$$\Leftrightarrow \text{there exists } \delta > 0 \text{ such that } \overline{\text{vec}}(X) + \overline{\text{vec}}(Y) \in \overline{\text{vec}}(S_{\geq}^n) \text{ whenever } Y \in S^n \text{ and } \|\text{vec}(Y)\| < \delta;$$

$$\Leftrightarrow \text{there exists } \delta > 0 \text{ such that } \overline{\text{vec}}(X) + z \in \overline{\text{vec}}(S_{\geq}^n) \text{ whenever } z \in R^r \text{ and } \|z\| < \delta \text{ (since, by Lemma 2.1(iii), } \{Y \in S^n: \|\text{vec}(Y)\| < \delta\} = \{\overline{\text{vec}}^{-1}(z): z \in R^r \text{ and } \|z\| < \delta\});$$

$$\Leftrightarrow \overline{\text{vec}}(X) \in \text{int } \overline{\text{vec}}(S_{\geq}^n),$$

so $\overline{\text{vec}}(S_{>}^n) = \text{int } \overline{\text{vec}}(S_{\geq}^n)$. Consequently, in view of Theorem 2.1(i), $\overline{\text{vec}}(S_{>}^n) = \text{int } \text{cone}[W'\Omega]$, which proves Theorem 2.1(ii).

(iii) Since U is compact, the set $\Psi = \{\text{vec}(B^2): B \in U\}$ is a compact subset of R^{n^2} so Ω , its convex hull, is also compact [7, Thm. 3.2.18].

Let $\tilde{\Sigma}_j$ denote $\sum_{j=1}^n$. For any $B \in U$

$$(2.8) \quad (B^2)_{jj} = \sum_{k=1}^n (b_{jk})^2 \geq 0, \quad \text{trace}[B^2] = \tilde{\Sigma}_j, \quad (B^2)_{jj} = 1.$$

Using representation (2.7) of any x from Ω

$$\|x\|^2 = \|\bar{\Sigma}_i \alpha_i \text{vec}(B_i^2)\|^2 = \|\bar{\Sigma}_i \alpha_i B_i^2\|_F^2 \geq \tilde{\Sigma}_j \{\bar{\Sigma}_i \alpha_i (B_i^2)_{jj}\}^2.$$

Hence, since [8, § 2.1] $\tilde{\Sigma}_j z_j^2 \geq [\tilde{\Sigma}_j |z_j|]^2/n$ for all $z \in R^n$, and in view of (2.8) and the fact that each B_i is in U of (2.4)

$$(2.9) \quad \|x\|^2 \geq [\tilde{\Sigma}_j \bar{\Sigma}_i \alpha_i (B_i^2)_{jj}]^2/n = [\bar{\Sigma}_i \alpha_i \tilde{\Sigma}_j (B_i^2)_{jj}]^2/n \\ = [\bar{\Sigma}_i \alpha_i]^2/n = n^{-1} \quad \forall x \in \Omega.$$

Take $B = n^{-1/2} I \in U$. Then $\text{vec}(B^2) \in \Omega$ and $\|\text{vec}(B^2)\| = n^{-1/2}$. Consequently, in view of (2.9): $\text{mindist}[0, \Omega] = n^{-1/2}$.

Since Ω is a nonempty, convex compact set and $\text{mindist}[0, \Omega] > 0$, it follows from [6, Cor. 9.6.1] that $\text{cone}[\Omega]$ is a nonempty convex closed cone. This completes the proof of Theorem 2.1(iii). \square

(iv) If $x \in LW'\Omega$, then $x = LW'\omega$ for some $\omega \in \Omega$ and $\|x\| \geq \sigma_{\min}[L] \|W'\omega\|$. From (2.1)–(2.2), WW' projects R^{n^2} orthogonally onto $R[W] = \text{vec}(S^n)$. Hence $\|W'\omega\|^2 = \omega' WW'\omega = \|\omega\|^2$ since, by Theorem 2.1(i), $\omega \in \text{vec}(S_{\geq}^n) \subset \text{vec}(S^n)$ when $\omega \in \Omega$. Hence,

by Theorem 2.1(iii)

$$\|x\| \geq \sigma_{\min}[L]\|\omega\| \geq \sigma_{\min}[L]n^{-1/2} \quad \forall x \in LW'\Omega,$$

which proves Theorem 2.1(iv).

(v) Put Z of (2.6) in the spectral form $Z = V\Lambda V'$, where $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n = \lambda_{\min}[Z]$. Then, since the minimization of $g'\gamma$ with respect to $\gamma \in LW'\Omega$ is equivalent to minimization of $\bar{g}'\omega$ with respect to $\omega \in \Omega$ for $\bar{g} = WL'g$, and since minimization of $\bar{g}'\omega$ with respect to $\omega \in \Omega$ is equivalent to minimization with respect to $\omega \in \Psi$ because, from (2.5), $\Omega = \text{conv}\{\Psi\}$

$$\begin{aligned} \min_{\gamma \in LW'\Omega} g'\gamma &= \min_{\omega \in \Omega} \bar{g}'\omega \\ &= \min_{B \in U} \bar{g}' \text{vec}(B^2) && \text{(from (2.5))} \\ &= \min_{B \in U} \text{trace}[BZB] && \text{(from Lemma 2.1(ii))} \\ (2.10) \quad &= \min_{B \in U} \text{trace}[V'BV\Lambda V'BV] && \text{(by Lemma 2.1(i) and since } Z = V\Lambda V') \\ &= \min_{C \in U} \text{trace}[C\Lambda C] && \text{(from Lemma 2.1(iv))} \\ &= \lambda_n = \lambda_{\min}[Z] && \text{(from Lemma 2.2).} \end{aligned}$$

Now consider $\hat{B} = \nu\nu'$, where ν is a normalized eigenvector of Z corresponding to the eigenvalue $\lambda_{\min}[Z]$. Then $\hat{B} \in U$, since $\hat{B}' = \hat{B}$ and since $\|\hat{B}\|_F^2 = \text{trace}[\hat{B}^2] = \|\nu\|^2 = 1$, so (from (2.5)) $\text{vec}(\hat{B}^2) \in \Omega$.

Let $\hat{\gamma} = LW' \text{vec}(\hat{B}^2)$. Then $\hat{\gamma} \in LW'\Omega$ and

$$\begin{aligned} g'\hat{\gamma} &= \bar{g}' \text{vec}(\hat{B}^2) = \text{trace}[\hat{B}Z\hat{B}] \quad \text{(by Lemma 2.1(ii))} \\ &= \lambda_{\min}[Z] \quad \text{(since } Z\nu = \lambda_{\min}[Z]\nu \text{ and } \|\nu\| = 1). \end{aligned}$$

Consequently, in view of (2.10), $\hat{\gamma}$ minimizes $g'\gamma$ with respect to $\gamma \in LW'\Omega$, which completes the proof of Theorem 2.1. \square

3. Minimization of $\|F - AG\|_F$ with respect to symmetric positive semidefinite A . Suppose $A \in R^{n \times n}$ and $F, G \in R^{n \times m}$. The optimization problem is

$$(3.1) \quad \text{P1} \quad \min_{A \in S_{\geq}^n} \|F - AG\|_F.$$

The conical hull characterization of the set of positive semidefinite matrices developed in § 2, $\text{vec}(S_{\geq}^n) = \text{cone}[W'\Omega]$, is used later to prove the following result.

THEOREM 3.1. *The minimum in P1 exists.* \square

Remark 3.1. There is actually a unique global minimizer \hat{A} for P1 when G has full rank (see Theorem 3.2) but \hat{A} is not necessarily unique when G does not have full rank (see, e.g., Theorem 3.5(i)). \square

When G has full rank, the conical hull characterization can be used to transform problem P1 into a problem, P2 of (3.3), for which the solution can be computed fairly easily. It will be seen shortly that the case with general G can be solved by employing the results for the full rank case which follow.

THEOREM 3.2. *Suppose $A \in R^{n \times n}$; $F, G \in R^{n \times m}$ with $m \geq n$ and $\text{rank}[G] = n$.*

Let $J = [I_{n \times n} \otimes G']W \in R^{mn \times r}$ where $r = n(n+1)/2$. Then J has rank r . Factorize J as $J = P[L'0']'$ for some invertible $L \in R^{r \times r}$, where $P \in R^{mn \times mn}$ is some orthogonal matrix. Using that P , factorize $f = \text{vec}[F] \in R^{mn}$ as $f = P[u'l']'$ with $u \in R^r$.

Then \hat{A} solves P1 if and only if

$$(3.2) \quad \hat{A} = \overline{\text{vec}}^{-1}(L^{-1}\hat{k}),$$

where \hat{k} is the unique solution of

$$(3.3) \quad \text{P2} \quad \min_{k \in K} \|u - k\|,$$

and

$$(3.4) \quad \|F - \hat{A}G\|_F^2 = \|u - \hat{k}\|^2 + \|l\|^2.$$

Here K is the convex cone

$$(3.5) \quad K = \text{cone}[LW'\Omega].$$

Further

$$(3.6) \quad \hat{k} = u \text{ iff } \overline{\text{vec}}^{-1}(L^{-1}u) \geq 0,$$

$$(3.7) \quad \hat{A} \text{ is not positive definite if } \hat{k} \neq u. \quad \square$$

Remark 3.2. Suppose $\overline{\text{vec}}^{-1}(L^{-1}u) \geq 0$. Then, from (3.6), $\hat{k} = u$, so that, from (3.2), $\hat{A} = \overline{\text{vec}}^{-1}(L^{-1}u)$. Hence \hat{A} is very easy to find in that case.

Now suppose $\overline{\text{vec}}^{-1}(L^{-1}u) \not\geq 0$. Then $\hat{k} \neq u$ and there does not seem to be a formula for \hat{k} ; however, an arbitrarily close approximation \bar{k} to \hat{k} can be obtained using an iterative procedure, Algorithm 4.1 of §4 (later). By Theorem 4.2, for any $\varepsilon > 0$, Algorithm 4.1 can be used to find an approximation \bar{k} which satisfies condition (3.8) below. The corresponding approximation \bar{A} to $\hat{A} = \overline{\text{vec}}^{-1}(L^{-1}\hat{k})$ is $\bar{A} = \overline{\text{vec}}^{-1}(L^{-1}\bar{k})$. Theorem 3.3, which follows, reveals that \bar{A} can be made an arbitrarily good approximation to \hat{A} by choosing ε to be small enough. Hence, when G has full rank, an arbitrarily accurate approximation \bar{A} to the solution \hat{A} for P1 of (3.1) can be found. \square

THEOREM 3.3. Suppose F, G, L, u and l are those of Theorem 3.2, with $\text{rank}[G] = n$. Consider any real $\varepsilon \geq 0$. If $\bar{k} \in K$ satisfies

$$(3.8) \quad [\|u - \bar{k}\|^2 + \|l\|^2] \leq (1 + \varepsilon)^2 [\|u - \hat{k}\|^2 + \|l\|^2]$$

and $\bar{A} = \overline{\text{vec}}^{-1}(L^{-1}\bar{k})$, then \bar{A} approximates \hat{A} of (3.2) in that

- (i) $\bar{A} \geq 0$,
- (ii) $\|\bar{A} - \hat{A}\|_F \leq (2\varepsilon + \varepsilon^2)^{1/2} \|L^{-1}\| \|F - \hat{A}G\|_F$,
- (iii) $\|F - \bar{A}G\|_F \leq (1 + \varepsilon) \|F - \hat{A}G\|_F. \quad \square$

Now consider the case when G does not have full column rank in that $\text{rank}[G] = q < n$. Then use of the next result, Theorem 3.4, enables an accurate, approximate solution \tilde{A} of P1 to be obtained easily from a matrix $F_2 \in R^{(n-q) \times m}$ and from an accurate approximation \tilde{B} to the global minimizer \hat{B} for

$$(3.9) \quad \text{P3} \quad \min_{B \in S_{\geq}^q} \|F_1 - BG_1\|_F,$$

where G_1 has full column rank, for suitable $F_1, G_1 \in R^{q \times m}$. Hence the above results for the case when G has full rank may be used to find a suitable \tilde{B} . The details are given next.

Suitable matrices F_1, F_2 and G_2 are obtained from F and G as follows. When $\text{rank}[G] = q < n$, there is an orthogonal $Q \in R^{n \times n}$ such that

$$(3.10) \quad QG = [G_1' 0']'$$

for some $G_1 \in R^{q \times m}$ of rank q . Using that Q , the matrices $F_1 \in R^{q \times m}$ and $F_2 \in R^{(n-q) \times m}$ are obtained by partitioning QF as

$$(3.11) \quad QF = [F_1' F_2']'.$$

The results which facilitate accurate approximate solution of P1 when G does not have full column rank are given next.

THEOREM 3.4. Suppose $\text{rank}[G] = q < n$; Q , F_1 , F_2 and G_1 are as specified by (3.10) and (3.11); \hat{A} solves P1 of (3.1) and \hat{B} solves P3 of (3.9). Then

$$(3.12) \quad (i) \quad \|F - \hat{A}G\|_F^2 = \|F_1 - \hat{B}G_1\|_F^2 + \|F_2(I - G_1^\dagger G_1)\|_F^2,$$

$$(3.13) \quad (ii) \quad \tilde{A} = Q' \begin{pmatrix} \tilde{B} & \tilde{C} \\ \tilde{C}' & \tilde{D} \end{pmatrix} Q$$

is a positive semidefinite approximate solution for P1 which satisfies

$$(3.14) \quad \|F - \tilde{A}G\|_F^2 \leq (1 + \varepsilon_1) \|F - \hat{A}G\|_F^2 + \varepsilon_2$$

for any prespecified $\varepsilon_1, \varepsilon_2 \in \mathbb{R}_\equiv$ if

(a) $\tilde{B} \in S_\equiv^q$ approximates the solution \hat{B} of P3 in that

$$(3.15) \quad \tilde{B} > 0 \quad \text{and} \quad \|F_1 - \tilde{B}G_1\|_F^2 \leq [(1 + \varepsilon_1) \|F_1 - \hat{B}G_1\|_F^2 + \varepsilon_2].$$

(b) \tilde{C} and \tilde{D} satisfy

$$(3.16) \quad \tilde{C}' = F_2 G_1^\dagger,$$

$$(3.17) \quad \tilde{D} \in S_\equiv^{n-q} \quad \text{with} \quad \tilde{D} \geq \tilde{C}' \tilde{B}^{-1} \tilde{C}.$$

Further, \tilde{A} of (3.13) is positive definite if, in addition,

$$(3.18) \quad \tilde{D} > \tilde{C}' \tilde{B}^{-1} \tilde{C}. \quad \square$$

An important point in connection with Theorem 3.4 is that the \hat{B} which solves P3 of (3.9) is not necessarily positive definite even though it is positive semidefinite. Hence even if \hat{B} were available, it might be necessary to perturb \hat{B} slightly to yield a positive definite \tilde{B} suitable for use with Theorem 3.4. The next result, Theorem 3.5, concerns the determination of such a positive definite \tilde{B} and the consequences of its use when finding an approximation \tilde{A} to \hat{A} for P1 of (3.1) using Theorem 3.4. In Theorem 3.5, $[B]_{(\alpha)}$ and N_B are as defined in § 1.

THEOREM 3.5. Let \tilde{L} , \tilde{u} and \tilde{l} be the L , u and l of Theorem 3.2 when applied to P3 of (3.9) instead of to P1 of (3.1). Consider the solution \hat{B} of P3, the matrices \tilde{A} , \tilde{B} , \tilde{C} and \tilde{D} of Theorem 3.4 and any solution \hat{A} of P1. There are three cases, depending on whether $\overline{\text{vec}}^{-1}(\tilde{L}^{-1}\tilde{u})$ is positive definite, positive semidefinite or not positive semidefinite.

(i) Suppose $\overline{\text{vec}}^{-1}(\tilde{L}^{-1}\tilde{u}) > 0$. Then $\hat{B} = \overline{\text{vec}}^{-1}(\tilde{L}^{-1}\tilde{u})$ and $\hat{B} > 0$ and if \tilde{B} is chosen to be \hat{B} , then \tilde{A} solves P1 (for any $\tilde{D} \in S_\equiv^{n-q}$ with $\tilde{D} \geq \tilde{C}' \tilde{B}^{-1} \tilde{C}$) and

$$(3.19) \quad \|F_1 - \hat{B}G_1\|_F = \|\tilde{l}\|, \quad \|F - \hat{A}G\|_F^2 = \|\tilde{l}\|^2 + \|F_2(I - G_1^\dagger G_1)\|_F^2.$$

(ii) Suppose $\overline{\text{vec}}^{-1}(\tilde{L}^{-1}\tilde{u}) \geq 0$ with $\overline{\text{vec}}^{-1}(\tilde{L}^{-1}\tilde{u}) \not> 0$. Then $\hat{B} = \overline{\text{vec}}^{-1}(\tilde{L}^{-1}\tilde{u}) \geq 0$ with $\hat{B} \not> 0$ and (3.19) applies.

For any $\alpha \in \mathbb{R}_\equiv$, if $\tilde{B} = [B]_{(\alpha)}$ then \tilde{B} is a positive definite approximation to \hat{B} and \tilde{A} is a positive semidefinite approximate minimizer for P1 which have the following absolute error properties if $\tilde{l} = 0$:

$$(3.20) \quad \|F_1 - \tilde{B}G_1\|_F \leq \|F_1 - \hat{B}G_1\|_F + \alpha(N_B)^{1/2} \|G_1\|_F,$$

$$(3.21) \quad \|F - \tilde{A}G\|_F^2 \leq \|F - \hat{A}G\|_F^2 + \alpha^2 N_B \|G_1\|_F^2$$

and which have the following relative error properties if $\tilde{l} \neq 0$:

$$(3.22) \quad \|F_1 - \tilde{B}G_1\|_F \leq (1 + \alpha[(N_B)^{1/2} \|G_1\|_F / \|\tilde{l}\|]) \|F_1 - \hat{B}G_1\|_F,$$

$$(3.23) \quad \|F - \tilde{A}G\|_F \leq (1 + \alpha[(N_B)^{1/2} \|G_1\|_F / \|\tilde{l}\|]) \|F - \hat{A}G\|_F.$$

(iii) Suppose $\overline{\text{vec}}^{-1}(\tilde{L}^{-1}\tilde{u}) \not\geq 0$. Then $\|F_1 - \hat{B}G_1\|_F > 0$. Further, for $\varepsilon \in \mathbb{R}_{\geq}$, suppose an approximation $\bar{B} \in S_{\geq}^q$ to \hat{B} has been found which satisfies

$$(3.24) \quad \|F_1 - \bar{B}G_1\|_F \leq (1 + \varepsilon)\|F_1 - \hat{B}G_1\|_F.$$

For $\alpha \in \mathbb{R}_{\geq}$, if \tilde{B} is chosen as $[\bar{B}]_{(\alpha)}$ then \tilde{B} is a positive definite approximation to \hat{B} , and \tilde{A} is a positive semidefinite approximate solution for P1, with the following relative error properties:

$$(3.25) \quad \|F_1 - \tilde{B}G_1\|_F \leq (1 + \varepsilon)(1 + \alpha[(N_B)^{1/2}\|G_1\|_F/\|F_1 - \bar{B}G_1\|_F])\|F_1 - \hat{B}G_1\|_F,$$

$$(3.26) \quad \|F - \tilde{A}G\|_F \leq (1 + \varepsilon)(1 + \alpha[(N_B)^{1/2}\|G_1\|_F/\|F_1 - \bar{B}G_1\|_F])\|F - \hat{A}G\|_F. \quad \square$$

Remark 3.3. For case (i), Theorem 3.5 reveals that an \tilde{A} which solves P1 of (3.1) exactly can be found explicitly.

Case (ii) of Theorem 3.5 reveals that an \tilde{A} giving an arbitrarily good approximation to the minimal value of P1 can be obtained easily by choosing α to be small enough.

Consider case (iii). The corresponding version of P2 of (3.3) is

$$\text{P2}' \quad \min_{k \in \text{cone}[\tilde{L}\tilde{W}'\tilde{\Omega}]} \|\tilde{u} - k\|$$

and the corresponding \hat{B} which minimizes $\|F_1 - BG_1\|_F$ is, from (3.2), $\hat{B} = \overline{\text{vec}}^{-1}(\tilde{L}^{-1}\hat{k})$ where \hat{k} solves P2'. Here \tilde{L} , \tilde{W} and $\tilde{\Omega}$ are L , W and Ω of Theorem 3.2, (2.1) and (2.5) for n replaced by q . Since $\overline{\text{vec}}^{-1}(\tilde{L}^{-1}\tilde{u}) \not\geq 0$ in case (iii), it follows from (3.6) that $\hat{k} \neq \tilde{u}$. Hence, by Theorem 4.2 (later), Algorithm 4.1 can be applied to obtain an accurate approximate solution \bar{k} for P2' which satisfies the corresponding version of (3.8), namely

$$[\|\tilde{u} - \bar{k}\|^2 + \|\tilde{l}\|^2] \leq (1 + \varepsilon)^2[\|\tilde{u} - \hat{k}\|^2 + \|\tilde{l}\|^2].$$

Then the corresponding approximation \bar{B} to \hat{B} is $\bar{B} = \overline{\text{vec}}^{-1}(\tilde{L}^{-1}\bar{k})$ and, by Theorem 3.3(iii),

$$\|F_1 - \bar{B}G_1\|_F \leq (1 + \varepsilon)\|F_1 - \hat{B}G_1\|_F$$

so that \bar{B} satisfies (3.24). It follows from (3.26) that use of $\tilde{B} = [\bar{B}]_{(\alpha)}$ yields an \tilde{A} which solves P1 arbitrarily accurately if α and ε are chosen to be small enough.

Hence, an \tilde{A} giving an arbitrarily accurate solution for P1 can be obtained, whatever the rank of G . \square

Proof of Theorem 3.1. By considering $\text{vec}(F - AG)$ and using Lemma 2.1(iii) and the fact [9, p. 42] that $\text{vec}(XYZ) = [X \otimes Z'] \text{vec}(Y)$, we find that

$$(3.27) \quad \|F - AG\|_F = \|f - Ha\|,$$

where $a = \text{vec}(A) \in \mathbb{R}^{n^2}$, $f = \text{vec}(F) \in \mathbb{R}^{mn}$ and $H = [I_{n \times n} \otimes G'] \in \mathbb{R}^{mn \times n^2}$.

In view of (3.27) and Theorem 2.1(i), problem P1 is equivalent to

$$\text{P4} \quad \min_{a \in K_1} \|f - Ha\|$$

where $K_1 = \text{cone}[\Omega]$, which is equivalent to

$$\text{P5} \quad \min_{b \in K_2} \|f - b\|$$

with $K_2 = HK_1$.

If $\text{rank}[G] = n$, then $N[G'] = \{0\}$ so $N[H] = \{0\}$. The recession cone 0^+K_1 of the convex cone K_1 , i.e., of cone $[\Omega]$, is K_1 itself [6, Cor. 8.3.2]. Therefore $N[H] \cap 0^+K_1 = \{0\}$. Consequently the nonempty convex cone $K_2 = HK_1$ is closed [6, Thm. 9.1], since Theorem 2.1(iii) reveals that K_1 is closed. Therefore the minimum in P5 exists [7, Lemma 3.3.1] and, since P5 and P1 are equivalent, the minimum in P1 exists. This proves Theorem 3.1 when $\text{rank}[G] = n$.

When $\text{rank}[G] < n$, the null space of H is nontrivial and consequently the fact that K_1 is closed does not reveal immediately that K_2 is closed [6, p. 73]. Therefore the existence of the minimum in P5, and hence of the minimum in P1, cannot be established so easily as for the case when $\text{rank}[G] = n$. The structure of P1 has to be considered in more detail when $\text{rank}[G] < n$, as follows.

There exists an orthogonal Q so that (3.10) and (3.11) are valid, where $q = \text{rank}[G] = \text{rank}[G_1]$. Any $A \in S_{\geq}^n$ can be written as $A = Q'M_{BCD}Q$, where

$$(3.28) \quad M_{BCD} = \begin{bmatrix} B & C \\ C' & D \end{bmatrix}$$

for $B \in R^{q \times q}$, $C \in R^{q \times (n-q)}$, $D \in R^{(n-q) \times (n-q)}$. By [9, p. 197] and the orthogonality of Q

$$(3.29) \quad S_{\geq}^n = \{M_{BCD} : (B, C, D) \in Y_{\geq}\} = \{Q'M_{BCD}Q : M_{BCD} \in S_{\geq}^n\}$$

where

$$(3.30) \quad Y_{\geq} = \{(B, C, D) : B \in S_{\geq}^q, R[C] \subset R[B], D' = D \geq C'B^+C\}.$$

Let

$$(3.31) \quad S_{BC0} = \left\{ \begin{bmatrix} B & C \\ C' & 0 \end{bmatrix} : \begin{bmatrix} B & C \\ C' & D \end{bmatrix} \in S_{\geq}^n \right\}.$$

Then

$$\begin{aligned} \min_{A \in S_{\geq}} \|F - AG\|_F^2 &= \min_{A \in S_{\geq}} \|Q[F - AG]\|_F^2 = \min_{M_{BCD} \in S_{\geq}} \|Q[F - Q'M_{BCD}QG]\|_F^2 \\ &= \min_{(B, C, D) \in Y_{\geq}} \left\| \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} - \begin{bmatrix} B & C \\ C' & D \end{bmatrix} \begin{bmatrix} G_1 \\ 0 \end{bmatrix} \right\|_F^2 \\ &= \min_{(B, C, D) \in Y_{\geq}} \left\| \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} - \begin{bmatrix} B & C \\ C' & 0 \end{bmatrix} \begin{bmatrix} G_1 \\ 0 \end{bmatrix} \right\|_F^2 \\ &= \min_{A \in S_{BC0}} \|\bar{F} - A\bar{G}\|_F^2 \end{aligned}$$

where $\bar{F} = [F_1' F_2']'$, $\bar{G} = [\bar{G}_1' 0']'$. The first equality is a consequence of the orthogonality of Q , and the second and third are from (3.29).

Hence P1 is equivalent to

$$\text{P6} \quad \min_{a \in K_3} \|\bar{f} - a\|_F^2$$

where $\bar{f} = \text{vec}(\bar{F})$ and $K_3 = \bar{H} \text{vec}(S_{BC0})$ for $\bar{H} = I_{n \times n} \otimes \bar{G}'$.

From Theorem 2.1(i)-(iii), $\text{vec}(S_{\geq}^n)$ is a nonempty, closed, convex cone. Therefore it follows from (3.31) that $\text{vec}(S_{BC0})$ is a nonempty, closed, convex cone. A consequence is that its recession cone, $0^+ \text{vec}(S_{BC0})$, is equal to $\text{vec}(S_{BC0})$ [6, Cor. 8.3.2]. Hence K_3 is closed if $\|\bar{H}x\| > 0$ whenever $0 \neq x \in \text{vec}(S_{BC0})$ [6, Thm. 9.1], i.e. (since $\text{vec}(A\bar{G}) = \bar{H} \text{vec}(A)$) if $\|A\bar{G}\|_F > 0$ whenever $0 \neq A \in S_{BC0}$. From (3.31), every nonzero A in S_{BC0} has the form of M_{BC0} of (3.28) with B and C not both zero. Further, in view of (3.29) and (3.30), $R[C] \subseteq R[B]$ so that C is zero if B is zero. Hence $A = M_{BC0}$ with $B \neq 0$ if $0 \neq A \in S_{BC0}$. Now $R[G_1] = R^q$, because $\text{rank}[G_1] = q$ and $G_1 \in R^{q \times m}$, so $BG_1 \neq 0$ when $B \neq 0$. Since $\|A\bar{G}\|_F^2 = \|BG_1\|_F^2 + \|C'G_1\|_F^2$ when $A = M_{BC0}$, it follows that $\|A\bar{G}\|_F > 0$ whenever $0 \neq A \in S_{BC0}$. Consequently, K_3 is closed.

Since K_3 is closed, the minimum in P6 exists. Hence, since P6 and P1 are equivalent, the minimum in P1 exists when $\text{rank}[G] < q$.

Therefore the minimum in P1 exists whatever the rank of G , which completes the proof of Theorem 3.1. \square

Proof of Theorem 3.2. Now $J = [I_{n \times n} \otimes G']W$, so that $R[J'] = R[W'(I_{n \times n} \otimes G)] = R[W'] = R'$, where the penultimate equality occurs because G has full rank and the final inequality is from (2.2). Hence $\text{rank}[J] = r$.

As in (3.27), $\|F - AG\|_F = \|f - [I_{n \times n} \otimes G']a\|$, where $a = \text{vec}(A)$, so, in view of (2.3)

$$(3.32) \quad \|F - AG\|_F = \|f - [I_{n \times n} \otimes G']W\bar{a}\| = \|f - J\bar{a}\|$$

where $\bar{a} = \overline{\text{vec}}(A)$. Use in (3.32) of Lemma 2.1(iii) and of the factorizations of J and f given in the statement of Theorem 3.2 yields

$$\|F - AG\|_F^2 = \|u - L\bar{a}\|^2 + \|l\|^2.$$

Hence solution of P1, i.e., minimization of $\|F - AG\|_F$ with respect to $A \in S_{\geq}^n$, can be achieved by solving

$$\text{P7} \quad \min_{a \in \text{vec}(S_{\geq}^n)} \|u - L\bar{a}\|$$

and then $\hat{A} = \overline{\text{vec}}^{-1}(\hat{\bar{a}})$ if $\hat{\bar{a}}$ solves P7.

Since, by Theorem 2.1(i), $\overline{\text{vec}}(S_{\geq}^n) = \text{cone}[W'\Omega]$, P7 becomes P2 of Theorem 3.2, namely

$$\text{P2} \quad \min_{k \in K} \|u - k\|$$

where $K = L \overline{\text{vec}}(S_{\geq}^n) = \text{cone}[LW'\Omega]$, a convex cone, and then $\hat{\bar{a}} = L^{-1}\hat{k}$ if \hat{k} solves P2.

It is known, from Theorem 3.1, that P1 has a solution. Since P1 is equivalent to P2 (via P7), because it is clear from above that the solution of P1 can be found from that of P2, and vice versa, P2 has a solution.

Hence the solution $\hat{\bar{a}}$ of P7 is $\hat{\bar{a}} = L^{-1}\hat{k}$ and consequently the solution of P1 is $\hat{A} = \overline{\text{vec}}^{-1}(\hat{\bar{a}}) = \overline{\text{vec}}^{-1}(L^{-1}\hat{k})$.

Clearly P2 is solved by $\hat{k} = u$ if and only if $u \in K$, i.e., (since $K = \overline{\text{vec}}(S_{\geq}^n)$) if and only if $u = L \overline{\text{vec}}(A)$ for some $A \in S_{\geq}^n$, i.e., if and only if $\overline{\text{vec}}^{-1}(L^{-1}u) \geq 0$. So, since $\hat{A} = \overline{\text{vec}}^{-1}(L^{-1}\hat{k})$, $\hat{A} = \overline{\text{vec}}^{-1}(L^{-1}u)$ if and only if $\overline{\text{vec}}^{-1}(L^{-1}u) \geq 0$.

If $\hat{k} \neq u$ then clearly $\hat{k} \in \partial K = \partial \text{cone}[LW'\Omega]$, and consequently, since L is invertible and $R[L] = R'$, $L^{-1}\hat{k} \in \partial \text{cone}[W'\Omega]$. Therefore, by Theorem 2.1(ii), then $\hat{A} = \overline{\text{vec}}^{-1}(L^{-1}\hat{k})$ is not positive definite. \square

Proof of Theorem 3.3. (i) Now $\bar{k} \in K = L \text{cone}[W'\Omega]$, so, by Theorem 2.1(i), $\bar{k} \in L \overline{\text{vec}}(S_{\geq}^n)$. Hence $\overline{\text{vec}}^{-1}(L^{-1}\bar{k}) \in S_{\geq}^n$, which proves part (i).

(ii) Clearly

$$(3.33) \quad \|u - \bar{k}\|^2 = \|u - \hat{k}\|^2 + 2(\hat{k} - u)'(\bar{k} - \hat{k}) + \|\bar{k} - \hat{k}\|^2.$$

Since \hat{k} minimizes $\|u - k\|^2$, with respect to k from the convex set K

$$2(\hat{k} - u)'(k - \hat{k}) \geq 0 \quad \forall k \in K.$$

Hence a consequence of (3.8) and (3.33) is that

$$\|\bar{k} - \hat{k}\|^2 \leq (2\varepsilon + \varepsilon^2)[\|u - \hat{k}\|^2 + \|l\|^2]$$

so, by (3.4)

$$\|\bar{k} - \hat{k}\| \leq (2\varepsilon + \varepsilon^2)^{1/2} \|F - \hat{A}G\|_F.$$

Therefore, by Lemma 2.1(iii)

$$\begin{aligned}\|\bar{A} - \hat{A}\|_F &= \|\overline{\text{vec}}^{-1}(L^{-1}\bar{k}) - \overline{\text{vec}}^{-1}(L^{-1}\hat{k})\|_F = \|L^{-1}\bar{k} - L^{-1}\hat{k}\| \\ &\leq (2\varepsilon + \varepsilon^2)^{1/2} \|L^{-1}\| \|F - \hat{A}G\|_F\end{aligned}$$

which establishes part (ii).

(iii) As in (3.32), and using the factorizations of Theorem 3.2

$$\|F - \bar{A}G\|_F^2 = \|f - J \overline{\text{vec}}(\bar{A})\|^2 = \|u - \bar{k}\|^2 + \|l\|^2.$$

So, in view of (3.4), the result of part (iii) is implied by (3.8). \square

Proof of Theorem 3.4. Much as in the proof of Theorem 3.1, let

$$(3.34) \quad \begin{aligned}Y_B &= \{C \in R^{q \times (n-q)}: R[C] \subset R[B]\} \quad \text{for } B \in S_{\geq}^q, \\ Y_{\geq} &= \{(B, C, D): B \in S_{\geq}^q, C \in Y_B, D \in S_{\geq}^{n-q}, D \geq C'B^+C\}_1\end{aligned}$$

$$(3.35) \quad A_{BCD} = Q \begin{bmatrix} B & C \\ C' & D \end{bmatrix} Q \in R^{n \times n}$$

where $B \in S^q$, $C \in R^{q \times (n-q)}$, $D \in S^{n-q}$ and where $Q \in R^{n \times n}$ is the orthogonal matrix specified in Theorem 3.4.

From p. 197 of [9]

$$(3.36) \quad [A_{BCD} \geq 0] \Leftrightarrow [(B, C, D) \in Y_{\geq}].$$

Further, since Q is orthogonal

$$(3.37) \quad \|F - A_{BCD}G\|_F^2 = \|Q[F - A_{BCD}G]\|_F^2 = \|F_1 - BG_1\|_F^2 + \|F_2 - C'G_1\|_F^2.$$

If we write \hat{A} as $A_{B^*C^*D^*}$ (3.37) gives

$$(3.38) \quad \begin{aligned}\|F - \hat{A}G\|_F^2 &= \|F_1 - B^*G_1\|_F^2 + \|F_2 - C^*G_1\|_F^2 \\ &\geq \|F_1 - \hat{B}G_1\|_F^2 + \|F_1(I - G_1^+G_1)\|_F^2\end{aligned}$$

since \hat{B} minimizes $\|F_1 - BG_1\|_F$ on S_{\geq}^n and since, from [9, p. 119], $F_2G_1^+$ minimizes $\|F_2 - C'G_1\|_F$ with respect to $C' \in R^{(n-q) \times q}$.

Consider now $A_{\bar{B}\bar{C}\bar{D}}$ when \bar{B} is $[\hat{B}]_{(\alpha)}$ (of § 1), $\bar{C}' = F_2G_1^+$, $\bar{D} \in S^q$ with $\bar{D} \geq \bar{C}'\bar{B}^+\bar{C}$. Since $[\hat{B}]_{(\alpha)} > 0$, for all $\alpha > 0$, it follows from (3.36) that $A_{\bar{B}\bar{C}\bar{D}} \in S_{\geq}^n$, for all $\alpha > 0$. So, since \hat{A} minimizes $\|F - AG\|_F$ on S_{\geq}^n

$$(3.39) \quad \|F - \hat{A}G\|_F^2 \leq \|F - A_{\bar{B}\bar{C}\bar{D}}G\|_F^2 = \|F_1 - [\hat{B}]_{(\alpha)}G_1\|_F^2 + \|F_2(I - G_1^+G_1)\|_F^2$$

where the equality is from (3.37). Hence, from (3.38), (3.39) and since $\|F - [\hat{B}]_{(\alpha)}G_1\|_F \leq \|F_1 - \hat{B}G_1\|_F + \|\hat{B} - [\hat{B}]_{(\alpha)}\|_F \|G_1\|_F$

$$(3.40) \quad \begin{aligned}\|F_1 - \hat{B}G_1\|_F^2 + \|F_2(I - G_1^+G_1)\|_F^2 &\leq \|F - \hat{A}G\|_F^2 \\ &\leq [\|F_1 - \hat{B}G_1\|_F + \alpha(N_{\hat{B}})^{1/2}\|G_1\|_F]^2 \\ &\quad + \|F_2(I - G_1^+G_1)\|_F^2, \quad \forall \alpha > 0.\end{aligned}$$

Consideration of (3.40) as α converges downwards to zero reveals that

$$\|F - \hat{A}G\|_F^2 = \|F_1 - \hat{B}G_1\|_F^2 + \|F_2(I - G_1^+G_1)\|_F^2$$

which proves part (i).

Now consider $\tilde{A} = A_{\tilde{B}\tilde{C}\tilde{D}}$ of (3.13) for the \tilde{B} , \tilde{C} and \tilde{D} of Theorem 3.4. From (3.36), $\tilde{A} \geq 0$ and from (3.37), since \tilde{C} of (3.16) minimizes $\|F_2 - C'G_1\|_F$ with respect to $C \in R^{q \times (n-q)}$

$$(3.41) \quad \|F - \tilde{A}G\|_F^2 = \|F_1 - \tilde{B}G_1\|_F^2 + \min_{C \in R^{q \times (n-q)}} \|F_2 - C'G_1\|_F^2.$$

So, from (3.41) and (3.15)

$$\begin{aligned}
 \|F - \tilde{A}G\|_F^2 &\leq (1 + \varepsilon_1) \min_{B \in S_{\Xi}^q} \|F_1 - BG_1\|_F^2 + \varepsilon_2 + \min_{C \in R^{q \times (n-q)}} \|F_2 - C'G_1\|_F^2 \\
 (3.42) \quad &\leq (1 + \varepsilon_1) \min_{B \in S_{\Xi}^q} [\|F_1 - BG_1\|_F^2 + \min_{C \in R^{q \times (n-q)}} \|F_2 - C'G_1\|_F^2] + \varepsilon_2.
 \end{aligned}$$

Now $Y_B \subset R^{q \times (n-q)}$ for all $B \in S_{\Xi}^q$. Therefore

$$\min_{C \in R^{q \times (n-q)}} \|F_2 - C'G_1\|_F^2 \leq \min_{C \in Y_B} \|F_2 - C'G_1\|_F^2 \quad \forall B \in S_{\Xi}^q.$$

Consequently, from (3.42)

$$\begin{aligned}
 \|F - \tilde{A}G\|_F^2 &\leq (1 + \varepsilon_1) \min_{B \in S_{\Xi}^q} [\|F_1 - BG_1\|_F^2 + \min_{C \in Y_B} \|F_2 - C'G_1\|_F^2] + \varepsilon_2 \\
 &= (1 + \varepsilon_1) \min_{(B, C, D) \in Y_{\Xi}} \{\|F_1 - BG_1\|_F^2 + \|F_2 - C'G_1\|_F^2\} + \varepsilon_2 \\
 &= (1 + \varepsilon_1) \min_{(B, C, D) \in Y_{\Xi}} \|F - A_{BCD}G\|_F^2 + \varepsilon_2 \\
 &= (1 + \varepsilon_1) \min_{A \in S_{\Xi}^n} \|F - AG\|_F^2 + \varepsilon_2
 \end{aligned}$$

where the first equality is from (3.34), the second is from (3.37) and the third is a consequence of the fact that, from (3.35), (3.36) and the orthogonality of Q

$$S_{\Xi}^n = \{A_{BCD} : (B, C, D) \in Y_{\Xi}\}.$$

This establishes (3.14).

From [9, p. 197], A_{BCD} of (3.35) is positive definite if and only if $B > 0$, $B - CD^{\dagger}C' > 0$ and $D - C'B^{-1}C > 0$. In fact, the first and last of these three conditions imply positive definiteness; the following proof is from [4]. The first and third conditions imply that A_{BCD} is positive semidefinite [9, p. 196]. Hence A_{BCD} is positive definite if it is invertible. Now the first and third conditions ensure that $|B| > 0$ and that $|D - C'B^{-1}C| > 0$. Hence A_{BCD} is invertible because $|A_{BCD}| = \|B\| |D - C'B^{-1}C| > 0$.

This completes the proof of Theorem 3.4. \square

Proof of Theorem 3.5. If $\text{vec}^{-1}(\tilde{L}^{-1}\tilde{u}) \geq 0$, it follows from (3.2) and (3.6) of Theorem 3.2 that the \tilde{B} solving P3 is $\hat{B} = \text{vec}^{-1}(\tilde{L}^{-1}\tilde{u})$. The formulae for $\|F_1 - \hat{B}G_1\|_F$ and $\|F - \hat{A}G\|_F$ of (3.19) are from (3.4) of Theorem 3.2 (since $\hat{k} = u$) and from Theorem 3.4(i).

If $\tilde{B} > 0$, then the choices $\tilde{B} = \hat{B}$ and $\varepsilon_1 = \varepsilon_2 = 0$ can be used in Theorem 3.4(ii) to show that \tilde{A} of (3.13) solves P1 exactly, which proves (i).

Now suppose $\text{vec}^{-1}(\tilde{L}^{-1}\tilde{u}) \geq 0$ with $\text{vec}^{-1}(\tilde{L}^{-1}\tilde{u}) \not\geq 0$. Then $\hat{B} \not\geq 0$ and the choice $\tilde{B} = [\tilde{B}]_{(\alpha)}$ for $\alpha > 0$ gives $\tilde{B} > 0$ and $\|\tilde{B} - \hat{B}\|_F = \alpha(N_B)^{1/2}$. Hence we have (3.20).

From (3.19), if $\tilde{I} = 0$ then $\|F_1 - \hat{B}G_1\|_F = 0$ in (3.20). Consequently (3.15) of Theorem 3.4 applies with $\varepsilon_1 = 0$ and $\varepsilon_2 = \alpha^2 N_B \|G_1\|_F^2$. Then (3.14) of Theorem 3.4 yields (3.21).

However if $\tilde{I} \neq 0$ then

$$\begin{aligned}
 \|F_1 - \tilde{B}G_1\|_F &\leq \|F_1 - \hat{B}G_1\|_F + \alpha(N_B)^{1/2} \|G_1\|_F \\
 &= \|F_1 - \hat{B}G_1\|_F + [\alpha(N_B)^{1/2} \|G_1\|_F / \|\tilde{I}\|] \|F_1 - \hat{B}G_1\|_F
 \end{aligned}$$

where the equality is from (3.19). Hence we have (3.22).

In view of (3.22), Theorem 3.4(ii) can be applied with ε_1 defined by $1 + \varepsilon_1 = (1 + \alpha(N_B)^{1/2} \|G_1\|_F / \|\tilde{I}\|)^2$ and with $\varepsilon_2 = 0$, in (3.15). Then (3.23) follows from (3.14).

Since it is assumed in Theorem 3.5(iii) that $\overline{\text{vec}}^{-1}(\tilde{L}^{-1}\tilde{u}) \not\equiv 0$, it follows from (3.6) that $\hat{k} \neq u$ so that, from (3.4), $\|F_1 - \hat{B}G_1\| > \|\hat{I}\|$. Hence $\|F_1 - \hat{B}G_1\| > 0$, as claimed. Therefore, since $\tilde{B} = [\tilde{B}]_{(\alpha)}$ in this case

$$\begin{aligned}\|F_1 - \tilde{B}G_1\|_F &\leq \|F_1 - \bar{B}G_1\|_F + \alpha(N_B)^{1/2}\|G_1\|_F \\ &\leq [(1 + \varepsilon) + \alpha(N_B)^{1/2}\|G_1\|_F/\|F_1 - \hat{B}G_1\|_F]\|F_1 - \hat{B}G_1\|_F \\ &\leq (1 + \varepsilon)[1 + \alpha(N_B)^{1/2}\|G_1\|_F/\|F_1 - \bar{B}G_1\|_F]\|F_1 - \hat{B}G_1\|_F\end{aligned}$$

where the second and third inequalities are from (3.24). Hence we have (3.25).

Therefore \tilde{B} satisfies (3.15) with ε_1 defined by $(1 + \varepsilon_1) = (1 + \varepsilon)^2[1 + \alpha(N_B)^{1/2}\|G_1\|_F/\|F_1 - \bar{B}G_1\|_F]^2$ and with $\varepsilon_2 = 0$. Then (3.26) follows from (3.14). \square

4. A proximal point algorithm for conical hulls. The problem considered here is P2 of (3.3), rewritten here as

$$(4.1) \quad \text{P2} \quad \min_{k \in \text{cone}[\Gamma]} v(k)$$

where

$$(4.2) \quad v(k) = \|u - k\|^2 \quad \forall k \in R^r, \quad \Gamma = LW'\Omega, \quad K = \text{cone}[\Gamma]$$

and where K , u and the invertible matrix L are those of Theorem 3.2, W is specified by (2.1) and Ω by (2.5). Algorithm 4.1, presented shortly, finds an approximation \bar{k} to the solution \hat{k} of P2 which is acceptable in that condition (3.8) of Theorem 3.3 is satisfied. The algorithm relies on the following properties of Γ :

(4.3) Γ is a nonempty closed convex set;

(4.4) There exists a $\pi \in R_>$ such that for any point x in Γ , there is a point $y \in \Gamma$ which is in the ray through 0 and x and for which $\|y\| \equiv \pi$;

(4.5) For each $g \in R^r$, a member of $\arg \min \{g'\gamma: \gamma \in \Gamma\}$ can be found.

Remark 4.1. From Theorem 2.1(iv), a suitable π is $\pi = \sigma_{\min}[L]n^{-1/2}$. From Theorem 2.1(v), the minimal value of $g'\gamma$ and a minimizing γ can actually be determined by solving an eigenvalue problem. \square

Algorithm 4.1 for approximating the minimizer \hat{k} of v on the unbounded set $\text{cone}[\Gamma]$ is based on the fact that \hat{k} is also the minimizer of v on a bounded set S , actually a suitably truncated version of $\text{cone}[\Gamma]$. Consequently \hat{k} may be approximated by approximately minimizing v on S . The result regarding S is stated next.

THEOREM 4.1. Let $\eta = \|u\|/\pi$; $S = \{\alpha\gamma: \alpha \in [0, \eta], \gamma \in \Gamma\}$. Then S is a convex set and $\text{minpoint}[u, S] = \text{minpoint}[u, \text{cone}[\Gamma]]$. \square

The algorithm for finding an acceptably accurate approximation \bar{k} to \hat{k} by minimizing v on S is stated next.

ALGORITHM 4.1.

0. *Select parameters by choosing*

$\varepsilon \in (0, \infty)$ (ε occurs in post-condition (4.12) and defines the degree of suboptimality acceptable in \bar{k}),

$\bar{k}_0 \in \text{cone}[\Gamma]$ (an initial estimate of \hat{k}).

I. *Initialize variables*

$k_0 := \text{minpoint}[u, \text{cone}[\{\bar{k}_0\}]]$ (the point nearest u in the ray through 0 and \bar{k}_0),

$\hat{b}_{-1} := 0$ (the best lower bound for $v(\hat{k})$ available so far),

$i := 0$ (the iteration index).

II. *Decide when to stop iterating*

Find a $y_i \in \arg \min_{y \in S} \nabla v(k_i)'(y_i - k_i)$,

by finding a

$$(4.6) \quad \gamma_i \in \arg \min_{\gamma \in \Gamma} \nabla v(k_i)' \gamma$$

and setting

$$(4.7) \quad y_i = \begin{cases} \eta \gamma_i & \text{if } \nabla v(k_i)' \gamma_i < 0, \\ k_i & \text{otherwise.} \end{cases}$$

Compute a lower bound b_i for $v(\hat{k})$

$$(4.8) \quad b_i := v(k_i) + \nabla v(k_i)'(y_i - k_i).$$

Compute \hat{b}_i , the best lower bound for $v(k)$ found so far

$$(4.9) \quad \hat{b}_i := \max \{\hat{b}_{i-1}, b_i\}.$$

If

$$(4.10) \quad [v(k_i) + \|l\|^2] \leq (1 + \varepsilon)^2 [\hat{b}_i + \|l\|^2] \quad (\text{where } l \text{ is that used in Theorem 3.2}), \text{ then set } \bar{k} = k_i \text{ and stop; else continue.}$$

III. *Choose the next iterand*

$$(4.11) \quad \begin{aligned} k_{i+1} &:= \text{minpoint}(u, \text{cone}[\text{line}\{k_i, y_i\}]), \\ i &:= i + 1. \end{aligned}$$

Go to II. \square

Remark 4.2. Minpoint $(u, \text{cone}[\text{line}\{k_i, y_i\}])$ may be found simply and exactly owing to the simple nature of the set $\text{cone}[\text{line}\{k_i, y_i\}]$. \square

THEOREM 4.2. *If $\hat{k} \neq u$, Algorithm 4.1 stops after a finite number of iterations with \bar{k} an approximate solution of P2 which satisfies*

$$(4.12) \quad \bar{k} \in \text{cone}[\Gamma], \quad \|\bar{k}\| \leq \|\hat{k}\|, \quad [\|u - \bar{k}\|^2 + \|l\|^2] \leq (1 + \varepsilon)^2 [\|u - \hat{k}\|^2 + \|l\|^2].$$

Further, if $\hat{k} \neq u$ and stopping condition (4.10) is omitted from Algorithm 4.1 so that it iterates indefinitely, then

$$k_i \in S \subset \text{cone}[\Gamma] \text{ and } \hat{b}_i \leq v(\hat{k}) \quad \forall i \geq 0, \quad k_i \rightarrow \hat{k}, \quad \hat{b}_i \rightarrow v(\hat{k}). \quad \square$$

Remark 4.3. In view of (3.6), $\hat{k} = u$ if and only if $\overline{\text{vec}}^{-1}(L^{-1}u) \geq 0$. Hence Theorem 4.2 reveals that Algorithm 4.1 may be used to find an approximate solution \bar{k} for P2 of (3.3), approximate in that condition (3.8) of Theorem 3.2 is satisfied, whenever $\overline{\text{vec}}^{-1}(L^{-1}u) \geq 0$. \square

Remark 4.4. It is desirable to use, in Algorithm 4.1, an initial approximation $\bar{k}_0 \in \text{cone}[\Gamma]$ which is a good approximation to \hat{k} . The connection between \hat{k} and \hat{A} is, from (3.2), $\hat{k} = L \overline{\text{vec}}(\hat{A})$. Also, $\text{cone}[\Gamma] = L \overline{\text{vec}}(S_{\geq}^n)$, in view of Theorem 2.1(i) and the fact that $\Gamma = LW'\Omega$. Hence a suitable \bar{k}_0 is $\bar{k}_0 = L \overline{\text{vec}}(A)$ for any $A \in S_{\geq}^n$ which approximates \hat{A} . If $\overline{\text{vec}}(L^{-1}u) \geq 0$, then (by (3.6)) $\hat{A} = \overline{\text{vec}}^{-1}(L^{-1}u)$ and we could use $\bar{k}_0 = L \overline{\text{vec}}(\hat{A})$ except that Algorithm 4.1 is not actually needed in that case. Algorithm 4.1 is needed only when $\overline{\text{vec}}^{-1}(L^{-1}u) \not\geq 0$. This suggests that when $\overline{\text{vec}}^{-1}(L^{-1}u) \not\geq 0$, the choice $\bar{k}_0 = L \overline{\text{vec}}(\hat{X})$ be used where \hat{X} minimizes $\|\overline{\text{vec}}^{-1}(L^{-1}u) - X\|_F$ with respect to X from S_{\geq}^n . It is easy to check that $\hat{X} = [\overline{\text{vec}}^{-1}(L^{-1}u)]_{(0)}$, where the notation $[\cdot]_{(0)}$

has the meaning assigned in § 1. Consequently a reasonable choice to use is $\bar{k}_0 = L[\overline{\text{vec}}(L^{-1}u)]_{(0)}$. \square

Proof of Theorem 4.1. The convexity of S may be verified easily.

Suppose $\hat{k} = \text{minpoint}[u, \text{cone}[\Gamma]]$. Then, since $S \subset \text{cone}[\Gamma]$, $\hat{k} = \text{minpoint}[u, S]$ if and only if $\hat{k} \in S$. It will be shown next that $\hat{k} \in S$. Clearly $\|\hat{k}\| \leq \|u\|$. Because $\hat{k} \in \text{cone}[\Gamma]$, \hat{k} may be written as $\hat{k} = \theta\tilde{\gamma}$ for some $\theta \in \mathbb{R}_\geq$ and for some $\tilde{\gamma} \in \Gamma$ such that (owing to property (4.4) of Γ) $\|\tilde{\gamma}\| \geq \pi$. Hence $\theta \leq \|\hat{k}\|/\pi \leq \|u\|/\pi = \eta$. Consequently $\hat{k} \in S$. Therefore $\hat{k} = \text{minpoint}[u, \text{cone}[S]]$. \square

Proof of Theorem 4.2. First it will be shown that $k_i \in S$, for all $i \geq 0$.

By the definition of k_0 in part I of Algorithm 4.1, and by the definition of k_{i+1} in part III: for all $i \geq 0$, k_i minimizes v along the ray through k_i . Consequently

$$(4.13) \quad \|k_i\| \leq \|u\| \quad \forall i \geq 0.$$

Now, by its definition in Algorithm 4.1, $y_i \in S \subset \text{cone}[\Gamma]$ for all $i \geq 0$. Therefore, from (4.11), if $k_i \in \text{cone}[\Gamma]$ then $k_{i+1} \in \text{cone}[\Gamma]$. Since $k_0 \in \text{cone}[\Gamma]$, by the way it is constructed in Algorithm 4.1:

$$k_i \in \text{cone}[\Gamma] \quad \forall i \geq 0.$$

Since $k_i \in \text{cone}[\Gamma]$, it follows that $k_i = \alpha\gamma$ for some $\alpha \in \mathbb{R}_\geq$ and for some $\gamma \in \Gamma$. Therefore, from (4.4), $\|k_i\| = \alpha\|\gamma\| \geq \alpha\pi$. Consequently, in view of (4.13), $\alpha \leq \|u\|/\pi = \eta$. So, owing to the definition of S in Theorem 4.1, $k_i \in S$, for all $i \geq 0$, as claimed in Theorem 4.2.

It will be shown next that the assertions made in the statement of Algorithm 4.1 regarding y_i and b_i are correct.

Consider the determination of y_i in step II.

Since each γ_i constructed by the algorithm is from Γ , and since each k_i is from S , it follows from (4.7) that $y_i \in S$, for all $i \geq 0$.

For $i \geq 0$, k_i minimizes v of (4.2) on the ray through k_i . So $\nabla v(k_i)'k_i = 0$, for all $i \geq 0$. Further, any $y \in S$ may be written as $y = \theta\gamma$ for some $\theta \in [0, \eta]$ and some $\gamma \in \Gamma$. Hence, for γ_i of (4.6)

$$\begin{aligned} \min_{y \in S} \nabla v(k_i)'(y - k_i) &= \min_{\theta \in [0, \eta]} \min_{\gamma \in \Gamma} \theta \nabla v(k_i)'\gamma \\ &= \begin{cases} \eta \nabla v(k_i)'\gamma_i & \text{if } \nabla v(k_i)'\gamma_i < 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The above minimal value for $\nabla v(k_i)'(y - k_i)$ is achieved by y_i of (4.7), which, since $y_i \in S$, proves that y_i minimizes $\nabla v(k_i)'(y - k_i)$ on S , as asserted in Algorithm 4.1.

Next, it will be shown that for b_i of (4.8), $b_i \leq v(\hat{k})$. Now

$$v(k) = v(k_i) + \nabla v(k_i)'(k - k_i) + \|k - k_i\|^2 \geq v(k_i) + \nabla v(k_i)'(k - k_i),$$

so, in view of Theorem 4.1,

$$(4.14) \quad \begin{aligned} v(\hat{k}) &= \min_{k \in S} v(k) \geq \min_{k \in S} \{v(k_i) + \nabla v(k_i)'(k - k_i)\} \\ &= v(k_i) + \nabla v(k_i)'(y_i - k_i) = b_i, \end{aligned}$$

which proves the assertion that the computed value of b_i is a lower bound for $v(\hat{k})$. A simple consequence is that \hat{b}_i of (4.9) is also a lower bound for $v(\hat{k})$, as claimed in Theorem 4.2.

The behaviour of Algorithm 4.1 will now be studied assuming that stopping condition (4.10) is omitted from the algorithm so that it iterates indefinitely.

Now, since Γ is bounded and η is finite, S is bounded and there is a finite number d so that $\|k - y\| \leq d, \forall k, y \in S$. Let

$$(4.15) \quad D = \|\nabla v(\hat{k})\|d + d^2.$$

Then

$$(4.16) \quad v(k) = v(\hat{k}) + \nabla v(\hat{k})'(k - \hat{k}) + \|k - \hat{k}\|^2 \leq v(\hat{k}) + D \quad \forall k \in S.$$

Consider the determination of k_{i+1} from k_i , using (4.11). Since $\text{line}\{k_i, y_i\} \subset \text{cone}[\text{line}\{k_i, y_i\}]$,

$$(4.17) \quad v(k_{i+1}) = \min_{k \in \text{cone}[\text{line}\{k_i, y_i\}]} v(k) \leq \min_{\alpha \in [0,1]} v(k_i + \alpha[y_i - k_i]).$$

From (4.14)

$$(4.18) \quad \nabla v(k_i)'(y_i - k_i) \leq [v(\hat{k}) - v(k_i)] \leq 0.$$

Suppose $y_i \neq k_i$. Let α_i denote the value of α which minimizes $v(k_i + \alpha[y_i - k_i])$ with respect to α from $[0, 1]$, and let $\tilde{\alpha}_i$ be the unconstrained global minimizer, so that

$$(4.19) \quad \tilde{\alpha}_i = -\nabla v(k_i)'(y_i - k_i) / [2\|y_i - k_i\|^2].$$

Then, from (4.18), $\tilde{\alpha}_i \geq 0$. The following two cases are possible:

- (i) $\tilde{\alpha}_i \in [0, 1]$;
- (ii) $\tilde{\alpha}_i > 1$.

For case (i), $\alpha_i = \tilde{\alpha}_i$ and consequently, from (4.17) and (4.18),

$$(4.20) \quad \begin{aligned} v(k_{i+1}) &\leq \min_{\alpha \in [0,1]} v(k_i + \alpha[y_i - k_i]) \\ &= v(k_i) - [\nabla v(k_i)'(y_i - k_i)]^2 / [4\|y_i - k_i\|^2] \\ &\leq v(k_i) - [v(k_i) - v(\hat{k})]^2 / [4D] \end{aligned}$$

where the last inequality is from (4.15) and the fact that $k_i, y_i \in S$, for all $i \geq 0$.

Now consider case (ii), for which $\alpha_i = 1$. From (4.19), the fact that $\tilde{\alpha}_i > 1$ implies that $\nabla v(k_i)'(y_i - k_i) < -2\|y_i - k_i\|^2$, and consequently

$$(4.21) \quad \begin{aligned} v(k_{i+1}) &\leq v(k_i + 1[y_i - k_i]) = v(k_i) + \nabla v(k_i)'(y_i - k_i) + \|y_i - k_i\|^2 \\ &< v(k_i) + [\nabla v(k_i)'(y_i - k_i)]/2 \\ &\leq v(k_i) - [v(k_i) - v(\hat{k})]/2 \\ &\leq v(k_i) - [v(k_i) - v(\hat{k})]^2 / [4D] \end{aligned}$$

where the penultimate inequality is from (4.18) and the last inequality is from (4.16) and the fact that $k_i \in S$, for all $i \geq 0$.

Hence, from (4.20)-(4.21), whether case (i) or case (ii) occurs, if $y_i \neq k_i$ then

$$[v(k_{i+1}) - v(\hat{k})] \leq [v(k_i) - v(\hat{k})] - [v(k_i) - v(\hat{k})]^2 / [4D]$$

and, from (4.18), if $y_i = k_i$ then

$$v(k_i) = v(\hat{k}).$$

Consequently $v(k_i) \rightarrow v(\hat{k})$ and, since v is strictly convex, $k_i \rightarrow \hat{k}$, as claimed in Theorem 4.2.

Also, $b_i \rightarrow v(\hat{k})$ since, because $k_i \rightarrow \hat{k}$, it follows [10, Thm. B.3.20] that

$$b_i = \min_{y \in S} \{v(k_i) + \nabla v(k_i)'(y - k_i)\} \rightarrow \min_{y \in S} \{v(\hat{k}) + \nabla v(\hat{k})'(y - \hat{k})\} = v(\hat{k}).$$

Therefore $\hat{b}_i \rightarrow v(\hat{k})$, as claimed.

It will be shown next that execution of the algorithm actually terminates after a finite number of iterations if stopping condition (4.10) is included in the algorithm.

Now $v(\hat{k}) > 0$ since the case $\hat{k} \neq u$ is being considered. Hence since $v(k_i)$, $\hat{b}_i \rightarrow v(\hat{k})$ and $v(\hat{k}) > 0$, stopping condition (4.10) will be satisfied eventually and iteration will cease after some finite number of iterations.

The proof of Theorem 4.2 will be concluded by showing that satisfaction of the stopping condition guarantees satisfaction of post-condition (4.12) of Theorem 4.2.

Now $k_i \in S \subseteq \text{cone}[\Gamma]$, for all $i \geq 0$, so it follows from the way \bar{k} is defined in part II of Algorithm 4.1 that $\bar{k} \in \text{cone}[\Gamma]$, which establishes the first part of (4.12).

Since k_i and \hat{k} are the closest points to u in the rays through k_i and \hat{k} , respectively: $\|k_i\|^2 = u'k_i$; $\|\hat{k}\|^2 = u'\hat{k}$. So

$$v(k_i) = \|u\|^2 - \|k_i\|^2, \quad v(\hat{k}) = \|u\|^2 - \|\hat{k}\|^2$$

and consequently, since $v(k_i) \geq v(\hat{k})$: $\|k_i\| \leq \|\hat{k}\|$ for all $i \geq 0$. Further, since $\hat{b} \leq v(\hat{k})$, satisfaction of (4.10) ensures that (4.12) is satisfied when \bar{k} is the last k_i which is generated by Algorithm 4.1. \square

5. Computational considerations and an example. The interesting issue is the actual behaviour of Algorithm 4.1 for the cases when it is required. By Theorem 3.5, the algorithm will be used when it is necessary to determine the solution \hat{A} for P1 of (3.1) when G has full rank and $\text{vec}^{-1}(L^{-1}u) \geq 0$, where L and u are those of Theorem 3.2. Recall that $r = n(n+1)/2$ when $F, G \in R^{n \times m}$.

The main computational requirements associated with the use of Algorithm 4.1 are outlined next. Procedures for computing the various matrix factorizations, singular values, etc, which are mentioned below are given in, for example, [8].

(i) *Determination of quantities associated with Theorem 3.2.* The main computational task involved in the evaluation of $L \in R^{r \times r}$, $u \in R^r$ and $l \in R^{mn-r}$ is the factorization of $J \in R^{mn \times r}$ as $P[L'0']$ for orthogonal P and for L of full rank. Any standard method for Q-R factorization could be used to perform that factorization of J .

(ii) *Specification of the set S of Theorem 4.1.* This requires evaluation of $\sigma_{\min}[L]$, where $L \in R^{r \times r}$.

(iii) *Steps 0 and I of Algorithm 4.1 (performed once).* The suggestion of Remark 4.4 is adopted here for the selection of the initial approximation \bar{k}_0 , so the evaluation of $X = \text{vec}^{-1}(L^{-1}u) = \text{vec}^{-1}(WL^{-1}u)$ is needed where $W \in R^{n^2 \times r}$. The determination of X just involves the assembly of the entries of the vector $WL^{-1}u \in R^{n^2}$ into the rows of $X \in R^{n \times n}$. Then $\bar{k}_0 = L[X]_{(0)}$, where the computation of $[X]_{(0)}$ requires evaluation of the result of changing all negative eigenvalues to 0 in the spectral form of X .

(iv) *Step II of Algorithm 4.1 (performed once per iteration).* The main burden here is the evaluation of γ_i which, in view of Theorem 2.1(v), can be done by finding a normalized eigenvector corresponding to the most negative eigenvalue of the corresponding symmetric $Z \in R^{n \times n}$ of (2.6).

(v) *Step III of Algorithm 4.1 (performed once per iteration).* This can be done by projecting u onto the subspace spanned by y_i and k_i and checking whether the result is in $\text{cone}[\text{line}\{k_i, y_i\}]$. If the projection is in that cone, it is the required k_{i+1} . Otherwise k_{i+1} is the closer to u of the closest points to u in the rays through k_i and

y_i , respectively. The main effort involved is the evaluation of five inner products in R' , the inversion of a 2×2 matrix and some logic.

The total amount of work involved depends very much on the matrices F and G which are considered and on the tolerance parameter ε chosen. To give an indication of the performance of Algorithm 4.1, some results follow for a case with

$$F = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -2 & 3 \\ 0 & 2 & 4 \end{bmatrix}, \quad G = \begin{bmatrix} 1 & 6 & 0 \\ 4 & 3 & 0 \\ 0 & 0 & -0.5 \end{bmatrix}.$$

The computations were coded in Fortran 77 using double precision for all real variables and were run on a VAX 11/750 with $\varepsilon = 10^{-9}$. Matrix inversion, Q-R factorization, singular-value decomposition and spectral decomposition were carried out using NAG subroutines F01AAF, F01QAF, F02WCF and F02ABF, respectively.

To make their interpretation easier, results are presented in Table 5.1 in the way described below.

TABLE 5.1

i	$\ F - A_i G\ _F$	$[\hat{b}_i + \ I\ ^2]^{1/2}$	ε_i
0	5.7824024346400	3.4277369606866	0.69
1	5.6385825638432	3.4277369606866	0.64
2	5.6180824578471	4.9565077042185	0.13
3	5.6041856642211	5.3160031568303	0.054
4	5.6029033847332	5.5351521480910	0.012
5	5.6016185784600	5.5433693128397	0.011
6	5.6013936067946	5.5867131439601	0.0026
7	5.6011842684349	5.5880547888253	0.0023
8	5.6011222214277	5.5972897504430	0.00068
9	5.6010076584435	5.6002056202548	0.00014
10	5.6010045000615	5.6008035555207	0.000036
11	5.6010016190206	5.6008195158433	0.000033
12	5.6010007751979	5.6009475009531	0.0000095
13	5.600992047102	5.6009865012541	0.0000023
14	5.600991545482	5.6009959775124	0.00000057
15	5.600991089662	5.6009962263921	0.00000052
16	5.600990956352	5.6009982523879	0.00000015
17	5.600990708254	5.6009988658921	0.000000037
18	5.600990700160	5.6009990174372	0.0000000094
19	5.600990692812	5.6009990174372	0.0000000093
20	5.600990690662	5.6009990646315	0.0000000079

Let $A_i = \overline{\text{vec}}^{-1}(L^{-1}k_i)$. Then A_i is the approximation to \hat{A} associated in a natural way with k_i of Algorithm 4.1. Much as in (3.4), it can be seen that

$$\|F - A_i G\|_F^2 = \|u - k_i\|^2 + \|I\|^2.$$

By Theorem 4.2, $k_i \rightarrow \hat{k}$ for Algorithm 4.1 so we would expect that

$$\|F - A_i G\|_F \rightarrow \|F - \hat{A} G\|_F.$$

Again by Theorem 4.2, $\hat{b}_i \leq \|u - \hat{k}\|^2$ and $\hat{b}_i \rightarrow \|u - \hat{k}\|^2$. Hence, in view of (3.4)

$$[\hat{b}_i + \|I\|^2]^{1/2} \leq \|F - \hat{A} G\|_F, \quad [\hat{b}_i + \|I\|^2]^{1/2} \rightarrow \|F - \hat{A} G\|_F.$$

The second and third columns of Table 5.1 are consistent with these claims.

The entries in the column headed ε_i give the smallest value of ε for which stopping condition (4.10) would be satisfied at iteration i . Consequently we find (without knowledge of the value of $\|F - \hat{A}G\|_F$) that $\|F - A_iG\|_F \leq (1 + \varepsilon_i)\|F - \hat{A}G\|_F$ at iteration i . The results show that ε_i decreases quite rapidly with increasing iterations i , indicating that $\|F - A_iG\|_F$ fairly rapidly becomes a good approximation to $\|F - \hat{A}G\|_F$.

Acknowledgment. Thanks are owed to K. G. Woodgate for many useful discussions.

REFERENCES

- [1] R. FLETCHER, *A nonlinear programming problem in statistics (educational testing)*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 257–267.
- [2] ———, *Semi-definite matrix constraints in optimization*, this Journal, 23 (1985), pp. 493–513.
- [3] N. J. HIGHAM, *Computing the polar decomposition—with applications*, Numerical Analysis Report 94, Dept. of Mathematics, Univ. of Manchester, England, 1984.
- [4] K. G. WOODGATE, *Optimization over positive semi-definite symmetric matrices with applications to quasi-Newton algorithms*, Ph.D. thesis, Dept. of Electrical Engineering, Imperial College, London, England, 1987.
- [5] J. C. ALLWRIGHT, *Positive semi-definite matrices: characterization via conical hulls and solution of matrix equations*, Preprint, 24th IEEE Conference on Decision and Control, Florida, December 11–13, 1985.
- [6] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.
- [7] J. STOER AND C. WITZGALL, *Convexity and Optimization in Finite Dimensions I*, Springer-Verlag, New York, 1970.
- [8] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, North Oxford-Academic, Oxford, 1983.
- [9] A. BEN-ISRAEL AND T. N. GREVILLE, *Generalized Inverses: Theory and Applications*, John Wiley, New York, 1974.
- [10] E. POLAK, *Computational Methods in Optimization; A Unified Theory*, Academic Press, New York, 1971.

STRUCTURAL STABILITY FOR SINGULAR SYSTEMS— A QUANTITATIVE APPROACH*

LIYI DAI†

Abstract. This paper studies the structural stability for singular systems. The problem is solved via the matrix perturbation analysis method. Allowed perturbation regions for structural stability are given.

Key words. matrix perturbation analysis method, singular system, structural stability, data point, perturbation, regulator

AMS(MOS) subject classification. 93D20

1. Introduction. In the design of control systems much attention has been paid to the problem of structural stability, for which many methods have been proposed in normal system theory (see, e.g., [1]). The reason for doing this is that when treating a practical system we always use a mathematical model to approximate it. The model, however, is generally only an approximation because of the influence of many external factors such as the nonaccuracy in system modeling and the aging of executive circuit elements. In this sense the structural parameters of a system model always have uncertainties. Here we will call them perturbations, defined as the deviations from the nominal points. The same problem exists for controllers: the realizations of controllers are also only approximations of the theoretical ones due to the same reasons, although they are accurately designed. In some cases, these reasons may cause the practical closed-loop system to be unstable, although it is well-designed on the basis of the nominal model. Then a problem naturally arises in the control system design: the designed controller should make the closed-loop system not only stable but also structurally stable. Only such controllers have practical interests and make the closed-loop system resistible to system parameter perturbations. Accurately speaking, a system is called structurally stable if it is stable and the stability is preserved in the presence of small arbitrary variations in system parameters. Papers [3] and [4] have studied qualitatively the existence and design methods of structurally stable compensators and regulators for singular systems. This paper deals with the same problem but from a completely new point of view—a quantitative approach. This is solved via the matrix perturbation analysis method.

The paper is divided into five sections. Section 2 introduces some notation that will be used. Section 3 reviews the related algebraic preliminaries. Section 4 analyzes the effects of parameter perturbations on system stability, and further presents the main results of this paper. Section 5 shows an application of the results to the structural stability analysis of closed-loop control systems.

2. Notation. The following notation will be used throughout this paper:

$\mathbb{C}^{n \times m} (R^{n \times m})$	Set of all $n \times m$ complex (real) constant matrices;
$\mathbb{C}^n (R^n)$	Set of all n dimension complex (real) vectors;
$\in (\notin)$	Belonging to (not belonging to);
$\equiv (\neq)$	Identically equal (not identically equal);
\subset	Included;

* Received by the editors October 6, 1986; accepted for publication (in revised form) July 28, 1987. This work was supported by the Science Foundation of Academia Sinica.

† Institute of Systems Science, Academia Sinica, Beijing 100080, People's Republic of China.

\mathbb{C}, \mathbb{C}^-	Complex plane, open left half \mathbb{C} , respectively;
$\ A\ $	Norm of matrix A , see definition;
A^*	Conjugate transpose of matrix A , i.e., $A^* = \bar{A}^T$;
$\text{tr}(A)$	Trace of matrix A ;
δA	Variation (perturbation) of matrix A ;
$P\{A\}$	Data point, which is a vector formed by listing the elements of A in arbitrary order;
$\sigma_i(\lambda_i)$	The (distinct) eigenvalue of a matrix;
\triangleq	Defined as, by definition;
$\text{Re}(\alpha)$	Real part of a complex constant α ;
$\sigma(E, A)$	Set of eigenvalues of matrix pencil (E, A) , i.e., $\sigma(E, A) = \{s sE - A = 0, s \in \mathbb{C}\}$;
$\sigma(A)$	$\sigma(A) = \sigma(I, A)$;
$\deg(\cdot)$	Degree of a polynomial;
$U_v(r)$	Neighborhood at the origin with radius r under norm v , i.e., $U_v(r) = \{x \ x\ _v < r, r > 0\}$.

3. Algebraic preliminaries. First we define the norm for matrices. For an arbitrary given matrix $A \in \mathbb{C}^{n \times m}$ the norms

$$\|A\|_2 = (\lambda_{\max}(A^*A))^{1/2}$$

and

$$\|A\|_F = (\text{tr}(A^*A))^{1/2}$$

are called the spectral norm and Frobenius norm, respectively. It is easy to see the following.

Property 1. $\|A\|_2 \leq \|A\|_F$.

Property 2. Let $B \in \mathbb{C}^{n \times r}$. Then $\|(AB)\|_v \geq \max\{\|A\|_v, \|B\|_v\}$, $v = 2, F$.

Property 3. Let $A \in \mathbb{C}^{n \times n}$, $\sigma(A) = \{\sigma_i\}$. Then

$$\|AA^* - A^*A\|_F^2 \leq 2 \left(\|A\|_F^4 - \left(\sum_{i=1}^n |\sigma_i|^2 \right)^2 \right).$$

LEMMA 3.1 [6]. Let $A, \delta A \in \mathbb{C}^{n \times n}$. Then for any $\mu \in \sigma(A + \delta A)$ there exists a $\lambda \in \sigma(A)$ such that

$$(3.1) \quad \frac{|\mu - \lambda|^m}{(H_A + |\mu - \lambda|)^{m-1}} \leq \Omega_A \|\delta A\|_2$$

where m is the highest order of the Jordan blocks of A and

$$(3.2) \quad \begin{aligned} \Omega_A &= (1 + H_A G_A)(1 + H_A G_A + \cdots + (H_A G_A)^{k-1}), \\ H_A &= \sqrt{\frac{4(n^3 - n)}{12}} \sqrt{\|AA^* - A^*A\|_F}, \\ G_A &= \begin{cases} 0 & \text{if } k = 1, \\ \max_{1 \leq i \leq k-1} \left(\frac{(G_i H_A)^{d_i} - 1}{G_i H_A - 1} G_i \right) & \text{if } k > 1, \end{cases} \\ G_i &= \frac{(\max_{i+1 \leq j \leq k} |\lambda_i - \lambda_j| + H_A)^{m_i-1}}{\prod_{j=i+1}^k |\lambda_i - \lambda_j|^{k_i}}, \end{aligned}$$

k is the number of distinct eigenvalues λ_i of A . $\lambda_i \neq \lambda_j$, $i \neq j$. k_i is the multiplicity of λ_i . d_i is the highest order of Jordan blocks belonging to λ_i and $m_i = k_{i+1} + \cdots + k_k$, $i = 1, 2, \dots, k-1$.

LEMMA 3.2 [8]. Let $A, \delta A \in \mathbb{C}^{n \times n}$. Then for any $\mu \in \sigma(A + \delta A)$ there exists a certain $\lambda \in \sigma(A)$ such that

$$(3.3) \quad |\lambda - \mu| \leq \Delta_F(A) / \xi_n(\eta)$$

where

$$\eta = \frac{\Delta_F(A)}{\|\delta A\|_F}, \quad \Delta_F(A) = \left(\|A\|_F^2 - \left(\sum_{i=1}^n |\sigma_i|^2 \right) \right)^{1/2}$$

and $\xi_n(\eta)$ is the unique nonnegative solution of the equation

$$f_n(\xi) = \xi^n + \dots + \xi^2 + \xi = \eta.$$

Generally, the eigenvalues of A and $B = A + \delta A$ ($\delta A = B - A$) are different. So we may ask the questions: "How far are the two eigenvalue sets?" and, in the presence of small perturbation δA , "How do the eigenvalues of $A + \delta A$ depend on it?" Lemmas 3.1 and 3.2 answer these questions and give an upper bound on the eigenvalue changes in the presence of perturbation δA . Obviously, if $\|\delta A\|$ is sufficiently small the matrices A and B are very close; so are their eigenvalues.

One more lemma is needed for future discussions.

LEMMA 3.3 [6]. For arbitrary matrix $A \in \mathbb{C}^{n \times n}$ there exists a nonsingular matrix T such that

$$(3.4) \quad T^{-1}AT = \text{block diag}(A_1, A_2, \dots, A_k);$$

here $A_i = \lambda_i I + L_i \in \mathbb{C}^{k_i \times k_i}$. k and k_i are as defined in Lemma 3.1. L_i is a strict upper-triangular matrix (whose diagonal elements are all zeros) and

$$(3.5) \quad \begin{aligned} \|T^{-1}\|_2 \|T\|_2 &\leq \Omega_A, \\ \|L_i\|_2 &\leq H_A, \quad i = 1, 2, \dots, k. \end{aligned}$$

The matrix T that transforms A into (3.4) is not unique. $\|T\| \|T^{-1}\|$ may change greatly for different choices of T . However, Lemma 3.3 assures us that an appropriate T can be chosen such that $\|T\| \|T^{-1}\|$ is bounded by Ω_A . $\|T\| \|T^{-1}\| \geq 1$ may be viewed as a deviation measure of A from a triangular form, the smaller the better. It reaches its lower bound 1 if A is upper triangular.

4. Structural stability for homogeneous singular systems. It is well known from classical linear system theory that if the homogeneous normal system

$$\dot{x} = Ax$$

is stable, it is structurally stable. This fact is not true for singular systems. The results of [4] show that a stable homogeneous singular system

$$(4.1) \quad E\dot{x} = Ax$$

where $x \in \mathbb{R}^n$ and rank $E < n$ may not be structurally stable. To be accurate we first give the following.

DEFINITION 4.1. Assume that system (4.1) is stable. If there exists a certain neighborhood at data point $P\{E, A\}$ such that the stability of system (4.1) is preserved in the presence of arbitrary variations of structural parameters in this neighborhood, we shall term the system (4.1) structurally stable at data point $P\{E, A\}$.

PROPOSITION 4.1 [4]. Let the system (4.1) be stable, i.e., $\sigma(E, A) \subset \mathbb{C}^-$. Then

- (1) (4.1) could not be structurally stable at $P\{E\}$;
- (2) (4.1) is structurally stable at $P\{A\}$ if and only if

$$(4.2) \quad \deg(|sE - A|) = \text{rank } E \triangleq d.$$

To have a further look at the results above, we now examine the following two examples.

Example 1. Consider the singular system

$$(4.3) \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \dot{x} = \begin{pmatrix} -2 & 1 & -1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} x$$

with parameter perturbations

$$\delta E = 0, \quad \delta A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\varepsilon \end{pmatrix}, \quad \varepsilon > 0.$$

Then

$$\sigma(E, A) = \{-2\}, \quad \sigma(E, A + \delta A) = \{-2, 1/\varepsilon\}.$$

The nominal system (4.3) is stable. But the perturbed system has an unstable pole $1/\varepsilon$ no matter how small ε (thus $\|\delta A\|$) is.

Example 2. The system

$$(4.4) \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \dot{x} = \begin{pmatrix} -1 & 0 & 10^4 \\ 0 & 0 & -2 \\ 0 & 1 & 0 \end{pmatrix} x$$

is stable. $\sigma(E, A) = \{-1, -2\}$. But if it has a small perturbation

$$\delta A = \begin{pmatrix} 0 & 0 & 0 \\ 3 \cdot 10^{-4} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

we will have $\sigma(E, A + \delta A) = \{-\sqrt{13}/2 - 3/2, \sqrt{13}/2 - 3/2\}$. The perturbed system has an unstable pole $\sqrt{13}/2 - 3/2$, although we can easily verify from Proposition 4.1 that system (4.4) is structurally stable at $P\{A\}$.

Example 1 shows that (4.2) is necessary to ensure the structural stability at data point $P\{A\}$. An equivalent expression of (4.2) is that for any sufficiently small $\|\delta A\|$, $\sigma(E, A + \delta A)$ does not create new unstable eigenvalues.

On the other hand, Example 2 reveals the fact that although the system (4.4) is structurally stable, the structural stability neighborhood depends on the matrix A and may be too small to be of any practical interest. Then, here naturally arises the question of how large the neighborhood is and of whether it could be estimated for any structurally stable system (4.1) using only E, A . The next section of this paper will provide an estimation via the matrix perturbation method.

From Lemmas 3.1 and 3.2 we know that for any matrix A and its perturbation δA , the eigenvalues of A and $A + \delta A$ will be very close provided $\|\delta A\|$ is small enough. Then by setting $\|\delta A\|$ small enough we will be certain that the eigenvalues of both A and $A + \delta A$ are in the left open-half complex plane. In the next step we need only find the upper bound in terms of E, A .

In view of the above results and to guarantee that our discussion proceeds, we hereafter assume that E is not perturbed, that $\sigma(E, A) \subset \mathbb{C}^-$ and that (4.2) holds.

Since the results are very obvious when $E = 0$, we also suppose hereafter that $E \neq 0$. First a helpful result is proved.

LEMMA 4.1. *Let $M \in \mathbb{C}^{n \times n}$ be nonsingular and*

$$M = \text{block diag}(M_1, M_2, \dots, M_k),$$

$$M_i = \lambda_i^{-1} I + L_i \in \mathbb{C}^{k_i \times k_i}, \quad i = 1, 2, \dots, k$$

where L_i is a strict upper-triangular matrix. $\lambda_i \neq \lambda_j$, $i \neq j$, $i, j = 1, 2, \dots, k$. Then for any $\mu \in \sigma(M^{-1} + \delta M)$ there exists a $\lambda \in \sigma(M^{-1})$ such that

$$(4.5) \quad \frac{|\lambda - \mu|^m}{(\bar{H}_M + |\lambda - \mu|)^{m-1}} \leq \|\delta M\|_2$$

where $m = \max_i \{k_i\}$ and

$$\bar{H}_M = \begin{cases} 0 & \text{when } m = 1, \\ \sum_{i=1}^{m-1} |\bar{\lambda}|^{i+1} H_M^i & \text{when } m > 1, \end{cases}$$

$$\bar{\lambda} = \max_i \{|\lambda_i|\}.$$

Proof. If $\mu \in \sigma(M^{-1})$, (4.5) obviously holds. So we assume that $\mu \notin \sigma(M^{-1})$. From our definition we know that $M^{-1} + \delta M - \mu I$ is singular, but $M^{-1} - \mu I$ is nonsingular. Therefore

$$I + (M^{-1} - \mu I)^{-1} \delta M = (M^{-1} - \mu I)^{-1} (M^{-1} - \mu I + \delta M)$$

is singular. Thus

$$\|(M^{-1} - \mu I)^{-1} \delta M\|_2 \geq 1,$$

i.e.,

$$(4.6) \quad (\|(M^{-1} - \mu I)^{-1}\|_2)^{-1} \leq \|\delta M\|_2.$$

Noticing the special form of M we know that there exists an i_0 such that

$$\|(M_{i_0}^{-1} - \mu I)^{-1}\|_2 = \|(M^{-1} - \mu I)^{-1}\|_2.$$

If we denote the eigenvalues of $(M_{i_0}^{-1} - \mu I)^*(M_{i_0}^{-1} - \mu I)$ by $s_{k_{i_0}}^2 \geq \dots \geq s_2^2 \geq s_1^2 > 0$, it follows that

$$\|(M_{i_0}^{-1} - \mu I)^{-1}\|_2 = \frac{1}{s_1}$$

and

$$(4.7) \quad s_i \leq \|(M_{i_0}^{-1} - \mu I)\|_2 \leq \|(\lambda_{i_0}^{-1} I + L_i)^{-1} - \mu I\|_2.$$

Since L_i is a strict upper-triangular matrix, it is easy to verify that

$$(\lambda_{i_0}^{-1} I + L_{i_0})^{-1} = \lambda_{i_0} (I - \lambda_{i_0} L_{i_0} + \dots + (-\lambda_{i_0} L_{i_0})^{k_{i_0}-1}).$$

Hence (4.7) becomes

$$\begin{aligned} s_i &\leq |\lambda_{i_0} - \mu| + \|\lambda_{i_0}^2 L_{i_0} + \lambda_{i_0}^3 L_{i_0}^2 + \dots + \lambda_{i_0} (-\lambda_{i_0} L_{i_0})^{k_{i_0}-1}\|_2 \\ &\leq |\lambda_{i_0} - \mu| + \sum_{j=1}^{k_{i_0}} |\lambda_{i_0}|^{j+1} \|L_{i_0}\|_2^j \\ &\leq |\lambda_{i_0} - \mu| + \bar{H}_M. \end{aligned}$$

On the other hand we also know that

$$s_1 s_2 \dots s_{k_{i_0}} = (|(M_{i_0}^{-1} - \mu I)^*(M_{i_0}^{-1} - \mu I)|)^{1/2} = |\lambda_{i_0} - \mu|^{k_{i_0}}.$$

Therefore

$$s_1 = \frac{|\lambda_{i_0} - \mu|^{k_{i_0}}}{s_2 \cdots s_{k_{i_0}}} \leq \frac{|\lambda_{i_0} - \mu|^{k_{i_0}}}{(\bar{H}_M + |\lambda_{i_0} - \mu|)^{k_{i_0}-1}} \leq \frac{|\lambda_{i_0} - \mu|^m}{(\bar{H}_M + |\lambda_{i_0} - \mu|)^{m-1}},$$

we can now obtain (4.5) by combining the above with (4.6). Q.E.D.

This lemma allows us to estimate the eigenvalue perturbations of M^{-1} in terms of M .

The following theorem is our main result.

THEOREM 4.1. *Consider the homogeneous singular system (4.1). Assume that $\sigma(E, A) \subset \mathbb{C}^-$ and (4.2) holds. Then the stability is preserved in the presence of arbitrary variation δA satisfying*

$$(4.8) \quad \delta A \in U_2(\varepsilon/\alpha),$$

i.e., $\sigma(E, A + \delta A) \subset \mathbb{C}^-$, where

$$(4.9) \quad \begin{aligned} \alpha &= \max\{\Omega_{A^{-1}E} \|A^{-1}E\|_2, 1\} \|A\|_2 \Omega_{A^{-1}E}, \\ \varepsilon &= \frac{\Delta_e^m}{(\beta + \Delta_e)^{m-1} + \Delta_e^m}, \Delta_e = \min_i \{|\operatorname{Re}(\lambda_i)| \mid \lambda_i \in \sigma(E, A)\} > 0, \\ \beta &= \begin{cases} 0 & \text{when } m = 1, \\ \sum_{j=1}^{m-1} \bar{\lambda}_e^{j+1} \hat{H}^j & \text{when } m > 1, \end{cases} \\ \bar{\lambda}_e &= \frac{1}{\min_i \{|\lambda_i| \mid \lambda_i \in \sigma(E, A)\}} > 0, \\ \hat{H} &= \sqrt{\frac{4(d^3-d)}{6}} \sqrt{\Omega_{A^{-1}E}^4 \|A^{-1}E\|_F^4 - \left(\sum_{i=1}^d |\sigma_i|^{-2}\right)^2}, \\ \sigma(E, A) &= \{\sigma_i\}. \end{aligned}$$

The other unspecified terms are the same as those defined in (3.2).

Proof. The proof is completed in four steps.

(a) Finding an appropriate decomposition of E, A . By the assumption $\sigma(E, A) \subset \mathbb{C}^-$ we know that A is nonsingular. Then it follows from Lemma 3.3 that there exists a nonsingular matrix $T_1 \in \mathbb{C}^{n \times n}$ such that

$$T_1^{-1} A^{-1} E T_1 = \text{block diag}(\tilde{E}_1, \tilde{E}_2, \dots, \tilde{E}_k, \tilde{N})$$

where $\tilde{E}_i = \lambda_i I + L_i \in \mathbb{C}^{k_i \times k_i}$, $\lambda_i \neq 0$, $i = 1, 2, \dots, k$, $\lambda_i \neq \lambda_j$, $i \neq j$. $\sum_{i=1}^k k_i = d < n$. \tilde{N} , L_i , $i = 1, 2, \dots, k$, are strict upper-triangular matrices. k_i is the multiplicity of the eigenvalue λ_i of $A^{-1}E$. And furthermore

$$\|T_1^{-1}\|_2 \|T_1\|_2 \leq \Omega_{A^{-1}E}$$

where $\Omega_{A^{-1}E}$ is as that in (3.2). It is easy to know from (4.2) that $\tilde{N} = 0$. Thus we have

$$(4.10) \quad T_1^{-1} A^{-1} E T_1 = \text{block diag}(\tilde{E}_1, \tilde{E}_2, \dots, \tilde{E}_k, 0).$$

Denote $E_1 = \text{block diag}(\tilde{E}_1, \tilde{E}_2, \dots, \tilde{E}_k) \in \mathbb{C}^{d \times d}$, which is nonsingular, and

$$(4.11) \quad Q = \text{block diag}(E_1^{-1}, I_{n-d}) T_1^{-1} A^{-1}, \quad P = T_1.$$

By summing the above results we easily get the following decomposition:

$$(4.12) \quad QEP = \text{block diag}(I_d, 0), \quad QAP = \text{block diag}(E_1^{-1}, I_{n-d}).$$

(b) Evaluating H_{E_1} . From our definition we know that

$$H_{E_1} = ((d^3 - d)/12)^{1/4} (\|E_1 E_1^* - E_1^* E_1\|_F)^{1/2}.$$

By Property 3 we obtain

$$(4.13) \quad H_{E_1} \leq ((d^3 - d)/12)^{1/4} \left(2 \left(\|E_1\|_F^4 - \left(\sum_{i=1}^d |\sigma_i(E_1)|^2 \right) \right) \right)^{1/4}.$$

Since

$$\sigma_i(E_1) = \sigma_i^{-1}(E_1^{-1}) = \sigma_{ei}^{-1}, \quad \sigma_{ei} \in \sigma(E, A),$$

according to (2.8) we have

$$\begin{aligned} \|E_1\|_F^2 &= \|\text{block diag}(E_1, 0)\|_F^2 = \text{tr}(T_1^{-1} A^{-1} E T_1 T_1^* E^* (A^{-1})^* (T_1^{-1})^*) \\ &\leq \|T_1\|_2^2 \text{tr}(E^* (A^{-1})^* (T_1^{-1})^* T_1^{-1} A^{-1} E) \\ &\leq \|T_1\|_2^2 \|T_1^{-1}\|_2^2 \|A^{-1} E\|_F^2 \leq \Omega_{A^{-1}E}^2 \|A^{-1} E\|_F^2. \end{aligned}$$

Thus (4.13) becomes

$$(4.14) \quad H_{E_1} \leq \hat{H}.$$

(c) Calculating the characteristic polynomial of the perturbed system. For arbitrary variation δA , the perturbed system is

$$(4.15) \quad E\dot{x} = (A + \delta A)x.$$

We denote

$$\delta A_0 = Q^{-1} \delta A P^{-1} = \begin{pmatrix} \delta A_{11} & \delta A_{12} \\ \delta A_{21} & \delta A_{22} \end{pmatrix};$$

then when condition (4.8) is satisfied, direct computation shows that

$$\begin{aligned} \|\delta A_0\|_2 &\leq \|Q^{-1}\|_2 \|P^{-1}\|_2 \|\delta A\|_2 \\ (4.16) \quad &\leq \|A T_1\|_2 \|\text{block diag}(E_1, I_{n-1})\|_2 \|T_1^{-1}\|_2 \|\delta A\|_2 \\ &\leq \|A\|_2 \Omega_{A^{-1}E} \max\{\|E_1\|_2, 1\} \|\delta A\|_2 \\ &\leq \alpha \|\delta A\|_2 < \varepsilon, \end{aligned}$$

i.e., $\|\delta A_0\|_2 < \varepsilon$. On the other hand, we also have by repeatedly using Property 2 that

$$\|\delta A_0\|_2 \geq \max\{\|\delta A_{11}\|_2, \|\delta A_{12}\|_2, \|\delta A_{21}\|_2, \|\delta A_{22}\|_2\}.$$

Combining this with (4.16) we finally obtain

$$(4.17) \quad \|\delta A_{ij}\|_2 < \varepsilon, \quad i, j = 1, 2.$$

Now we return to the characteristic polynomial of (4.15). The calculation shows us that

$$\begin{aligned} |Q\|P\|sE - (A + \delta A)| &= |sQEP - Q(A + \delta A)P| \\ &= \left| \begin{bmatrix} sI - (E_1^{-1} + \delta A_{11}) & -\delta A_{12} \\ -\delta A_{21} & -(I + \delta A_{22}) \end{bmatrix} \right|. \end{aligned}$$

Noting the fact that $\varepsilon < 1$, from (4.17) we see that $\|\delta A_{22}\|_2 < 1$. Then $(I + \delta A_{22})^{-1}$ exists. The above equation becomes

$$\begin{aligned} |Q\|P\|sE - (A + \delta A)| &= \left| \begin{bmatrix} sI - (E_1^{-1} + \delta A_{11}) & -\delta A_{12} \\ -\delta A_{21} & -(I + \delta A_{22}) \end{bmatrix} \begin{bmatrix} I & 0 \\ -(I + \delta A_{22})^{-1}\delta A_{21} & I \end{bmatrix} \right| \\ &= |-(I + \delta A_{22})| \cdot |sI - (E_1^{-1} + \delta A_{11} - \delta A_{12}(I + \delta A_{22})^{-1}\delta A_{21})|, \end{aligned}$$

i.e.,

$$(4.18) \quad \sigma(E, A + \delta A) = \sigma(E_1^{-1} + \delta A_{11} - \delta A_{12}(I + \delta A_{22})^{-1}\delta A_{21}).$$

(d) Under the assumption of (4.8) we analyze the stability of the system (2.15). Note the fact that E_1 is nonsingular and has the same form as that of (3.4). So we know E_1^{-1} has the same property. Therefore from Lemma 4.1 we deduce that for any $\mu \in \sigma(E_1^{-1} + \delta A_{11} - \delta A_{12}(I + \delta A_{22})^{-1}\delta A_{21})$ there exists a $\lambda \in \sigma(E_1^{-1}) = \sigma(E, A)$ satisfying

$$(4.19) \quad \frac{|\lambda - \mu|^m}{(\bar{H}_{E_1} + |\lambda - \mu|)^{m-1}} \leq \|\delta A_{11} - \delta A_{12}(I + \delta A_{22})^{-1}\delta A_{21}\|_2$$

where m is the highest order of Jordan blocks of E_1^{-1} (or (E, A)). Furthermore, the result of (b) indicates that

$$\begin{aligned} \bar{H}_{E_1} &= \begin{cases} 0 & \text{when } m = 1, \\ \sum_{j=1}^{m-1} |\bar{\lambda}_e|^{j+1} H_{E_1}^j & \text{when } m > 1 \end{cases} \\ &\leq \beta. \end{aligned}$$

Combining this with (4.17) and also noting that $\|\delta A_{22}\|_2 < \varepsilon < 1$, we obtain the following from (4.19):

$$\begin{aligned} \frac{|\lambda - \mu|^m}{(\beta + |\lambda - \mu|)^{m-1}} &\leq \frac{|\lambda - \mu|^m}{(\bar{H}_{E_1} + |\lambda - \mu|)^{m-1}} \\ &\leq \|\delta A_{11}\|_2 + \|\delta A_{12}\|_2 \|\delta A_{21}\|_2 \|(I + \delta A_{22})^{-1}\|_2 \\ &\leq \varepsilon + \varepsilon^2(1 - \varepsilon)^{-1} = \varepsilon(1 - \varepsilon)^{-1} \\ &= \frac{\Delta_e^m}{(\beta + \Delta_e)^{m-1} + \Delta_e^m}. \end{aligned}$$

Since the function $x^m(\beta + x)^{-(m-1)}$ is strictly monotone increasing on $x \geq 0$, we thus have

$$\lambda - \mu \in U_2(\Delta_e).$$

Hence $\operatorname{Re}(\mu) < \operatorname{Re}(\lambda) + \Delta_e \leq 0$. We know from (4.18) that $\sigma(E, A + \delta A) \subset \mathbb{C}^-$. The perturbed singular system (4.15) is stable. Q.E.D.

This theorem gives us an estimation of the structural stability region in terms of E and A . We would like to make some comments on it.

Remark 1. It may be seen from the proof procedure of Theorem 4.1 that if H_{E_1} is used in lieu of \hat{H} in (4.9) (sometimes) a much more accurate estimation may be obtained. On the other hand, the matrix E_1 is not unique and cannot be determined in advance in this case. So this method presents great difficulty in determining an appropriate E_1 .

In the proof of Theorem 4.1 it is also indicated that if we use (3.3) instead of (3.1) to estimate the bound of $(\delta A_{11} - \delta A_{12}(I + \delta A_{22})^{-1}\delta A_{21})$, the same procedure can be used to prove Theorem 4.2.

THEOREM 4.2. *Consider the homogeneous singular system (4.1). Assume that $\sigma(E, A) \subset \mathbb{C}^-$ and (4.2) holds. Then the stability of (4.1) is preserved in the presence of δA satisfying*

$$(4.20) \quad \delta A \in U_F(\bar{\varepsilon}/\alpha)$$

where

$$(4.21) \quad \bar{\varepsilon} = \frac{\bar{\eta}\Delta_e}{f_d(\bar{\eta}) + \bar{\eta}\Delta_e}, \quad \bar{\eta} = \frac{\nu}{\Delta_e},$$

$$\nu = \left(\sum_{i=1}^k \sum_{j=1}^{k_i} |\lambda_{e_i}|^{j+1} \Omega_{A^{-1}E}^j \|A^{-1}E\|_F \right)^{1/2}.$$

The other unspecified terms are the same as those defined in (4.9).

Proof. From Lemma 3.2 we know that for any $\mu \in \sigma(E_1^{-1} + \delta A_{11} - \delta A_{12} \cdot (I + \delta A_{22})^{-1} \delta A_{21})$ there exists a $\lambda \in \sigma(E_1^{-1}) = \sigma(E, A)$ such that

$$(4.22) \quad |\lambda - \mu| \leq \frac{\Delta_F(E_1^{-1})}{\xi_d(\eta)}, \quad \eta = \frac{\Delta_F(E_1^{-1})}{\|\delta A_{11} - \delta A_{12}(1 + \delta A_{22})^{-1} \delta A_{21}\|_F}.$$

Furthermore, by Lemma 3.2, we have

$$(4.23) \quad \Delta_F(E_1^{-1}) = \left(\|E_1^{-1}\|_F^2 - \sum_{i=1}^d |\sigma_{e_i}|^2 \right)^{1/2}, \quad \sigma_{e_i} \in \sigma(E_1^{-1}) = \sigma(E, A).$$

Thus from the definition of E_1 we obtain

$$\begin{aligned} \|E_1^{-1}\|_F^2 &= \sum_{i=1}^k \|\tilde{E}_i^{-1}\|_F^2 = \sum_{i=1}^d \|(\lambda_i I + L_i)^{-1}\|_F^2 \\ &\leq \sum_{i=1}^k \sum_{j=0}^{k_i} |\lambda_{e_i}|^{2(j+1)} \|L_i^{2j}\|_F \quad \left(\lambda_{e_i} = \frac{1}{\lambda_i} \right) \\ &\leq \sum_{i=1}^k \sum_{j=1}^{k_i} |\lambda_{e_i}|^{2(j+1)} \|T_1^{-1} A^{-1} E T_1\|_F^{2j} + \sum_{i=1}^k k_i |\lambda_{e_i}|^2 \\ &\leq \sum_{i=1}^k \sum_{j=1}^{k_i} |\lambda_{e_i}|^{2(j+1)} \|A^{-1} E\|_F^{2j} \Omega_{A^{-1}E}^{2j-1} + \sum_{i=1}^k k_i |\lambda_{e_i}|^2. \end{aligned}$$

The substitution of the above inequality into (4.23) immediately yields

$$(4.24) \quad \Delta_F(E_1^{-1}) \leq \gamma.$$

Using a procedure similar to that in the proof of (4.17), it is not difficult to verify that

$$(4.25) \quad \delta A_{ij} \in U_F(\bar{\varepsilon}), \quad i, j = 1, 2.$$

Hence

$$\delta A_{11} - \delta A_{12}(I + \delta A_{22})^{-1} \delta A_{21} \in U_F(\bar{\varepsilon} + \bar{\varepsilon}^2(1 - \bar{\varepsilon})^{-1}) = U_F(\bar{\varepsilon}(1 - \bar{\varepsilon})^{-1}) = U_F(\bar{\eta}\Delta_e/f_d(\bar{\eta})).$$

Note that the function $1/\bar{\eta}f_d(\bar{\eta}) = \bar{\eta}^{d-1} + \dots + \bar{\eta} + 1$ is strictly monotone increasing on $\bar{\eta} \geq 0$. The above shows that

$$f_d(\Delta_F(E_1^{-1})/\Delta_e) < \bar{\eta}.$$

This is equivalent to

$$(4.26) \quad \xi_d(\bar{\eta}) > \Delta_F(E_1^{-1})/\Delta_e.$$

We thus get from (4.22) and (4.26) that

$$|\lambda - \mu| \in U_2(\Delta_e),$$

$\operatorname{Re}(\mu) < 0$, i.e., $\sigma(E, A + \delta A) \subset \mathbb{C}^-$. Q.E.D.

Remark 2. Generally speaking, Δ_e is very small. Therefore $\bar{\eta}$ is often very large. Consequently we know $\bar{\varepsilon}$ is very small. This means that the structural stability region given by Theorem 4.2 is conservative.

Theorems 4.1 and 4.2 include the worst-case perturbation—we need to know only the norm bound of perturbation δA . However, in practice, it is usually the case that much more may be known about parameter perturbations. For example, δA may take the form of

$$\delta A = F\delta DG.$$

In such cases, the same procedure may be applied to the analysis of a perturbation region for the structural stability of δD . Careful analysis may result in a much larger region.

Remark 3. Suppose that the matrix pencil (E, A) has only simple eigenvalues (with multiplicity of one); so does E_1 . Then $m = k_i = 1$, $i = 1, 2, \dots, k = d$. In this special case, we have the following corollary.

COROLLARY 4.1. *Suppose that the matrix pencil (E, A) has only simple eigenvalues. Then the perturbed system (4.15) is stable if either of the two following conditions is satisfied:*

(1)

$$(4.27) \quad \begin{aligned} \delta A &\in U_2(\varepsilon/\alpha), \\ \varepsilon &= \Delta_e/(1 + \Delta_e), \quad \alpha = \max\{\|A^{-1}E\|_2 \Omega_{A^{-1}E}, 1\} \|A\|_2 \Omega_{A^{-1}E}; \end{aligned}$$

(2)

$$(4.28) \quad \begin{aligned} \delta A &\in U_F(\bar{\varepsilon}/\alpha), \\ \bar{\varepsilon} &\text{ is the same as in (4.21).} \end{aligned}$$

5. Structural stability for regulation systems. In the previous section we have analysed quantitatively the structural stability region for homogeneous singular systems. Now, by utilizing these results, we consider the structural stability for general regulation systems. Such systems have the property of being able to resist small parameter perturbations and external disturbances. Our aim here is to find a perturbation region for the nominal system or its compensator so that the perturbed system is still internally stable, and output regulation (for definition, cf. [4]) provided the perturbation is within the above region.

Consider the following singular system:

$$(5.1) \quad \begin{aligned} E\dot{x} &= Ax + Bu + A_1f, \\ \dot{f} &= A_2f, \\ y &= C_1x + C_2f, \\ z &= D_1x + D_2f \end{aligned}$$

where $x \in R^n$ is its state, $u \in R^r$ is its control input, $y \in R^h$ is its measure output, $f \in R^p$ is its external disturbance (or reference signal, or any other signals to be tracked) and $z \in R^q$ is the vector to be regulated. $E, A, B, A_1, A_2, C_1, C_2, D_1, D_2$ are constant

matrices of appropriate dimensions. We also suppose that $\text{rank } E < n$, $E \neq 0$, and $|sE - A| \neq 0$.

DEFINITION 5.1. For the system (5.1) if there exists a dynamic compensator

$$(5.2) \quad \dot{x}_c = A_c x_c + B_c y, \quad u = F_c x_c + Fy$$

where $x_c \in R^{n_c}$, A_c , B_c , F_c , F are constant matrices of appropriate dimensions, such that the closed-loop system is internally stable (i.e., it is stable when $f \equiv 0$) and output regulation (i.e., $\lim_{t \rightarrow \infty} z = 0$), we shall term (5.2) an output regulator for the system (5.1). Furthermore, if there exists a certain neighborhood at data point $P\{E, A, B, C_1\}$ such that (5.2) is an output regulator for (5.1) with arbitrary variations of structural parameters in this neighborhood, we shall call (5.2) a structurally stable output regulator (SSOR) for the system (5.1) at $P\{E, A, B, C_1\}$.

Dai and Wang [4] have shown qualitatively the following theorem.

THEOREM 5.1. Suppose that $\text{rank } E < n$. Then we have that

- (1) System (5.2) could not be an SSOR for system (5.1) at $P\{E\}$;
- (2) System (5.2) is an SSOR for system (5.1) at $P\{A, B, C_1, A_1, B_c, F_c, F\}$ if and only if

- (i) (E, A, B) is stabilizable and (E, A, C_1) is detectable;
- (ii) z is readable from y , i.e., $z = Gy$;
- (iii) The closed-loop system is internally stable;
- (iv) System (5.2) incorporates an internal model of A_2 . This model is observable about u and controllable about z ;

$$(v) \quad \deg \left(\left| s \begin{bmatrix} E & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} A + BFC_1 & BF_c \\ B_c C_1 & A_c \end{bmatrix} \right| \right) = n_c + \text{rank } E.$$

In view of the above results, we now consider the structural stability for regulation systems.

First we suppose that only the nominal system (5.1) is perturbed. Let

$$\bar{E} = \begin{bmatrix} E & 0 \\ 0 & I \end{bmatrix}, \quad \bar{A} = \begin{bmatrix} A + BFC_1 & BF_c \\ B_c C_1 & A_c \end{bmatrix}.$$

Theorems 4.1 and 4.2 give us Theorem 5.2.

THEOREM 5.2. Consider the system (5.1) and its output regulator (5.2). Assume that conditions (i)–(v) in Theorem 5.1 are satisfied. Then (5.2) is an output regulator for the system (5.1) in the presence of δA and δB , provided one of the following conditions is satisfied:

$$(1) \quad [\delta A \quad \delta B] \in U_2(\varepsilon/\alpha a_1), \quad a_1 = \left\| \begin{bmatrix} I & 0 \\ FC_1 & F_c \end{bmatrix} \right\|_2;$$

$$(2) \quad [\delta A \quad \delta B] \in U_F(\bar{\varepsilon}/\alpha a_1).$$

Here the unspecified values are as defined in (4.9) and (4.21), where E, A are replaced by \bar{E}, \bar{A} .

In another case, let only the compensator (5.2) be perturbed. We have Theorem 5.3.

THEOREM 5.3. Assume that conditions (i)–(v) in Theorem 5.1 are satisfied. When the perturbation δB_c , δF_c , δF satisfy either of the following conditions:

$$(a) \quad \begin{bmatrix} \delta F & \delta F_c \\ \delta B_c & 0 \end{bmatrix} \in U_2(\varepsilon/\alpha a_2), \quad a_2 = \left\| \begin{bmatrix} B_1 & 0 \\ 0 & I \end{bmatrix} \right\|_2 \left\| \begin{bmatrix} C_1 & 0 \\ 0 & I \end{bmatrix} \right\|_2;$$

$$(b) \quad \begin{bmatrix} \delta F & \delta F_c \\ \delta B_c & 0 \end{bmatrix} \in U_F(\bar{\varepsilon}/\alpha a_2).$$

Formula (5.2) is still an SSOR for the system (5.1). Here the unspecified terms are similar to those in Theorem 5.2.

6. Conclusion. This paper studies the robust region for structurally stable singular systems using the matrix perturbation analysis method. This is a completely new method based on state space description. The results are applied to analyzing the structural stability of singular regulation systems. But it is worthwhile to point out that the robust regions given here are conservative. Therefore, further studies are needed to find a more useful and better method to estimate the robust region.

REFERENCES

- [1] B. R. BARMISH, M. CORLESS AND G. LEITMANN, *A new class of stabilizing controllers for uncertain dynamical systems*, this Journal, 21 (1983), pp. 246–255.
- [2] F. L. BAUER AND C. T. FIKE, *Norms and exclusion theorem*, Numer. Math., 2 (1960), pp. 137–141.
- [3] S. L. CAMPBELL, *Singular Systems of Differential Equations II*, Pitman, New York, 1982.
- [4] L. DAI AND C. WANG, *Structurally stable normal compensators for singular systems*, 5th National Conference on Control Theory and Applications, Vol. 1, 1985, pp. 31–35, preprint; *J. Systems Sci. Math. Sci.*, 7 (1987), pp. 89–93.
- [5] ———, *Stable and structurally stable regulators for singular systems*, Acta Math. Appl. Sinica (1987), to appear.
- [6] E. JIANG, *On spectral variation of a nonnormal matrix*, Linear Algebra Appl., 42 (1982), pp. 223–241.
- [7] P. HENRICI, *Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices*, Numer. Math., 4 (1962), pp. 24–40.
- [8] F. R. GANTMACHER, *The Theory of Matrices*, Vol. 2, Chelsea, New York, 1974.
- [9] A. M. OSTROWSKI, *Über die Stetigkeit von charakteristischen Wurzeln in Abhängigkeit von den Matrizenelementen*, Jahresber. Deutsch. Math.-Verein., 60 (1957), pp. 40–42.

NEARLY OPTIMAL SINGULAR CONTROLS FOR WIDEBAND NOISE DRIVEN SYSTEMS*

H. J. KUSHNER† AND K. M. RAMACHANDRAN†

Abstract. Singular stochastic control problems arise in many applications, for example, in storage, inventory, finite fuel, consumption and investment and limits of impulsive control problems. Here, the increment of the control effort is not of the usual form $u(t) dt$, but is the differential of a nondecreasing and suitably adapted process. The diffusion process models used are only approximations to some “physical” process—which might be a “wideband” noise driven system or a suitably scaled discrete parameter process. Since the optimal controls for the “physical” processes are usually impossible to obtain, it is of interest to know whether “nearly” optimal controls for the diffusion model are “nearly” optimum when applied to the physical problem. It is shown that this is true under broad conditions, for discounted and average cost per unit time problems. The usual weak convergence analysis via the Skorokhod topology on $D[0, \infty)$ is not appropriate here, due to the nature of the singular controls, and it is necessary to use a combination of the Skorokhod and “pseudopath” topology.

Key words. singular stochastic control, approximately optimal control, weak convergence, pseudopath topology, control of wideband noisy systems, modeling of physical systems by singular control processes

AMS(MOS) subject classifications. 93E20, 93E25, 60F17

1. Introduction. Let $Y_i(\cdot)$, $i=0, 1$, be nondecreasing processes with $Y_i(0)=0$, which are nonanticipative with respect to a Wiener process $w(\cdot)$. Define $x(\cdot)$, $Y(\cdot)$ and $Z(\cdot)$ by $x(0)=x$, $Z(0)=0$, $Y(\cdot)=Y_0(\cdot)-Y_1(\cdot)$ and

$$(1.1) \quad dx = [b(x) dt + \sigma(x) dw] + [dY_0 - dY_1] \equiv dZ + dY.$$

We assume that there is a $\bar{B} \in (0, \infty)$ (given as part of the problem statement) such that we are obliged to keep $x(t) \in [0, \bar{B}]$. Unless otherwise mentioned, we always assume that the $Y(\cdot)$ process is such that $x(t) \in [0, \bar{B}]$. The process (1.1) has been widely used as a model of storage and dam processes, both with and without control [1]–[5], [16]. The $Y_1(\cdot)$ might denote the “withdrawal” process, whereby actual use is made of the system’s contents. $Y_0(\cdot)$ might simply denote a process which is used solely as a modeling device to guarantee that $x(t) \geq 0$, (see, e.g., [4]). The process $z(\cdot)$ might denote the difference between the “natural” inputs and “natural” demand. See the discussion in [4] on this point. Other interpretations are discussed in the references.

Let $k_0 > k_1 > 0$ and let $k(\cdot)$ be a bounded continuous function. Define the two types of costs (E_x denotes the expectation under initial condition $x(0)=x$):

$$(1.2) \quad \begin{aligned} V_0(x, Y_0, Y_1) &= E_x \int_0^\infty e^{-\beta t} [k_0 dY_0(t) - k_1 dY_1(t) + k(x(t)) dt], \\ V(x) &= \inf_{Y_0, Y_1} V_0(x, Y_0, Y_1) \end{aligned}$$

* Received by the editors October 12, 1986; accepted for publication (in revised form) July 28, 1987. The research of the first author was supported in part by the National Science Foundation under grant ECS-8505674, the Air Force Office of Scientific Research under contract AFOSR-85-0315 and the Office of Naval Research under contract N00014-83-K-0542. The research of the second author was supported in part by the Army Research Office under contract DAAG-29-84-K-0082 and the Office of Naval Research under contract N00014-85-K-0607.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

and

$$(1.3) \quad \begin{aligned} \gamma_0(x, Y_0, Y_1) &= \overline{\lim}_T E_x \int_0^T [k_0 dY_0(t) - k_1 dY_1(t) + k(x(t)) dt] / T, \\ \gamma_0(x) &= \inf_{Y_0, Y_1} \gamma_0(x, Y_0, Y_1). \end{aligned}$$

In (2.1), the inf is over all *admissible* $Y_i(\cdot)$, namely over all nonanticipative, nondecreasing $Y_i(\cdot)$ such that $x(t) \in [0, \bar{B}]$. The class over which the inf is taken in (1.3) will be described in § 7.

Reference [1] gives an elegant presentation of the optimal control problem (1.1), (1.2) and of the properties of the associated Bellman equation. Most of the other current literature seems to concern the case where $Z(\cdot)$ is a Wiener process (perhaps with drift). Since there are likely to be few applications which are perfectly modeled by (1.1), we must look at the model in the sense that it approximates in some way an actual physical problem. Since the model (1.1) would be simpler than the approximated "physical" process, it is attractive to use it for calculating a control for the actual physical process. But the question arises concerning how good it is in comparison with the optimal control for the physical process.

Models such as (1.1) also arise as limits of suitably interpolated discrete parameter processes. Consider one type: For each $\varepsilon > 0$, let $(i=0, 1)$ $Y_i^\varepsilon(\cdot)$ be nondecreasing processes, piecewise constant on the intervals $[n\varepsilon, n\varepsilon + \varepsilon)$, and define $\delta Y_i^\varepsilon(n\varepsilon) \equiv Y_i^\varepsilon(n\varepsilon + \varepsilon) - Y_i^\varepsilon(n\varepsilon)$. For appropriate functions F and G , define $\{X_n^\varepsilon, Z_n^\varepsilon\}$ by $X_0^\varepsilon = x$, $Z_0^\varepsilon = 0$, and

$$(1.4) \quad \begin{aligned} X_{n+1}^\varepsilon &= X_n^\varepsilon + (Z_{n+1}^\varepsilon - Z_n^\varepsilon) + \delta Y_0^\varepsilon(n\varepsilon) - \delta Y_1^\varepsilon(n\varepsilon), \\ Z_{n+1}^\varepsilon - Z_n^\varepsilon &= \varepsilon G(X_n^\varepsilon, \xi_n^\varepsilon) + \sqrt{\varepsilon} F(X_n^\varepsilon, \xi_n^\varepsilon), \\ EF(x, \xi_n^\varepsilon) &\equiv 0, \end{aligned}$$

where $\{\xi_n^\varepsilon\}$ is a random sequence. We say that the controls with values $\delta Y_i^\varepsilon(n\varepsilon)$ in (1.4) at time n are admissible if they depend on the "full information" $\{X_i^\varepsilon, i \leq n, Y_j^\varepsilon(\varepsilon i), j=0, 1, i \leq n, \xi_j^\varepsilon, j < n\}$ available at time n . Define the interpolated processes $X^\varepsilon(t) = X_n^\varepsilon$ and $Z^\varepsilon(t) = Z_n^\varepsilon$ on $[n\varepsilon, n\varepsilon + \varepsilon)$ and the costs

$$(1.5) \quad \begin{aligned} V_0^\varepsilon(x, Y_0^\varepsilon, Y_1^\varepsilon) &= E_x \int_0^\infty e^{-\beta t} [k_0 dY_0^\varepsilon(t) - k_1 dY_1^\varepsilon(t) + k(X^\varepsilon(t)) dt], \\ V^\varepsilon(x) &= \inf_{Y_0^\varepsilon, Y_1^\varepsilon} V_0^\varepsilon(x, Y_0^\varepsilon, Y_1^\varepsilon). \end{aligned}$$

In general, we know virtually nothing about the optimal or δ -optimal policies for (1.4), (1.5). Suppose that, for reasonable $Y_i^\varepsilon(\cdot)$, the set $\{X^\varepsilon(\cdot), Y_0^\varepsilon(\cdot), Y_1^\varepsilon(\cdot)\}$ converges weakly in some sense to a solution of (1.1), thus providing justification for use of (1.1). It is of considerable interest to know just how good (compared to the optimal controls for (1.4)) are the optimal (or δ -optimal) policies which we obtain for (1.1), (1.2), when suitably adapted to and applied to the system (1.4), (1.5). Consider the following example.

Let $\bar{Y}_i(\cdot)$, $i=0, 1$, denote the optimal or (for some given $\delta > 0$) δ -optimal controls for (1.1), (1.2). Frequently [1], [4] they are of the barrier form: There are $0 \leq L^* < U^* < \infty$ such that $\bar{Y}_0(\cdot)$ is used only to keep $x(\cdot)$ from "falling below" L^* and $\bar{Y}_1(\cdot)$ is used only to keep $x(\cdot)$ from going above U^* . The policy $\bar{Y}_i(\cdot)$ adapted to (1.4) (call it $\bar{Y}_i^\varepsilon(\cdot)$) is a policy which returns $X^\varepsilon(\cdot)$ to L^* or to U^* immediately if it ever drops below L^* or exceeds U^* .

From the point of view of optimal control, we wish to show that the costs $V^\varepsilon(x)$ and $V^\varepsilon(x, \bar{Y}_0^\varepsilon, \bar{Y}_1^\varepsilon)$ are close for small ε , whether or not the optimal controls for (1.1), (1.2) are of the barrier policy type. This is the class of problems dealt with here. The basis tools are those of weak convergence theory. The work here extends that of [6] and [7]. The methodologies of these references require considerable modification to be of use here.

Owing to lack of knowledge of the properties of the optimal $Y_i^\varepsilon(\cdot)$, we cannot generally prove tightness or weak convergence of $\{X^\varepsilon(\cdot), Y_0^\varepsilon(\cdot), Y_1^\varepsilon(\cdot)\}$ in the Skorokhod topology on $D^3[0, \infty)$. A weaker topology must be used, and it is described in § 2. Section 3 concerns a weak convergence result for a continuous time model, and § 4 extends it to the discrete time case. Section 5 contains some auxiliary results which are needed later. The optimal control problem for the discounted cost case is dealt with in § 6, where it is shown that the suitably adapted optimal policy for the limit is indeed “nearly” optimal for the actual physical process. Section 7 concerns the average cost per unit time problem. Here, owing to the natural requirement of stationarity, we impose a Markov structure on the problem.

The basic methods work just as well for many nonscalar models—and several extensions to such models are discussed in § 8. There are extensions of the results to cases where the dynamical terms are not smooth or the noise is state dependent. We would then adapt the weak convergence technique and assumptions of [8, Chaps. 5.3, 5.5, 5.8] to the problem here. The results also extend to the case where there is also a “nonsingular” control component $b(x, u) dt$ (as in [16]), via a combination of the methods developed here and those used in [6] for the nonsingular case). Another extension would be for the problem in [9], where the singular control problem appears as a limit of impulsive control problems: Let δ_ε denote the fixed cost per impulse and (for some $\bar{k} > 0$) $\bar{k}|Y^\varepsilon(t) - Y^\varepsilon(t^-)|$ the variable cost, for an impulse at time t . Then we can let $\delta_\varepsilon \rightarrow 0$ as the noise bandwidth goes to ∞ , to get an approximation theorem for small fixed impulsive cost and wide bandwidth simultaneously.

2. The pseudopath topology. In this section, we discuss the topology on $D[0, \infty)$ (replacing the Skorokhod topology) which will allow us to obtain the desired weak convergence results.

Let the physical system be modeled by

$$X^\varepsilon(t) = Z^\varepsilon(t) + Y^\varepsilon(t)$$

where $X^\varepsilon(0) = x$, $Z^\varepsilon(0) = 0$, $Y^\varepsilon(\cdot) = Y_0^\varepsilon(\cdot) - Y_1^\varepsilon(\cdot)$ and define $\bar{X}^\varepsilon(\cdot) = (X^\varepsilon(\cdot), Y_0^\varepsilon(\cdot), Y_1^\varepsilon(\cdot), Z^\varepsilon(\cdot))$. The $\bar{X}^\varepsilon(\cdot)$ can be viewed either as a continuous parameter interpolation of a discrete parameter system as discussed in § 1, or it might be an actual physical model for a continuous parameter system. We always take the paths of $X^\varepsilon(\cdot)$, $Z^\varepsilon(\cdot)$ and $Y_i^\varepsilon(\cdot)$ to be right continuous and $Y^\varepsilon(\cdot)$ nondecreasing.

We suppose that the $\bar{X}^\varepsilon(\cdot)$ take values in $D^4[0, \infty)$, the space of R^4 -valued functions which are right continuous and have left-hand limits. The appropriate topology for our purposes on $D^4[0, \infty)$ is what Dellacherie and Meyer [10] and Meyer and Zheng [11] call the *pseudopath topology*. For completeness, we state some definitions and results from [11] which will be needed in the sequel. The results are stated for a real-valued process, but the natural extensions for the R^r -valued case should be obvious and are used below.

Let $y(\cdot) \in D[0, \infty)$ and define the measure $\lambda(\cdot)$ on the Borel subsets of $[0, \infty)$ by $\lambda(dt) = e^{-t} dt$. Let \bar{P} denote the compact space of probability measures (with the weak topology) on the compactified space $[0, \infty] \times \bar{R}$, where \bar{R} is the closure of the real line. The *pseudopath* of $y(\cdot)$ is defined to be the probability measure on the Borel subsets

of $[0, \infty) \times \bar{R}$ which is the image of $\lambda(\cdot)$ under the map $t \rightarrow (t, y(t))$ of $[0, \infty]$ into $[0, \infty) \times \bar{R}$ (i.e., it is a point in \bar{P}). Let ψ denote the map which takes $y(\cdot)$ into its pseudopath, the corresponding point in \bar{P} . If we write $P = \psi(y(\cdot))$, then the pseudopath P is the measure defined by $P(A \times B) = \int_A e^{-t} I_{\{y(t) \in B\}} dt$, where A is a Borel subset of $[0, \infty]$ and B is a Borel subset of \bar{R} .

ψ is 1:1 on $D[0, \infty)$, since it identifies all paths which are equal almost everywhere (Lebesgue measure). The topology which \bar{P} induces on $D[0, \infty)$ via ψ is called the *pseudopath topology*. The associated σ -algebra on $D[0, \infty)$ is the same as we get with the Skorokhod topology. In fact [11, Lemma 1 and comment after its proof], the *pseudopath topology* on $D[0, \infty)$ is the *topology of convergence in measure*. The last assertion remains true if R and $D[0, \infty)$ are replaced by R' and $D'[0, \infty)$, where ψ then maps points $y(\cdot) \in D'[0, \infty)$ into a measure on the Borel subsets of $[0, \infty) \times \bar{R}'$. Let \bar{P}_r denote the space of probability measures on the Borel subsets of $[0, \infty) \times \bar{R}'$.

The process $\bar{X}^e(\cdot)$ induces a measure (which we denote by \bar{P}_e) on \bar{P}_4 via the pseudopath mapping ψ . The set $\{\bar{P}_e\}$ is obviously tight since \bar{P}_4 is compact. If \bar{P} is a limit measure of any weakly convergent subsequence, then for the convergence to be useful we need at least that \bar{P} be supported by $\psi(D^4[0, \infty))$, since then the limit \bar{P} would correspond to some process $\bar{X}(\cdot) = (X(\cdot), Y_0(\cdot), Y_1(\cdot), Z(\cdot))$ with paths in $D^4[0, \infty)$, via the mapping ψ . A convenient criterion for this is given by Meyer and Zheng [11], and will now be described.

Let τ denote a finite partition $\{t_i, i \leq n\}$: $0 = t_0 < t_1 < \dots < t_n = \infty$. Let $U(\cdot)$ denote a process with paths in $D[0, \infty)$ and adapted to a nondecreasing sequence of σ -algebras $\{\mathcal{F}_t\}$, and with $E|U(t)| < \infty$ for each $t < \infty$. For convenience in comparing with [11], let $U(t) = 0$ for large t . Define the conditional variations

$$\begin{aligned} \text{var}_\tau(U) &= \sum_{i \leq n} E|E_{\mathcal{F}_{t_i}} U(t_{i+1}) - U(t_i)|, \\ (2.1) \quad \text{var}(U) &= \sup_\tau \text{var}_\tau(U). \end{aligned}$$

If $\text{var}(U) < \infty$, then $U(\cdot)$ is said to be a *quasimartingale*.

For $u < v$, let $N^{u,v}(U)$ denote the number of upcrossings of $U(\cdot)$ on $[0, \infty)$ between the levels u and v . If $U(\cdot)$ is a quasimartingale, then

$$(2.2) \quad EN^{uv}(U) \leq \frac{|u| + \text{var}(U)}{v - u},$$

an extension of the usual result for martingales [11, Lemma 3]. The main result is [11, Thm. 4].

THEOREM 2.1. *For each $n = 1, 2, \dots$, let P_n be a probability law on the Borel¹ subsets (with the pseudopath topology) of $D[0, \infty)$ with the associated process $U_n(\cdot)$ being a quasimartingale with $\sup_n \text{var}(U_n) < \infty$. Then there is a subsequence $\{P_{n_k}\}$ of $\{P_n\}$ which converges weakly on $D[0, \infty)$ (with the pseudopath topology) to a law P , and the associated process $U(\cdot)$ is a quasimartingale. (Alternatively, let \bar{P}_n be the measure induced on \bar{P} by the map ψ acting on $U_n(\cdot)$. Then $\{\bar{P}_n\}$ is tight on \bar{P} and there is a weakly convergent subsequence $\{\bar{P}_{n_k}\}$ with limit denoted by \bar{P} . \bar{P} is supported on $D[0, \infty)$ and the associated process is a quasimartingale.)*

Combining this with the previous results, we have the following theorem.

THEOREM 2.2. *Assume the conditions and terminology of Theorem 2.1 and let $h(\cdot)$ be any bounded real valued function on $D[0, \infty)$ which is continuous (with probability 1*

¹ The Skorokhod topology and the pseudopath topology generate the same σ -algebra on $D[0, \infty)$.

with respect to P) when the topology of convergence in measure is used on $D[0, \infty)$. Then there is a subsequence $\{n_k\}$ of the integers such that $Eh(U_{n_k}(\cdot)) \rightarrow Eh(U(\cdot))$. Also [11, Thm. 5] there is a further subsequence $\{m_k\} \subset \{n_k\}$ and a set I of full measure (depending on P) such that the finite-dimensional distributions of $\{U_m(t), t \in I\}$ converge to those of $\{U(t), t \in I\}$. Let $f(\cdot)$ be bounded and continuous on $[0, \infty)$. Then [11, Thm. 6] the function $(t_1, \dots, t_q) \rightarrow Ef(U_{n_k}(t_1), \dots, U_{n_k}(t_q))$ converges in measure to the function $(t_1, \dots, t_q) \rightarrow Ef(U(t_1), \dots, U(t_q))$.

Notation. To simplify the exposition, we often abuse terminology and (in the terminology of Theorem 2.1) speak of the $\{U_n\}$ (rather than the associated measures) as being tight or converging weakly to some $U(\cdot)$ on $D[0, \infty)$ in the pseudopath topology.

3. The quasimartingale property and weak convergence of $\bar{X}^\varepsilon(\cdot) = (X^\varepsilon(\cdot), Z^\varepsilon(\cdot), Y_0^\varepsilon(\cdot), Y_1^\varepsilon(\cdot))$. A continuous parameter case will be discussed in this section. The discrete parameter case requires only a few modifications and is discussed in the next section. For concreteness, a specific model which is of a widely used form [8], [12], [13] for representing wide bandwidth noise driven systems will be treated. The techniques are usable for a much broader class of systems—just as for the case where $Y_i^\varepsilon(\cdot) \equiv 0$ dealt with in [6] or the various continuous parameter models in [8]. The model to be used is

$$(3.1) \quad dX^\varepsilon = G(X^\varepsilon, \xi^\varepsilon) dt + F(X^\varepsilon, \xi^\varepsilon) dt/\varepsilon + dY^\varepsilon(t),$$

where $\xi^\varepsilon(t) = \xi(t/\varepsilon^2)$, $\xi(\cdot)$ is a right continuous random process and $Y^\varepsilon = Y_0^\varepsilon - Y_1^\varepsilon$.

Let E_t^ε denote the expectation conditioned on $\{X^\varepsilon(s), \xi^\varepsilon(s), Y_0^\varepsilon(s), Y_1^\varepsilon(s), s \leq t\}$, and E_t the expectation conditioned on $\{\xi(s), s \leq t\}$. Define

$$Z^\varepsilon(t) \equiv \int_0^t G(X^\varepsilon(s), \xi^\varepsilon(s)) ds + \frac{1}{\varepsilon} \int_0^t F(X^\varepsilon(s), \xi^\varepsilon(s)) ds.$$

We will use the following assumptions. Various extensions (vector case, discontinuous dynamics, state dependent noise) are possible, as was discussed in the Introduction.

(A3.1) $G(\cdot, \cdot)$, $F(\cdot, \cdot)$, and $F_x(\cdot, \cdot)$ are bounded continuous functions and the latter two are continuous in x uniformly in ξ .

(A3.2) For each scalar x , $EF(x, \xi(s)) \equiv 0$ and $\xi(\cdot)$ is right continuous and sufficiently mixing such that there is a $K < \infty$ for which for each $T < \infty$

$$(3.2) \quad \sup_{x, t \leq T} \left| \int_t^T E_t g(x, \xi(s)) ds \right| \leq K \quad \text{w.p.1,}$$

where $g(\cdot, \cdot)$ represents either $F(\cdot, \cdot)$ or $F_x(\cdot, \cdot)$.

Remark. Owing to the “with probability 1” statement of (A3.2) the inequalities which use (A3.2) will hold uniformly in ω , with probability 1. In particular, the $o(\varepsilon)$, $O(\varepsilon)$ and $O(\Delta)$ of the sequel will not depend on ω .

DEFINITION. The $Y_i^\varepsilon(\cdot)$ are said to be *admissible* if for each t , the $Y_i^\varepsilon(t)$ are nondecreasing, measurable and adapted to the σ -algebra measuring $\{\xi^\varepsilon(s), s \leq t, X^\varepsilon(s), s < t\}$.

Write $Y_{ic}^\varepsilon(\cdot)$, $Y_{id}^\varepsilon(\cdot)$ and $Y_c^\varepsilon(\cdot)$, $Y_d^\varepsilon(\cdot)$ for the continuous and jump components of $Y_i^\varepsilon(\cdot)$ and $Y^\varepsilon(\cdot)$, respectively. By our convention on the right continuity of the $Y_i^\varepsilon(\cdot)$, we use the “left” differential $dY_{id}^\varepsilon(u) = Y_{id}^\varepsilon(u) - Y_{id}^\varepsilon(u^-)$ in the integrals below. The same applies for the discontinuous components of the $Y_i(\cdot)$.

THEOREM 3.1. Assume (A3.1) and (A3.2) and let $Y_i^\varepsilon(\cdot)$ be admissible with $\sup_\varepsilon [EY_0^\varepsilon(t) + EY_1^\varepsilon(t)] < \infty$ for each t . Then (possibly having to add to each component a process whose maximum value goes to zero as $\varepsilon \rightarrow 0$), $\{X^\varepsilon(\cdot), Y_0^\varepsilon(\cdot), Y_1^\varepsilon(\cdot), Z^\varepsilon(\cdot)\}$ are quasimartingales whose conditional variation is bounded uniformly in ε (with probability 1) on each bounded time interval.

Remark. We need not assume that $X^\varepsilon(t) \in [0, \bar{B}]$ in this theorem.

Proof. Since $\sup_\varepsilon EY_i^\varepsilon(t) < \infty$, the $Y_i^\varepsilon(\cdot)$ are obviously quasimartingales with uniformly (in ε) bounded conditional variation on each interval $[0, t]$. Thus, we need only work with the $Z^\varepsilon(\cdot)$. We will use the so-called perturbed test function method described in [8], [12], [14] but adapted to our present needs. For some sufficiently large T , define the process $Z_1^\varepsilon(\cdot)$ on $[0, T]$ by

$$Z_1^\varepsilon(t) = \frac{1}{\varepsilon} \int_t^T E_t^\varepsilon F(X^\varepsilon(s), \xi^\varepsilon(s)) ds = \varepsilon \int_{t/\varepsilon^2}^{T/\varepsilon^2} E_t^\varepsilon F(X^\varepsilon(s), \xi^\varepsilon(s)) ds.$$

To get the second integral, we used the change of variable $s/\varepsilon^2 \rightarrow s$, which will be used frequently in the averaging and bounding in the sequel, when working with integrals such as $z_1^\varepsilon(\cdot)$. (Perturbations such as $Z_1^\varepsilon(\cdot)$ play a basic role in the “averaging” of the noise. See the cited references.) By (A3.2)

$$\sup_{t \leq T} |Z_1^\varepsilon(t)| = O(\varepsilon).$$

We will show that the function defined by $f^\varepsilon(t) = Z^\varepsilon(t) + Z_1^\varepsilon(t)$ is a quasimartingale whose conditional variation is bounded uniformly in ε (with probability 1) on each interval $[0, T]$. The calculations will be done in a slightly indirect way so that they can be reused later. Let $f(\cdot)$ denote a “test” function with bounded and continuous derivatives up to order three, and define $f^\varepsilon(t) = f(Z^\varepsilon(t)) + f_1^\varepsilon(t)$, where $f_1^\varepsilon(\cdot)$ is the “perturbation” defined by

$$\begin{aligned} f_1^\varepsilon(t) &= \frac{1}{\varepsilon} \int_t^T f_z(Z^\varepsilon(s)) E_s^\varepsilon F(X^\varepsilon(s), \xi^\varepsilon(s)) ds \\ &= \varepsilon \int_{t/\varepsilon^2}^{T/\varepsilon^2} f_z(Z^\varepsilon(s)) E_s^\varepsilon F(X^\varepsilon(s), \xi^\varepsilon(s)) ds. \end{aligned}$$

By integrating the derivative of $f(Z^\varepsilon(\cdot))$,

$$\begin{aligned} (3.3) \quad & E_t^\varepsilon f(Z^\varepsilon(t+\Delta)) - f(Z^\varepsilon(t)) \\ &= E_t^\varepsilon \int_t^{t+\Delta} f_z(Z^\varepsilon(u)) \left[G(X^\varepsilon(u), \xi^\varepsilon(u)) + \frac{F(X^\varepsilon(u), \xi^\varepsilon(u))}{\varepsilon} \right] du. \end{aligned}$$

Similarly, by evaluating $[E_u^\varepsilon f_1^\varepsilon(u+\Delta) - f_1^\varepsilon(u)]/\Delta$ and letting $\Delta \rightarrow 0$, we get

$$\begin{aligned} (3.4) \quad & E_t^\varepsilon f_1^\varepsilon(t+\Delta) - f_1^\varepsilon(t) = -\frac{1}{\varepsilon} \int_t^{t+\Delta} f_z(Z^\varepsilon(u)) E_u^\varepsilon F(X^\varepsilon(u), \xi^\varepsilon(u)) du \\ &+ \int_t^{t+\Delta} du \frac{1}{\varepsilon} E_t^\varepsilon \int_u^T ds f_{zz}(Z^\varepsilon(u)) E_u^\varepsilon \frac{F(X^\varepsilon(u), \xi^\varepsilon(s))}{\varepsilon} \\ &\cdot \left[\frac{F(X^\varepsilon(u), \xi^\varepsilon(u))}{\varepsilon} + G(X^\varepsilon(u), \xi^\varepsilon(u)) \right] \\ &+ E_t^\varepsilon \int_t^{t+\Delta} dY_c^\varepsilon(u) \frac{1}{\varepsilon} \int_u^T f_z(Z^\varepsilon(u)) E_u^\varepsilon F_x(X^\varepsilon(u), \xi^\varepsilon(s)) ds \end{aligned}$$

$$\begin{aligned}
& + \int_t^{t+\Delta} E_t^\varepsilon \left[\frac{F(X^\varepsilon(u), \xi^\varepsilon(u))}{\varepsilon} + G(X^\varepsilon(u), \xi^\varepsilon(u)) \right] du \\
& \cdot \int_u^T E_u^\varepsilon f(Z^\varepsilon(u)) \frac{F(X^\varepsilon(u), \xi^\varepsilon(s))}{\varepsilon} ds \\
& + \sum_{t < u \leq t+\Delta} \frac{1}{\varepsilon} E_t^\varepsilon \int_u^T f_z(Z^\varepsilon(u)) E_u^\varepsilon [F(X^\varepsilon(u^-) \\
& \quad + dY_d^\varepsilon(u), \xi^\varepsilon(s)) - F(X^\varepsilon(u^-), \xi^\varepsilon(s))] ds.
\end{aligned}$$

By a change of scale $s/\varepsilon^2 \rightarrow s$ and the use of (A3.1) and (A3.2), the second and fourth terms on the right-hand side of (3.4) are seen to be $O(\Delta)$, with probability 1. By a similar scale change, the third term is seen to be $O(\varepsilon)E_t^\varepsilon(Y_c^\varepsilon(t+\Delta) - Y_c^\varepsilon(t))$, with probability 1. The first term of (3.4) is the negative of the “ $1/\varepsilon$ ” term in (3.3). For the evaluation of the last term in (3.4), first use the law of the mean to rewrite it as

$$(3.5) \quad \sum_{t < u \leq t+\Delta} \frac{1}{\varepsilon} E_t^\varepsilon \int_u^T ds f_z(Z^\varepsilon(u)) E_u^\varepsilon \int_0^1 d\tau [F_x(X^\varepsilon(u^-) + \tau dY_d^\varepsilon(u), \xi^\varepsilon(s))] dY_d^\varepsilon(u).$$

Now, by a change of scale $s/\varepsilon^2 \rightarrow s$ and the use of (A3.1) and (A3.2) again, we see that this term is $O(\varepsilon)E_t^\varepsilon[Y_d^\varepsilon(t+\Delta) - Y_d^\varepsilon(t)]$, with probability 1.

Putting all the estimates together and canceling the “ $1/\varepsilon$ ” term on the right-hand side of (3.3) and the first term on the right-hand side of (3.4), we get (with probability 1)

$$(3.6) \quad \begin{aligned} E_t^\varepsilon f^\varepsilon(t+\Delta) - f^\varepsilon(t) &= O(\Delta) + O(\varepsilon)E_t^\varepsilon(Y_c^\varepsilon(t+\Delta) - Y_c^\varepsilon(t)) \\ &\quad + O(\varepsilon)E_t^\varepsilon(Y_d^\varepsilon(t+\Delta) - Y_d^\varepsilon(t)). \end{aligned}$$

Equation (3.6) yields the quasimartingale and the uniformly (in ε) bounded conditional variation properties on each interval $[0, T]$ for $f^\varepsilon(\cdot)$. By letting $f(z) = z$ and noting that $z_1^\varepsilon(t) = O(\varepsilon)$, with probability 1, we see that the theorem holds for the $\{Z^\varepsilon(\cdot)\}$ component. Hence, it also holds for $\{X^\varepsilon(\cdot)\}$, since $\sup_\varepsilon E(Y_0^\varepsilon(t) + Y_1^\varepsilon(t)) < \infty$ for each t and

$$X^\varepsilon(t) = (Z^\varepsilon(t) + Z_1^\varepsilon(t)) + Y^\varepsilon(t) - Z_1^\varepsilon(t). \quad \text{Q.E.D.}$$

We summarize (3.3) to (3.6) for future use:

$$\begin{aligned}
(3.7) \quad & E_t^\varepsilon f^\varepsilon(t+\Delta) - f^\varepsilon(t) \\
&= \int_t^{t+\Delta} E_t^\varepsilon f_z(Z^\varepsilon(u)) G(X^\varepsilon(u), \xi^\varepsilon(u)) du \\
&\quad + \frac{1}{\varepsilon^2} E_t^\varepsilon \int_t^{t+\Delta} du \int_u^T f_{zz}(Z^\varepsilon(u)) E_u^\varepsilon F(X^\varepsilon(u), \xi^\varepsilon(s)) ds F(X^\varepsilon(u), \xi^\varepsilon(u)) \\
&\quad + \frac{1}{\varepsilon^2} E_t^\varepsilon \int_t^{t+\Delta} du E_u^\varepsilon \int_u^T f_z(Z^\varepsilon(u)) F_x(X^\varepsilon(u), \xi^\varepsilon(s)) ds F(X^\varepsilon(u), \xi^\varepsilon(u)) \\
&\quad + O(\Delta) + O(\varepsilon)E_t^\varepsilon(Y_c^\varepsilon(t+\Delta) - Y_c^\varepsilon(t)) + O(\varepsilon)E_t^\varepsilon(Y_d^\varepsilon(t+\Delta) - Y_d^\varepsilon(t)).
\end{aligned}$$

Theorem 3.1 implies that $\{X^\varepsilon(\cdot) - Z_1^\varepsilon, Y_0^\varepsilon(\cdot), Y_1^\varepsilon(\cdot), Z^\varepsilon(\cdot) + Z_1^\varepsilon(\cdot)\}$ are quasimartingales with uniformly bounded conditional variation on each interval $[0, T]$ and their probability laws are tight on $D^4[0, \infty]$ in the pseudopath topology. Hence, the same tightness in the pseudopath topology on $D^4[0, \infty]$ holds for the laws of $\{\bar{X}^\varepsilon(\cdot)\}$. In the next theorem, we choose and work with a weakly convergent subsequence, also indexed by ε and with the limit denoted by $\bar{X}(\cdot) = (X(\cdot), Y_0(\cdot), Y_1(\cdot), Z(\cdot))$. Clearly, the sample functions $Y_i(\cdot)$ can be taken to be nondecreasing elements of $D[0, \infty)$.

Although $Y_i^\varepsilon(0) = 0$, the limits $Y_i(\cdot)$ might not have value zero at $t = 0$. To account for this possible jump in the integrals, we use the normalization $Y_i(0^-) = 0$ in defining integrals with respect to the $Y_i(\cdot)$, so that $dY_i(0)$ includes this possible jump.

Since the pseudopath topology is equivalent to convergence in measure, for almost all ω, t ,

$$(3.8) \quad X(t) = Z(t) + Y_0(t) - Y_1(t).$$

In fact, (3.8) holds also at all t at which the functions are continuous. In the next theorem, we obtain the stronger and more useful result that there is a Wiener process $w(\cdot)$ such that $\bar{X}(\cdot)$ is nonanticipative with respect to $w(\cdot)$ and $(X(\cdot), Y_0(\cdot), Y_1(\cdot))$ satisfies (1.1) for that $w(\cdot)$. The limits $Y_i(\cdot)$ would not be too useful were this not the case. We impose the following "ergodic" assumptions.

(A3.3) There is a continuous function $\bar{G}(\cdot)$ such that for each x ,

$$\frac{1}{N} \int_u^{u+N} E_u G(x, \xi(s)) ds \xrightarrow{P} \bar{G}(x)$$

as u and N go to ∞ .

(A3.4) For g equal to either F or F_x and $T > u + N$,

$$E \sup_x \left| \int_{u+N}^T E_u g(x, \xi(s)) ds \right| \rightarrow 0,$$

as u, N, T go to ∞ .

(A3.5) There is a continuous function $\sigma(\cdot)$ such that for each x

$$\frac{1}{T_1} \int_u^{u+T_1} E_u F(x, \xi(\tau)) d\tau \int_\tau^{T+\tau} F(x, \xi(s)) ds \xrightarrow{P} \sigma^2(x)/2$$

as T, u and T_1 go to ∞ . Also, there is a continuous function $\bar{G}_0(\cdot)$ such that for each x

$$\frac{1}{T_1} \int_u^{u+T_1} E_u F(x, \xi(\tau)) d\tau \int_\tau^{T+\tau} F_x(x, \xi(s)) ds \xrightarrow{P} \bar{G}_0(x).$$

Remark. If $\xi(\cdot)$ is stationary, then

$$\sigma^2(x) = \int_{-\infty}^{\infty} EF(x, \xi(0))F(x, \xi(s)) ds,$$

$$\bar{G}_0(x) = \int_0^{\infty} EF(x, \xi(0))F_x(x, \xi(s)) ds.$$

The requirements in (A3.4) and (A3.5) are simply conditions on the rate of convergence as $\tau - u \rightarrow \infty$ of the conditional expectation of functions $g(\tau)$, of the noise data after time τ , given the data up to time u .

THEOREM 3.2. Assume (A3.1)–(A3.5) and let $\sup_\varepsilon E_x[Y_0^\varepsilon(t) + Y_1^\varepsilon(t)] < \infty$ for each $t < \infty$. Then $\{Z^\varepsilon(\cdot)\}$ is tight in the Skorokhod topology on $D[0, \infty)$ and any weak limit process is continuous w.p.1. Let $\{X^{\varepsilon_n}(\cdot), Y_0^{\varepsilon_n}(\cdot), Y_1^{\varepsilon_n}(\cdot), Z^{\varepsilon_n}(\cdot)\}$ converge weakly in $D^4[0, \infty)$, where we use the product topology obtained by using the pseudopath topology on the first three components and the Skorokhod topology on the last. Denote the limit

by $\bar{X}(\cdot) = (X(\cdot), Y_0(\cdot), Y_1(\cdot), Z(\cdot))$. There is a standard Wiener process $w(\cdot)$ such that $\bar{X}(\cdot)$ is nonanticipative with respect to $w(\cdot)$ and with probability 1

$$(3.9) \quad Z(t) = \int_0^t \bar{G}(X(s)) ds + \int_0^t \bar{G}_0(X(s)) ds + \int_0^t \sigma(X(s)) dw(s).$$

Also, for all t , with probability 1,

$$(3.10) \quad X(t) = Z(t) + Y_0(t) - Y_1(t).$$

Remark. We need not require that $X^\varepsilon(t) \in [0, \bar{B}]$ in this theorem. When a convergent subsequence indexed by ε is referred to below, we let ε range over a countable set.

Proof. We will do the proof under the assumption that the x -support of all functions is compact, and that the $Y_i^\varepsilon(\cdot)$ are uniformly bounded. The general case follows from this by taking appropriate limits on the bounds. This assumption implies that the $Z^\varepsilon(\cdot)$ are uniformly bounded. Hence we can also assume that the z -support of all functions is compact.

(a) *Tightness of $\{Z^\varepsilon(\cdot)\}$ in the Skorokhod topology.* Let $f(\cdot)$, $f_1^\varepsilon(\cdot)$ and $f^\varepsilon(\cdot)$ be defined as in Theorem 3.1. We use the perturbed test function method of [8] or [14] for proving tightness. Since $f_1^\varepsilon(t) = O(\varepsilon)$, with probability 1, Theorem 3.4 of [8] or Lemma 1 of [14] applied to the perturbed test function $f^\varepsilon(\cdot)$ yields the tightness of $\{f(Z^\varepsilon(\cdot))\}$ in $D[0, \infty)$ with the Skorokhod topology used, for each smooth $f(\cdot)$. Hence $\{Z^\varepsilon(\cdot)\}$ is tight on $D[0, \infty)$ in the Skorokhod topology.

(b) *The limit of $\{Z^\varepsilon(\cdot)\}$.* We will show that $Z(\cdot)$ is the solution to a particular martingale problem. Fix and work with a weakly convergent subsequence of $\{X^\varepsilon(\cdot), Y_0^\varepsilon(\cdot), Y_1^\varepsilon(\cdot), Z^\varepsilon(\cdot)\}$, also indexed by ε for simplicity (and not by ε_n). The first three components converge in the pseudopath topology and the last in the Skorokhod topology. Let $h(\cdot)$ be a bounded and continuous function. For $0 < k < \infty$ and $\Delta_j > 0$, $t_j \geq 0$, let us define the function $H(\cdot)$ with "averaged" arguments by

$$H(\varepsilon, \Delta_j, t_j, j \leq k) = h \left(\frac{1}{\Delta_j} \int_{t_j - \Delta_j}^{t_j} Y_i^\varepsilon(s) ds, i=0,1, \frac{1}{\Delta_j} \int_{t_j - \Delta_j}^{t_j} X^\varepsilon(s) ds, \frac{1}{\Delta_j} \int_{t_j - \Delta_j}^{t_j} Z^\varepsilon(s) ds, j \leq k \right).$$

Let t and s be such that $t_j \leq t < t + s$, for all j . We have from (3.7)

$$(3.11) \quad \lim_{\varepsilon} EH(\varepsilon, \Delta_j, t_j, j \leq k) \left[f^\varepsilon(t+s) - f^\varepsilon(t) - \int_t^{t+s} E_t^\varepsilon(T_1^\varepsilon(u) + T_2^\varepsilon(u) + T_3^\varepsilon(u)) du \right] = 0,$$

where

$$(3.12) \quad \begin{aligned} T_1^\varepsilon(u) &= f_z(Z^\varepsilon(u))G(X^\varepsilon(u), \xi^\varepsilon(u)), \\ T_2^\varepsilon(u) &= \frac{1}{\varepsilon^2} \int_u^T f_z(Z^\varepsilon(u))E_u^\varepsilon F_x(X^\varepsilon(u), \xi^\varepsilon(s)) ds F(X^\varepsilon(u), \xi^\varepsilon(u)), \\ T_3^\varepsilon(u) &= \frac{1}{\varepsilon^2} \int_u^T f_{zz}(Z^\varepsilon(u))E_u^\varepsilon F(X^\varepsilon(u), \xi^\varepsilon(s)) ds F(X^\varepsilon(u), \xi^\varepsilon(u)). \end{aligned}$$

Fix s . Let $\delta_\varepsilon \rightarrow 0$ such that $\varepsilon^2/\delta_\varepsilon \rightarrow 0$ as $\varepsilon \rightarrow 0$. Write $s = m_\varepsilon \delta_\varepsilon$ and suppose (without loss of generality) that the m_ε are integers. The limits of the terms in (3.11) are the same if the $f_1^\varepsilon(\cdot)$ components of the $f^\varepsilon(\cdot)$ are dropped (since they are $O(\varepsilon)$ with probability 1). We next replace the integrals $\int_t^{t+s} E_t^\varepsilon T_i^\varepsilon(u) du$ by the equivalent expression

$$(3.13) \quad E_t^\varepsilon \sum_{j=0}^{m_\varepsilon-1} \frac{1}{\delta_\varepsilon} \cdot \delta_\varepsilon \int_{t+j\delta_\varepsilon}^{t+(j+1)\delta_\varepsilon} E_{t+j\delta_\varepsilon}^\varepsilon T_i^\varepsilon(u) du.$$

We evaluate only the limit of (3.13) with $T_3^\varepsilon(\cdot)$, since the others are treated in essentially the same way.

The limit as $\varepsilon \rightarrow 0$ will be obtained by a sequence of substitutions, each being more convenient to evaluate. In the substitutions, we use the scale change, the ergodic assumptions (A3.3)–(A3.5), the tightness of $X^\varepsilon(\cdot)$ in the pseudopath topology (and the consequent upcrossing bounds), and the tightness of $Z^\varepsilon(\cdot)$ in the Skorokhod topology.

Rewrite (3.13) (with $i = 3$) as

$$(3.14) \quad E_t^\varepsilon \sum_{j=0}^{m_\varepsilon-1} \delta_\varepsilon \cdot \frac{1}{\delta_\varepsilon \cdot \varepsilon^2} \int_{t+j\delta_\varepsilon}^{t+j\delta_\varepsilon+\delta_\varepsilon} E_{t+j\delta_\varepsilon}^\varepsilon f_{zz}(Z^\varepsilon(u)) du \cdot \int_0^T F(X^\varepsilon(u), \xi^\varepsilon(s)) ds F(X^\varepsilon(u), \xi^\varepsilon(u)).$$

Via a scale change, write (3.14) as

$$(3.15) \quad E_t^\varepsilon \sum_{j=0}^{m_\varepsilon-1} \delta_\varepsilon \cdot \frac{\varepsilon^2}{\delta_\varepsilon} \int_{s_j^\varepsilon}^{s_{j+1}^\varepsilon} E_{t+j\delta_\varepsilon}^\varepsilon f_{zz}(Z^\varepsilon(\varepsilon^2 u)) F(X^\varepsilon(\varepsilon^2 u), \xi(u)) du \cdot \int_u^{T/\varepsilon^2} F(X^\varepsilon(\varepsilon^2 u), \xi(s)) ds,$$

where $s_j^\varepsilon = (t - j\delta_\varepsilon)/\varepsilon^2$. Note that $E_{t+j\delta_\varepsilon}^\varepsilon$ is the expectation conditioned on $\{Y_i^\varepsilon(s), i = 0, 1, s \leq t + j\delta_\varepsilon, \xi(s), s \leq s_j^\varepsilon\}$.

Let $\{B_i^\delta\}$ be disjoint intervals covering the range of $X^\varepsilon(\cdot)$ and with diameter less than $\delta > 0$, and let x_i denote an arbitrary point in B_i^δ . Recall that $X^\varepsilon(\cdot) + Z_1^\varepsilon(\cdot)$ is a quasimartingale with uniformly bounded (with probability 1) conditional variation. Thus the uncrossings result (2.2), and the fact that $Z_1^\varepsilon(\cdot) = O(\varepsilon)$, w.p.1 imply that the fraction of the number of intervals in the set of intervals $\{[t + j\delta_\varepsilon, t + j\delta_\varepsilon + \delta_\varepsilon], j \leq m_\varepsilon\}$ for which $\sup_{u \leq \delta_\varepsilon} |X^\varepsilon(t + j\delta_\varepsilon + u) - X^\varepsilon(t + j\delta_\varepsilon)| \geq \delta/2$ holds goes to zero in probability as $\varepsilon \rightarrow 0$. The facts above in this paragraph, (A3.4) and the tightness of $Z^\varepsilon(\cdot)$ in the Skorokhod topology imply that the limit of (3.16) as $\varepsilon \rightarrow \infty$ and then $\delta \rightarrow 0$ is the same as the limit of (3.15) as $\varepsilon \rightarrow 0$:

$$(3.16) \quad E_t^\varepsilon \sum_{j=0}^{m_\varepsilon-1} \delta_\varepsilon \sum_i I_{\{X^\varepsilon(t+j\delta_\varepsilon) \in B_i^\delta\}} \cdot \frac{\varepsilon^2}{\delta_\varepsilon} \int_{s_j^\varepsilon}^{s_{j+1}^\varepsilon} du f_{zz}(Z^\varepsilon(t + j\delta_\varepsilon)) E_{t+j\delta_\varepsilon}^\varepsilon F(x_i, \xi(u)) \int_u^{T/\varepsilon^2} F(x_i, \xi(s)) ds.$$

Assumption (A3.5) implies that the limits of (3.17) as $\varepsilon \rightarrow 0$ and then $\delta \rightarrow 0$ are the same as those of (3.16):

$$(3.17) \quad E_t^\varepsilon \sum_{j=0}^{m_\varepsilon-1} \delta_\varepsilon \sum_i I_{\{X^\varepsilon(t+j\delta_\varepsilon) \in B_i^\delta\}} \frac{\sigma^2(x_i)}{2} f_{zz}(Z^\varepsilon(t + j\delta_\varepsilon)).$$

The upcrossing bounds for $X^\varepsilon(\cdot)$ and the tightness of $Z^\varepsilon(\cdot)$ in the Skorokhod topology imply that the limit of (3.18) as $\varepsilon \rightarrow 0$ is also the limit of (3.17), as $\varepsilon \rightarrow 0$, then $\delta \rightarrow 0$:

$$(3.18) \quad E_t^\varepsilon \int_t^{t+s} \frac{\sigma^2(X^\varepsilon(u))}{2} f_{zz}(Z^\varepsilon(u)) du.$$

Define the operator $A(X)$ (acting on twice differentiable functions with compact support) by

$$(3.19) \quad A(x)f(z) = f_z(z)\bar{G}(x) + f_z(z)\bar{G}_0(x) + f_{zz}(z)\sigma^2(x)/2.$$

Repeating the procedure leading to (3.18) for the terms in (3.13) with $T_1^\varepsilon(\cdot)$ and $T_2^\varepsilon(\cdot)$ yields

$$(3.20) \quad \lim_{\varepsilon} EH(\varepsilon, \Delta_j, t_j, j \leq k) \left[f(Z^\varepsilon(t+s)) - f(Z^\varepsilon(t)) - \int_t^{t+s} A(X^\varepsilon(u))f(Z^\varepsilon(u)) du \right] = 0.$$

The sequence $Z^\varepsilon(\cdot)$ converges in the Skorokhod topology on $D[0, \infty)$, but we have not yet proved the continuity of the limit $Z(\cdot)$. Thus, there might be a set (at most countable) \hat{I} of t -points for which $P\{Z(t) \neq Z(t^-)\} > 0$. Let t and $t+s$ not be in \hat{I} . Recall that $\{X^\varepsilon(\cdot), Y_0^\varepsilon(\cdot), Y_1^\varepsilon(\cdot)\}$ converge in the pseudopath topology, and that convergence in this topology is equivalent to convergence in measure. Thus, taking limits in (3.20), we have

$$(3.21) \quad EH(\Delta_j, t_j, j \leq k) \left[f(Z(t+s)) - f(Z(t)) - \int_t^{t+s} A(X(u))f(Z(u)) du \right] = 0$$

where the function $H(\Delta_j, t_j, j \leq k)$ is defined to be just $H(\varepsilon, \Delta_j, t_j, j \leq k)$ with all arguments replaced by their limits as $\varepsilon \rightarrow 0$.

Owing to the arbitrariness of k, t_j, Δ_j and $h(\cdot)$ and of the points s and $t+s$ (not in \hat{I}), (3.21) implies that for each smooth $f(\cdot)$, the process defined by

$$f(Z(t)) - \int_0^t A(X(u))f(Z(u)) du \equiv M_f(t)$$

is a martingale with respect to the sequence of σ -algebras generated by $\{X(s), Y_0(s), Y_1(s), Z(s), s \leq t\}$. The fact that the operator $A(x)$ is "local" implies the continuity of $Z(\cdot)$. (See a proof of a related continuity result in [8], [14].)

If $f(z) = z$, then the quadratic variation of $M_f(\cdot)$ is $\int_0^t \sigma^2(X(u)) du$. Owing to these facts we can construct a standard Wiener process $w(\cdot)$ such that $X(\cdot), Z(\cdot), Y_0(\cdot)$ and $Y_1(\cdot)$ are nonanticipative with respect to $w(\cdot)$ and

$$(3.22) \quad Z(t) = \int_0^t [\bar{G}(X(u)) + \bar{G}_0(X(u))] du + \int_0^t \sigma(X(u)) dw(u).$$

It follows from the continuity of $Z(\cdot)$ and the nondecreasing property of the $Y_i(\cdot)$ that we can define the limit $X(\cdot)$ of $\{X^\varepsilon(\cdot)\}$ by (3.10). Q.E.D.

4. The discrete parameter problem. The discrete parameter analogue of Theorems 3.1 and 3.2 is obtained in a way very similar to the schemes used in those theorems, and we discuss only a few of the details for one discrete parameter form. Just as for the continuous parameter case, the general ideas are applicable to a much broader class of processes than used here. Define $\{X_n^\varepsilon\}$ by $X_0^\varepsilon = x$ and

$$(4.1) \quad X_{n+1}^\varepsilon = X_n^\varepsilon + \varepsilon G(X_n^\varepsilon, \xi_n^\varepsilon) + \sqrt{\varepsilon} F(X_n^\varepsilon, \xi_n^\varepsilon) + \delta Y_n^\varepsilon,$$

where we define $\delta Y_n^\varepsilon = \delta Y_{0n}^\varepsilon - \delta Y_{1n}^\varepsilon$ and $\delta Y_{in}^\varepsilon \geq 0$. Let E_n^ε denote the expectation conditioned on $\{X_j^\varepsilon, j \leq n, \delta Y_{0j}^\varepsilon, \delta Y_{1j}^\varepsilon, \xi_j^\varepsilon, j < n\}$. Define the processes $Y_i^\varepsilon(\cdot)$ by $Y_i^\varepsilon(t) = \sum_{j=0}^{n-1} \delta Y_{ij}^\varepsilon$, $i=0, 1$, and $X^\varepsilon(t) = X_n^\varepsilon$ for $t \in [n\varepsilon, (n+1)\varepsilon)$. Define $Z_0^\varepsilon = 0$ and $Z_{n+1}^\varepsilon = Z_n^\varepsilon + \varepsilon G(X_n^\varepsilon, \xi_n^\varepsilon) + \sqrt{\varepsilon} F(X_n^\varepsilon, \xi_n^\varepsilon)$. Let $Z^\varepsilon(\cdot)$ denote the continuous parameter interpolation. We will use the following.

(A4.1) $\sup_{\varepsilon} E(Y_0^\varepsilon(t) + Y_1^\varepsilon(t)) < \infty$ for each t . $G(\cdot, \cdot)$, $F(\cdot, \cdot)$ and $F_x(\cdot, \cdot)$ are bounded and measurable and the latter two functions are continuous in x , uniformly in ξ .

(A4.2) For each x , $EF(x, \xi_n^e) = 0$. There is a $K < \infty$ such that for all N and $n \leq N$,

$$(4.2) \quad \sup_{n, \varepsilon, x} \left| \sum_{j=n}^N E_n^e g(x, \xi_j^e) \right| \leq K \quad \text{with probability 1,}$$

where g equals either F or F_x .

THEOREM 4.1. Assume (A4.1) and (A4.2). Then (possibly changing their values by at most $O(\sqrt{\varepsilon})$) $\{X^\varepsilon(\cdot), Y_i^\varepsilon(\cdot), i = 0, 1\}$ is a quasimartingale with conditional variation uniformly bounded (with probability 1) in ε on each interval $[0, T]$.

Remarks on the proof. The proof is very similar to that of Theorem 3.1, and only a few remarks will be made. For a smooth test function $f(\cdot)$ and sufficiently large N , define the "perturbation"

$$f_{1n}^\varepsilon = \sqrt{\varepsilon} \sum_{j=n}^N E_n^e f_z(Z_n^\varepsilon) F(X_n^\varepsilon, \xi_j^\varepsilon) = O(\sqrt{\varepsilon}) \quad \text{with probability 1.}$$

Define $f_n^\varepsilon = f(Z_n^\varepsilon) + f_{1n}^\varepsilon$.

It can be shown that

$$E_n^\varepsilon f_{n+1}^\varepsilon - f_n^\varepsilon = O(\varepsilon) + O(\sqrt{\varepsilon}) E_n^\varepsilon |\delta Y_n^\varepsilon|.$$

Letting $f(z) = z$ yields the desired conditional variation result, since $f_{1n}^\varepsilon = O(\sqrt{\varepsilon})$ and $\sup_\varepsilon E(Y_0^\varepsilon(T) + Y_1^\varepsilon(T)) < \infty$.

Theorem 3.2 can also be carried over to the discrete parameter case. We will use the following conditions.

(A4.3) $G(\cdot, \xi)$ is continuous in x , uniformly in ξ . There is a continuous $\bar{G}(\cdot)$ such that for each x

$$\frac{1}{N} \sum_n^{n+N} E_n^\varepsilon G(x, \xi_j^\varepsilon) \xrightarrow{P} \bar{G}(x)$$

as n and N go to ∞ .

(A4.4) There are continuous $R(j, x)$ and $R_0(j, x)$ such that for each x

$$\begin{aligned} \frac{1}{N} \sum_{n=m}^{m+N} E_m^\varepsilon F(x, \xi_{n+j}^\varepsilon) F(x, \xi_n^\varepsilon) &\xrightarrow{P} R(j, x), \\ \frac{1}{N} \sum_{n=m}^{m+N} E_m^\varepsilon F_x(x, \xi_{n+j}^\varepsilon) F(x, \xi_n^\varepsilon) &\xrightarrow{P} R_0(j, x), \end{aligned}$$

as m, N and $n - m$ go to ∞ .

(A4.5) For g equal to either F or F_x ,

$$E \sup_x \left| \sum_{n=N_1}^N E_n^\varepsilon g(x, \xi_j^\varepsilon) \right| \rightarrow 0$$

as N, n and N_1 go to ∞ (with $N > n + N_1$).

Define

$$\sigma^2(x) = R(0, x) + 2 \sum_1^\infty R(j, x) = \sum_{-\infty}^\infty R(j, x),$$

$$\bar{G}_0(x) = \sum_1^\infty R_0(j, x).$$

A proof parallel to that of Theorem 3.2 yields Theorem 4.2.

THEOREM 4.2. Assume (A4.1)–(A4.5). Then the conclusions of Theorem 3.2 hold for the model (4.1).

5. Auxiliary results. In this section, we obtain some estimates which will be useful in § 6, for the proofs of the convergence of the costs $V_0^\varepsilon(x, Y_0^\varepsilon, Y_1^\varepsilon)$ to either $V_0(x, Y_0, Y_1)$ or $V(x)$. We will show, for several reasonable classes of control policies, that $\sup_\varepsilon E|Y_i^\varepsilon(t)|^k < \infty$ for each $k > 0$ and $t < \infty$. This implies the uniform integrability property needed in the next section.

The symbol τ_ε will denote a stopping time with respect to either of the “data” σ -algebras $B\{\xi^\varepsilon(s), s \leq t\} \equiv B_t^\varepsilon$ or $B\{\xi_n^\varepsilon, \varepsilon n < t\} \equiv B_t^\varepsilon$ depending on the case, and we write $E_{\tau_\varepsilon}^\varepsilon$ and $P_{\tau_\varepsilon}^\varepsilon$ for the expectation and probability, conditioned on the data up to time τ_ε .

THEOREM 5.1. Assume either (A3.1), (A3.2) or (A4.1), (A4.2). Let $Q_\varepsilon(\cdot)$ and $Q_{\varepsilon n}$ be bounded and B_t^ε measurable (for $\varepsilon n < t$, in the latter case). Define $X^\varepsilon(\cdot)$ and X_n^ε by

$$(5.1a) \quad dX^\varepsilon = [G(X^\varepsilon, \xi^\varepsilon) + F(X^\varepsilon, \xi^\varepsilon)/\varepsilon + Q_\varepsilon/\varepsilon] dt,$$

$$(5.1b) \quad X_{n+1}^\varepsilon = X_n^\varepsilon + \varepsilon G(X_n^\varepsilon, \xi_n^\varepsilon) + \sqrt{\varepsilon} F(X_n^\varepsilon, \xi_n^\varepsilon) + \sqrt{\varepsilon} Q_{\varepsilon n}.$$

Define $Z^\varepsilon(\cdot)$ as in § 3 or 4 (continuous and discrete parameters case, respectively). For integer k and $t < \infty$, there are $\infty > K_k(t) \rightarrow 0$ as $t \rightarrow 0$ such that (for small $\varepsilon > 0$)

$$(5.2) \quad E \sup_{s \leq t} |Z^\varepsilon(\tau_\varepsilon + s) - Z^\varepsilon(\tau_\varepsilon)|^{2k} \leq K_{2k}(t)$$

for all finite (with probability 1) τ_ε .

Remark. The proof will treat only the continuous parameter case, since the “discrete parameter” details are similar.

Proof. A perturbed test function method will be used. For arbitrary $T < \infty$ and $t \leq T$, define the processes

$$(5.2') \quad f_{2k}^\varepsilon(t) = \int_t^T 2kZ^\varepsilon(t)^{2k-1} E_t^\varepsilon F(X^\varepsilon(t), \xi^\varepsilon(s)) ds / \varepsilon = O(\varepsilon) |Z^\varepsilon(t)|^{2k-1}.$$

The right-hand equality is a consequence of (A3.2) and a scale change. We carry out the proof only for $\tau_\varepsilon = 0$, $G(\cdot) = 0$ for simplicity. The proof of the other cases is essentially the same.

By analogy with the procedure that led to (3.4) and (3.7), we have

$$(5.3) \quad E_t^\varepsilon[Z^\varepsilon(t+s)^{2k} + f_{2k}^\varepsilon(t+s)] - [Z^\varepsilon(t)^{2k} + f_{2k}^\varepsilon(t)] = E_t^\varepsilon \int_t^{t+s} c_{2k}^\varepsilon(u) du,$$

where

$$\begin{aligned} c_{2k}^\varepsilon(u) &= \frac{1}{\varepsilon^2} \int_u^T 2k[(2k-1)Z^\varepsilon(u)^{2k-2} E_u^\varepsilon F(X^\varepsilon(u), \xi^\varepsilon(s)) ds] F(X^\varepsilon(u), \xi^\varepsilon(u)) \\ &\quad + \frac{1}{\varepsilon^2} \int_u^T 2kZ^\varepsilon(u)^{2k-1} E_u^\varepsilon F_x(X^\varepsilon(u), \xi^\varepsilon(s)) ds [Q^\varepsilon(u) + F(X^\varepsilon(u), \xi^\varepsilon(u))] \\ &= O(1) \int_{u/\varepsilon^2}^{T/\varepsilon^2} Z^\varepsilon(u)^{2k-2} E_u^\varepsilon F(X^\varepsilon(u), \xi(s)) ds \\ &\quad + O(1) \int_{u/\varepsilon^2}^{T/\varepsilon^2} Z^\varepsilon(u)^{2k-1} E_u^\varepsilon F_x(X^\varepsilon(u), \xi(s)) ds. \end{aligned}$$

By (A3.2), we can write this expression as

$$(5.4) \quad Z^\varepsilon(u)^{2k-2} \hat{C}_{2k}^\varepsilon(u) + Z^\varepsilon(u)^{2k-1} C_{2k}^\varepsilon(u) = O(1)(|Z^\varepsilon(u)|^{2k-2} + |Z^\varepsilon(u)|^{2k-1}),$$

where the \bar{C}_{2k}^ε and C_{2k}^ε are defined in the obvious way and are bounded.

By (5.3), (5.4) and the bound on $f_{2k}^\varepsilon(\cdot)$, there are $\infty > K_{2k}(\tau) \rightarrow 0$ as $\tau \rightarrow 0$ and constants K'_{2k} such that, for $t \leq \tau$,

$$E[Z^\varepsilon(t)^{2k} + f_{2k}^\varepsilon(t)] \leq K'_{2k} \int_0^t E[1 + |Z^\varepsilon(s)|^{2k-1}] ds \leq K_{2k}(\tau).$$

Define

$$M_{2k}^\varepsilon(t) = [Z^\varepsilon(t)^{2k} + f_{2k}^\varepsilon(t)] - \int_0^t [Z^\varepsilon(s)^{2k-2} \hat{C}_{2k}^\varepsilon(s) + Z^\varepsilon(s)^{2k-1} C_{2k}^\varepsilon(s)] ds.$$

By (5.3), $M_{2k}^\varepsilon(\cdot)$ is a martingale. From the above estimates there are functions $K''_{2k}(t) \rightarrow 0$ as $t \rightarrow 0$ such that (use Doob's inequality [15, Thm. 7.3.4])

$$(5.5) \quad E \sup_{s \leq t} |M_{2k}^\varepsilon(s)|^2 \leq 4E|M_{2k}^\varepsilon(t)|^2 \leq K''_{2k}(t).$$

By the bound on $f_{2k}^\varepsilon(\cdot)$ and (5.5), we get (5.2) for $\tau_\varepsilon = 0$. Q.E.D.

THEOREM 5.2. Assume the conditions of Theorem 5.1, except with $Q_\varepsilon(t) = 0$ (or $Q_{\varepsilon n} = 0$) for $t \geq \tau_\varepsilon$. Given $\Delta_0 > 0$, there are $\delta_0 > 0$ and $T_0 > 0$ such that for all small ε

$$(5.6) \quad P_{\tau_\varepsilon}^\varepsilon \{ \sup_{t \leq T_0} |Z^\varepsilon(\tau_\varepsilon + t) - Z^\varepsilon(\tau_\varepsilon)| \geq \Delta_0 \} \leq 1 - \delta_0.$$

Proof. The result follows from Theorem 5.1.

Recall the definition of \bar{B} in § 1. We now describe some classes of controls and obtain some estimates of path excursions under the controls. Let L and U be numbers such that $0 \leq L < U \leq \bar{B}$. Define

$$(5.7) \quad \begin{aligned} dY_0^\varepsilon(t) &= [F(X^\varepsilon(t), \xi^\varepsilon(t))/\varepsilon + G(X^\varepsilon(t), \xi^\varepsilon(t))]^- dt I_{\{X^\varepsilon(t)=L\}}, \\ dY_1^\varepsilon(t) &= [F(X^\varepsilon(t), \xi^\varepsilon(t))/\varepsilon + G(X^\varepsilon(t), \xi^\varepsilon(t))]^+ dt I_{\{X^\varepsilon(t)=U\}}. \end{aligned}$$

For obvious reasons, we call this the (L, U) -barrier control (following the usage in [4]). Define the discrete parameter barrier policy in the analogous way: the $dY_i^\varepsilon(\cdot)$ are just large enough to keep $X^\varepsilon(\cdot)$ in the set $[L, U]$. The dY_i^ε/dt will be one of the candidates for the Q_ε in Theorem 5.1.

Let $\Delta_0 < \bar{B}/2$. We define a specific control policy—called the (\bar{B}, Δ_0) -policy—as follows. Continuous parameter case: If $X^\varepsilon(t^-) = \bar{B}$, immediately set $Y_1^\varepsilon(t) = Y_1^\varepsilon(t^-) + \Delta_0$ and $X^\varepsilon(t) = \bar{B} - \Delta_0$. Also, $Y_0^\varepsilon(\cdot)$ will increase just fast enough to keep $X^\varepsilon(t) \geq 0$, i.e., $Y_0^\varepsilon(\cdot)$ is given by (5.7) for $L = 0$. Use the analogous definitions for the discrete parameter process. The (\bar{B}, Δ_0) -control has some nice properties which render it useful for the discussion in the next section.

THEOREM 5.3. Let the $Y_i^\varepsilon(\cdot)$ be the control values given by the (Δ_0, \bar{B}) -policy and assume either (A3.1), (A3.2) or (A4.1), (A4.2). For each t and integer k (the τ_ε are random times),

$$(5.8) \quad \sup_{\varepsilon, \tau_\varepsilon} E|Y_i^\varepsilon(\tau_\varepsilon + t) - Y_i^\varepsilon(\tau_\varepsilon)|^k < \infty.$$

Remark. Owing to the conditioning in (5.6), the estimates for Y_i^ε are proved almost as if the “return” process from the point $(\bar{B} - \Delta_0)$ to (either \bar{B} or $\bar{B} - 2\Delta_0$), then back to $\bar{B} - \Delta_0$, etc., were constructed from a Bernoulli sequence.

Proof. We prove the continuous parameter case only, and $i = 1$. The case $i = 0$ is treated by an argument based on Theorems 5.2 and 5.4. Without loss of generality set $\tau_\varepsilon = 0$.

Define the stopping times: $\sigma_0^\varepsilon = \min \{t \geq 0: X^\varepsilon(t) = \bar{B} - \Delta_0\}$ and for $i > 0$, $\rho_i^\varepsilon = \min \{t > \sigma_{i-1}^\varepsilon: |X^\varepsilon(t) - (\bar{B} - \Delta_0)| \geq \Delta_0\}$, $\sigma_i^\varepsilon = \min \{t \geq \rho_i^\varepsilon: X^\varepsilon(t) = \bar{B} - \Delta_0\}$. We will estimate the k th moment of $N^\varepsilon(t) = \max \{i: \sigma_i^\varepsilon \leq t\}$. Define the $\{0, 1\}$ valued random variable U_i^ε as follows. With T_0 as in Theorem 5.2, set

$$U_i^\varepsilon = \begin{cases} 1 & \text{if } \rho_i^\varepsilon - \sigma_{i-1}^\varepsilon \geq T_0 \quad (\text{"success"}), \\ 0 & \text{if } \rho_i^\varepsilon - \sigma_{i-1}^\varepsilon < T_0 \quad (\text{"failure"}). \end{cases}$$

Let N_i^ε denote the number of successive passages of $X^\varepsilon(\cdot)$ from $\bar{B} - \Delta_0$ to either \bar{B} or $\bar{B} - 2\Delta_0$ which are failures, after the i th success. Then

$$N^\varepsilon(t) \leq \frac{t}{T_0} + \sum_0^{t/T_0-1} N_i^\varepsilon.$$

There are $K_k < \infty$ such that

$$N^\varepsilon(t)^k \leq K_k(t/T_0)^k + K_k(t/T_0)^k \sum_0^{t/T_0-1} (N_i^\varepsilon)^k.$$

We will bound $E(N_i^\varepsilon)^k$. Let $\sigma_{s_i}^\varepsilon$ denote the return time of $X^\varepsilon(\cdot)$ to $\bar{B} - \Delta_0$ immediately after the i th success. Then, by Theorem 5.2,

$$\begin{aligned} P_{\sigma_{s_i}^\varepsilon}^\varepsilon \{N_i^\varepsilon \geq n\} &= P_{\sigma_{s_i}^\varepsilon}^\varepsilon \{\sigma_{s_i+j+1}^\varepsilon - \sigma_{s_i+j}^\varepsilon < T_0, \text{ for all } j < n\} \\ &\leq (1 - \delta_0)^{n-1}. \end{aligned}$$

This yields $E(N_i^\varepsilon)^k \leq (\text{constant})/\delta_0$, and the proof is concluded, since $Y_1^\varepsilon(t) \leq \Delta_0 N^\varepsilon(t)$. Q.E.D.

THEOREM 5.4. Assume either (A3.1), (A3.2) or (A4.1), (A4.2) and let the $Y_i^\varepsilon(\cdot)$ be the control values associated with the (L, U) -barrier policy. Then for each t

$$(5.9) \quad \sup_{\varepsilon, T} [E(Y_0^\varepsilon(t+T) - Y_0^\varepsilon(T))^2 + E(Y_1^\varepsilon(t+T) - Y_1^\varepsilon(T))^2] < \infty.$$

Proof. Again, we do only some of the details for $Y_0^\varepsilon(\cdot)$, and for the continuous parameter case. For notational simplicity, we only present the case $G(\cdot) = 0$. Denote the initial time by t_0 and let $\Delta_0 < (U - L)/2$. Define the stopping times

$$\sigma_0^\varepsilon = \min \{t \geq t_0: X^\varepsilon(t) = L\}$$

and, for $i = 1, 2, \dots$,

$$\sigma_i^\varepsilon = \min \{t > \rho_i^\varepsilon: X^\varepsilon(t) = L\}, \quad \rho_i^\varepsilon = \min \{t > \sigma_{i-1}^\varepsilon: X^\varepsilon(t) = L + \Delta_0\},$$

with the convention that the min over an empty set is infinite. All the needed estimates can be shown to be uniform in t_0 and we set $t_0 = 0$ for simplicity.

We can write (and simultaneously define $\hat{Z}^\varepsilon(\cdot)$)

$$(5.10) \quad \begin{aligned} \sum_i [X^\varepsilon(\rho_{i+1}^\varepsilon \cap t) - X^\varepsilon(\sigma_i^\varepsilon \cap t)] &= \sum_i [Z^\varepsilon(\rho_{i+1}^\varepsilon \cap t) - Z^\varepsilon(\sigma_i^\varepsilon \cap t)] + Y_0^\varepsilon(t) \\ &= \hat{Z}^\varepsilon(t) + Y_0^\varepsilon(t). \end{aligned}$$

The mean square value of the term on the left of (5.10) is bounded above by Δ_0^2 times the second moment of the number of i for which $\rho_i^\varepsilon \leq t$. By an argument very similar to that used in Theorem 5.3, this can be shown to be bounded uniformly in ε for each t .

Define $M^\varepsilon(t) = Z_1^\varepsilon(t) - \int_0^t C_1^\varepsilon(s) ds$, where $Z_1^\varepsilon(\cdot)$ is defined in Theorem 3.1. Equivalently, it is the $f_1^\varepsilon(\cdot)$ of (5.2'). The $C_1^\varepsilon(\cdot)$ is defined in (5.4). The $\hat{C}_1^\varepsilon(\cdot)$ defined in (5.4) does not appear here, since $2k = 1$. Define $N^\varepsilon(\cdot)$ as in the proof of Theorem 5.3. Then, since $M^\varepsilon(\cdot)$ is a martingale, on the "interval where $dY_1^\varepsilon(\cdot) = 0$ " we have

$$\begin{aligned} E \left(\sum_i [M^\varepsilon(\rho_{i+1}^\varepsilon \cap t) - M^\varepsilon(\sigma_i^\varepsilon \cap t)] \right)^2 &= \sum_i E |M^\varepsilon(\rho_{i+1}^\varepsilon \cap t) - M^\varepsilon(\sigma_i^\varepsilon \cap t)|^2 \\ (5.11) \quad &= O(1) E (\sup_{s \leq t} |Z^\varepsilon(s)|^2 + 1) N^\varepsilon(t) \\ &= O(1) E^{1/2} (\sup_{s \leq t} |Z^\varepsilon(s)|^4 + 1) E^{1/2} |N^\varepsilon(t)|^2 \\ &\leq K_1 < \infty. \end{aligned}$$

The last inequality follows from Theorems 5.1 and 5.3. Since $C_1^\varepsilon(\cdot) = O(1)$, $Z_1^\varepsilon(\cdot) = O(\varepsilon)$, and $\sup_\varepsilon E(N^\varepsilon(t))^2 < \infty$, there are $K_2 < \infty$, $K_3 < \infty$, such that the left side of (5.11) can be bounded below by

$$K_2 E \left(\sum_i [Z^\varepsilon(\rho_{i+1}^\varepsilon \cap t) - Z^\varepsilon(\sigma_i^\varepsilon \cap t)] \right)^2 - K_3 = K_2 E |\hat{Z}^\varepsilon(t)|^2 - K_3.$$

The proof that $\sup E |\hat{Z}^\varepsilon(t)|^2 < \infty$ follows from these inequalities. Q.E.D.

6. Convergence of the costs and controls. In [1], it is shown that there are $0 \leq L^* < U^* < \infty$ such that (under appropriate conditions) the optimal control for (1.1) is an (L^*, U^*) -barrier control. We assume that \bar{B} is large enough so that $U^* \leq \bar{B}$. Let $\bar{Y}_i(\cdot)$, $i = 0, 1$, denote this optimal control. The set of increments of the "local time" control processes $\{\bar{Y}_i(n+1) - \bar{Y}_i(n), i = 0, 1, n < 1\}$ are uniformly integrable. Let $\bar{Y}_i^\varepsilon(\cdot)$, $i = 0, 1$, denote the (L^*, U^*) -barrier control for $X^\varepsilon(\cdot)$ (continuous or discrete time). The following theorem says that the optimal control for $X(\cdot)$ is "nearly" optimal for $X^\varepsilon(\cdot)$.

THEOREM 6.1. Assume either (A3.1)–(A3.5) or (A4.1)–(A4.5). Let (1.1) have a unique weak sense solution for the (L^*, U^*) -barrier policy, and let this policy be optimal. Then $\{X^\varepsilon(\cdot), \bar{Y}_0^\varepsilon(\cdot), \bar{Y}_1^\varepsilon(\cdot)\} \Rightarrow (X(\cdot), \bar{Y}_0(\cdot), \bar{Y}_1(\cdot))$ in the pseudopath topology, and there is a Wiener process $w(\cdot)$ such that $(X(\cdot), \bar{Y}_0(\cdot), \bar{Y}_1(\cdot))$ is nonanticipative with respect to $w(\cdot)$, and (1.1) holds. Also, as $\varepsilon \rightarrow 0$,

$$(6.1) \quad V_0^\varepsilon(x, \bar{Y}_0^\varepsilon, \bar{Y}_1^\varepsilon) \rightarrow V_0(x, \bar{Y}_0, \bar{Y}_1) = V(x).$$

In addition, for $\delta > 0$, let $\hat{Y}_0^\varepsilon(\cdot), \hat{Y}_1^\varepsilon(\cdot)$ be δ -optimal policies for $X^\varepsilon(\cdot)$ such that the set (6.2)

$$(6.2) \quad \{\hat{Y}_i^\varepsilon(n+1) - \hat{Y}_i^\varepsilon(n), \varepsilon > 0, n < \infty\}$$

is uniformly integrable. Then

$$(6.3) \quad \delta + \liminf_{\varepsilon} V^\varepsilon(x) \geq \liminf_{\varepsilon} V_0^\varepsilon(x, \hat{Y}_0^\varepsilon, \hat{Y}_1^\varepsilon) \geq V(x).$$

Remark on (6.2). The uniform integrability condition on (6.2) is used basically to assure that the cost associated with the limit process is the limit of the costs associated with $X^\varepsilon(\cdot)$. We have not been able to prove the theorem without this condition, unless all cost terms are positive (see Theorem 6.2).

The uniform integrability condition holds for a wide variety of control processes, and the methods of Theorems 5.3 and 5.4 can be used to show that it holds for the following cases: (1) Let there be numbers $L^0, U^0, \Delta_0, \Delta_1$ where $\Delta_0 + \Delta_1 < (U^0 - L^0)/2$ and $Y_0^\varepsilon(\cdot)$ changes only when $X^\varepsilon(\cdot)$ is in $[L^0, L^0 + \Delta_0]$, $Y_1^\varepsilon(\cdot)$ changes only when

$X^\varepsilon(\cdot)$ is in $[U^0 - \Delta_0, U^0]$, and the maximum jump is no greater than Δ_1 ; (2) Let $Y_i^\varepsilon(\cdot)$ denote any admissible policy and fix N . Define

$$\tau_n^\varepsilon = \min \{t > n: (Y_0^\varepsilon(t) - Y_0^\varepsilon(n)) + (Y_1^\varepsilon(t) - Y_1^\varepsilon(n)) \geq N\} \cap (n+1).$$

Now, choose the control policy on the interval $[n, n+1)$ as follows. Use $Y_i^\varepsilon(\cdot)$ on $[n, \tau_n^\varepsilon)$, then switch to a barrier or (\bar{B}, Δ_0) policy on $[\tau_n^\varepsilon, n+1)$.

In Theorem 6.2, it is shown that the uniform integrability condition is not needed if $-k_1 dY_1(t)$ is replaced by the positive cost increment $k_1 dY_1(t)$.

Proof. We deal only with the continuous parameter case. Let $X^\varepsilon(\cdot)$ denote the process with the $\bar{Y}^\varepsilon(\cdot)$ used. By Theorem 5.4, $\{\bar{Y}_i^\varepsilon(n+1) - \bar{Y}_i^\varepsilon(n), \varepsilon > 0, n < \infty\}$ is uniformly integrable. Extract a weakly convergent subsequence of $\{X^\varepsilon(\cdot), \bar{Y}_0^\varepsilon(\cdot), \bar{Y}_1^\varepsilon(\cdot)\}$ (pseudopath topology) and denote the limit by $(X(\cdot), Y_0(\cdot), Y_1(\cdot))$. By Theorem 3.2, this triple satisfies (1.1) for some $w(\cdot)$.

It follows from the weak convergence and the fact that the $\bar{Y}_0^\varepsilon(\cdot), \bar{Y}_1^\varepsilon(\cdot)$ is a (L^*, U^*) -barrier policy that the limit $Y_0(\cdot), Y_1(\cdot)$ is a (L^*, U^*) -barrier policy for $X(\cdot)$. Thus $Y_i(\cdot) = \bar{Y}_i(\cdot)$. This is true for any weakly convergent subsequence. By hypothesis, the solution to (1.1) associated with any (L^*, U^*) -barrier policy is unique (in the sense of distributions). Thus, the (distribution of the) limit does not depend on the chosen subsequence.

By the uniform integrability asserted in the above paragraph and the weak convergence, we have

$$\begin{aligned} \lim_{\varepsilon} V_0^\varepsilon(x, \bar{Y}_0^\varepsilon, \bar{Y}_1^\varepsilon) &= \lim_{\varepsilon} E \int_0^\infty e^{-\beta t} [k_0 d\bar{Y}_0^\varepsilon(t) - k_1 d\bar{Y}_1^\varepsilon(t) + k(X^\varepsilon(t))] dt \\ (6.4) \qquad &= E \int_0^\infty e^{-\beta t} [k_0 d\bar{Y}_0(t) - k_1 d\bar{Y}_1(t) + k(X(t))] dt \\ &= V_0(x, \bar{Y}_0, \bar{Y}_1) = V(x). \end{aligned}$$

To get (6.3), repeat the procedure with the controls $\hat{Y}_0^\varepsilon(\cdot), \hat{Y}_1^\varepsilon(\cdot)$. With these controls, the limit $(X(\cdot), Y_0(\cdot), Y_1(\cdot))$ might depend on the chosen subsequence. But, if $\{\varepsilon_n\}$ indexes any weakly convergent subsequence, with limit $(X(\cdot), Y_0(\cdot), Y_1(\cdot))$ we have $\lim_n V_0^\varepsilon(x, Y_0^n, Y_1^n) \rightarrow V_0(x, Y_0, Y_1) \geq V(x)$. Hence, by the definition of δ -optimality and the weak convergence,

$$\begin{aligned} \delta + \lim_{\varepsilon} V^\varepsilon(x) &\geq \lim_{\varepsilon} V_0^\varepsilon(x, \hat{Y}_0^\varepsilon, \hat{Y}_1^\varepsilon) \\ &\geq \inf_{Y_0, Y_1} V_0(x, Y_0, Y_1) = V(x). \end{aligned} \quad \text{Q.E.D.}$$

THEOREM 6.2. Assume the conditions of Theorem 6.1, except for the uniform integrability of the set (6.2), but for $k_i > 0$ let the cost be

$$E_x \int_0^\infty e^{-\beta t} [k_0 dY_0(t) + k_1 dY_1(t) + k(x(t))] dt = V_0(x, Y_0, Y_1),$$

and similarly define the cost $V_0^\varepsilon(x, Y_0^\varepsilon, Y_1^\varepsilon)$. Then the conclusions of Theorem 6.1 (with the δ in (6.3) replaced by 2δ) hold.

Proof. Let $\hat{Y}_i^\varepsilon(\cdot), i=0,1$, denote a δ -optimal policy. We can suppose that $\sup_{\varepsilon} [E_x Y_0^\varepsilon(t) + E_x Y_1^\varepsilon(t)] < \infty$ for each $t < \infty$ and that

$$\lim_{T \rightarrow \infty} \sup_{\varepsilon} \int_T^\infty e^{-\beta t} [k_0 dY_0^\varepsilon(t) + k_1 dY_1^\varepsilon(t) + k(X^\varepsilon(t))] dt = 0,$$

since this holds for any barrier policy. In fact, there is an $N_\delta < \infty$ such that if we switch to the (L^*, U^*) -barrier policy (or to any barrier policy) once the $Y_i^\varepsilon(t)$ exceeds N_δ , we change the cost by less than δ . But, then the set in (6.2) is uniformly integrable, and Theorem 6.1 holds. Q.E.D.

7. Average cost per unit time. The methods of §§ 1–5 can be used to adjust the proof of Theorem 8 in [6] to get the result which is analogous to Theorem 6.1 for the average cost per unit time problem. Only an outline of the method will be given. The reader is referred to [6] for more details on the structure of the approximation for the average cost problem for the nonsingular case (and which can be carried over to our case).

For the average cost per unit time problem, we wish to work with feedback controls and, hence, use only $Y_i^\varepsilon(\cdot)$, $i = 0, 1$, or $Y_i(\cdot)$, $i = 0, 1$, for which the associated processes $\xi^\varepsilon(\cdot)$ and $(X^\varepsilon(\cdot), \xi^\varepsilon(\cdot))$ or $X(\cdot)$, respectively, are bounded Markov-Feller processes. Also, we let $\{\xi^\varepsilon(t), \varepsilon > 0, t < \infty\}$ be bounded. The cost criteria are:

$$\begin{aligned} \overline{\lim}_T E \frac{1}{T} \int_0^T [k_0 dY_0(t) - k_1 dY_1(t) + k(X(t))] dt &\equiv \gamma(Y_0, Y_1), \\ \overline{\lim}_T E \frac{1}{T} \int_0^T [k_0 dY_0^\varepsilon(t) - k_1 dY_1(t) + k(X^\varepsilon(t))] dt &= \gamma^\varepsilon(Y_0, Y_1). \end{aligned}$$

For simplicity, we discuss only the continuous parameter case. The discrete parameter case uses very similar assumptions and proof. Let PM (PM^ε , respectively) denote the class of feedback control processes for which $X(\cdot)$ ($(X^\varepsilon(\cdot), \xi^\varepsilon(\cdot))$, respectively) is a Markov-Feller process. Let NA (NA^ε , respectively) denote the class of nonanticipative controls. We will use the following assumptions.

(A7.1) There is an $\varepsilon_0 > 0$ such that for each $\delta > 0$ and $\varepsilon \leq \varepsilon_0$, there are δ -optimal controls $\in PM^\varepsilon$ of the form

$$(7.1) \quad dY_i^\varepsilon = Q_i^\varepsilon(x, \xi) dt, \quad i = 0, 1,$$

where the $Q_i^\varepsilon(\cdot, \cdot)$ are continuous.

Note. If the $Q_i^\varepsilon(x, \xi)$ are Lipschitz continuous in x , uniformly in ξ , then $Y_i^\varepsilon(\cdot)$, $i = 0, 1$, is in PM^ε . See the remark below where it is shown that the barrier policy can be smoothed to yield a continuous $Q_i^\varepsilon(\cdot, \cdot)$.

(A7.2) A (L^*, U^*) -barrier control $\bar{Y}_i(\cdot)$, $i = 0, 1$, is optimal for (1.1), and with this control (1.1) has a unique invariant measure. This control is in PM and its adaptation $\bar{Y}_i^\varepsilon(\cdot)$, $i = 0, 1$, to $X^\varepsilon(\cdot)$ is in PM^ε . When applied to PM^ε , $(X^\varepsilon(\cdot), \xi^\varepsilon(\cdot))$ has a unique solution and invariant measure.

$$(A7.3) \quad \inf_{Y_i \in PM} \gamma(Y_0, Y_1) = \inf_{Y_i \in NA} \gamma(Y_0, Y_1).$$

Remark. The (L^*, U^*) -barrier control can be approximated for $X^\varepsilon(\cdot)$ in such a way that it is of the form in (A7.1). In particular, let $\Delta_\varepsilon \rightarrow 0$ as $\varepsilon \rightarrow 0$ and define

$$(7.2) \quad d\tilde{Y}_1^\varepsilon(t) = dt[F(X^\varepsilon(t), \xi^\varepsilon(t))/\varepsilon + G(X^\varepsilon(t), \xi^\varepsilon(t))]^+ I_{\{X^\varepsilon(t) \in [U^* - \Delta_\varepsilon, U^*]\}} \frac{[X^\varepsilon(t) - U^* + \Delta_\varepsilon]}{\Delta_\varepsilon}$$

and similarly for $\tilde{Y}_0^\varepsilon(\cdot)$. It can be shown that

$$(7.3) \quad \gamma^\varepsilon(\tilde{Y}_0^\varepsilon, \tilde{Y}_1^\varepsilon) \xrightarrow{\Delta_\varepsilon \rightarrow 0} \gamma^\varepsilon(\bar{Y}_0^\varepsilon, \bar{Y}_1^\varepsilon),$$

where the \bar{Y}_i^ε is the (L^*, U^*) -barrier policy for $X^\varepsilon(\cdot)$. Clearly, the $\tilde{Y}_i^\varepsilon(\cdot)$ are of the form used in (A7.1). By (7.3), for each ε , we can choose Δ_ε so that the left and right sides of (7.3) are as close as desired. By using the techniques in § 5, it can be shown that $\{\tilde{Y}_i^\varepsilon(n+1) - \tilde{Y}_i^\varepsilon(n), i=0, 1, n < \infty, \varepsilon > 0, X^\varepsilon(0)=x, \xi^\varepsilon(0)=\xi\}$ is uniformly integrable.

THEOREM 7.1. Assume (A3.1)–(A3.5) and (A7.1)–(A7.3). Let $G(x, \xi)$ be Lipschitz continuous in x , uniformly in ξ . Let $\delta > 0$, and let $Y_i^\varepsilon(\cdot), i=1, 2$, be the δ -optimal controls defined in (A7.1). Suppose that the set

(7.4) $\{Y_i^\varepsilon(n+1) - Y_i^\varepsilon(n), \varepsilon > 0, n < \infty, X^\varepsilon(0)=x, \xi^\varepsilon(0)=\xi, \text{ for all } x \in [0, \bar{B}], \text{ for all } \xi\}$ is uniformly integrable. Then

$$\delta + \lim_{\varepsilon} \gamma^\varepsilon(Y_0^\varepsilon, Y_1^\varepsilon) \geq \lim_{\varepsilon} \gamma^\varepsilon(\bar{Y}_0^\varepsilon, \bar{Y}_1^\varepsilon) = \gamma(\bar{Y}_0, \bar{Y}_1).$$

Proof. For each ε, δ, T , define the measure

$$P_T^{\varepsilon, \delta}(\cdot) = \frac{1}{T} E \int_0^T P^{\varepsilon, \delta}(X^\varepsilon(0), \xi^\varepsilon(0), t, \cdot) dt,$$

where $P^{\varepsilon, \delta}$ is the transition function for $(X^\varepsilon(\cdot), \xi^\varepsilon(\cdot))$, under the δ -optimal control $Y_i^\varepsilon(\cdot)$ (or Q_i^ε) of (A7.1). Then

$$\gamma^\varepsilon(Y_0^\varepsilon, Y_1^\varepsilon) = \overline{\lim}_T \int P_T^{\varepsilon, \delta}(dx d\xi) [k_0 Q_0^\varepsilon(x, \xi) - k_1 Q_1^\varepsilon(x, \xi) + k(x)].$$

Choose a subsequence $T_k \rightarrow \infty$ such that both the $\overline{\lim}$ is attained and $P_{T_k}^{\varepsilon, \delta}$ converges weakly (with limit denoted by $\mu^{\varepsilon, \delta}$). Then, by the Markov–Feller property of $(X^\varepsilon(\cdot), \xi^\varepsilon(\cdot))$ for $(Y_0^\varepsilon(\cdot), Y_1^\varepsilon(\cdot)) \in PM^\varepsilon$, $\mu^{\varepsilon, \delta}$ is an invariant measure for $(X^\varepsilon(\cdot), \xi^\varepsilon(\cdot))$ and, by the continuity of the Q_i^ε and the weak convergence,

$$(7.5) \quad \gamma^\varepsilon(Y_0^\varepsilon, Y_1^\varepsilon) = \int \mu^{\varepsilon, \delta}(dx d\xi) [k_0 Q_0^\varepsilon(x, \xi) - k_1 Q_1^\varepsilon(x, \xi) + k(x)].$$

Let $(\hat{X}^\varepsilon(\cdot), \hat{\xi}^\varepsilon(\cdot))$ denote the stationary process associated with the control function $Q_i^\varepsilon(\cdot, \cdot)$ and measure $\mu^{\varepsilon, \delta}(\cdot)$, and let $\hat{Y}_i^\varepsilon(\cdot)$ denote the corresponding stationary control processes. Then we can write (7.5) as:

$$(7.6) \quad \begin{aligned} \gamma^\varepsilon(Y_0^\varepsilon, Y_1^\varepsilon) &= E \int_0^1 dt [k_0 Q_0^\varepsilon(\hat{X}^\varepsilon(t), \hat{\xi}^\varepsilon(t)) - k_1 Q_1^\varepsilon(\hat{X}^\varepsilon(t), \hat{\xi}^\varepsilon(t)) + k(\hat{X}^\varepsilon(t))] \\ &= E \int_0^1 k(\hat{X}^\varepsilon(t)) dt + Ek_0 \hat{Y}_0^\varepsilon(1) - Ek_1 \hat{Y}_1^\varepsilon(1). \end{aligned}$$

By the uniform integrability (7.4), $\{\hat{Y}_0^\varepsilon(1), \hat{Y}_1^\varepsilon(1), \varepsilon > 0\}$ is uniformly integrable.

Now, choose a weakly convergent subsequence of $\{\hat{X}^\varepsilon(\cdot), Y_0^\varepsilon(\cdot), Y_1^\varepsilon(\cdot)\}$, with limit denoted by $(\hat{X}(\cdot), \hat{Y}_0(\cdot), \hat{Y}_1(\cdot))$. The limit is stationary, satisfies (1.1) and (indexing the subsequence by ε also), we have

$$\gamma^\varepsilon(Y_0^\varepsilon, Y_1^\varepsilon) \rightarrow \gamma(\hat{Y}_0, \hat{Y}_1) \geq \gamma(\bar{Y}_0, \bar{Y}_1),$$

where the optimality of $\bar{Y}_0(\cdot), \bar{Y}_1(\cdot)$ is used.

The proof is concluded by applying the same procedure to $\tilde{Y}_0^\varepsilon(\cdot), \tilde{Y}_1^\varepsilon(\cdot)$, where the “smoothing interval” Δ_ε (see remark above the theorem where \tilde{Y}_i^ε and Δ_ε are defined) goes to zero fast enough as $\varepsilon \rightarrow 0$. Q.E.D.

8. The vector case. Formulation, quasimartingale estimates and the approximation theorem. Most of the foregoing analysis and results can be carried over to the case of vector ($x \in R^r$, Euclidean r -space) valued G, F in (1.4) or (3.1). Since the details of the proofs are essentially the same as in the foregoing sections, only an outline will be given. Only the continuous parameter case will be discussed, but under the obvious changes in assumptions (A3.1)–(A3.4) and (A8.1) used below, the discrete parameter results also extend to the vector case.

We use the model (vector F, G)

$$(8.1) \quad dX^\varepsilon = [G(X^\varepsilon, \xi^\varepsilon) + F(X^\varepsilon, \xi^\varepsilon)/\varepsilon] dt + dY^\varepsilon(t),$$

with cost

$$(8.2) \quad V_0^\varepsilon(x, Y^\varepsilon) = \int_0^\infty e^{-\beta t} [k_0 |dY^\varepsilon(t)| + k(X^\varepsilon(t))] dt, \quad \beta > 0.$$

The results of § 7 can also be extended to the vector case.

THEOREM 8.1. Assume (A3.1), (A3.2) with vector G, F used, and let $\sup_\varepsilon E \int_0^T |dY^\varepsilon(t)| < \infty$ for each $T < \infty$. Then (possibly having to add to each component a process whose maximum value goes to zero as $\varepsilon \rightarrow 0$) $\{X^\varepsilon(\cdot), Z^\varepsilon(\cdot), Y^\varepsilon(\cdot)\}$ are quasimartingales with uniformly (in ε) bounded conditional variation on each bounded time interval.

Remark. The proof is essentially identical to that of Theorem 3.1. Similarly, the proof of Theorem 8.2 below is essentially identical to that of Theorem 3.2.

We will next use (A8.1), the vector form of (A3.5).

(A8.1) There is a matrix $\Sigma(\cdot)$ with a continuous and bounded square root $\sigma(\cdot)$ such that for each x ,

$$\begin{aligned} & \frac{1}{T_1} \int_u^{u+T_1} E_u F(x, \xi(\tau)) d\tau \int_\tau^{T+\tau} F'(x, \xi(s)) ds \\ & + \frac{1}{T_1} \left[\int_u^{u+T_1} E_u F(x, \xi(\tau)) d\tau \int_\tau^{T+\tau} F'(x, \xi(s)) ds \right]' \xrightarrow{P} \Sigma(x), \end{aligned}$$

as T, u and T_1 go to ∞ . There is a continuous function $\bar{G}_0(\cdot)$ with components $\bar{G}_{0i}(\cdot)$, $i \leq r$, such that for each x ,

$$\frac{1}{T_1} \int_u^{u+T_1} \sum_j E_u F_j(x, \xi(\tau)) d\tau \int_\tau^{T+\tau} F_{ix_j}(x, \xi(s)) ds \xrightarrow{P} \bar{G}_{0i}(x)$$

as T, u and T_1 go to ∞ .

THEOREM 8.2. Assume (A3.1)–(A3.4) and (A8.1), and let $\sup_\varepsilon E_x \int_0^T |dY^\varepsilon(s)| < \infty$ for each $T < \infty$. Then $\{Z^\varepsilon(\cdot)\}$ is tight in the Skorokhod topology on $D^r[0, \infty)$ and any weak limit process is continuous with probability 1. Let $\{X^\varepsilon(\cdot), Y^\varepsilon(\cdot), Z^\varepsilon(\cdot)\}$ be a weakly convergent subsequence in $D^{3r}[0, \infty)$, with the pseudopath topology used on the first two components and the Skorokhod topology on the last. Let $\bar{X}(\cdot) = (X(\cdot), Y(\cdot), Z(\cdot))$ denote the limit of a weakly convergent subsequence. Then the conclusions of Theorem 3.2 continue to hold, with the limit $Y(\cdot)$ replacing $Y_0(\cdot) - Y_1(\cdot)$. In particular, the limit satisfies

$$\begin{aligned} (8.3) \quad & X(t) = Z(t) + Y(t) + X(0), \\ & dZ(t) = [\bar{G}(X(t)) + \bar{G}_0(X(t))] dt + \sigma(X(t)) dw, \\ & Z(0) = 0. \end{aligned}$$

Definition and assumptions. Below, $v(\cdot)$ will be a continuous vector field on R^n with $|v(x)| \equiv 1$, and S a compact set with a piecewise differential boundary and with the following properties: (a) There is a truncated cone such that the apex can be placed at any point on ∂S such that the cone is in S . (b) There is a $\Delta_0 > 0$ such that for $x \in \partial S$ and $\Delta \leq \Delta_0$, the points $x + \Delta v(x)$ are interior to S . Define the $(S, \Delta, v(\cdot))$ -reflecting policy for $X^\varepsilon(\cdot)$ as the (admissible) policy which sets $X^\varepsilon(t) = x + \Delta v(x)$, if $X^\varepsilon(t^-) = x \in \partial S$. Then, of course, $dY^\varepsilon(t) = \Delta v(x)$. We define similarly the $(S, \Delta, v(\cdot))$ -reflecting policy for the $X(\cdot)$ of (8.3).

A policy $Y(\cdot)$ for (8.3) is called a $(S, v(\cdot))$ -reflecting policy if the associated process $X(\cdot)$ is a reflected diffusion in S , with continuous reflection direction $v(\cdot)$ on ∂S , and there is a $\Delta_0 > 0$ such that for $\Delta \leq \Delta_0$, the policy which sets $X(t) = x + \Delta v(x)$ if $X(t^-) = x \in \partial S$ is an admissible $(S, \Delta, v(\cdot))$ -reflecting policy.

The conditions on $S, v(\cdot)$, are motivated by the optimal control for the "finite fuel" problem [2].

THEOREM 8.3. Assume (A3.1) and (A3.2) (vector case), and let $Y^{\varepsilon, \Delta}(\cdot)$ denote a $(S, \Delta, v(\cdot))$ -reflecting policy for $X^\varepsilon(\cdot)$. Then for each $T < \infty$ and integer k ,

$$(8.4) \quad \sup_{\substack{\varepsilon, \Delta \leq \Delta_0 \\ x \in S}} E_x \left(\int_0^T |dY^{\varepsilon, \Delta}(s)| \right)^k < \infty.$$

THEOREM 8.4. Let $X(\cdot)$ satisfy (8.3) with bounded and continuous $\bar{G}(\cdot)$, $\bar{G}_0(\cdot)$ and $\sigma(\cdot)$. Let $Y^\Delta(\cdot)$ be a $(S, \Delta, v(\cdot))$ -reflecting policy for $x(\cdot)$. Then, for each $T < \infty$ and integer k ,

$$\sup_{\substack{\Delta \leq \Delta_0 \\ x \in S}} E_x \left[\int_0^T |dY^\Delta(s)| \right]^k < \infty.$$

We will prove Theorem 8.4 only. The proof of Theorem 8.3 is similar and uses the (vector case) estimate (5.6) for the process $Z^\varepsilon(t) + Z_1^\varepsilon(t)$, where $Z_1^\varepsilon(\cdot)$ here is the appropriate "vector" case form of the $Z_1^\varepsilon(\cdot)$ used in §§ 3-5.

Proof of Theorem 8.4. Let $Y^\Delta(\cdot)$ denote the $(S, \Delta, v(\cdot))$ -reflecting policy and $X^\Delta(\cdot)$ the associated solution to (8.3). For small $\alpha > 0$ let $N_\alpha(x)$ denote the α -neighborhood of x . There are x_1, \dots, x_q on ∂S such that $\bigcup_{i=1}^q N_\alpha(x_i) \supset \partial S$ and

$$(8.5) \quad \sup_{x, y \in N_{2\alpha}(x_i)} |v(x) - v(y)| < \alpha.$$

Let σ_0^m denote the first time of entry of $X^\Delta(\cdot)$ into $\bar{N}_\alpha(x_m)$. Define

$$\rho_i^m = \min \{t > \sigma_{i-1}^m: X^\Delta(t) \notin N_{2\alpha}(x_m)\},$$

$$\sigma_i^m = \min \{t > \rho_i^m: X^\Delta(t) \in \bar{N}_\alpha(x_m)\}.$$

Define $N_m^\Delta = \max \{i: \sigma_i^m \leq T\}$, $Y_i^{\Delta, m}(T) = Y^\Delta(\rho_{i+1}^m \cap T) - Y^\Delta(\sigma_i^m \cap T)$ and $Y^{\Delta, m}(T) = \sum_i Y_i^{\Delta, m}(T)$.

Owing to (8.5) and the smallness of α , there is a $K_1 < \infty$ (depending on α , but not on Δ) such that

$$\int_0^T |dY^\Delta(s)| \leq K_1 \sum_{m=1}^q |Y^{\Delta, m}(T)|.$$

Hence we need only evaluate $E[Y^{\Delta, m}(T)]^k$. We have

$$(8.6) \quad Y^{\Delta, m} = \sum_{i=1}^{N_m^\Delta} [X^\Delta(\rho_i^m \cap T) - X^\Delta(\sigma_{i-1}^m \cap T)] - \sum_{i=1}^{N_m^\Delta} [Z^\Delta(\rho_i^m \cap T) - Z^\Delta(\sigma_{i-1}^m \cap T)].$$

The absolute value of the first term on the right-hand side of (8.6) is $\leq \alpha K_2 N_m^\Delta$, for some constant $K_2 < \infty$. The absolute value of the last term on the right-hand side of (8.6) is bounded above by

$$(8.7) \quad N_m^\Delta \cdot \sup_{s, t \leq T} |Z^\Delta(t) - Z^\Delta(s)|.$$

Since $Z^\Delta(\cdot)$ is just the sum of an ordinary integral and a stochastic integral whose integrands are bounded uniformly in Δ , all the moments of the last factor in (8.7) are bounded uniformly in Δ . Hence it is enough to show that $\sup_{x \in S, \Delta} E_x |N_m^\Delta|^p < \infty$ for all integers p .

This last problem is similar to that dealt with in Theorem 5.3. The structure of the $(S, \Delta, v(\cdot))$ -reflecting policy implies that there is an $\alpha' > 0$ (not depending on Δ) such that in order for $X^\Delta(\cdot)$ to move from the exterior of $N_{2\alpha}(x_m)$ at time ρ_i^m to $\bar{N}_\alpha(x_m)$ at time σ_i^m , we must have $\sup_{\rho_i^m \leq t \leq \sigma_i^m} |Z^\Delta(t) - Z^\Delta(\rho_i^m)| \geq \alpha'$. Let τ ($< \infty$ with probability 1) be a stopping time. For each $\delta_0 \in (0, 1)$, there is a $T_0 > 0$ such that for all small $\Delta > 0$,

$$(8.8) \quad \sup_{x \in S, \tau} P_x \left\{ \sup_{t \leq T_0} |Z^\Delta(t + \tau) - Z^\Delta(\tau)| \geq \alpha'/2 \right\} \leq 1 - \delta_0.$$

Inequality (8.8) and an argument like that used in Theorem 5.3 (to get the upper estimate on $E|N_i^\varepsilon|^k$ there) completes our proof. Q.E.D.

DEFINITION. We now add an additional condition on the control problem. It is supposed that there is a compact set S_1 with a piecewise differentiable boundary satisfying the cone condition (a) above and such that $X^\varepsilon(\cdot)$ and $X(\cdot)$ are to be confined to S_1 . For small enough Δ and some continuous $v_1(\cdot)$, let there exist a $(S_1, \Delta, v_1(\cdot))$ -reflecting policy. In Theorem 5.1, we assume that the optimal control for $X(\cdot)$ is $(S, v(\cdot))$ -reflecting for some compact S and continuous $v(\cdot)$. The result can be extended to cover (for example) combined singular and nonsingular controls or true impulsive controls.

Remark. In many applications S_1 is not compact, but the special structure of the optimal policy (as for the "finite fuel" problems which have been investigated to date) yields $\delta/2$ -optimal policies for which we have the required uniform integrability.

THEOREM 8.5. Assume (A3.1)–(A3.4) and (A8.1), and the condition on S_1 in the above paragraph. Suppose that there are $S, v(\cdot)$ such that the optimal policy $\bar{Y}(\cdot)$ for $X(\cdot)$ is a $(S, v(\cdot))$ -reflecting policy. Let $\bar{Y}^\Delta(\cdot)$ denote the $(S, \Delta, v(\cdot))$ reflecting policy obtained from $\bar{Y}(\cdot)$. Let (8.3) have a unique weak sense solution under both policies, for all small Δ . Let $\hat{Y}^{\varepsilon, \Delta}(\cdot)$ denote the $(S, \Delta, v(\cdot))$ -reflecting policy adapted to $X^\varepsilon(\cdot)$. Given $\delta > 0$, there is a $\Delta > 0$ such that $\hat{Y}^{\varepsilon, \Delta}(\cdot)$ is δ -optimal for $X^\varepsilon(\cdot)$ and small ε in the sense that

$$(8.9) \quad \delta + \lim_{\varepsilon} V^\varepsilon(x) \geq \lim_{\varepsilon} V^{\varepsilon_0}(x, \hat{Y}^{\varepsilon, \Delta}) \geq \inf_{Y \text{ adm.}} V_0(x, Y) = V_0(x, \bar{Y}) = V(x).$$

Proof. The method is that of Theorems 6.1 and 6.2, but is simpler since all cost terms are nonnegative. Let $Y^\varepsilon(\cdot)$ denote the optimal (or $\delta/2$ -optimal, if there is no optimal policy) policy for $X^\varepsilon(\cdot)$ and let $X^{\varepsilon, \Delta}(\cdot)$ denote the process corresponding to $\hat{Y}^{\varepsilon, \Delta}(\cdot)$. By an argument of the type used in Theorem 6.2, there is no loss of generality if we suppose that the second set of the pair of sets

$$\left\{ \int_n^{n+1} |d\hat{Y}^{\varepsilon, \Delta}(s)|, \varepsilon > 0, n < \infty \right\}, \quad \left\{ \int_n^{n+1} |dY^\varepsilon(s)|, \varepsilon > 0, n < \infty \right\}$$

is uniformly integrable. The first set is uniformly integrable by Theorem 8.3. Choose a weakly convergent subsequence of $\{X^{\varepsilon, \Delta}(\cdot), \hat{Y}^{\varepsilon, \Delta}(\cdot)\}, \{X^\varepsilon(\cdot), Y^\varepsilon(\cdot)\}$ with the

subsequence also indexed by ε and with the limit denoted by $(X^\Delta(\cdot), \hat{Y}^\Delta(\cdot)), (X(\cdot), Y(\cdot))$, respectively. Then $X^\Delta(\cdot)$ is the $(S, \Delta, v(\cdot))$ -reflecting diffusion and $\hat{Y}^\Delta(\cdot) = \bar{Y}^\Delta(\cdot)$. Thus, by the weak convergence and the optimality of $\bar{Y}(\cdot)$, we have

$$(8.10) \quad \frac{\delta}{2} + \lim_{\varepsilon} V^\varepsilon(x) \geq \lim_{\varepsilon} V_0^\varepsilon(x, Y^\varepsilon) \geq \inf_{Y \text{ adm.}} V(x, Y) \geq V_0(x, \bar{Y}) = V(x).$$

Also by the weak convergence

$$(8.11) \quad \lim_{\varepsilon} V^\varepsilon(x, \hat{Y}^{\varepsilon, \Delta}) = V(x, \bar{Y}^\Delta).$$

Another weak convergence argument and the uniqueness assumption on the reflecting diffusion $X(\cdot)$ under policy $\bar{Y}(\cdot)$ yields the weak convergence of $\{X^\Delta(\cdot), \bar{Y}^\Delta(\cdot)\}$ to $(X(\cdot), \bar{Y}(\cdot))$ as $\Delta \rightarrow 0$. Also, the set $\{\int_n^{n+1} |d\bar{Y}^\Delta(s)|, \Delta \leq \Delta_0, n < \infty\}$ is uniformly integrable. This yields

$$(8.12) \quad \frac{\delta}{2} + V(x) \geq V_0(x, \bar{Y}^\Delta)$$

for small Δ . Now, (8.9) follows by combining (8.10)–(8.12). Q.E.D.

REFERENCES

- [1] S. E. SHREVE, J. P. LEHOSZKY AND D. P. GAVAR, *Optimal consumption for general diffusion with absorbing and reflecting barriers*, this Journal, 22 (1984), pp. 55–75.
- [2] I. KARATZAS AND S. E. SHREVE, *Equivalent models for finite fuel stochastic control*, Stochastics, 18 (1986), pp. 245–276.
- [3] I. KARATZAS, *A class of singular stochastic control problems*, Adv. in Appl. Prob., 15 (1983), pp. 225–254.
- [4] J. M. HARRISON, *Brownian Motion and Stochastic Flow Systems*, John Wiley, New York, 1985.
- [5] J. M. HARRISON AND M. I. TAKSAR, *Instantaneous control of Brownian motion*, Math. Oper. Res., 8 (1983), pp. 454–466.
- [6] H. KUSHNER AND W. Runggaldier, *Nearly optimal state feedback controls for stochastic systems with wideband noise disturbances*, Lefschetz Center for Dynamical Systems, Report 85–23, Brown University, Providence, RI, 1985; this Journal, 25, (1987), pp. 289–315.
- [7] ———, *Filtering and control for wide bandwidth noise driven systems*, Lefschetz Center for Dynamical Systems, Report 86–8, Brown Univ., Providence, RI, January 1986; IEEE Trans. Automat. Control, 32, (1987), pp. 123–133.
- [8] H. J. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes; with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.
- [9] J. L. MENALDI AND M. ROBIN, *On some cheap control problems for diffusion processes*, Trans. Amer. Math. Soc., 278 (1983), pp. 771–802.
- [10] C. DELLACHERIE AND P. A. MEYER, *Probabilités et Potentiel*, Hermann, Paris; North-Holland, Amsterdam.
- [11] P. A. MEYER AND W. A. ZHENG, *Tightness criteria for laws of semimartingales*, Ann. Inst. H. Poincaré, 20 (1984), pp. 353–372.
- [12] 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.
- [13] G. C. PAPANICOLAOU AND W. KOHLER, *Asymptotic theory of mixing ordinary differential equations*, Comm. Pure. Appl. Math., 27 (1974), pp. 641–668.
- [14] H. KUSHNER, *Jump diffusion approximations for ordinary differential equations with wideband random right-hand sides*, this Journal, 17 (1979), pp. 729–744.
- [15] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1952.
- [16] J. P. LEHOCZKY AND S. E. SHREVE, *Absolutely continuous and singular stochastic control*, Stochastics, 17 (1986), pp. 91–110.

NECESSARY CONDITIONS FOR OPTIMAL CONTROL PROBLEMS: CONJUGATE POINTS*

V. ZEIDAN† AND P. ZEZZA‡

Abstract. In this paper we introduce a definition of “normality” and of “conjugate points” for a general optimal control problem. Using these concepts we obtain new second-order necessary conditions for optimality. In the special case when the control set U is the whole space or in the classical setting of calculus of variations, our conditions reduce to known results, namely, the Jacobi condition and the existence of a solution to a certain Riccati equation.

Key words. optimal control, conjugate points, normality, Riccati equation, necessary conditions

AMS(MOS) subject classification. 49B10

Introduction. In the classical calculus of variations, conjugate point theory plays an important role in establishing second-order necessary and sufficient conditions for optimality (as reference, see among others [2], [4]–[6]). It is well known that if \hat{x} is optimal and the “strengthened Legendre” condition holds, then there is no point in (a, b) conjugate to b . On the other hand, the nonexistence of conjugate points to b in $[a, b)$ is a condition that belongs to a set of sufficient conditions. It is also known as the Jacobi condition and could be formulated in terms of the existence of a solution to a certain Riccati equation.

For optimal controls, the Jacobi condition in a form of Riccati equation has been used in developing sufficient conditions when the control set U is open [9]. Then, it was extended to the general case in [12]–[15]. Dealing with the question of necessary conditions, the accessory problem was obtained in [8] for the case when U is open; then it was generalized, in [10] for the case when U is closed, and in [7] for the abnormal problems. However, the results on necessary conditions using conjugate points in the calculus of variations are not yet transferred to the optimal control context, particularly when the control set U is not trivial. The books [8] and [11], for instance, stop short of conjugate points in their discussion of optimal control theory.

There have been several attempts to define when a point c in $[a, b)$ is conjugate to b and then to show that the nonexistence of conjugate points in (a, b) is necessary for optimality. For instance, the first formulation for the case when $U = R^m$ was introduced in [3]. There, the definition of a conjugate point is not the one needed, and moreover, in [16] we provided a counterexample to the necessity theorem therein. A recent attempt was made in [1] to treat the case when U is given by smooth functions. Here again the results are incorrect, as is shown by two counterexamples in [16].

It is then clear that a mathematically rigorous proof of the necessity theorem is needed for the case when $U = R^m$. Moreover, an extension of the concepts and the results to the general case is desired.

In this paper we use the results involving the accessory problem [10] or [7] to, first, formulate definitions of “conjugate points” and of “normality” for general optimal control problems. Naturally, they must involve the control set U . Second, we show

* Received by the editors April 14, 1986; accepted for publication (in revised form) March 18, 1987.

† Department of Mathematics, University of Alberta, Edmonton, Alberta, Canada T6G 2G1. The research of this author was supported by the Natural Sciences and Engineering Research Council of Canada under grant NSERC 8570.

‡ Dipartimento di Sistemi e Informatica, Università di Firenze, 50139 Florence, Italy.

that, under the “normality” assumption, the nonexistence of points conjugate to b in (a, b) is necessary for optimality. Then, we formulate our new necessary conditions in terms of the existence of a solution to a certain Riccati-type equation, Corollary 3.2, which reduces to the classical form used for sufficiency when $U = R^m$ or when the optimal control \hat{u} is in the int $L^\infty[a, b]$ (see for instance [9]). We present an example illustrating that the “normality” assumption cannot be eliminated. Then we study the special case when we have a free endpoint. There, we show that our results can be strengthened. Finally, we give an example to which we apply our new necessary conditions.

1. Statement of the problem. Given a compact interval $I = [a, b]$, $A \in R^n$, open sets $X \subset R^n$, $V \subset R^m$, a set $U \subset V$ and functions f, g, K and Φ

$$\begin{aligned} g &: I \times X \times V \rightarrow R, \\ f &: I \times X \times V \rightarrow R^n, \\ K &: X \rightarrow R, \\ \Phi &: X \rightarrow R^k, \end{aligned}$$

the optimal control problem can be stated as

$$(P) \quad \text{Minimize } J(x, u) := K(x(b)) + \int_a^b g(t, x(t), u(t)) dt$$

over all absolutely continuous functions x , $x \in AC(I, X)$, and measurable functions u satisfying

$$\begin{aligned} (1.1) \quad \dot{x}(t) &= f(t, x(t), u(t)) \quad \text{a.e. } t \in I, \\ u(t) &\in U \quad \text{a.e. } t \in I, \\ x(a) &= A, \quad \Phi(x(b)) = 0. \end{aligned}$$

DEFINITION 1.1. A pair (x, u) is said to be feasible for (P) if $x \in AC(I, X)$, u is measurable and (x, u) satisfies the constraints (1.1).

DEFINITION 1.2. A feasible pair (\hat{x}, \hat{u}) is a weak local minimum for (P) if for some positive ε , (\hat{x}, \hat{u}) minimizes $J(x, u)$ over all feasible pairs (x, u) satisfying

$$\begin{aligned} x(t) &\in \hat{x}(t) + \varepsilon B_n, \quad t \in I, \\ u(t) &\in \hat{u}(t) + \varepsilon B_m \quad \text{a.e. } t \in I, \end{aligned}$$

where B_N is the unit ball in R^N .

2. Preliminary results. For $F := (f, g)$, the following smoothness assumptions will be recalled.

Assumption (A). For all $t \in I$, $F(t, \cdot, \cdot)$ is C^1 , for all $(x, u) \in X \times V$, $F(\cdot, x, u)$, $F_x(\cdot, x, u)$ and $F_u(\cdot, x, u)$ are measurable on I ; there exists an integrable function $\alpha: I \rightarrow R$ such that

$$(2.1) \quad |F(t, x, u)| + |F_x(t, x, u)| + |F_u(t, x, u)| \leq \alpha(t), \quad (t, x, u) \in I \times X \times V;$$

and K and Φ are C^1 on X .

Assumption (B). For all $t \in I$, $F(t, \cdot, \cdot)$ is C^2 , for all $(x, u) \in X \times V$, $F_{xx}(\cdot, x, u)$, $F_{xu}(\cdot, x, u)$ and $F_{uu}(\cdot, x, u)$ are measurable on I ; there exists an integrable function $\beta: I \rightarrow R$ such that

$$|F_{xx}(t, x, u)| + |F_{xu}(t, x, u)| + |F_{uu}(t, x, u)| \leq \beta(t), \quad (t, x, u) \in I \times X \times V;$$

K and Φ are C^2 on X .

Given $\hat{u}(\cdot)$ and $\varepsilon > 0$, we denote the ε -tube in U around $\hat{u}(t)$ by

$$U_\varepsilon(t) = \{u \in R^m : u \in U \cap (\hat{u}(t) + \varepsilon B_m)\}.$$

In what follows the notation $\hat{\alpha}(t)$ means $\alpha(t, \hat{x}(t), \hat{u}(t))$.

THEOREM 2.1 (Pontryagin's Principle). *Let (\hat{x}, \hat{u}) be a weak local minimum for (P) with a corresponding $\varepsilon > 0$. If Assumption (A) holds then there exist an absolutely continuous function $p : I \rightarrow R^n$, a number λ_0 , and a vector $\gamma \in R^k$ such that*

- (1) $\lambda_0 \geq 0$, $\lambda_0 + |\gamma| \neq 0$,
- (2) $-\dot{p}(t) = \hat{f}_x^T(t)p(t) + \lambda_0 \hat{g}_x(t)$ a.e. $t \in I$,
- (3) $p(b)^T = \gamma^T D\Phi(\hat{x}(b)) + \lambda_0 DK(\hat{x}(b))$,
- (4) $\min_{t \in I} \{\langle p(t), f(t, \hat{x}(t), u) \rangle + \lambda_0 g(t, \hat{x}(t), u) : u \in U_\varepsilon(t)\} = \langle p(t), \hat{f}(t) \rangle + \lambda_0 \hat{g}(t)$ a.e.

If in addition U is convex, then (4) implies

$$(2.2) \quad \langle \hat{f}_u^T(t)p(t) + \lambda_0 \hat{g}_u(t), u - \hat{u}(t) \rangle \geq 0, \quad u \in U \quad \text{a.e. } t \in I.$$

When Assumption (A) holds and U is convex we will use the following.

DEFINITION 2.1. An extremal for problem (P) is a feasible pair (\hat{x}, \hat{u}) such that there exist an absolutely continuous function $p : I \rightarrow R^n$, a number λ_0 and a vector $\gamma \in R^k$ that satisfy conditions (1)–(3) of Pontryagin's principle and condition (2.2).

Let (\hat{x}, \hat{u}) be a feasible pair. To identify the set of admissible directions define:

$$\mathcal{D} = \{(\eta(\cdot), w(\cdot)) \in AC(I, R^n) \times \mathcal{U} \text{ satisfying (2.3)–(2.5)}\},$$

where

$$\mathcal{U} = \{w(\cdot) \in L^\infty(I, R^m) : w(t) \in U - \hat{u}(t) \text{ a.e. } t \in I\},$$

$$(2.3) \quad \dot{\eta}(t) = \hat{f}_x(t)\eta(t) + \hat{f}_u(t)w(t) \text{ a.e. } t \in I,$$

$$(2.4) \quad \eta(a) = 0, \quad D\Phi(\hat{x}(b))\eta(b) = 0,$$

$$(2.5) \quad DK(\hat{x}(b))\eta(b) + \int_a^b [\hat{g}_x(t)\eta(t) + \hat{g}_u(t)w(t)] dt \leq 0.$$

Define

$$(2.6) \quad H(t, x, p, u, \lambda) = \langle p, f(t, x, u) \rangle + \lambda g(t, x, u).$$

The following result can be derived from Theorem 1.1 [10] or of Theorem 3.1 [7] upon writing our problem in Mayer form.

THEOREM 2.2. *Let (\hat{x}, \hat{u}) be a weak local minimum for (P). If U is convex, Assumptions (A) and (B) hold, and $\hat{u} \in L^\infty(I, R^m)$, then for each $(\eta, w) \in \mathcal{D}$ there exist p , λ_0 and γ that can depend on (η, w) , satisfying conditions (1)–(3) of Theorem 2.1, condition (2.2) and*

$$(2.7) \quad \begin{aligned} J_2(\eta, w) := & \frac{1}{2} \eta^T(b) \{ \lambda_0 D^2 K(\hat{x}(b)) + [D^2 \Phi(\hat{x}(b))]^T \gamma \} \eta(b) \\ & + \frac{1}{2} \int_a^b \{ \eta^T(t) \hat{H}_{xx}(t, p(t), \lambda_0) \eta(t) + 2\eta^T(t) \hat{H}_{xu}(t, p(t), \lambda_0) w(t) \\ & + w^T(t) \hat{H}_{uu}(t, p(t), \lambda_0) w(t) \} dt \geq 0, \end{aligned}$$

where

$$\hat{H}(t, p(t), \lambda_0) = H(t, \hat{x}(t), p(t), \hat{u}(t), \lambda_0).$$

The following proposition provides conditions under which $\lambda_0 \neq 0$ and hence the set of admissible directions for the control variable can be described in terms of the Hamiltonian.

PROPOSITION 2.1. We are given an extremal (\hat{x}, \hat{u}) such that $D\Phi(\hat{x}(b))$ is of full rank and the only solution of the system

$$\begin{aligned} -\dot{p}(t) &= \hat{f}_x^T(t)p(t) \quad \text{a.e. } t \in I, \\ p(b)^T &= l^T D\Phi(\hat{x}(b)) \quad \text{for some } l \in R^k, \\ \langle \hat{f}_u^T(t)p(t), u - \hat{u}(t) \rangle &\geq 0 \quad \text{a.e. } t \in I \quad \forall u \in U \end{aligned}$$

is $p \equiv 0$. Then λ_0 can be taken to be 1 and if, moreover, (\hat{x}, \hat{u}) is a weak local minimum then (2.5) reduces to equality and is equivalent to:

$$\langle \hat{H}_u(t, p(t), 1), w(t) \rangle = 0 \quad \text{a.e. } t \in I.$$

Proof. Any λ_0, γ, p satisfying conditions (1)–(3) of Theorem 2.1 and (2.2) must have $\lambda_0 \neq 0$. Otherwise, we get

$$\begin{aligned} -\dot{p}(t) &= \hat{f}_x^T(t)p(t) \quad \text{a.e. } t \in I, \\ p(b)^T &= \gamma^T D\Phi(\hat{x}(b)), \\ \langle \hat{f}_u^T(t)p(t), u - \hat{u}(t) \rangle &\geq 0 \quad \text{a.e. } t \in I \quad \forall u \in U. \end{aligned}$$

Thus $p \equiv 0$ and $\gamma^T D\Phi(\hat{x}(b)) = 0$, and hence $\gamma = 0$, contradicting condition (1) of Pontryagin's principle.

Let (\hat{x}, \hat{u}) be a weak local minimum. From Theorem 2.2 we know that there exist p, λ_0, γ satisfying conditions (1)–(3) of Theorem 2.1 and (2.2). Since $\lambda_0 \neq 0$, then $\lambda_0 = 1$ and (2.2) implies that, for $(\eta, w) \in \mathcal{D}$,

$$0 \leq \langle \hat{H}_u(t, p(t), 1), w(t) \rangle = \langle \hat{f}_u^T(t)p(t) + \hat{g}_u(t), w(t) \rangle \quad \text{a.e. } t \in I.$$

Thus

$$\begin{aligned} 0 &\leq \int_a^b \langle \hat{f}_u^T(t)p(t) + \hat{g}_u(t), w(t) \rangle dt \\ &= \int_a^b \{ \langle p(t), \dot{\eta}(t) - \hat{f}_x(t)\eta(t) \rangle + \langle \hat{g}_u(t), w(t) \rangle \} dt \\ &= \int_a^b \{ \langle p(t), \dot{\eta}(t) \rangle - \langle \hat{f}_x^T(t)p(t), \eta(t) \rangle + \langle \hat{g}_u(t), w(t) \rangle \} dt \\ &= \int_a^b \{ \langle p(t), \dot{\eta}(t) \rangle - \langle -\dot{p}(t) - \hat{g}_x(t), \eta(t) \rangle + \langle \hat{g}_u(t), w(t) \rangle \} dt \\ &= p(b)\eta(b) + \int_a^b \{ \langle \hat{g}_x(t), \eta(t) \rangle + \langle \hat{g}_u(t), w(t) \rangle \} dt \\ &= \langle \gamma^T D\Phi(\hat{x}(b)), \eta(b) \rangle + DK(\hat{x}(b))\eta(b) \\ &\quad + \int_a^b \{ \langle \hat{g}_x(t), \eta(t) \rangle + \langle \hat{g}_u(t), w(t) \rangle \} dt \leq 0. \end{aligned}$$

Therefore (2.5) can be replaced by equality, in other words, by

$$\langle \hat{H}_u(t, p(t), 1), w(t) \rangle = 0 \quad \text{a.e. } t \in I. \quad \text{Q.E.D.}$$

Let (\hat{x}, \hat{u}) be an extremal for (P) and let $\hat{p}(\cdot), \lambda_0, \gamma$ satisfy conditions (1)–(3) of Theorem 2.1 and condition (2.2). We can then take for the control variable admissible directions belonging to the set:

$$(2.8) \quad \tilde{U}(t) = \{ w \in R^n : w \in U - \hat{u}(t) \text{ and } \langle \hat{H}_u(t, \hat{p}(t), \lambda_0), w \rangle = 0 \}.$$

In order to use the previous results the following assumption will be made:

(H) The matrix $D\Phi(\hat{x}(b))$ is of full rank, and the pair (\hat{x}, \hat{u}) is strongly normal on $[a, b]$, i.e., the only solution of

$$\begin{aligned} -\dot{p}(t) &= \hat{f}_x^T(t)p(t) \quad \text{a.e. } t \in I, \\ p(b)^T &= l^T D\Phi(\hat{x}(b)) \quad \text{for some } l \in R^k, \\ \langle \hat{f}_u^T(t)p(t), w \rangle &\geq 0 \quad \text{a.e. } t \in I \quad \forall w \in \tilde{U}(t) \end{aligned}$$

is $p \equiv 0$.

Remark 2.1. Assumption (H) is not merely a technical condition; it is related to the controllability of the linearized system. In fact, it can be easily proven that assumption (H) is equivalent to the following.

Denote by \mathcal{R} the reachable set for the system

$$\begin{aligned} \dot{y}(t) &= \hat{f}_x(t)y(t) + \hat{f}_u(t)w(t), \\ y(a) &= 0, \\ w(t) &\in \tilde{U}(t) \quad \text{a.e. } t \in I; \end{aligned}$$

then $0 \in \text{rel int } D\Phi(\hat{x}(b))\mathcal{R}$.

PROPOSITION 2.2. Let (\hat{x}, \hat{u}) be an extremal for (P), and $\hat{p}(\cdot)$, λ_0 , γ satisfying conditions (1)–(3) of Theorem 2.1 and condition (2.2). If (H) is satisfied, then $\lambda_0 \neq 0$. Set $\lambda_0 = 1$, then $\lambda_0 = 1$, $\hat{p}(\cdot)$, and γ are unique.

Proof. The proof is by contradiction. Suppose there exists another triple $\bar{p}(\cdot)$, $\bar{\gamma}$, λ_0 corresponding to (\hat{x}, \hat{u}) . Since $U - \hat{u}(t) \supset \tilde{U}(t)$ then assumption (H) implies that $\lambda_0 \neq 1$; set $\lambda_0 = 1$ and we get

$$\begin{aligned} -(\dot{\bar{p}}(t) - \dot{\hat{p}}(t)) &= \hat{f}_x^T(t)(\bar{p}(t) - \hat{p}(t)) \quad \text{a.e. } t \in I, \\ (\bar{p} - \hat{p})^T(b) &= (\bar{\gamma} - \gamma)^T D\Phi(\hat{x}(b)), \\ \langle \hat{f}_u^T(t)\bar{p}(t) + \hat{g}_u(t), w \rangle &\geq 0 \quad \text{a.e. } t \in I \quad \forall w \in \tilde{U}(t), \\ \langle \hat{f}_u^T(t)\hat{p}(t) + \hat{g}_u(t), w \rangle &= 0 \quad \text{a.e. } t \in I \quad \forall w \in \tilde{U}(t). \end{aligned}$$

Thus

$$\langle \hat{f}_u^T(t)(\bar{p}(t) - \hat{p}(t)), w \rangle \geq 0 \quad \text{a.e. } t \in I \quad \forall w \in \tilde{U}(t),$$

and by (H) we get $\bar{p} \equiv \hat{p}$. Since $D\Phi\hat{x}(b)$ is of full rank we obtain $\bar{\gamma} = \gamma$. Q.E.D.

Let (\hat{x}, \hat{u}) be an extremal. From the preceding we obtain: let $\hat{p}(\cdot)$, λ_0 , γ satisfying (1)–(3) of Theorem 2.1 and condition (2.2). Define $\tilde{U}(t)$ by (2.8) and

$$(2.9) \quad \tilde{\mathcal{U}} = \{w(\cdot) \in L^\infty(I, R^m); w(t) \in \tilde{U}(t) \text{ a.e. } t \in I\}.$$

Assume (H); then λ_0 can be taken to be 1. Consider the so-called accessory problem to (P), i.e.,

$$\begin{aligned} \text{(AP)} \quad \text{Minimize} \quad J_2(\eta, w) &= \frac{1}{2} \eta^T(b) \Gamma \eta(b) \\ &+ \frac{1}{2} \int_a^b \{ \eta^T(t) \hat{H}_{xx}(t) \eta(t) + 2\eta^T(t) \hat{H}_{xu}(t) w(t) \\ &+ w^T(t) \hat{H}_{uu}(t) w(t) \} dt \end{aligned}$$

over all $(\eta, w) \in AC(I, R^n) \times \text{cone } \tilde{\mathcal{U}}$, satisfying

$$\begin{aligned} \dot{\eta}(t) &= \hat{f}_x(t)\eta(t) + \hat{f}_u(t)w(t) \quad \text{a.e. } t \in I, \\ \eta(a) &= 0, \quad \Psi\eta(b) = 0, \end{aligned}$$

where

$$\begin{aligned}\Psi &= D\Phi(\hat{x}(b)), \\ \Gamma &= D^2K(\hat{x}(b)) + D^2\Phi(\hat{x}(b))^T\gamma, \\ \hat{H}(t) &= H(t, \hat{x}(t), \hat{p}(t), \hat{u}(t), 1).\end{aligned}$$

When (\hat{x}, \hat{u}) is a weak local minimum, then $(0, 0)$ solves (AP), i.e.,

$$J_2(\eta, w) \geq 0.$$

3. Main results. Discussion. In the classical calculus of variations the definition of conjugate point is derived by computing the second variation of the original problem which yields what is known as the accessory problem, whose Euler-Lagrange equation is the Jacobi equation. However, in the optimal control setting, the uniqueness of the adjoint function $p(t)$ (which is guaranteed by assumption (H)) is required to obtain an accessory problem out of the second variation, as we have done in § 2. The application of the maximum principle, Theorem 2.1, to the accessory problem (AP) motivates the definition of conjugate point below.

Throughout this section we will suppose that Assumptions (A) and (B) and (H) are satisfied and that (\hat{x}, \hat{u}) is an extremal for (P) with $\hat{u} \in L^\infty(I, R^m)$.

DEFINITION 3.1. A point $c \in [a, b]$ is said to be conjugate to b if there exists nonzero $(\lambda, \eta, w) \in AC(I, R^n) \times AC(I, R^n) \times \text{cone } \tilde{\mathcal{U}}$ such that, for some $l_1 \in R^k$,

$$\begin{aligned}(3.1) \quad \dot{\eta}(t) &= \hat{f}_x(t)\eta(t) + \hat{f}_u(t)w(t) \quad \text{a.e. } t \in I, \\ -\dot{\lambda}(t) &= \hat{H}_{xx}(t)\eta(t) + \hat{H}_{xu}(t)w(t) + \hat{f}_x^T(t)\lambda(t) \quad \text{a.e. } t \in I,\end{aligned}$$

$$\begin{aligned}(3.2) \quad \Psi\eta(b) &= 0, \quad \eta(c) = 0, \\ \lambda(b) &= \Psi^T l_1 + \Gamma\eta(b),\end{aligned}$$

and

$$(3.3) \quad \langle \hat{H}_{uu}(t)w(t) + \hat{H}_{ux}(t)\eta(t) + \hat{f}_u^T(t)\lambda(t), z \rangle = 0 \quad \text{a.e. } t \in I \quad \forall z \in \text{span } \tilde{U}(t).$$

Remark 3.1. Since Assumption (H) holds, then a straightforward application of the Pontryagin principle to the accessory problem yields that whenever (η, w) is its solution, there exists λ such that (λ, η, w) satisfies (3.1), (3.2) with $c = a$, and

$$\langle \hat{H}_{uu}(t)w(t) + \hat{H}_{ux}(t)\eta(t) + \hat{f}_u^T(t)\lambda(t), z \rangle \geq 0 \quad \text{a.e. } t \in I \quad \forall z \in \tilde{U}(t),$$

instead of the equality in (3.3).

Remark 3.2. If U is convex with nonempty interior, then

$$\text{span } \tilde{U}(t) = \hat{H}_u(t)^\perp := \{w \in R^m : \langle \hat{H}_u(t), w \rangle = 0\}.$$

Remark 3.3. Some geometrical assumptions on the original problem may simplify Definition 3.1. For instance, if U is open or $\hat{u}(t) \in \text{int } U$ almost everywhere $t \in I$ then, for almost all t in I , $\hat{H}_u(t) = 0$, $\text{span } \tilde{U}(t) = R^m$, and $\text{cone } \tilde{\mathcal{U}} = \text{cone } \mathcal{U}$. But $\text{cone } \mathcal{U} = L^\infty(I, R^m)$ if and only if $\hat{u} \in \text{int } L^\infty(I, U)$.

Solving (3.3) for w allows us to replace (3.1)–(3.3) by a $2n$ -system of linear differential equations in terms of η and λ only.

From now on all the functions appearing in any formula have to be considered evaluated at t .

PROPOSITION 3.1. Assume that U is convex with nonempty interior. If $\hat{H}_{uu}(t) > 0$ almost everywhere $t \in I$, then (η, λ, w) satisfies Definition 3.1 if and only if $0 \neq (\lambda, \eta) \in AC(I, \mathbb{R}^n) \times AC(I, \mathbb{R}^n)$ and

$$(3.4) \quad \dot{\eta} = [\hat{f}_x - \hat{f}_u K \hat{H}_{uu}^{-1} \hat{H}_{ux}] \eta - \hat{f}_u K \hat{H}_{uu}^{-1} \hat{f}_u^T \lambda \quad \text{a.e. } t \in I,$$

$$(3.5) \quad -\dot{\lambda} = [\hat{H}_{xx} - \hat{H}_{xu} K \hat{H}_{uu}^{-1} \hat{H}_{ux}] \eta + [\hat{f}_x^T - \hat{H}_{xu} K \hat{H}_{uu}^{-1} \hat{f}_u^T] \lambda \quad \text{a.e. } t \in I$$

with

$$\Psi \eta(b) = 0, \quad \eta(c) = 0,$$

$$\lambda(b) = \Psi^T l_1 + \Gamma \eta(b) \quad \text{for some } l_1 \in \mathbb{R}^k,$$

and

$$w \in \mathcal{U},$$

where

$$(3.6) \quad w := -K \hat{H}_{uu}^{-1} [\hat{H}_{ux} \eta + \hat{f}_u^T \lambda],$$

and

$$(3.7) \quad K(t) = \begin{cases} I_m & \text{if } \hat{H}_u(t) = 0, \\ -\frac{\hat{H}_{uu}^{-1} \hat{H}_u \hat{H}_u^T}{\hat{H}_u^T \hat{H}_{uu}^{-1} \hat{H}_u} + I_m & \text{if } \hat{H}_u(t) \neq 0. \end{cases}$$

Proof. If c is conjugate to b , then there exists $(\lambda, \eta, w) \neq 0$ satisfying Definition 3.1. The function w can be taken in $\tilde{\mathcal{U}}$. Since $\text{span } \tilde{U}(t) = \hat{H}_u^\perp(t)$, from (3.3) it follows that

$$\hat{H}_{uu}(t)w(t) + \hat{H}_{ux}(t)\eta(t) + \hat{f}_u^T(t)\lambda(t) = 0 \quad \text{if } \hat{H}_u(t) = 0$$

or

$$\hat{H}_{uu}(t)w(t) + \hat{H}_{ux}(t)\eta(t) + \hat{f}_u^T(t)\lambda(t) = \alpha(t)\hat{H}_u^T(t) \quad \text{if } \hat{H}_u(t) \neq 0,$$

where $\alpha(t) \in \mathbb{R}$.

Thus, when $\hat{H}_u(t) = 0$,

$$w(t) = -\hat{H}_{uu}^{-1}(t)[\hat{H}_{ux}(t)\eta(t) + \hat{f}_u^T(t)\lambda(t)].$$

On the other hand, when $\hat{H}_u(t) \neq 0$,

$$w(t) = -\hat{H}_{uu}^{-1}(t)[\hat{H}_{ux}(t)\eta(t) + \hat{f}_u^T(t)\lambda(t)] + \alpha(t)\hat{H}_{uu}^{-1}(t)\hat{H}_u(t).$$

Multiply both sides of the above equation by $\hat{H}_u^T(t)$. Note that $w \in \text{cone } \tilde{\mathcal{U}}$ to get

$$0 = -\hat{H}_u^T(t)\hat{H}_{uu}^{-1}(t)[\hat{H}_{ux}(t)\eta(t) + \hat{f}_u^T(t)\lambda(t)] + \alpha(t)\hat{H}_u^T(t)\hat{H}_{uu}^{-1}(t)\hat{H}_u(t)$$

and hence

$$\alpha(t) = \frac{\hat{H}_u^T(t)\hat{H}_{uu}^{-1}(t)}{\hat{H}_u^T(t)\hat{H}_{uu}^{-1}(t)\hat{H}_u(t)} [\hat{H}_{ux}(t)\eta(t) + \hat{f}_u^T(t)\lambda(t)];$$

substituting $\alpha(\cdot)$ in the above equation it is easy to get

$$w(t) = \left[\frac{\hat{H}_{uu}^{-1}(t)\hat{H}_u(t)\hat{H}_u^T(t)}{\hat{H}_u^T(t)\hat{H}_{uu}^{-1}(t)\hat{H}_u(t)} - I \right] \hat{H}_{uu}^{-1}(t)(\hat{H}_{ux}(t)\eta(t) + \hat{f}_u^T(t)\lambda(t)).$$

Therefore (3.3) yields

$$(3.8) \quad w(t) = -K(t)\hat{H}_{uu}^{-1}(t)[\hat{H}_{ux}(t)\eta(t) + \hat{f}_u^T(t)\lambda(t)] \quad \text{a.e. } t \in I,$$

where $K(t)$ is given by (3.7). Now replace (3.8) in each of (3.1) and (3.2). We obtain that (λ, η) satisfies (3.4)–(3.6) and the boundary conditions.

Conversely, assume that there exist (λ, η) satisfying the conditions of Proposition 3.1. From (3.6) and (3.7) the following result:

$$\left. \begin{array}{l} \text{If } \hat{H}_u(t) = 0, \text{ then} \\ \text{If } \hat{H}_u(t) \neq 0, \text{ then} \end{array} \right\} \hat{H}_u(t)w(t) = 0;$$

$$\hat{H}_u^T(t)w(t) = \left[\frac{\hat{H}_u^T(t)\hat{H}_{uu}^{-1}(t)\hat{H}_u(t)\hat{H}_u^T(t)}{\hat{H}_u^T(t)\hat{H}_{uu}^{-1}(t)\hat{H}_u(t)} - \hat{H}_u^T(t) \right] \hat{H}_{uu}^{-1}(t)(\hat{H}_{ux}(t)\eta(t) + \hat{f}_u^T(t)\lambda(t)) = 0.$$

Thus w , defined by (3.6), is in \mathcal{Q} . If we replace (3.6) in (3.4) and (3.5) we obtain that (3.1) and (3.2) are satisfied by (λ, η) and w . It remains to show (3.3). If $\hat{H}_u(t) \neq 0$, (3.6) and (3.7) give

$$\begin{aligned} & \hat{H}_{uu}(t)w(t) + \hat{H}_{ux}(t)\eta(t) + \hat{f}_u^T(t)\lambda(t) \\ &= \frac{\hat{H}_u(t)\hat{H}_u^T(t)\hat{H}_{uu}^{-1}(t)}{\hat{H}_u^T(t)\hat{H}_{uu}^{-1}(t)\hat{H}_u(t)} (\hat{H}_{ux}(t)\eta(t) + \hat{f}_u^T(t)\lambda(t)) = \alpha(t)\hat{H}_u(t), \end{aligned}$$

where $\alpha(t)$ is defined as above. Therefore, for all $z \in \hat{H}_u^\perp(t)$,

$$\langle \hat{H}_{uu}(t)w(t) + \hat{H}_{ux}(t)\eta(t) + \hat{f}_u^T(t)\lambda(t), z \rangle = 0 \quad \text{a.e. } t \in I.$$

If $\hat{H}_u(t) = 0$, then (3.3) follows immediately from (3.6). Q.E.D.

Remark 3.4. Some regularity assumptions on H imply that the control defined by (3.6) is in L^∞ . For instance, if for some $\alpha > 0$, $\hat{H}_{uu}^{(t)} \geq \alpha I_m$, for almost all $t \in I$, then $\hat{H}_{uu}^{-1}(t) \in L^\infty(I, R^{m \times m})$. If, in addition, $H_{ux}(\cdot, x, u)$ and $f_u(\cdot, x, u)$ are essentially bounded, so is the function w given by (3.6). Since (\hat{x}, \hat{u}) is essentially bounded, thus, it remains to show that, when $\hat{H}_u(t) \neq 0$, $-\hat{K}\hat{H}_{uu}^{-1}$ is essentially bounded. Let $0 < \lambda_1(t) \leq \dots \leq \lambda_m(t)$ be the eigenvalues of $\hat{H}_{uu}^{-1}(t)$; then

$$|\lambda_i(t)| \leq \|\hat{H}_{uu}^{-1}\|_\infty =: \rho, \quad i = 1, 2, \dots, m.$$

Thus the λ_i 's are in $L^\infty(I, R)$. Since \hat{H}_{uu}^{-1} is symmetric, then

$$\hat{H}_{uu}^{-1}(t) = P^T(t)\Lambda(t)P(t),$$

where P is such that $P^T P = I$ and

$$\Lambda = \begin{bmatrix} \lambda_1 & & & 0 \\ & \ddots & & \\ 0 & & & \lambda_m \end{bmatrix}.$$

Hence, when $\hat{H}_u(t) \neq 0$, (3.7) yields

$$-\hat{K}\hat{H}_{uu}^{-1} + \hat{H}_{uu}^{-1} = \frac{\hat{H}_{uu}^{-1}\hat{H}_u\hat{H}_u^T\hat{H}_{uu}^{-1}}{\hat{H}_u^T\hat{H}_{uu}^{-1}\hat{H}_u} = \frac{P^T\Lambda P\hat{H}_u\hat{H}_u^T P^T\Lambda P}{\hat{H}_u^T P^T\Lambda P\hat{H}_u} = P^T \left(\frac{\lambda_i\lambda_j\alpha_i\alpha_j}{\sum_{i=1}^m \alpha_k^2\lambda_k} \right) P,$$

where

$$P\hat{H}_u = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}.$$

But for any i, j

$$\left| \frac{\lambda_i\lambda_j\alpha_i\alpha_j}{\sum_{i=1}^m \alpha_k^2\lambda_k} \right| \leq \left| \frac{\lambda_i\lambda_j\alpha_i\alpha_j}{\alpha_i^2\lambda_i + \alpha_j^2\lambda_j} \right| \leq \frac{1}{2} \sqrt{\lambda_i} \sqrt{\lambda_j} \leq \frac{1}{2} \rho.$$

Therefore w is essentially bounded.

Our main result concerning the nonexistence of conjugate points corresponding to a weak local minimum requires the following two normality assumptions:

(H₁) $\Psi = D\Phi(\hat{x}(b))$ is of full rank and there exists a unique solution of

$$-\dot{p}(t) = \hat{f}_x^T(t)p(t) \quad \text{a.e. } t \in [c, b],$$

$$p(b)^T = l^T \Psi \quad \text{for some } l \in \mathbb{R}^k$$

and

$$\langle \hat{f}_u^T(t)p(t), w \rangle \geq 0 \quad \text{a.e. } t \in [c, b] \quad \forall w \in \tilde{U}(t)$$

on any subinterval $[c, b]$ of I .

(H₂) There exists a unique solution of

$$-\dot{p}(t) = \hat{f}_x^T(t)p(t) \quad \text{a.e. } t \in [a, c],$$

$$\langle \hat{f}_u^T(t)p(t), w \rangle \geq 0 \quad \text{a.e. } t \in [a, c] \quad \forall w \in \tilde{U}(t)$$

on any subinterval $[a, c]$ of I .

Our main result is given by the following theorem.

THEOREM 3.1. *Let (\hat{x}, \hat{u}) be a weak local minimum for (P) with $\hat{u} \in L^\infty(I, \mathbb{R}^m)$. Assume that (H₁) and (H₂) hold and that U is convex with nonempty interior. If $\hat{H}_{uu}(t) \geq \alpha I_m$, for $\alpha > 0$ almost everywhere $t \in I$, then there is no point in (a, b) conjugate to b .*

Before proving Theorem 3.1, we will present an example which illustrates the indispensability of the strong normality assumptions (H₁) and (H₂).

Example 1. Consider the following optimal control problem:

$$\text{Minimize } J(u) = \frac{1}{2} \int_0^1 u^2(t) dt$$

subject to

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{pmatrix} = B(t)u(t) \quad \text{a.e. } t \in [0, 1],$$

$$x(0) = x(1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where

$$B(t) = \begin{cases} \begin{pmatrix} 1 \\ 0 \end{pmatrix} & \text{on } \left[0, \frac{1}{2}\right] \\ \begin{pmatrix} 0 \\ 1 \end{pmatrix} & \text{on } \left(\frac{1}{2}, 1\right] \end{cases},$$

and $U = \mathbb{R}$.

It is clear that the solution of the problem is $(\hat{x}, \hat{u}) \equiv \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, 0\right)$. Let us first show that Assumption (H) holds. Let $p = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$ satisfying on $[0, 1]$, $\dot{p}(t) = 0$ and $B^T(t)p(t) = 0$; then $p \equiv \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$ and $c_1 = c_2 = 0$. Thus (H) is satisfied.

Now take $p \equiv \begin{pmatrix} 0 \\ 1 \end{pmatrix}$; it is clear that, on an interval of the form $[0, c]$, when $0 < c < \frac{1}{2}$, such p satisfies $\dot{p}(t) = 0$ and $B^T(t)p(t) = 0$. Thus Assumption (H₂) is violated. On the other hand, for $p \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ we have $\dot{p}(t) = 0$ and $B^T(t)p(t) = 0$ on any interval of the form $[c, 1]$, where $\frac{1}{2} < c < 1$. Thus (H₁) does not hold; therefore Theorem 3.1 does not apply.

In fact, take

$$\eta(t) = \begin{cases} \begin{pmatrix} -t + \frac{1}{2} \\ 0 \end{pmatrix} & \text{on } [0, \frac{1}{2}], \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \text{on } [\frac{1}{2}, 1] \end{cases}$$

and $\lambda(t) \equiv \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Then (η, λ) satisfies (3.4) and (3.5), that is,

$$\begin{aligned} \dot{\eta}(t) &= -B(t)B^T(t)\lambda(t) \quad \text{a.e. } t \in [0, 1], \\ -\dot{\lambda}(t) &= 0, \end{aligned}$$

with

$$\eta(1) = \eta(c) = 0 \quad \text{for any } c \in [\frac{1}{2}, 1].$$

This implies that any point in $[\frac{1}{2}, 1]$ is a conjugate point to 1.

Proof of Theorem 3.1. Suppose that there exists a point $c \in (a, b)$ conjugate to b ; then there exists $(\lambda, \eta, w) \neq 0$ satisfying Definition 3.1. We can assume that w satisfies $2w \in \tilde{\mathcal{U}}$. Define

$$(\bar{\lambda}(t), \bar{\eta}(t), \bar{w}(t)) = (\lambda(t), \eta(t), w(t))\chi_{(c,b]}(t) \quad \text{for } t \in I$$

($\chi_A(\cdot)$ is the characteristic function of the set A). Then: $(\bar{\eta}, \bar{w}) \in AC(I, R^n) \times \tilde{\mathcal{U}}$ and

$$\begin{aligned} \dot{\bar{\eta}}(t) &= \hat{f}_x(t)\bar{\eta}(t) + \hat{f}_u(t)\bar{w}(t) \quad \text{a.e. } t \in I, \\ (3.9) \quad \bar{\eta}(a) &= 0, \quad \Psi\bar{\eta}(b) = 0. \end{aligned}$$

Thus $(\bar{\eta}, \bar{w})$ is feasible for the accessory problem. By using (3.1)–(3.3) on $[c, b]$ we get

$$\begin{aligned} J_2(\bar{\eta}, \bar{w}) &= \frac{1}{2} \eta^T(b) \Gamma \eta(b) \\ &\quad + \frac{1}{2} \int_c^b \{ \eta^T(t) \hat{H}_{xx}(t) \eta(t) + 2 \eta^T(t) \hat{H}_{xu}(t) w(t) + w^T(t) \hat{H}_{uu}(t) w(t) \} dt \\ &= \frac{1}{2} \eta^T(b) \Gamma \eta(b) + \frac{1}{2} \int_c^b \{ -\dot{\lambda}^T(t) \eta(t) - \lambda^T(t) \hat{f}_x(t) \eta(t) - \eta^T(t) \hat{H}_{xu}(t) w(t) \\ &\quad + 2 \eta^T(t) \hat{H}_{xu}^T(t) w(t) + w^T(t) \hat{H}_{uu}(t) w(t) \} dt \\ &= \frac{1}{2} \eta^T(b) \Gamma \eta(b) + \frac{1}{2} \int_c^b \{ -\dot{\lambda}^T(t) \eta(t) - \lambda^T(t) \hat{f}_x(t) \eta(t) + \eta^T(t) \hat{H}_{xu}(t) w(t) \\ &\quad - \eta^T(t) \hat{H}_{xu}(t) w(t) - \lambda^T(t) \hat{f}_u(t) w(t) \} dt \\ &= \frac{1}{2} \eta^T(b) \Gamma \eta(b) + \frac{1}{2} \int_c^b \{ -\dot{\lambda}^T(t) \eta(t) - \lambda^T(t) \dot{\eta}(t) \} dt \\ &= \frac{1}{2} \eta^T(b) \Gamma \eta(b) - \frac{1}{2} \eta^T(b) \lambda(b) \\ &= \frac{1}{2} \eta^T(b) \Gamma \eta(b) - \frac{1}{2} \eta^T(b) [\Psi^T l_1 + \Gamma \eta(b)] \\ &= -\frac{1}{2} \eta^T(b) \Psi^T l_1 = 0. \end{aligned}$$

Thus $(\bar{\eta}, \bar{w})$ minimizes $J_2(\eta, w)$ over $(\eta, w) \in \mathcal{U}$ with $\bar{w} \in \text{cone } \tilde{\mathcal{U}}$. In particular, $(\bar{\eta}, \bar{w})$ provides a minimum for the accessory problem (AP), with the additional condition $w \in \tilde{\mathcal{U}}$. By the Pontryagin principle we get the following: there exist a number α_0 , a vector $v \in R^k$ and an absolutely continuous function p such that

$$(3.10) \quad \alpha_0 \geq 0, \quad |v| + \alpha_0 \neq 0,$$

$$(3.11) \quad -\dot{p}(t) = \hat{f}_x^T(t)p(t) + \alpha_0[\hat{H}_{xx}(t)\bar{\eta}(t) + \hat{H}_{xu}(t)\bar{w}(t)] \quad \text{a.e. } t \in I,$$

$$(3.12) \quad p(b) = \Psi^T v + \alpha_0 \Gamma \bar{\eta}(b),$$

and

$$\min \{ \langle \hat{f}_u^T(t)p(t), w \rangle + \frac{1}{2}\alpha_0[2\bar{\eta}(t)\hat{H}_{xu}(t)w + w^T\hat{H}_{uu}(t)w] : \forall w \in \tilde{U}(t) \}$$

is attained at $\bar{w}(t)$ almost everywhere $t \in I$.

This last condition implies

$$\langle \hat{f}_u^T(t)p(t) + \alpha_0[\hat{H}_{ux}(t)\bar{\eta}(t) + \hat{H}_{uu}(t)\bar{w}(t)], w - \bar{w}(t) \rangle \geq 0 \quad \text{a.e. } t \in I \quad \forall w \in \tilde{U}(t).$$

Since 0 and $2\bar{w}(t)$ are in $\tilde{U}(t)$, almost everywhere $t \in I$, then

$$(3.13) \quad \langle \hat{f}_u^T(t)p(t) + \alpha_0[\hat{H}_{ux}(t)\bar{\eta}(t) + \hat{H}_{uu}(t)\bar{w}(t)], w \rangle \geq 0 \quad \text{a.e. } t \in I \quad \forall w \in \tilde{U}(t);$$

and equality in (3.13) holds at $\bar{w}(t)$.

If $\alpha_0 = 0$, then

$$\begin{aligned} -\dot{p}(t) &= \hat{f}_x^T(t)p(t) \quad \text{a.e. } t \in I, \\ p(b) &= \Psi^T v, \end{aligned}$$

and

$$\langle \hat{f}_u^T(t)p(t), w \rangle \geq 0 \quad \text{a.e. } t \in I \quad \forall w \in \tilde{U}(t).$$

By hypothesis (H), $p \equiv 0$ and hence $v = 0$, which yields a contradiction. Thus we can take α_0 to be 1.

On $[a, c]$, (3.11) and (3.13) yield

$$-\dot{p}(t) = \hat{f}_x^T(t)p(t) \quad \text{a.e. } t,$$

and

$$\langle \hat{f}_u^T(t)p(t), w \rangle \geq 0 \quad \text{a.e. } t \quad \forall w \in \tilde{U}(t).$$

Thus by (H₂) we obtain

$$p \equiv 0 \quad \text{on } [a, c].$$

Now, if we use (3.1)–(3.3) with (3.11)–(3.13) we get for $t \in [c, b]$,

$$\begin{aligned} -(\dot{p}(t) - \dot{\lambda}(t)) &= \hat{f}_x^T(t)[p(t) - \lambda(t)] \quad \text{a.e. } t, \\ [p(b) - \lambda(b)] &= \Psi^T(v - l_1) \end{aligned}$$

and

$$\langle \hat{f}_u^T(t)[p(t) - \lambda(t)], w \rangle \geq 0 \quad \text{a.e. } t \quad \forall w \in \tilde{U}(t).$$

Using (H₁) the following results:

$$p \equiv \lambda \quad \text{on } [c, b];$$

hence $\lambda(c) = 0$. But from Proposition 3.1 we know that (λ, η, w) satisfies (3.4)–(3.6). Since $\lambda(c) = \eta(c) = 0$ it follows that

$$\lambda \equiv \eta \equiv w \equiv 0 \quad \text{on } [a, b],$$

which contradicts the fact that $(\lambda, \eta, w) \neq 0$. Q.E.D.

Let us now consider the matrix system corresponding to (3.4) and (3.5):

$$(3.14) \quad \begin{aligned} \dot{X} &= [\hat{f}_x - \hat{f}_u K \hat{H}_{uu}^{-1} \hat{H}_{ux}] X - \hat{f}_u K \hat{H}_{uu}^{-1} \hat{f}_u^T \Lambda, \\ -\dot{\Lambda} &= [\hat{H}_{xx} - \hat{H}_{xu} K \hat{H}_{uu}^{-1} \hat{H}_{ux}] X + [\hat{f}_x^T - \hat{H}_{xu} K \hat{H}_{uu}^{-1} \hat{f}_u^T] \Lambda \end{aligned}$$

with

$$(3.15) \quad \begin{aligned} \Psi X(b) &= 0, \\ \Lambda(b) &= \Psi^T N + \Gamma X(b) \quad \text{for some } N. \end{aligned}$$

Define

$$(3.16) \quad Z = \{\alpha \in \mathbb{R}^n: -K(t) \hat{H}_{uu}^{-1}(t) [\hat{H}_{ux}(t) X(t) + \hat{f}_u^T(t) \Lambda] \alpha \in U - \hat{u}(t) \text{ a.e. } t \in I\}.$$

From Proposition 3.1, Remark 3.4 and Theorem 3.1 we can deduce the following corollary.

COROLLARY 3.1. *Assume that $H_{ux}(\cdot, x, u)$ and $f_u(\cdot, x, u)$ are essentially bounded. Then under the conditions of Theorem 3.1, we have:*

$$X(t) \alpha \neq 0 \quad \forall \alpha \in Z \text{ and } \forall t \in (a, b) \quad \text{where } X(\cdot) \text{ solves (3.14) and (3.15).}$$

Remark 3.5. When $\hat{u} \in \text{int } L^\infty(I, U)$, then cone $Z = \mathbb{R}^n$ and hence $X(t)$ is invertible on (a, b) .

Let (X, Λ) be a solution of (3.14) and (3.15) and let S be any subspace: $S \subset (\text{cone } Z) \cup (-\text{cone } Z)$. Let r be its dimension and Y be an $n \times r$ matrix whose columns form an orthonormal basis of S .

COROLLARY 3.2. *Under the conditions of Corollary 3.1 there exists a locally Lipschitz matrix function Q on (a, b) such that:*

$$(3.17) \quad Y^T X^T(t) [Q(t) - Q^T(t)] X(t) Y = 0 \quad \text{on } (a, b),$$

and

$$(3.18) \quad L(Q(t)) X(t) Y = 0 \quad \text{a.e. } t \in (a, b),$$

where

$$\begin{aligned} L(Q(t)) &:= \dot{Q}(t) - Q(t) \hat{f}_u(t) K(t) \hat{H}_{uu}^{-1}(t) \hat{f}_u^T(t) Q(t) \\ &\quad + Q(t) [\hat{f}_x(t) - \hat{f}_u(t) K(t) \hat{H}_{uu}^{-1}(t) \hat{H}_{ux}(t)] \\ &\quad + [\hat{f}_x^T(t) - \hat{H}_{xu}(t) \hat{H}_{uu}^{-1}(t) K^T(t) \hat{f}_u^T(t)] Q(t) \\ &\quad + [\hat{H}_{xx}(t) - \hat{H}_{xu}(t) K(t) \hat{H}_{uu}^{-1}(t) \hat{H}_{ux}(t)]. \end{aligned}$$

Remark 3.6. If $\hat{u} \in \text{int } L^\infty(I, U)$, then $X(t)$ and Y are invertible and (3.17) and (3.18) reduce to Q being symmetric and satisfying the Riccati equation

$$L(Q(t)) = 0 \quad \text{a.e. } t \in (a, b).$$

Proof of Corollary 3.2. We first notice that any (X, Λ) satisfying (3.14) and (3.15) also satisfies

$$X^T(t) \Lambda(t) = \Lambda^T(t) X(t).$$

From Corollary 3.1 it follows that the $r \times r$ matrix $Y^T X^T(t) X(t) Y$ is invertible on (a, b) . Define

$$(3.19) \quad Q(t) := \Lambda(t) Y [Y^T X^T(t) X(t) Y]^{-1} Y^T X^T(t).$$

Equation (3.17) can now be easily deduced.

From Remark 3.4 and (3.14) it follows that $Q(t)$ is locally Lipschitz on (a, b) . Now using (3.14) and (3.19) we can verify that $Q(t)$ satisfies (3.18) on (a, b) . Q.E.D.

4. Special cases. In this section we study some special cases in which the results can be strengthened and the hypotheses can be simplified.

Consider the optimal control problem (P) where $\Phi \equiv 0$, that is, we have a free final state. In this case the problem is automatically normal on any $[c, b] \subset [a, b]$, that is Hypothesis (H_1) , and hence (H), holds. Proposition 2.2 is then valid and if (\hat{x}, \hat{u}) is a weak local minimum, then

$$\forall (\eta, w) \in AC(I, R^n) \times \text{cone } \tilde{\mathcal{U}}$$

with

$$\dot{\eta}(t) = \hat{f}_x(t)\eta(t) + \hat{f}_u(t)w(t) \quad \text{a.e. } t \in I,$$

$$\eta(a) = 0$$

we have

$$J_2(\eta, w) = \frac{1}{2} \eta^T(b) D^2 K(\hat{x}(b)) \eta(b) + \frac{1}{2} \int_a^b \{ \eta^T(t) \hat{H}_{xx}(t) \eta(t) + 2 \eta^T(t) \hat{H}_{xu}(t) w(t) + w^T(t) \hat{H}_{uu}(t) w(t) \} dt,$$

where

$$\begin{aligned} \tilde{\mathcal{U}} = \{ w(\cdot) \in L^\infty(I, R^m) : w(t) \in U - \hat{u}(t) \text{ a.e. } t \in I \text{ and} \\ \langle w(t), \hat{H}_u(t) \rangle = 0 \text{ a.e. } t \in I \}. \end{aligned}$$

THEOREM 4.1. Let (\hat{x}, \hat{u}) be a weak local minimum for (P) with $\Phi \equiv 0$ and $\hat{u} \in L^\infty(I, R^m)$, and let U be closed and convex. Then

$$J_2(\eta, w) \geq 0 \quad \forall (\eta, w) \in AC(I, R^n) \times W$$

with

$$\dot{\eta}(t) = \hat{f}_x(t)\eta(t) + \hat{f}_u(t)w(t) \quad \text{a.e. } t \in I,$$

$$\eta(a) = 0,$$

where

$$(4.1) \quad \begin{aligned} W = \{ w(\cdot) \in L^\infty(I, R^m) : w(t) \in \text{cone } \{ U - \hat{u}(t) \} \text{ a.e. } t \in I \text{ and} \\ \langle \hat{H}_u(t), w(t) \rangle = 0 \text{ a.e. } t \in I \}. \end{aligned}$$

(In general $W \supset \text{cone } \tilde{\mathcal{U}}$.)

Proof. Let (η, w) satisfy the conditions of the theorem, and define

$$\begin{aligned} A_n &= \left\{ t \in I : d \left(\frac{1}{n} w(t), U - \hat{u}(t) \right) = 0 \right\} \\ &= \left\{ t \in I : \frac{1}{n} w(t) \in U - \hat{u}(t) \right\} \\ &= \left\{ t \in I : \frac{1}{n} w(t) + \hat{u}(t) \in U \right\}; \end{aligned}$$

then A_n is measurable. Since $0 \in U - \hat{u}(t)$ and $U - \hat{u}(t)$ is convex, then $A_n \subset A_{n+1}$. From (4.1) we get

$$\exists \Omega \subset I: \mu(\Omega) = 0 \quad \text{and} \quad \forall t \in \Omega^C, \quad w(t) \in \text{cone} \{U - \hat{u}(t)\},$$

where $\Omega^C = \{t \in I: t \notin \Omega\}$. Thus $\Omega^C \subset \bigcup_{n=1}^{\infty} A_n$, and hence from the inclusion $A_n \subset A_{n+1}$ we get

$$\lim_{n \rightarrow \infty} \mu(A_n^C) = 0.$$

Define

$$w_n(t) = w(t) \chi_{A_n}(t) \quad \text{where } A_0 = [a, b]$$

and let $\eta_n(t)$ be the solution of

$$\begin{aligned} \dot{\eta}(t) &= \hat{f}_x(t) \eta(t) + \hat{f}_u(t) w_n(t), \\ \eta(a) &= 0. \end{aligned}$$

Let $E(t, s)$ be the evolution operator associated to

$$\dot{\eta}(t) = \hat{f}_x(t) \eta(t).$$

We get

$$\eta(t) - \eta_n(t) = \int_{[a, t] \cap A_n^C} E(t, s) \hat{f}_u(s) w(s) ds;$$

thus

$$\begin{aligned} |\eta(t) - \eta_n(t)| &\leq \int_{[a, t] \cap A_n^C} |E(t, s)| |\hat{f}_u(s)| |w(s)| ds \\ &\leq M \|w\|_{\infty} \int_{A_n^C} |\hat{f}_u(s)| ds \quad \forall t \in I, \end{aligned}$$

where

$$M = \max_{\substack{t \in I \\ s \in I}} |E(t, s)|.$$

On the other hand,

$$\begin{aligned} |\eta(t) + \eta_n(t)| &\leq 2M \|w\|_{\infty} \|\hat{f}_u\|_1 \quad \forall t \in I, \\ |w(t) - w_n(t)| &= |w(t)| \chi_{A_n^C}(t) \quad \text{a.e. } t \in I, \\ |w(t) + w_n(t)| &\leq 2\|w\|_{\infty} \quad \text{a.e. } t \in I. \end{aligned}$$

Since J_2 can be written as

$$J_2(\eta, w) = \frac{1}{2} \eta^T(b) D^2 K(\hat{x}(b)) \eta(b) + \frac{1}{2} \int_a^b (\eta^T(t), w^T(t)) D_{x,u}^2 \hat{H}(t) \begin{pmatrix} \eta(t) \\ w(t) \end{pmatrix} dt$$

and $D^2K(\hat{x}(b))$ and $D_{x,u}^2\hat{H}(t)$ are symmetric, then

$$\begin{aligned}
 2|J_2(\eta, w) - J_2(\eta_n, w_n)| &\leq |(\eta(b) + \eta_n(b))^T D^2K(\hat{x}(b))(\eta(b) - \eta_n(b))| \\
 &\quad + \left| \int_a^b (\eta^T(t) + \eta_n^T(t), w^T(t) \right. \\
 &\quad \left. + w_n^T(t)) D_{x,u}^2\hat{H}(t) \begin{pmatrix} \eta(t) - \eta_n(t) \\ w(t) - w_n(t) \end{pmatrix} dt \right| \\
 &\leq |\eta(b) + \eta_n(b)| |D^2K(\hat{x}(b))| |\eta(b) - \eta_n(b)| \\
 &\quad + \int_a^b \left\{ (|(\eta + \eta_n)(t)| + |(w + w_n)(t)|) |D_{x,u}^2\hat{H}(t)| \right. \\
 &\quad \left. \cdot \left[M \|w\|_\infty \int_{A_n^C} |\hat{f}_u(s)| ds + |w(t)| \chi_{A_n^C}(t) \right] \right\} dt \\
 &\leq 2M^2 \|w\|_\infty^2 \|\hat{f}_u\|_1 |D^2K(\hat{x}(b))| \|\hat{f}_u\|_\infty \mu(A_n^C) \\
 &\quad + (2\|w\|_\infty + 2M\|w\|_\infty \|\hat{f}_u\|_1) |D_{x,u}^2\hat{H}(t)|_\infty \\
 &\quad \cdot [M\|w\|_\infty \|\hat{f}_u\|_1 \mu(A_n^C) + \|w\|_\infty \mu(A_n^C)](b-a).
 \end{aligned}$$

Since $\mu(A_n^C) \rightarrow 0$, as $n \rightarrow \infty$, we get

$$\lim_{n \rightarrow \infty} J_2(\eta_n, w_n) = J_2(\eta, w).$$

But $w_n \in n(U - \hat{u}(t))$ or $w_n \in \text{cone } \tilde{\mathcal{U}}$ almost everywhere $t \in I$. Therefore $J_2(\eta_n, w_n) \geq 0$ for all n , which implies that

$$J_2(\eta, w) \geq 0. \quad \text{Q.E.D.}$$

Let us make the following hypothesis:

(H₂)^{*} There exists a unique solution of

$$\begin{aligned}
 -\dot{p}(t) &= \hat{f}_x^T(t)p(t) \quad \text{a.e. } t \in I, \\
 \langle \hat{f}_u^T(t)p(t), w \rangle &\geq 0 \quad \text{a.e. } t \in I \quad \forall w \in \text{cone } \tilde{\mathcal{U}}(t)
 \end{aligned}$$

on each subinterval $[a, c]$ of $[a, b]$.

In Definition 3.1 and in the definition of \mathcal{U} , replace cone $\tilde{\mathcal{U}}$ and $U - \hat{u}(t)$ by W and cone $\{U - \hat{u}(t)\}$, respectively. If this is done we obtain the following corollary.

COROLLARY 4.1. *When $\Phi \equiv 0$, then Theorem 3.1 and Corollaries 3.1 and 3.2 remain valid if (H₂)^{*} is assumed instead of (H₁) and (H₂) and if in (3.16) we replace $U - \hat{u}(t)$ by cone $\{U - \hat{u}(t)\}$. If in addition we have $\hat{u}(t) \in \text{int } U$, almost everywhere $t \in I$, then*

$$W = L^\infty(I, R^m), \quad S = R^n$$

and

$$L(Q(t)) = 0 \quad \text{a.e. } t \in (a, b).$$

Remark 4.1. When $\Phi \neq 0$, we know that the results of the second part of Corollary 4.1 hold if and only if $\hat{u} \in \text{int } L^\infty(I, U)$. Hence the condition $\hat{u} \in \text{int } U$, almost everywhere $t \in I$, is not enough to get the results and does not give a special case unless $\Phi \equiv 0$.

5. An example. Consider the following optimal control problem:

$$(P) \quad \text{Minimize} \quad J(u) = \int_0^{2\pi} \left\{ x(t) e^{-t} + tu(t) e^{-t} \sin \frac{1}{t(2\pi-t)} + u^2(t) \right\} dt$$

$$\text{subject to} \quad \dot{x}(t) = x(t) + u(t) \sin \frac{1}{t(2\pi-t)},$$

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

$$u(t) \in U = [0, 1] \quad \text{a.e.}$$

Here $f(t, x, u) = x + u \sin (1/t(2\pi - 1))$ and $g(t, x, u) = xe^{-t} + tu e^{-t} \sin (1/t(2\pi - t)) + u^2$. Let us first show that the solution of the above problem is $(\hat{x}, \hat{u}) \equiv 0$.

The solution of the differential equation satisfying the boundary condition is

$$(5.1) \quad x(t) = e^t \int_0^t e^{-s} u(s) \sin \frac{1}{s(2\pi-s)} ds,$$

with

$$(5.2) \quad \int_0^{2\pi} e^{-s} u(s) \sin \frac{1}{s(2\pi-s)} ds = 0.$$

Thus,

$$J(u) = \int_0^{2\pi} \left\{ \int_0^t e^{-s} u(s) \sin \frac{1}{s(2\pi-s)} ds + tu(t) e^{-t} \sin \frac{1}{t(2\pi-t)} + u^2(t) \right\} dt.$$

Integrating the first term by parts and using (5.2), we get

$$\begin{aligned} J(u) &= \left\{ t \int_0^t e^{-s} u(s) \sin \frac{1}{s(2\pi-s)} ds \right\}_{t=0}^{t=2\pi} - \int_0^{2\pi} t e^{-t} u(t) \sin \frac{1}{t(2\pi-t)} dt \\ &\quad + \int_0^{2\pi} tu(t) e^{-t} \sin \frac{1}{t(2\pi-t)} dt + \int_0^{2\pi} u^2(t) dt \\ &= \int_0^{2\pi} u^2(t) dt. \end{aligned}$$

Since $(\hat{x}, \hat{u}) = (0, 0)$ is admissible for the problem (P), then it is optimal.

Define

$$\hat{p}(t) = -t e^{-t};$$

then it satisfies

$$-\dot{\hat{p}}(t) = \hat{f}_x^T(t) \hat{p}(t) + \hat{g}_x(t) = \hat{p}(t) + e^{-t}$$

and

$$\hat{f}_u^T(t) \hat{p}(t) + \hat{g}_u(t) = 2\hat{u}(t) \equiv 0.$$

Thus

$$\tilde{U}(t) = U = [0, 1].$$

Now we show that (H_1) and (H_2) hold. If $p(t)$ solves the system

$$-\dot{p}(t) = p(t),$$

$$\sin \frac{1}{t(2\pi-t)} p(t) \geq 0,$$

then $p(t) = ce^{-t}$ and $ce^{-t} \sin(1/t(2\pi - t)) \geq 0$. Since $\sin(1/t(2\pi - t))$ oscillates near $t = 0$ and near $t = 2\pi$, then $p \equiv 0$ on any interval of the form $[0, c]$ or $[c, 2\pi]$. Therefore, Hypotheses (H_1) and (H_2) are satisfied.

Since $\hat{H}_{uu}(t) \equiv 2$ for this problem then, by Theorem 3.1, there is no point in $(0, 2\pi)$ conjugate to 2π , i.e., there exists no $(\lambda, \eta) \neq 0$ such that

$$\begin{aligned}\dot{\eta}(t) &= \eta(t) - \frac{1}{2} \left[\sin^2 \frac{1}{t(2\pi - t)} \right] \lambda(t) \quad \text{a.e. } t \in [0, 1], \\ -\dot{\lambda}(t) &= \lambda(t) \quad \text{a.e. } t \in [0, 1], \\ -\frac{\lambda}{2}(t) \left(\sin \frac{1}{t(2\pi - t)} \right) &\in [0, 1],\end{aligned}$$

and $\eta(2\pi) = \eta(c) = 0$ for some $c \in (0, 2\pi)$.

Acknowledgment. We thank Professor J. Macki, who made it possible for Professor Zezza to spend one year at the University of Alberta during which time this research was performed.

REFERENCES

- [1] P. BERNHARD, *La théorie de la seconde variation et le problème linéaire quadratique*, in *Advance in Hamiltonian Systems*, J. P. Aubin, A. Bensoussan and I. Ekeland, eds., Birkhäuser, Boston, MA, 1983, pp. 109-142.
- [2] G. A. BLISS, *Lectures on the Calculus of Variations*, Univ. of Chicago Press, Chicago, IL, 1946.
- [3] J. V. BREAKWELL AND Y. C. HO, *On the conjugate point condition for the control problem*, *Internat. J. Engrg. Sci.*, 2 (1965), pp. 565-579.
- [4] C. CARATHEODORY, *Calculus of Variations and Partial Differential Equations of the First Order, Part II*, Holden-Day, San Francisco, CA, 1967.
- [5] G. M. EWING, *Calculus of Variations with Applications*, W. W. Norton, New York, 1969.
- [6] I. M. GELFAND AND S. V. FOMIN, *Calculus of Variations* (R. A. Silverman, trans.), Prentice-Hall, Englewood Cliffs, NJ, 1963.
- [7] E. G. GILBERT AND D. S. BERNSTEIN, *Second-order necessary conditions in optimal control: accessory-problem results without normality conditions*, *J. Optim. Theory Appl.*, 41 (1983), pp. 75-106.
- [8] M. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
- [9] D. Q. MAYNE, *Sufficient conditions for a control to be strong minimum*, *J. Optim. Theory Appl.*, 21 (1977), pp. 339-351.
- [10] J. WARGA, *A second-order Lagrangian condition for restricted control problems*, *J. Optim. Theory Appl.*, 24 (1978), pp. 475-483.
- [11] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, New York, 1969.
- [12] V. ZEIDAN, *Conjugate points criterion for optimal control*, *Proc. IEEE Conference on Decision and Control*, 1983, pp. 379-383.
- [13] ———, *Sufficient conditions for the generalized problem of Bolza*, *Trans. Amer. Math. Soc.*, 275 (1983), pp. 561-583.
- [14] ———, *Extended Jacobi criterion for optimal control*, *this Journal*, 22 (1984), pp. 294-301.
- [15] ———, *First and second order sufficient conditions for optimal control and the calculus of variations*, *J. Appl. Math. Optim.*, 11 (1984), pp. 209-226.
- [16] V. ZEIDAN AND P. ZEZZA, *Conjugate points and optimal control: counter-examples*, to appear.

FILTERING WITH OBSERVATIONS ON A RIEMANNIAN SYMMETRIC SPACE*

MONIQUE PONTIER† AND JACQUES SZPIRGLAS‡

Abstract. The point under discussion in this paper refers to a problem of filtering in which observation Y of a system state X is a process with values in a symmetric space M . The observation is a Brownian motion transformed by an isometry of M depending on the state. It takes its values in manifold M and its multiplicative formulation is nonstandard. In many physical situations, e.g., mechanics, robotics, spatial fields, the filtering problems are naturally set up in manifolds as well for the signal and the observation. The reference probability method is used to construct the model. Then filtering equations are deduced; these comply with the conditional law according to its observation. Unique characterization of this conditional law is given. Last, two examples are investigated. First the multivariate case: the observation Y is in \mathbb{R}^d so that $Y_t = \Sigma_t(X) W_t + H_t(X)$, where W is a multivariate Brownian motion; Σ (a rotation) and H (a translation) are absolutely continuous with respect to time t . The second example is filtering on the sphere; observation Y is of the following form: $Y_t = g_t(X) \cdot W_t$, where W is a Brownian motion on sphere S_2 and $g_t(X)$ is a rotation of M absolutely continuous with respect to time t .

Key words. filtering, manifolds, symmetric space

AMS (MOS) subject classifications. 93E11, 60G35, 60J60

1. Introduction. This paper is devoted to a filtering model with observations taking their values in a manifold, and the multiplicative formulation of which is nonstandard. In many physical situations, e.g., mechanics, robotics, spatial fields, the filtering problems are naturally set up in manifolds as well for the signal and the observation (for instance, satellite tracking). The model is as follows: the observation Y of a Markovian signal X with generator $(A, \mathcal{D}(A))$ takes its values in a symmetric space M . Briefly speaking, M is such that there exists a subgroup G of the Lie group $I(M)$ of all isometries of M , which acts transitively on M . Process Y_t is assumed to be

$$(1.1) \quad Y_t = g_t(X) \cdot W_t$$

where W_t is a Brownian motion on M and $g_t(X)$ is a G -valued functional of signal X such that $t \rightarrow g_t(X)$ is absolutely continuous on $[0, T]$:

$$(1.2) \quad dg_t = (R_{g_t})_* \psi(t, X_s; s \leq t) dt, \quad g_0 = e; \quad t \in [0, T],$$

where ψ is a bounded semimartingale functional of t and the sample path of X , taking values in Lie algebra \mathfrak{G} , and $(R_g)_* \psi$ is the right invariant vector field associated to ψ .

This paper is devoted to the construction of the filtering model and the computation of the stochastic differential equation which is satisfied by the conditional law of signal X , given its observation Y , filter π of X given Y . We construct this model by the reference probability method [5], [21], [22], using an important result by Shigekawa [18] about equivalence of the laws of $g \cdot W$ and W : it can be seen that the problem is finally reduced to a nontrivial multivariate problem with a nonbounded signal. Then we get the filtering equation for a bounded function f in domain $\mathcal{D}(A)$ of X_t :

$$(1.3) \quad \pi_t(f) = \mu(f) + \int_0^t \pi_s(Af) ds + \int_0^t \omega_{\pi_s}(\bar{f}\psi^*) \circ dY_s - \frac{1}{2} \int_0^t \pi_s(\bar{f}(\kappa_s + \|\psi_{Y_s}^*\|^2)) ds$$

* Received by the editors July 1, 1985; accepted for publication (in revised form) August 4, 1987.

† Université d'Orléans, Département de Mathématiques et d'Informatique, 45046 Orleans Cedex, France.

‡ Institut National des Télécommunications, 9, rue Charles Fourier, 91011 Evry Cedex, France.

where $\tilde{f}_t = f(X_t) - \pi_t(f)$, ψ^* is a vector field on M deduced from ψ and the stochastic integral is this of the one-form associated to vector field $\pi_s(\tilde{f}\psi^*)$ (defined in [19]); κ is a real process, due to symmetric structure of space M .

Then, applying Kurtz and Ocone's results [11], we give a unique characterization of the conditional law of signal X given its observation Y .

Our results are similar to those of Ng and Caines [15] for the general Riemannian case. But our restriction to a symmetric space allows a more general formulation of the filtering problem. Duncan [5] was also concerned with filtering on symmetric spaces, but he used this last structure to ensure the nonexplosion of the Brownian motion and studied the same problem as Ng and Caines. Furthermore, Duncan's construction of the Brownian motion on a manifold is quite different from that of Ikeda and Watanabe [8]. The relations between those two constructions are studied in [4]. For more details about diffusions on manifolds, see [14] or [17].

A few necessary results about symmetric spaces and Brownian motion on Riemannian symmetric space are given in § 2. Section 3 is devoted to the construction of the filtering model and filtering equations, which have, under slight conditions, unique solution as measure-valued equations. We conclude with examples about multivariate and spherical cases.

Notation. All filtrations in this paper are assumed to be complete and right continuous. For any process Z defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, \mathcal{G}^Z will denote the natural filtration generated by Z , completed with respect to probability \mathbb{P} , and right continuous.

Moreover, the Stratonovitch integral is denoted by $\circ : \int Y \circ dB$ and Ito's integral by $: \int Y \cdot dB$.

Last, $(L_g)_*$ and $(R_g)_*$ denote the differential of the left and right translations on G ; so $(R_g)_*\psi$ is the right invariant vector field associated to any ψ in \mathfrak{G} . Except in (1.2) we are concerned essentially with left invariant vector fields on G defined from vectors ξ of \mathfrak{G} by $(L_g)_*\xi$, denoted by $\xi(g)$, and we generally omit the point where they are taken.

2. Stochastic calculus on symmetric spaces. In this section, some definitions about symmetric spaces are first recalled. Then Brownian motions on symmetric spaces are introduced according to different points of view. This induces some results about their natural filtrations which obviously have an importance in our applications to filtering. Some rules about stochastic calculus for semimartingales on Lie groups [18] are recalled in § 2.3.

2.1. Symmetric spaces. A d -dimensional symmetric space M is a C^∞ d -dimensional Riemannian manifold such that for each $p \in M$, the geodesic symmetry s_p with respect to p is an isometry and p is an isolated fixed point of s_p . The following properties will be useful in the sequel (for more details, see [3] or [7]). Let $I(M)$ be the isometric transformation group of M and G be a subgroup of $I(M)$, which is a connected component containing identity e ; it is well known that the Lie group G acts on M on the left transitively. Take a point O of M and let K be the isotropy subgroup of G at O (i.e., for all $k \in K$, $k \cdot O = O$); K is a compact subgroup of G ; furthermore G/K and M are diffeomorphic.

Let \mathfrak{G} and \mathfrak{K} denote, respectively, Lie algebras of G and K . Let σ be an automorphism of G defined by

$$(2.1) \quad \sigma(g) = s_0 g s_0, \quad g \in G.$$

Because σ is involutive, its differential σ_* on \mathfrak{G} has ± 1 for eigenvalues. The eigenspace

for 1 is \mathfrak{K} . Let \mathfrak{M} denote the eigenspace for -1 . Then we have the direct sum: $\mathfrak{G} = \mathfrak{M} \oplus \mathfrak{K}$. Space \mathfrak{M} is d -dimensional. This can be proved as follows. Define mapping $i: G \rightarrow M$ by

$$(2.2) \quad i(g) = g \cdot O, \quad g \in G;$$

its differential at point e , i_* , is an isomorphism between \mathfrak{M} and $T_0(M)$, the tangent space at point O .

As in Shigekawa's paper for a better understanding we use indices I, J, K, \dots for vectors $x = (x_I; I = 1, \dots, n)$ in algebra \mathfrak{G} (n -dimensional); indices i, j, \dots for vectors $x = (x_i; i = 1, \dots, d)$ in \mathfrak{M} (d -dimensional); indices α, β, \dots for vectors $x = (x_\alpha; \alpha = 1, \dots, p)$ in \mathfrak{K} ($p = n - d$ dimensional).

Now take a point (O, u_0) in $O(M)$, the bundle of orthonormal frames. We extend mapping i to $O(M)$ as follows:

$$(2.3) \quad i(g) = g \cdot u_0 = \{(g_*)_e u_0^i; i = 1, \dots, d\}$$

where the u_0^i are the basic vectors of frame u_0 . To any vector C of \mathfrak{G} is associated a vector field C^* on M such that

$$(2.4) \quad C^* f(x) = \frac{d}{dt} f(\exp tC \cdot x), \quad x \in M.$$

Then we define a linear map $L: \mathfrak{M} \rightarrow T_0(M)$ by

$$(2.5) \quad L(C) = C^*(O)$$

and an isomorphism A from \mathbb{R}^d to \mathfrak{M}

$$(2.6) \quad \xi \rightarrow u_0(\xi) \rightarrow L^{-1}(u_0(\xi)) = A(\xi).$$

Hence from a standard basis (e_i) of \mathbb{R}^d , a basis of \mathfrak{M} can be constructed $(A_i; i = 1, \dots, d)$:

$$(2.7) \quad A_i = A(e_i) = L^{-1}(u_0(e_i)).$$

Take any basis $(A_{d+1}, \dots, A_{d+p})$ of \mathfrak{K} ; then $(A_1, \dots, A_d, A_{d+1}, \dots, A_{d+p})$ form a basis of \mathfrak{G} .

Let us recall that adjoint representation Adg denotes the differential at e of the following automorphism of G :

$$(2.8) \quad h \rightarrow ghg^{-1}.$$

Mapping Ad is a homomorphism of group from G to $GL(\mathfrak{G})$, the linear group of \mathfrak{G} . On the other hand, it is known that \mathfrak{K} is the Lie algebra of K and \mathfrak{M} is invariant under $Ad(K)$, i.e.,

$$(2.9) \quad Adk(\mathfrak{M}) \subset \mathfrak{M}, \quad k \in K.$$

Furthermore, $Adk|_{\mathfrak{M}}$ belongs to $O(d)$. For all C in \mathfrak{G} with decomposition $C_M + C_{\mathfrak{M}}$ in $\mathfrak{K} \oplus \mathfrak{M}$, the decomposition of $Adk(C)$ is given in the same way:

$$(2.10) \quad Adk(C_M) + Adk(C_{\mathfrak{M}}).$$

2.2. Brownian motions on Lie groups and symmetric spaces. Let B be a standard d -dimensional Brownian motion defined on its canonical space $(\Omega^B, \mathcal{G}^B, \mathbb{P}^B)$. There are two classical ways to construct a Brownian motion Y taking values in a symmetric space M : Shigekawa's construction [18] and that of Ikeda and Watanabe [8]. In [8],

process Y is defined as the canonical projection from a bundle of orthonormal frames $O(M)$ to M :

$$(2.11) \quad Y_t = \tau(U_t).$$

Process U is the solution of the stochastic differential equation

$$(2.12) \quad dU_t = B(e_i)(U_t) \circ dB_t^i, \quad U_0 = (O, u_0)$$

where $B(e_i)$ are the canonical horizontal vector fields of $O(M)$, associated to the Riemannian connection of M [7].

Shigekawa defines Brownian motion Y as the image obtained by mapping i of a Brownian motion h taking values in G , which is easier to deal with because h is a diffusion on G and stochastic calculus on Lie groups is well developed (see Proposition 2.2 below, and [1], [18]):

$$(2.13) \quad Y_t = i(h_t) = h_t \cdot O.$$

Process h is the solution of the stochastic differential equation

$$(2.14) \quad dh_t = A_j(h_t) \circ dB_t^j; \quad h_0 = e$$

where the A_j s are vector fields defined by (2.7). It can easily be proved that the two definitions coincide and, moreover, that

$$(2.15) \quad h_t \cdot u_0 = u(t).$$

We have the following result, asserted in [15].

LEMMA 2.1. *Let Y be the Brownian motion on M defined by (2.11) and (2.12). The natural filtration of Y coincides with those of U , h and B .*

Proof. On one hand, process Y is \mathcal{G}^U -adapted because of definition (2.11); hence $\mathcal{G}^Y \subset \mathcal{G}^U$. On the other hand, process U is locally defined by the system [8]:

$$(2.16) \quad \begin{aligned} dy^i(t) &= u_j^i(t) \circ dB_t^j, \\ du_j^i(t) &= -\Gamma_{mk}^i(Y(t)) u_j^k(t) \circ dy^m(t) \end{aligned}$$

where Γ is the Riemannian connection of M . Now a system $(\tilde{U}_\alpha, \tilde{\phi}_\alpha)$ of local coordinates in $O(M)$ is easily deduced from that of M , (U_α, ϕ_α) , as follows:

$$(2.17) \quad \begin{aligned} \tilde{U}_\alpha &= \{U = (x, u) / x \in U_\alpha, u \text{ orthonormal frame at } x\}, \\ \tilde{\phi}_\alpha(x, u) &= (\phi_\alpha(x), (u_j^i)), \quad u_j^i \text{ local coordinates of } u. \end{aligned}$$

Thus the sequence of stopping times defining U_t are \mathcal{G}^Y -stopping times. Then, the pathwise uniqueness of system (2.16) between two stopping times implies as in [16] that $u(t)$ is \mathcal{G}_t^Y -measurable, and this proves $\mathcal{G}^Y = \mathcal{G}^U$. Now, from (2.13) we get

$$(2.18) \quad \mathcal{G}^Y \subset \mathcal{G}^h$$

and from (2.14) and the unicity of its solution,

$$(2.19) \quad \mathcal{G}^h \subset \mathcal{G}^B.$$

In [19], Shigekawa expresses B with respect to U . Let ω be the \mathbb{R}^d -valued 1-form defined by

$$(2.20) \quad \omega_u(\xi) = u^{-1}(\pi_*)_u(\xi), \quad u \in O(M), \quad \xi \in T_u O(M).$$

Then B satisfies

$$(2.21) \quad B_t = \int_0^t \omega_{u_s} \circ dU_s.$$

This implies that $\mathcal{G}^B \subset \mathcal{G}^U$. This inclusion and (2.18)–(2.19) conclude the proof.

2.3. Stochastic calculus on Lie groups. To enlighten the next section we give some stochastic calculus rules for semimartingales on Lie groups [18]. Let k and l be the solution of the following stochastic differential equations on G :

$$(2.22) \quad dk_t = \xi_{i,t} \circ dB_t^i + \xi_{0,t} dt, \quad k_0 = k,$$

$$(2.23) \quad dl_t = \eta_{i,t} \circ dB_t^i + \eta_{0,t} dt, \quad l_0 = l$$

where $\xi_i, \eta_i, i = 1, \dots, d$ are \mathcal{G} -valued continuous semimartingales and ξ_0, η_0 are \mathcal{G} -valued locally integrable processes. From [18] we have the following proposition

PROPOSITION 2.2. *Let us denote m_t the product $k_t l_t$; then process m satisfies the following stochastic differential equation on G :*

$$(2.24) \quad \begin{aligned} dm_t &= (Adl_t^{-1}(\xi_{i,t}) + \eta_{i,t}) \circ dB_t^i + (Adl_t^{-1}(\xi_{0,t}) + \eta_{0,t}) dt, \\ m_0 &= kl. \end{aligned}$$

Let $(Adl_t^{-1})_J^I$ denote the elements of matrix Adl_t^{-1} expressed in basis $(A_I; I = 1, \dots, n)$; then we have

$$(2.25) \quad \begin{aligned} d(Adl_t^{-1})_J^I &= C_{KL}^I (Adl_t^{-1})_J^K (\eta_{i,t}^L \circ dB_t^i + \eta_{0,t}^L dt), \\ (Adl_0^{-1})_J^I &= (Adl^{-1})_J^I \end{aligned}$$

where C_{KL}^I 's are the structure constants defined by

$$(2.26) \quad [A_K, A_L] = C_{KL}^I A_I.$$

Moreover process l^{-1} satisfies the following equation:

$$(2.27) \quad \begin{aligned} dl_t^{-1} &= -Adl_t(\xi_{i,t}) \circ dB_t^i - Adl_t(\xi_{0,t}) dt, \\ l_0^{-1} &= 0. \end{aligned}$$

3. Construction of the model-filtering equations. In this section we first construct the filtering model by the reference probability method [5], [21], [22]. We derive the filtering equations with respect to Brownian motion B (which is observable from Lemma 2.1). Then we express those equations with respect to the observation process Y . This implies Stratonovitch calculus and the extension of some projections theorems to Stratonovitch integrals.

3.1. Construction of the filtering model. Let us construct the filtering model by the reference probability method; we use some of Shigekawa's results [18]. Let signal X be a Markov process valued in a Lusin space E with generator $(A, \mathcal{D}(A))$ and initial law μ . We consider X on probability space $(\Omega^X, \mathcal{G}^X, \mathbb{P}_\mu^X)$. Let observation Y be a Brownian motion valued in M , defined in § 2.2 by means of a standard d -dimensional Brownian motion B on probability space $(\Omega^B, \mathcal{G}^B, \mathbb{P}^B)$. Notice that $\mathcal{G}^B = \mathcal{G}^Y$ (Lemma 2.1).

Let us define

$$\Omega = (\Omega^X \times \Omega^B), \quad \mathcal{G} = \mathcal{G}^X \otimes \mathcal{G}^Y, \quad \mathbb{P}_\mu = \mathbb{P}_\mu^X \otimes \mathbb{P}^B.$$

Processes X, B, Y can be extended trivially to $(\Omega, \mathcal{G}, \mathbb{P}_\mu)$. Let us notice that B and Y are still \mathcal{G} -Brownian motions because of the \mathbb{P}_μ -independence of \mathcal{G}^X and \mathcal{G}^Y .

The probability space modelizing the system signal-observation (1.1)–(1.2) will be constructed by a change of equivalent probability measure.

Let a transformation g_t be given on G with the following assumptions:

$$(3.1) \quad \begin{aligned} dg_t &= (R_{g_t})_* \psi(t, X_s; s \leq t) dt, \quad t \in [0, T], \\ g_0 &= e \end{aligned}$$

where ψ is a \mathcal{G}^X -continuous semimartingale (hence a \mathcal{G} -semimartingale), uniformly bounded on $[0, T] \times \Omega^X$, taking values in \mathfrak{G} . For example, ψ can be taken in $\mathcal{D}(A)$ as in § 3.3. Because of independence of X and Y , process X can be considered as a parameter, and so (3.1) is deterministic and has a unique solution.

We have the following proposition, thanks to the results of Shigekawa.

PROPOSITION 3.1. *With assumption (3.1) a process L can be defined by*

$$(3.2) \quad L_t = \exp \left(\int_0^t \phi_s^i dB_s^i - \frac{1}{2} \int_0^t (\phi_s^i)^2 ds \right)$$

where ϕ_s is defined in \mathfrak{G} by

$$(3.3) \quad \phi_s = (Adh_s^{-1} \psi_s).$$

Moreover, process L is a strictly positive uniformly integrable $(\Omega, \mathcal{G}, \mathbb{P}_\mu)$ -martingale which is \mathcal{G}^Y -locally square integrable. Therefore, let us define: $\mathbb{Q}_\mu = L_T \cdot \mathbb{P}_\mu$; \mathbb{Q}_μ and \mathbb{P}_μ are equivalent probability measures. Furthermore, there exists a $(\Omega, \mathcal{G}, \mathbb{Q}_\mu)$ Brownian motion W valued in M such that

$$(3.4) \quad Y_t = g_t(X) \cdot W_t.$$

Proof. Process X is considered as a parameter and so it is possible to use Shigekawa's results working on $(\Omega^Y, \mathcal{G}^Y, \mathbb{P}^Y)$.

The idea of Shigekawa's proof is to find the following decomposition for h , with a Brownian motion h' on G :

$$(3.5) \quad h_t = g_t \cdot h'_t \cdot P_t$$

where p is a semimartingale on K (hence $p_t \cdot 0 = 0$), the solution of the following differential equation:

$$(3.6) \quad dp_t = -(Adh_t^{-1} \cdot \psi_t)_{\mathfrak{M}} dt, \quad p_0 = e.$$

So, let us define process h' as

$$(3.7) \quad h'_t = g_t^{-1} \cdot h_t \cdot p_t^{-1}.$$

If we apply results recalled from § 2.3, we get

$$(3.8) \quad \begin{aligned} dh'_t &= Adp_t(A_t) \circ dB_t^i - Adp_t(Adh_t^{-1} \cdot \psi_t)_{\mathfrak{M}} dt, \\ h'_0 &= e. \end{aligned}$$

Then we define the d -dimensional process Z :

$$(3.9) \quad \begin{aligned} dZ_t^i &= (Adp_t)_j^i dB_t^j - (Adp_t)_j^i (Adh_t^{-1} \psi_t)_{\mathfrak{M}}^j dt, \\ Z_0^i &= 0. \end{aligned}$$

The results of [18] ((2.7)–(2.10)) for semimartingales valued in G prove easily that (3.8) is equivalent to

$$(3.10) \quad dh'_t = A_t(h'_t) \circ dZ_t^i, \quad h'_0 = e.$$

Thus, h' is a Brownian motion on G as soon as Z is a d -dimensional Brownian motion.

This is the case, if we apply a Liptser and Shyrayev result [13] on absolute continuity of probability measures.

Shigekawa's result is concerned with left invariant vector fields. So we get an expression of g^{-1} with respect to such a vector field. Define process ξ taking values in \mathcal{G} :

$$(3.11) \quad \xi(t, X_s; s \leq t) = \text{Ad}g_t^{-1}(X) \cdot \psi_t$$

where g is the unique solution of (3.1). So we have

$$(3.12) \quad (R_g)_*(\psi_t) = (R_g)_* \circ \text{Ad}g_t(\xi_t) = (L_g)_*(\xi_t).$$

So, the *right* invariant vector field associated to ψ_t and the left invariant vector field associated to ξ_t coincide and (3.1) becomes

$$(3.13) \quad dg_t = \xi(t, X_s; s \leq t)(g_t) dt, \quad g_0 = e$$

and because of (2.27), we get

$$(3.14) \quad dg_t^{-1} = \text{Ad}g_t(\xi_t) dt, \quad g_0 = e;$$

that is to say, because of definition (3.11)

$$(3.15) \quad dg_t^{-1} = (L_{g^{-1}})_*(-\psi_t) dt, \quad g_0 = e.$$

The \mathcal{G} -valued continuous semimartingale ψ is bounded, so by Shigekawa [18, Thm. 4.1] $g_t^{-1}Y_t$ and Y_t have equivalent measures. Thus, $W_t = g_t^{-1}Y_t$ is a Brownian motion associated to G -valued process h' because

$$(3.16) \quad h'_t \cdot 0 = g_t^{-1}h_t p_t^{-1} \cdot 0 = g_t^{-1}h_t \cdot 0 = g_t^{-1} \cdot Y_t = W_t.$$

At last, let us notice that process L defined by (3.2) is $((d\nu'/d\nu)(W \cdot))^{-1}$ where ν and ν' are the measures of Y and W under probability \mathbb{P}_μ ; so L is a Radon-Nicodym derivative and it is a strictly positive uniformly integrable \mathbb{P}_μ -martingale. On the other hand, Girsanov's Theorem and (3.9) show that Z is a $\mathbb{Q}_\mu = L \cdot \mathbb{P}_\mu$ -Brownian motion.

It is left to prove that L is \mathcal{G}^Y locally square integrable. Process ψ is bounded on $[0, T]$ by assumption (3.1) and $\text{Ad}h^{-1}$ is a continuous process which is \mathcal{G}^Y -adapted by Lemma 2.1. Hence, ϕ is \mathcal{G}^Y -locally bounded and as a consequence, it is classical (see, for example, [11]) that L is \mathcal{G}^Y -locally square integrable and thus satisfies the conclusions of Proposition 3.1.

3.2. Filtering equations. However, we assume, as in the standard filtering model,

$$\psi_t = \psi(X_t)$$

where ψ is a \mathcal{G} -valued bounded continuous application on E .

Now, let us define filter π and unnormalized filter $\tilde{\pi}$. It is easy to derive filtering equations with respect to B . We shall insist on the derivation of equations with respect to observation Y , which are perhaps less standard.

Filter π of X given Y is defined as the unique measure valued process such that, for all bounded function f on E , $\pi_t(f)$ is the \mathbb{Q}_μ -optional projection with respect to filtration \mathcal{G}^Y of process $f(X_t)$. Briefly speaking, $\pi_t(f)$ is a "smooth version" of the conditional mean $E_{\mathbb{Q}_\mu}(f(X_t)/\mathcal{G}_t^Y)$. The reference probability method allows simple computations, thanks to the existence of probability \mathbb{P}_μ equivalent to \mathbb{Q}_μ such that, under \mathbb{P}_μ , \mathcal{G}_t and \mathcal{G}_∞^Y are independent conditionally on \mathcal{G}_t^Y , i.e., if

$$(3.17) \quad Z_t \in \mathcal{G}_t, \quad E_{\mathbb{P}_\mu}(Z_t/\mathcal{G}_t^Y) = E_{\mathbb{P}_\mu}(Z_t/\mathcal{G}_\infty^Y).$$

Here, this property is implied by the \mathbb{P}_μ -independence of X and Y . This allows us to express a complex operator of optional projections by means of simpler operators of conditional means:

$$(3.18) \quad E_{\mathbb{Q}_\mu}(Z_t/\mathcal{G}_t^Y) = \frac{E_{\mathbb{P}_\mu}(L_t Z_t/\mathcal{G}_\infty^Y)}{E_{\mathbb{P}_\mu}(L_t/\mathcal{G}_\infty^Y)}.$$

This induces the definition of a new measure valued process: the unnormalized filter $\tilde{\pi}_t$, such that for all bounded function f on E , $\tilde{\pi}_t(f)$ is the conditional mean of process $L \cdot f(X)$ with respect to \mathcal{G}_∞^Y :

$$(3.19) \quad \pi_t(f) = \tilde{\pi}_t(f) / \tilde{\pi}_t(1).$$

We get classically the filtering equations by projecting, with respect to \mathcal{G}_t^Y , the product $L_t \cdot f(X_t)$ and its expression obtained from Itô formula. Martingale L is \mathcal{G}^Y -locally square integrable and B is \mathcal{G} -adapted; the classical derivation of the filtering equations works locally. Process $\tilde{\pi}$ and π are trivially extended to \mathcal{G} -adapted and \mathcal{G}^Y -locally bounded functionals of X_t and process Y :

$$\pi_t f(\cdot, Y) = E_{Q_\mu}(f(X_t, Y) / \mathcal{G}_t^Y).$$

Hence we get the following proposition.

PROPOSITION 3.2. *Under the assumptions of § 3.1, for all f in $\mathcal{D}(A)$, we have*

$$(3.20) \quad \tilde{\pi}_t(f) = \mu(f) + \int_0^t \tilde{\pi}_s(Af) ds + \int_0^t \tilde{\pi}(f\phi^i) dB_s^i,$$

$$(3.21) \quad \pi_t(f) = \mu(f) + \int_0^t \pi_s(Af) ds + \int_0^t \pi_s(\bar{f}\phi^i)(dB_s^i - \pi_s(\phi^i) ds)$$

where $\phi_t^i = (Adh_t^{-1})_t^i(\psi_t)^i$ and $\bar{f}_t = f(X_t) - \pi_t(f)$.

In Proposition 3.2, the equations are expressed with respect to B . Although B is observable, we prefer to derive them with respect to observation Y . It could have been possible to use (3.20), (3.21) to get new equations. However, a new proof is given, using Stratonovitch calculus.

Let us define the integral of 1-form ω_C (associated to vector field C on M) along the path of Brownian motion Y (see [19]):

$$(3.22) \quad \int_0^t \omega_C \circ dY_s = \int_0^t c_s^i g_{ij}(Y_s) \circ dy_s^j = \int_0^t c_s^i (u_s^{-1})_i^j \circ dB_s^j$$

where c^i and y^i are local coordinates of C and Y ; and the y^i 's satisfy (2.13).

Let us define the following processes: ϕ, κ , such that ϕ is valued in \mathcal{G} and κ is a real process:

$$(3.23) \quad \phi_t = Adh_t^{-1} \cdot \psi_t,$$

$$(3.24) \quad \kappa_t = C_{Ki}^i (Adh_t^{-1})_t^K (\psi_t)^i = C_{Ki}^i \phi_t^K.$$

Then we get filtering equations depending only on the observation.

PROPOSITION 3.3. *Let us consider the model constructed in § 3.1; for all f in $\mathcal{D}(A)$, we have*

$$(3.25) \quad \tilde{\pi}_t(f) = \mu(f) + \int_0^t \tilde{\pi}_s(Af) ds + \int_0^t \omega_{\tilde{\pi}_s(f\psi^*)} \circ dY_s - \frac{1}{2} \int_0^t \tilde{\pi}_s(f(\kappa_s + \|\psi_{Y_s}^*\|_g^2)) ds,$$

$$(3.26) \quad \pi_t(f) = \mu(f) + \int_0^t \pi_s(Af) ds + \int_0^t \omega_{\pi_s(\bar{f}\psi^*)} \circ dY_s - \frac{1}{2} \int_0^t \pi_s(\bar{f}(\kappa_s + \|\psi_{Y_s}^*\|_g^2)) ds$$

where $\bar{f}_t = f(X_t) - \pi_t(f)$, κ is defined above and vector field ψ^* in (2.4).

Proof. The first step of the proof is to express process L with respect to Y ; the second is to compute product $L_t f(X_t)$, thanks to the formula of integration by parts. The next is to prove a projection lemma of Stratonovitch integrals which is applied

to the previous formula. Last, we derive (3.26) from (3.25) with a Stratonovitch calculus of ratio (3.18).

(1) From [18, p. 513] and (2.15) we have that

$$(3.27) \quad \begin{aligned} \phi_t^i &= (Adh_t^{-1})_i^j \psi_t^j = [(h_t \cdot u_0)^{-1} \psi^*(Y_t)]^i \\ &= (u(t)^{-1})_j^i (\psi^*(Y_t))^j. \end{aligned}$$

Hence, from definition (3.22) it follows that

$$(3.28) \quad \sum_i \int_0^t \phi_s^i \circ dB_s^i = \int_0^t \omega_{\psi_s^*} \circ dY_s.$$

Furthermore, by the definition of the Stratonovitch integral, we get

$$(3.29) \quad \int \phi^i \cdot dB^i = \int \phi^i \circ dB^i - \frac{1}{2} \langle \phi^i, B^i \rangle.$$

From (2.24), (3.23) and the independence of X and B , we get

$$(3.30) \quad \frac{d}{dt} \langle \phi^i, B^i \rangle_t = C_{\kappa i}^i (Adh_t^{-1})_i^K \psi_t^K = \kappa_i.$$

From (3.27) and because $u(t)$ belongs to $O(d)$, we have

$$(3.31) \quad \|(\phi_t)_\mathbb{M}\|^2 = \|\psi^*(Y_t)\|^2.$$

Then, with (3.28), (3.31), process L can be written as follows:

$$(3.32) \quad L_t = \exp \left(\int_0^t \omega_{\psi_s^*} \circ dY_s - \frac{1}{2} \int_0^t (\kappa_s + \|\psi^*(Y_s)\|)^2 ds \right).$$

(2) Process L is a solution of the following "Stratonovitch" equation:

$$(3.33) \quad dL_t = \omega_{L, \psi_t^*} \circ dY_t - \frac{1}{2} L_t (\kappa_t + \|\psi^*(Y_t)\|)^2 dt.$$

Let f be in $\mathcal{D}(A)$:

$$(3.34) \quad M_t(f) = f(X_t) - f(X_0) - \int_0^t Af(X_s) ds$$

is a \mathcal{G}^X (hence a \mathcal{G})-semimartingale.

Then we compute product $L_t \cdot f(X_t)$ with the formula of integration by parts:

$$(3.35) \quad \begin{aligned} L_t f(X_t) &= f(X_0) + \int_0^t L_s Af(X_s) ds + \int_0^t L_s \circ dM_s f + \int_0^t \omega_{(L f \psi)_s^*} \circ dY_s \\ &\quad - \frac{1}{2} \int_0^t L_s f(X_s) (\kappa_s + \|\psi^*(Y_s)\|)^2 ds. \end{aligned}$$

(3) Let us notice that the bracket of processes L and $M(f)$ vanishes; then process $\int_0^t L_s \circ dM_s f$ is a \mathcal{G}^X -martingale whose projection, with respect to $(\mathcal{G}^Y, \mathbb{P}^\mu)$, classically vanishes. The other terms, except the integral of the 1-form along the path of Y , can be easily projected.

LEMMA 3.4. *Let ω_C be the 1-form associated to \mathcal{G} -adapted vector field C as defined in (3.22) the local coordinates of which are continuous \mathcal{G} -semimartingales, \mathcal{G}^Y -locally square integrable; then*

$$(3.36) \quad E_{\mathbb{P}_\mu} \left(\int_0^t \omega_{C_s} \circ dY_s / \mathcal{G}_\infty^Y \right) = \int_0^t \omega_{\tilde{C}_s} \circ dY_s$$

where \tilde{C}_s is the vector field on M defined by $E_{\mathbb{P}_\mu}(C_s / \mathcal{G}_\infty^Y)$.

Proof. There exists a sequence $(T_n; n \in N)$ of \mathcal{G}^Y -stopping times such that

$$(3.37) \quad N_t = \int_0^t \omega_{C_s} \circ dY_s = \sum_{n,i} \int_{T_n \wedge t}^{T_{n+1} \wedge t} c_s^{n,i} g_{ij}^n(Y_s) \circ dy_s^{n,j}$$

where T_{n+1} is defined recursively as the last exit time of (U^n, ϕ^n) , neighbourhood of Y_{T_n} in M (cf. [16]), and $c^{n,i}$ and $y^{n,i}$ are the local coordinates of C and Y in (U^n, ϕ^n) . Thus, we shall work only on real Stratonovitch integrals

$$(3.38) \quad N_t = \int_0^t c_s g(Y_s) \circ dy_s.$$

(For the sake of simplicity, we now omit indices n and i , and E denotes $E_{\mathbb{P}_\mu}$ and we note dy instead of $g(Y) \cdot dy$.) This can be written in Itô's form

$$(3.39) \quad N_t = \int_0^t c_s \cdot dy_s + \frac{1}{2} \langle c, y \rangle_t.$$

The Itô integral is projected as follows:

$$(3.40) \quad E \left(\int_0^t c_s \cdot dy_s / \mathcal{G}_\infty^Y \right) = \int_0^t E(c_s / \mathcal{G}_\infty^Y) \cdot dy_s$$

because local coordinates c are \mathcal{G}^Y -locally square integrable. We now have to show that

$$(3.41) \quad E(\langle c, y \rangle_t / \mathcal{G}_\infty^Y) = \langle E(c / \mathcal{G}_\infty^Y), y \rangle_t.$$

Let Z be a \mathcal{G}^Y - (hence a \mathcal{G} -) bounded martingale:

$$(3.42) \quad E(Z_t \cdot \langle c, y \rangle_t) = E \left(\int_0^t Z_s d\langle c, y \rangle_s \right) = E \left(\left\langle c, \int Z \cdot dy \right\rangle_t \right)$$

using [9, Thm. 2.51].

Let M denote the martingale part of c as a \mathcal{G} semimartingale. Then we get

$$(3.43) \quad E \left\langle c, \int Z \cdot dy \right\rangle_t = E \left(M_t \cdot \int_0^t Z_s \cdot dy_s \right) = E \left(E(M_t / \mathcal{G}_\infty^Y) \cdot \int_0^t Z_s dy_s \right).$$

From property (K) [21] the \mathcal{G} -martingale part of $E(c / \mathcal{G}_\infty^Y)$ —which is a \mathcal{G}^Y -martingale—coincides with $E(M_t / \mathcal{G}_\infty^Y)$. Hence, coming back to (3.42), it follows that

$$(3.44) \quad E(Z_t \cdot \langle c, y \rangle_t) = E(Z_t \cdot \langle E(c / \mathcal{G}_\infty^Y), y \rangle_t),$$

and this achieves the proof of Lemma 3.4.

Now, as L is \mathcal{G}^Y -locally square integrable, Lemma 3.4 can be applied to the vector field $C_t = L_t f(X_t) \psi^*$, and we get easily the unnormalised filtering equation (3.25).

(4) We have seen that $\pi_t(f) = \tilde{\pi}_t(f) / \tilde{\pi}_t(1)$. We get formula (3.26) by computing this ratio as a Stratonovitch integral:

$$(3.45) \quad \frac{u_t}{v_t} = \frac{u_0}{v_0} + \int_0^t \frac{1}{v_s} \circ du_s - \int_0^t \frac{u_s}{(v_s)^2} \circ dv_s.$$

Remark. Equations (3.25), (3.26) are actually the Stratonovitch form of (3.20), (3.21) as a consequence of the following equality:

$$\sum_i \frac{d}{dt} \langle \tilde{\pi}(f \phi^i), B^i \rangle_t = \tilde{\pi}_t(f \cdot (\kappa + \|\psi_Y^*\|))$$

proved by means of (2.25) and (3.27).

3.3. Uniqueness of the solution of the filtering equations. In this section, Kurtz' and Ocone's results [11] are applied to show that, under slight hypotheses, filtering equations (3.21)–(3.26) have a unique \mathcal{G}^Y -optional measure valued solution which is the filter. In that sense, these results are not a strict extension of Szpirglas' results [20], but they apply to more general situations, for example, unbounded nonregular signal functions. Our result is a new application of the work of Kurtz and Ocone [11] with an unbounded signal function that depends on the observation process.

Recall that X is an E -valued Markov process with infinitesimal generator $(A, \mathcal{D}(A))$. Let us assume the following hypotheses.

- (H1) E is a locally compact, separable, metric space (for instance, a Riemannian manifold).
- (H2) $\mathcal{D}(A)$ is a sense (in the sup norm) subalgebra in $C_0(E)$ (space of continuous functions that vanish at infinity).
- (H3) Generator A applies $\mathcal{D}(A)$ in $C_0(E)$: $A(\mathcal{D}(A)) \subset C_0(E)$.
- (H4) Uniqueness holds for the martingale problem associated to $(A, \mathcal{D}(A))$, i.e., a unique family of probability measures \mathbb{P}_x^X exists on canonical space $(\Omega^X, \mathcal{G}^X, X_t)$ of continuous functions taking values in E such that coordinate process X satisfies

$$(3.46) \quad \begin{aligned} &X_0 = x \quad \mathbb{P}_x^X \text{ a.e.}, \\ &\forall f \in \mathcal{D}(A), \\ &f(X_t) - f(x) - \int_0^t Af(X_s) ds \text{ is a } (\Omega^X, \mathcal{G}^X, \mathbb{P}_x^X)\text{-martingale.} \end{aligned}$$

- (H5) Function $\psi(x)$ belongs to $\mathcal{D}(A)$.

So, we have the following theorem (which can be transposed to the other filtering equations).

THEOREM 3.5. *Let μ be a cadlag adapted process taking its values in the space of probability measures on E (i.e., for all $C_0(E)$ -function f , $t \rightarrow \mu_t(f)$ is a cadlag \mathcal{G}^Y -adapted process). Let μ be such that, for each function f of $\mathcal{D}(A)$*

$$(3.47) \quad \mu_t(f) = \mu(f) + \int_0^t \mu_s(Af) ds + \int_0^t \omega_{\mu_s}(\bar{f}\psi^*) \circ dY_s - \frac{1}{2} \int_0^t \mu_s(\bar{f}(\kappa_s + \|\psi_s^*\|^2)) ds$$

where the notation is the same as in (3.26). Then, processes μ and π are indistinguishable.

Remark 1. Here we have to extend μ (as for π) to functions on E depending on Y as a parameter; first, to functions as product $f \cdot g$, f over E , g over M , so

$$(3.48) \quad \mu_t(f \cdot g) = \mu_t(f)g(Y_t)$$

or more general functions on $E \times M$:

$$(3.49) \quad \mu_t(f(\cdot, Y_t)) = \int_E f(x, Y_t) \mu_t(dx).$$

Remark 2. The uniqueness here is only among \mathcal{G}^Y -adapted measure processes. In this sense, the result is weaker than that of Szpirglas [20], and it is not a real pathwise uniqueness.

Before giving the proof, let us recall the results of Kurtz and Ocone [11]. First, a filtered martingale problem has to be defined with respect to a given generator $(A', \mathcal{D}(A'))$.

DEFINITION. Let $(A', \mathcal{D}(A'))$ an operator on a product space $E_1 \times E_2$; (μ, Z) is a solution of the *filtered martingale problem* (F.M.P.) associated to $(A', \mathcal{D}(A'))$ if (μ, Z)

is a cadlag process in product space of probability measures on E_1 by E_2 , i.e., for each f in $\mathcal{D}(A')$, $(\mu_t(f), Z_t)$ is a cadlag process in $\mathbb{R} \times E_2$, such that

$$(3.50) \quad \mu_t(f) \text{ is } \mathcal{G}_t^Z\text{-measurable and } \mu_t f(\cdot, Z_t) - \int_0^t \mu_s A f(\cdot, Z_s) ds \text{ is a } \mathcal{G}^Z\text{-martingale.}$$

Then, Kurtz and Ocone prove the following.

THEOREM 3.6. *Let a filtering problem with signal X take its values in E_1 and observation Y in E_2 ; E_i are locally compact separable metric spaces; (X, Y) is a solution of a martingale problem associated to generator $(A', \mathcal{D}(A'))$ verifying (H2) and (H3). Then, there is uniqueness for the filtered martingale problem associated to $(A', \mathcal{D}(A'))$, namely, two solutions (μ, Y) and (μ', Y') have the same law.*

COROLLARY. *If (μ, Y) and (π, Y) are solutions of the F.M.P. $(A', \mathcal{D}(A'))$, for each f in $\mathcal{D}(A')$, we get $\mu_t(f) = \pi_t(f)$ almost everywhere.*

We now sketch the proof of Theorem 3.5. First, the generator $(A', \mathcal{D}(A'))$ of system (X, Y) has to be defined; then it is sufficient to verify the hypotheses of Theorem 3.6: hypotheses (H1) to (H5) imply that $(A', \mathcal{D}(A'))$ verifies (H1) to (H4). Thus we get four steps.

Step 1. Let $\mathcal{D}(A')$ be defined as

$$(3.51) \quad \mathcal{D}(A') = \text{linear span } \{f \cdot g, f \in \mathcal{D}(A), g \in C_k^\infty(O(M))\},$$

i.e., the vector space generated by product of elements of $\mathcal{D}(A)$ and functions on $O(M)$, C^∞ and with compact support.

Let f belong to $\mathcal{D}(A)$ and g to $C_k^\infty(O(M))$; A' is defined by

$$(3.52) \quad A'(f \cdot g)(x, U) = A f(x) g(U) + f(x) (\frac{1}{2} \Delta g(U) + (u^{-1} \psi^*(x, y))^i B_i g(U))$$

where x belongs to E , $U = (y, u)$ belongs to $O(M)$, Δ is the horizontal Laplacian of Bochner (cf. [8]) and B_i are canonical horizontal vector fields on $O(M)$ used to define diffusion U (cf. 2.12).

Expression (3.52) is just a consequence of the Itô formula.

Step 2. As was already remarked, μ is extended to functions of $\mathcal{D}(A')$, so we get a new relation, similar to (3.47) with f in $\mathcal{D}(A')$ and A' instead of A .

Step 3. A point of the proof that is not obvious is the following: if μ satisfies (3.47), can (μ, Y) be a solution of the filtered martingale problem associated to $(A', \mathcal{D}(A'))$? That is to say: a probability measure \mathbb{Q}_μ such that

$$(3.53) \quad \mu_t f(\cdot, U_t) - \int_0^t \mu_s A' f(\cdot, U_s) ds \text{ is a } (\mathbb{Q}_\mu, \mathcal{G}^Y)\text{-martingale}$$

is to be exhibited. Actually, there exists a sequence of probability measures \mathbb{Q}^n , solution of the F.M.P. $(A', \mathcal{D}(A'))$:

$$\mathbb{Q}^n = L^n \cdot \mathbb{P} \quad \text{where } L_t^n = L_{t \wedge Z_n},$$

$$(3.54) \quad dL_t = L_t \mu_t(\phi^i) dB_t^i \quad (\phi_t \text{ is the signal function}),$$

$$Z_n = \inf \left\{ t / \int_0^t |\mu_s(\phi) - \pi_s(\phi)|^2 ds \geq n \text{ or } : \int_0^t |\mu_s(\phi)|^2 ds \geq n \right\}.$$

Notice that Z_n is chosen such that $L_{t \wedge Z_n}$ is an uniformly integrable strictly positive martingale; Z_n can be proved to increase up to infinity. Then, (3.47) is developed with the Stratonovitch integral written with local coordinates of ψ^* and Y and transformed

in an Itô integral with respect to \mathbb{P} -Brownian motion B (cf. (2.12)–(2.14)). After tedious computations, we get

$$(3.55) \quad \mu_t f(\cdot, U_t) - \int_0^t \mu_s A' f(\cdot, U_s) ds = \int_0^t \mu_s (\bar{f} \phi^i) (dB_s^i - \mu_s(\phi^i) ds),$$

so \mathbb{Q}_n is a solution, whatever n .

Step 4. Finally, it remains to prove that hypotheses (H1)–(H5) imply (H1)–(H4) for $(A', \mathcal{D}(A'))$. The construction of $\mathcal{D}(A')$ and (H2) prove that $\mathcal{D}(A')$ is a dense subalgebra of $C_0(E \times O(M))$, by applying the Stone–Weierstrass theorem to the one-point compactification of $E \times O(M)$. Then, $\mathcal{D}(A')$ is the set of finite linear combinations of product $f \cdot g$, f in $\mathcal{D}(A)$, g in $C_k^\infty(O(M))$, so the image by A' of $\mathcal{D}(A')$ is in $C_0(E \times O(M))$; indeed, let us remark that in (3.52) the last term can be written (cf. [18, p. 513]):

$$(3.56) \quad f(x)(u^{-1}\psi^*(x, y))^i B_i g(U) = f(x)\psi^i(x)(Adh^{-1})^i B_i g(U)$$

where h is such that $U = h \cdot U_0$; Adh^{-1} is an unbounded function on G , but $B_i g(U)$ is with compact support, so is the product and (3.56) belongs to $C_0(E \times M)$, and hypothesis (H2) is verified by $(A', \mathcal{D}(A'))$.

Now we are concerned with hypothesis (H4). We adapt the proof of Lemma 4.4 of Kurtz and Ocone [11] when the signal function is dependent on the observation. First, we prove that uniqueness holds for the martingale problem deduced from the filtering model under the reference probability. Indeed, the martingale problem associated to $(\frac{1}{2}\Delta, C_k^\infty(O(M)))$ has a unique solution: the Brownian motion on $O(M)$ [8]. So, hypotheses (H2) and (H4) and Lemma 4.3 of [11] show the following:

(3.57) Uniqueness holds for the martingale problem associated to $(B, \mathcal{D}(A'))$ with B defined by linear extension of

$$B(f \cdot g)(x, U) = Af(x)g(U) + \frac{1}{2}f(x)\Delta g(U), \quad f \cdot g \in \mathcal{D}(A').$$

Besides, let $(\Omega^{\tilde{X}, \tilde{U}}, \mathcal{G}^{\tilde{X}, \tilde{U}}, (\tilde{X}_t, \tilde{U}_t), \tilde{\mathbb{Q}})$ be any solution of the martingale problem associated to $(A', \mathcal{D}(A'))$ and let process $V_t = g_t^{-1} \cdot \tilde{U}_t$ such that

$$(3.58) \quad dg_t^{-1} = g_t^{-1}(-\psi(\tilde{X}_t)) dt, \quad g_0 = e$$

(that is to say $g_t^{-1}(-\psi)$ is a left invariant vector field on \mathcal{G}). We want to prove, for each $f \cdot b$ in $\mathcal{D}(A')$ and s lesser than t ,

$$(3.59) \quad E(f(\tilde{X}_t)b(V_t)/\mathcal{G}_s) = f(\tilde{X}_s)b(V_s) + E\left(\int_s^t B(f \cdot b)(\tilde{X}_u, V_u) du / \mathcal{G}_s\right).$$

So, $(\Omega^{\tilde{X}, \tilde{U}}, \mathcal{G}^{\tilde{X}, \tilde{U}}, (\tilde{X}_t, V_t), \tilde{\mathbb{Q}})$ will be a solution of martingale problem $(B, \mathcal{D}(A'))$ as is the initial system under reference probability $(\Omega, \mathcal{G}, (X_t, U_t), \mathbb{P}_{\mu_0})$. Then, (3.57) proves that \tilde{X} and V are independent and V is a Brownian motion on $O(M)$ as (X, U) are under P_{μ_0} ; furthermore, the laws of X and \tilde{X} are the same: so \tilde{U}_t is the transformation of Brownian motion V_t by g_t , solution of $dg_t = \psi(\tilde{X}_t)g_t dt$ (consequence of (3.58); see [18, 2.16]) and uniqueness holds for the martingale problem associated to $(A', \mathcal{D}(A'))$.

Now, let us prove (3.59). Thanks to derivation rules on a manifold ((3.58), (3.56)), if we define $R(s, t)$ as $b(g_s^{-1}U_t)$, we get

$$(3.60) \quad R(s, t) - R(0, t) = \int_0^s \frac{\partial R}{\partial v}(v, t) dv = - \int_0^s (u_t^{-1})^i \psi^{*j}(X_v, Y_t)(B_i b)(g_v^{-1}U_t) dv.$$

For instance, if s equals t , we get

$$(3.61) \quad f(X_t)b(V_t) = f(X_t)b(U_t) - \int_0^t f(X_t)(u_t^{-1})^i \psi^{*j}(X_v, Y_t)(B_i b)(g_v^{-1}U_t) dv.$$

Expression $f(X_t)(\partial R/\partial v)(v, t)$ —as a function of (X_t, U_t) —is a bounded function of $\mathcal{D}(A')$ thanks to (H5) and (3.51), when v is fixed. Thus, to compute \mathcal{G}_s -conditional mean of (3.55), we cut the integration at $v = s$. After s , we apply \mathcal{G}_v -conditional mean before \mathcal{G}_s -conditional mean:

$$(3.62) \quad \begin{aligned} & E \left(\int_0^t f(X_t) \frac{\partial R}{\partial v}(v, t) dv / \mathcal{G}_s \right) \\ &= \int_0^s f(X_s) \frac{\partial R}{\partial v}(v, s) dv + E \left(\int_s^t f(X_v) \frac{\partial R}{\partial v}(v, v) dv / \mathcal{G}_s \right) \\ &\quad + E \left(\int_0^t \int_{s \vee v}^t A' \left(f \frac{\partial R}{\partial v}(v, \cdot) \right) (X_u, U_u) du dv / \mathcal{G}_s \right). \end{aligned}$$

This last term, by exchanging the order of u and v -integration, becomes

$$(3.63) \quad E \left(\int_s^t \int_0^u A' \left(f \frac{\partial R}{\partial v}(v, \cdot) \right) (X_u, U_u) dv du / \mathcal{G}_s \right).$$

Let u be fixed and consider $f \cdot R$ as a function on $E \times O(M) \times [0, T]$:

$$(3.64) \quad f \cdot R : (x, U, v) \rightarrow f(x)b(g_v^{-1} \cdot U) = (f \cdot b \circ L_{g_v} - 1)(x, U).$$

Since $f \cdot \partial R/\partial v$ belongs to $\mathcal{D}(A')$ and $A'(f \cdot R(v, \cdot))$ is time-differentiable, operators A' and $\partial/\partial v$ are commuting, so we get

$$(3.65) \quad \int_0^u A' \left(f \frac{\partial R}{\partial v}(v, \cdot) \right) (X_u, U_u) dv = (A'(f \cdot R(u, \cdot)) - A'(f \cdot R(0, \cdot)))(X_u, U_u).$$

Definition (3.58) shows that the first term is

$$(3.66) \quad A'(f \cdot R(u, \cdot))(X_u, U_u) = A'(f \cdot b \circ L_{g_u}^{-1})(X_u, U_u).$$

We use the fact that $B_i(b \circ L_g)(U) = (L_g)_*(B_i)(b)(gU) = (B_i b)(g \cdot U)$ because $B_i = (i)_*(A_i)$, $L_g \circ i = i \circ L_g$ and A_i is left invariant; the same is true for Δ which equals $\sum_i B_i(B_i)$; so we get

$$(3.67) \quad A'(f \cdot b \circ L_g)(x, U) = B(f \cdot b)(x, g \cdot U) + (u^{-1}\psi^*(x, y))^i (B_i b)(g \cdot U) f(x).$$

Thus, we can write (3.59) as

$$(3.68) \quad \begin{aligned} & (A'(f \cdot R(v, \cdot)) - A'(f \cdot R(0, \cdot)))(X_v, U_v) \\ &= B(f \cdot b)(X_v, V_v) - B(f \cdot b)(X_v, U_v) \\ &\quad + [(u_v^{-1}\psi^*(X_v, Y_v))(B_i b)(V_v) - (u_v^{-1}\psi^*(X_v, Y_v))^i (B_i b)(U_v)] f(X_v). \end{aligned}$$

Hence, if we apply \mathcal{G}_s -conditional mean to (3.55) and use (3.56), definition (3.54) of $\partial R/\partial v(v, v)$ and (3.57)–(3.62) lead to (3.53). This concludes the proof.

4. Examples. In this part, we apply the previous results to filtering with observations taking values in space \mathbb{R}^d and in sphere S_2 considered as symmetric spaces. The first example which can be directly solved is however a nonstandard filtering problem. Observation Y is the transformation of an \mathbb{R}^d -valued Brownian motion by a space displacement (translation plus rotation). This simple model leads to an unbounded signal function of the observation.

The second example is a more interesting application of the previous theory. Observation Y is then the transformation of a spherical Brownian motion by a rotation. When sphere S_2 is embedded in \mathbb{R}^3 , the dynamics of Y can be expressed by a three-dimensional diffusion. But this diffusion is degenerate, so this example cannot be solved by classical multivariate filtering methods.

4.1. Multivariate case. Observation Y is such that

$$(4.1) \quad Y_t = \Sigma_t(X) W_t + H_t(X)$$

where W_t is a d -dimensional Brownian motion independent of X ; $\Sigma_t(X)$ is a matrix in $SO(d)$; $H_t(X)$ is a vector in \mathbb{R}^d . Thus Y_t is the transformation of W_t by space displacement (Σ, H) , rotation Σ and translation H being functional of X . Assume that

$$(4.2) \quad \dot{\Sigma}_t = \sigma(X_t) \Sigma_t, \quad \dot{H}_t = \sigma(X_t) H_t + h(X_t)$$

where σ is a bounded skew matrix and h a bounded d -dimensional vector. Hypothesis (4.2) implies (3.1) for $M = \mathbb{R}^d$ endowed with a structure of symmetric space as follows: we identify it to the affine hyperplane $(\mathbb{R}^d, 1)$ of \mathbb{R}^{d+1} . The isometric transformation group G of M can be modeled by the set of matrices as follows:

$$(4.3) \quad g = \begin{pmatrix} \Sigma & H \\ 0 & 1 \end{pmatrix}$$

where Σ is a matrix in $SO(d)$ and H is in \mathbb{R}^d . The corresponding Lie algebra \mathfrak{G} of G is the set of matrices ψ such that

$$(4.4) \quad \psi = \begin{pmatrix} \sigma & h \\ 0 & 0 \end{pmatrix}$$

where σ is a (d, d) skew matrix and h is in \mathbb{R}^d . Let ψ_t be defined by

$$(4.5) \quad \psi_t = \psi(X_t) = \begin{pmatrix} \sigma(X_t) & h(X_t) \\ 0 & 0 \end{pmatrix}.$$

So by (3.3) we get

$$(4.6) \quad \phi_t = \phi(X_t, Y_t) = \sigma(X_t) Y_t + h(X_t)$$

and application of Proposition 3.3 leads to the filtering equation for a bounded function f in domain $\mathcal{D}(A)$ of X :

$$\pi_t(f) = \mu(f) + \int_0^t \pi_s(Af) ds + \int_0^t (\pi_s(f\phi) - \pi_s(f)\pi_s(\phi), (dY_s - \pi_s(\phi) ds)).$$

4.2. Spherical case. We now apply the previous results to the sphere S_2 considered as a two-dimensional symmetric space. Sphere S_2 is, embedded in \mathbb{R}^3 , defined as the set of $x = (x_1, x_2, x_3)$ such that $\sum_i x_i^2 = 1$. As a manifold S_2 is endowed with the following open covering associated to spherical coordinates:

$$(4.7) \quad U = \left\{ x \left/ \sum_{i=1}^3 x_i^2 = 1 \right. \right\} - \{x/x_2 = 0; x_1 \leq 0\},$$

$$\psi: U \rightarrow]-\pi, +\pi[\times]-\frac{\pi}{2}, +\frac{\pi}{2}[,$$

$$\psi: x \rightarrow (\theta, \phi) \quad / \quad x = (\cos \phi \cos \theta, \cos \phi \sin \theta, \sin \phi),$$

$$\begin{aligned}
 U' &= \left\{ x / \sum_{i=1}^3 x_i^2 = 1 \right\} - \{x / x_3 = 0; x_1 \geq 0\}, \\
 (4.8) \quad \psi' : U' &\rightarrow]-\pi, +\pi[\times]-\frac{\pi}{2}, +\frac{\pi}{2}[, \\
 \psi' : x &\rightarrow (\theta', \phi') \quad / \quad x = (\cos \phi' \sin \theta', \sin \phi', \cos \phi' \cos \theta').
 \end{aligned}$$

Then, the Jacobian matrix J is

$$(4.9) \quad J = \begin{pmatrix} \frac{\partial \theta'}{\partial \theta} & \frac{\partial \theta'}{\partial \phi} \\ \frac{\partial \phi'}{\partial \theta} & \frac{\partial \phi'}{\partial \phi} \end{pmatrix} = \begin{pmatrix} -\frac{\cos \phi \sin \phi \sin \theta}{1 - \cos^2 \phi \sin^2 \theta} & -\frac{\cos \theta}{1 - \cos^2 \phi \sin^2 \theta} \\ \frac{\cos \phi \cos \theta}{\sqrt{1 - \cos^2 \phi \sin^2 \theta}} & \frac{-\sin \phi \sin \theta}{\sqrt{1 - \cos^2 \phi \sin^2 \theta}} \end{pmatrix}.$$

Let us now define S_2 as a symmetric space.

Fix point $0 = (1, 0, 0)$ in S_2 and orthonormal frame u_0 in $O(S_2)$ such that $\tau(u_0) = 0$ and defined by (e_1, e_2) , with tangent vectors at 0 : $e_1 = (0, 1, 0)$ and $e_2 = (0, 0, 1)$. On the other hand, for any p in S_2 the involutive symmetry s_p associates to any x its symmetric point on the great circle which contains points p and x . For instance, if $p = 0$, symmetry s_0 is done by the matrix

$$(4.10) \quad s_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

Let us consider Lie group $G = \text{SO}(3)$, the isometric transformation group of S_2 ; G naturally acts on S_2 . Let be given g in $\text{SO}(3)$ and $x = (x_1, x_2, x_3)$ in S_2 ; the action $g \cdot x$ is defined by

$$(4.11) \quad (g \cdot x)_i = g_i^j x_j.$$

The isotropy subgroup of G at 0 , K , is the group of rotations around the first axis, that is to say the set of matrices k

$$(4.12) \quad k = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & \sin \theta \\ 0 & -\sin \theta & \cos \theta \end{pmatrix}.$$

The Lie algebra $\mathcal{SO}(3)$ of $\text{SO}(3)$ is the set of $(3, 3)$ skew matrices. The sub-Lie algebra \mathfrak{K} of K is generated by the following matrix A_3 :

$$(4.13) \quad A_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Let A_1 and A_2 be the following matrices of $\mathcal{SO}(3)$:

$$(4.14) \quad A_1 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

It is easy to show that (A_1, A_2) is a basis of \mathfrak{M} , the eigenspace of mapping σ_* associated to eigenvalue -1 (see § 2.1). The structure constants are now simple to compute:

$$(4.15) \quad C_{KI}^I = 0, \quad I = 1, 2, 3 \quad \text{and} \quad C_{13}^2 = -C_{12}^3 = -C_{23}^1 = 1.$$

Let us now consider the filtering model constructed in § 3.1, applied to the spherical case. Let $g_t(X)$ be defined as in (3.1) where $\psi_t = \psi(X_t)$ is a matrix in $\mathcal{SO}(3)$ such that

$$(4.16) \quad \psi(X_t) = \psi^I(X_t) A_I \quad (I = 1, 2, 3)$$

and $\psi^I(X_t)$ are bounded continuous real semimartingales. We define on $(\Omega, \mathcal{G}, \mathbb{P}_\mu)$ a standard two-dimensional Brownian motion B and a Brownian motion Y valued in S_2 , which can be defined by the following system:

$$(4.17) \quad \begin{aligned} dh_t &= h_t A_1 \circ dB_t^1 + h_t A_2 \circ dB_t^2, & h_0 &= e, \\ Y_t &= h_t \cdot 0 \end{aligned}$$

or by the other expressed in $O(S_2)$:

$$(4.18) \quad \begin{aligned} dY_t &= -u^1(t) \circ dB_t^1 - u^2(t) \circ dB_t^2, & Y_0 &= 0, \\ du^1(t) &= Y_t \circ dB_t^1, & u^1(0) &= e_1, \\ du^2(t) &= Y_t \circ dB_t^2, & u^2(0) &= e_2. \end{aligned}$$

To apply Proposition (3.3) we need vector field $(\psi(X_t))^*$. Let us remark [10] that mapping $C \rightarrow C^*$ (2.4) is a Lie algebra homomorphism. Hence it follows that

$$(4.19) \quad (\psi(x))^*(y) = \psi^I(x) A_I^*(y)$$

and we have to compute $A^*(y)$. Returning to its definition,

$$(4.20) \quad A^* f(y) = \frac{d}{dt} f(\exp t A \cdot y)|_{t=0},$$

we easily get

$$\begin{aligned} \exp t A_1 &= \begin{pmatrix} \cos t & -\sin t & 0 \\ \sin t & \cos t & 0 \\ 0 & 0 & 0 \end{pmatrix}, & \exp t A_2 &= \begin{pmatrix} \cos t & 0 & -\sin t \\ 0 & 1 & 0 \\ \sin t & 0 & \cos t \end{pmatrix}, \\ \exp t A_3 &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos t & -\sin t \\ 0 & \sin t & \cos t \end{pmatrix}, \end{aligned}$$

thus we get the local coordinates of A_I^* in (U, ψ)

$$(4.21) \quad \begin{aligned} A_1^*(\theta, \phi) &= \frac{\partial}{\partial \theta}; & A_2^*(\theta, \phi) &= \operatorname{tg} \phi \sin \theta \frac{\partial}{\partial \theta} + \cos \theta \frac{\partial}{\partial \phi}, \\ A_3^*(\theta, \phi) &= -\operatorname{tg} \phi \cos \theta \frac{\partial}{\partial \theta} + \sin \theta \frac{\partial}{\partial \phi}, \end{aligned}$$

and in (U', ψ)

$$(4.22) \quad \begin{aligned} A_1^*(\theta', \phi') &= -\operatorname{tg} \phi' \cos \theta' \frac{\partial}{\partial \theta'} + \sin \theta' \frac{\partial}{\partial \phi'}, & A_2^*(\theta', \phi') &= -\frac{\partial}{\partial \theta'}, \\ A_3^*(\theta', \phi') &= -\operatorname{tg} \phi' \sin \theta' \frac{\partial}{\partial \theta'} - \cos \theta' \frac{\partial}{\partial \phi'}. \end{aligned}$$

Then, it only remains to apply Proposition (3.3) to this spherical filtering model. In this case, for all f in $\mathcal{D}(A)$, we have

$$(4.23) \quad \pi_t(f) = \mu(f) + \int_0^t \pi_s(Af) ds + \int_0^t \omega_{\pi_s(\bar{f}\psi^I)A_I^*} \circ dY_s - \frac{1}{2} \int_0^t \pi_s(\bar{f} \|\psi_{Y_s}^*\|^2) ds$$

because of (4.19) and (4.15). Finally, we can give an explicit formulation with respect to the local coordinates (θ_t, ϕ_t) or (θ'_t, ϕ'_t) whenever process Y_t belongs to U or U' :

$$(4.24) \quad d\pi_t(f) = \pi_t(Af) dt + (\pi_t(\bar{f}\psi^1) + \pi_t(\bar{f}\psi^2) \operatorname{tg} \phi_t \sin \theta_t - \pi_t(\bar{f}\psi^3) \operatorname{tg} \phi_t \cos \theta_t) \circ d\theta_t \\ + (\pi_t(\bar{f}\psi^2) \cos \theta_t + \pi_t(\bar{f}\psi^3) \sin \theta_t) \circ d\phi_t - \frac{1}{2} \pi_t(\bar{f}l_t) dt$$

where

$$l_t(x) = \|\psi_x^*(Y_t)\|^2 = (\psi^1(x) + (\psi^2(x) \sin \theta_t - \psi^3(x) \cos \theta_t) \operatorname{tg} \phi_t)^2 \\ + (\psi^2(x) \operatorname{csc} \theta_t + \psi^3(x) \sin \theta_t)^2$$

if Y_t belongs to U and

$$(4.25) \quad d\pi_t(f) = \pi_t(Af) dt + (-\pi_t(\bar{f}\psi^1) \operatorname{tg} \phi'_t \cos \theta'_t - \pi_t(\bar{f}\psi^2) - \pi_t(\bar{f}\psi^3) \operatorname{tg} \phi'_t \sin \theta'_t) \circ d\theta'_t \\ + (\pi_t(\bar{f}\psi^1) \sin \theta'_t - \pi_t(\bar{f}\psi^2) \cos \theta'_t) \circ d\phi'_t - \frac{1}{2} \pi_t(\bar{f}l_t) dt$$

where

$$l_t(x) = (\psi^2(x) + (\psi^1(x) \cos \theta'_t + \psi^2(x) \sin \theta'_t) \operatorname{tg} \phi'_t)^2 + (\psi^1(x) \sin \theta'_t - \psi^2(x) \cos \theta'_t)^2$$

if Y_t belongs to U' .

5. Conclusion. We have studied a nonstandard filtering model. Observation process Y takes values in a symmetric space M . This particular assumption allows us to consider process Y under a multiplicative form, and then generalizes the previous cases of [15] and [16]. More precisely, process Y depends on signal X by means of a stochastic isometric transformation $Y_t = g_t(X) \cdot W_t$, where W_t is a Brownian motion taking its values in M . We get intrinsical filtering equations under both Stratonovitch and Itô forms which are unique characterizations of conditional distribution of X and Y .

Acknowledgment. We thank the reviewer for suggesting that we study uniqueness of the solution of filtering equations.

REFERENCES

- [1] J. M. BISMUT, *Mécanique aléatoire*, Lecture Notes in Math., 866, Springer-Verlag, Berlin, New York, 1981.
- [2] P. BREMAUD AND M. YOR, *Changes of filtration and of probability measures*, Z. Wahrsch. Verw. Gebiete, 45 (1978), pp. 269–295.
- [3] W. M. BOOTHBY, *An Introduction to Differential Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.
- [4] R. W. R. DARLING, *On the convergence of Gangolli processes to Brownian motion on a manifold*, Stochastics, 12 (1984), pp. 277–301.
- [5] T. E. DUNCAN, *Stochastic filtering in manifolds*, Proc. IFAC World Congress, Pergamon Press, New York, Oxford, 1981.
- [6] F. R. GANTMACHER, *Théorie des matrices*, Tome 2, Dunod, Paris, 1966.
- [7] S. HELGASON, *Differential Geometry and Symmetric Spaces*, Academic Press, New York, 1962.
- [8] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, 1981.
- [9] J. JACOD, *Calcul stochastique et problème de martingales*, Lecture Notes in Math., 715, Springer-Verlag, Berlin, New York, Heidelberg, 1979.
- [10a] S. KOBAYASHI AND K. NOMIZU, *Foundations of Differential Geometry*, Interscience, New York, 1963.
- [10b] ———, *Foundations of Differential Geometry*, II, Interscience, New York, 1969.
- [11] T. G. KURTZ AND D. L. OCONE, *Unique characterization of conditional distributions in nonlinear filtering*, preprint, 1986.

- [12] D. LEPINGLE AND J. MEMIN, *Sur l'intégrabilité uniforme des martingales exponentielles*, Z. Wahrsch. Verw. Gebiete, 42 (1978), pp. 175–203.
- [13] R. C. LIPTSER AND A. SHYRAYEV, *Statistics of Random Processes*, I, II, Springer-Verlag, Berlin, New York, Heidelberg, 1974.
- [14] P. A. MEYER, *Géométrie différentielle stochastique*, Séminaire de Probabilités XVI, Lecture Notes in Math., 921, Springer-Verlag, Berlin, New York, Heidelberg, 1982, pp. 165–207.
- [15] S. K. NG AND P. E. CAINES, *Nonlinear filtering in Riemannian manifolds*, IMA J. Math. Control Inform. (1985), pp. 25–36.
- [16] M. PONTIER AND J. SZPIRGLAS, *Filtrage non linéaire avec observation sur une variété*, Stochastics, 15 (1985), pp. 121–148.
- [17] L. SCHWARTZ, *Construction directe d'une diffusion sur une variété*, Note interne du Centre de Mathématiques de l'Ecole Polytechnique, 1984.
- [18] I. SHIGEKAWA, *Transformations of the Brownian motion on a Riemannian symmetric space*, Z. Wahrsch. Verw. Gebiete, 65 (1984), pp. 493–522.
- [19] ———, *On stochastic horizontal lifts*, Z. Wahrsch. Verw. Gebiete, 59 (1982), pp. 211–222.
- [20] J. SZPIRGLAS, *Sur l'équivalence d'équations différentielles stochastiques*, Ann. Inst. H. Poincaré, 14 (1978), pp. 33–59.
- [21] J. SZPIRGLAS AND G. MAZZIOTTO, *Modèle général de filtrage non linéaire et équations différentielles stochastiques associées*, Ann. Inst. H. Poincaré, 15 (1979), pp. 147–173.
- [22] M. ZAKAI, *On the optimal filtering of diffusion processes*, Z. Wahrsch. Verw. Gebiete, 11 (1969), pp. 230–249.

SENSITIVITY ANALYSIS OF NONLINEAR PROGRAMS AND DIFFERENTIABILITY PROPERTIES OF METRIC PROJECTIONS*

ALEXANDER SHAPIRO†

Abstract. This paper is concerned with a study of differentiability properties of the optimal value function and an associated optimal solution of a parametrized nonlinear program. Second order analysis is presented essentially under the Mangasarian–Fromovitz constraint qualification when the corresponding vector of Lagrange multipliers is not necessarily unique. It is shown that under certain regularity conditions the optimal value function possesses second order directional derivatives and the optimal solution mapping is directionally differentiable. The results obtained are applied to an investigation of metric projections in finite-dimensional spaces.

Key words. nonlinear programming, sensitivity analysis, optimal value, parametric programming, metric projections, second order directional derivatives

AMS(MOS) subject classification. 90C31

1. Introduction. In this paper we consider the mathematical programming problem

$$(\mathcal{P}_y) \quad \underset{x}{\text{minimize}} \ f(x, y) \quad \text{subject to } x \in \Omega(y),$$

where the objective function $f(x, y)$ as well as the set $\Omega(y)$ of feasible solutions depends on the parameter vector $y \in \mathbb{R}^n$. It will be assumed that the point-to-set mapping (multifunction) $\Omega: \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is defined by equality and inequality constraints

$$(1.1) \quad \Omega(y) = \{x: g_i(x, y) = 0, i = 1, \dots, q; g_i(x, y) \leq 0, i = q+1, \dots, p\}$$

and that all involved functions f and $g_i, i = 1, \dots, p$, are at least C^1 smooth on $\mathbb{R}^m \times \mathbb{R}^n$. The optimal value function

$$\varphi(y) = \inf \{f(x, y): x \in \Omega(y)\}$$

is associated with program (\mathcal{P}_y) and the corresponding set $M(y)$ of optimal solutions. By convention $\varphi(y)$ is defined to be $+\infty$ if the set $\Omega(y)$ is empty.

The aim of this paper is to investigate first- and second-order differential properties of $\varphi(y)$ and differential properties of an optimal solution $\bar{x}(y) \in M(y)$ in vicinity of a given point y_0 . In the case $\Omega(y)$ is defined by (1.1), this is the standard problem of sensitivity analysis in nonlinear programming which has been discussed extensively [4], [6], [7], [10], [11], [16], [25], [28], [30], [33], [34], [37]. The approach we adopt here is to try to approximate the nonlinear program (\mathcal{P}_y) by a simpler one rather than to “linearize” equations representing optimality conditions of (\mathcal{P}_y) (cf. [4], [7], [16], [27], [28]). Thus the present investigation can be considered as an extension and generalization of an approach suggested in [34]. The regularity conditions ensuring uniqueness of Lagrange multipliers [18], which have been employed in [34], will be replaced by the Mangasarian–Fromovitz constraint qualification (MF-condition) [20]. Since the MF-condition is necessary and sufficient for the set of Lagrange multipliers to be bounded [9], this appears to be a natural assumption we should use in order to investigate (\mathcal{P}_y) by means of the differential information at the point $(x_0, y_0), x_0 \in M(y_0)$.

* Received by the editors April 14, 1986; accepted for publication (in revised form) August 10, 1987.

† Department of Mathematics and Applied Mathematics, University of South Africa, Pretoria, 0001 South Africa.

When the objective function $f(x, y)$ is taken to be $f(x, y) = \|y - x\|$, where $x, y \in \mathbb{R}^n$ and $\|\cdot\|$ is a chosen norm, the optimal set valued multifunction $M(y)$ becomes the set-valued metric projection onto the (moving) set $\Omega(y)$. We consider a corresponding selection mapping $P_\Omega: \mathbb{R}^n \rightarrow \mathbb{R}^n$, $P_\Omega(y) \in M(y)$. Thus $P_\Omega(y)$ is a closest point of the set $\Omega(y)$ to the point y . In the case of the constant (independent of y) and convex set $\Omega(y) = \Omega_0$ differentiability properties of the associated metric projection P_Ω have been studied in a number of publications (see [1], [8], [12], [14], [23], [39] and references therein). Recently Malanowski [19] applied some results from sensitivity analysis of nonlinear programs to an investigation of differentiability of metric projections. On the basis of our study of program (\mathcal{P}_y) we will be able to strengthen some results due to Haraux [12], Holmes [14], Malanowski [19] and others. We show that under certain regularity conditions P_Ω is directionally differentiable at $y_0 \in \mathbb{R}^n \setminus \Omega(y_0)$, although the corresponding directional derivative $P'_\Omega(y_0; v)$ is not necessarily continuous in v . It should be mentioned that, in general, directional differentiability of P_Ω is not guaranteed even in the case of constant and convex set $\Omega(y) = \Omega_0$ and the Euclidean norm (see [17], [24, p. 696]).

Throughout the paper we assume that the feasible set $\Omega(y_0)$ corresponding to the point y_0 is nonempty. For a multifunction Ω we write $\text{gph } \Omega$ for its graph,

$$\text{gph } \Omega = \{(y, x) : x \in \Omega(y)\}.$$

The multifunction Ω is called closed (convex) if $\text{gph } \Omega$ is a closed (convex) subset of $\mathbb{R}^n \times \mathbb{R}^m$. A multifunction $\Sigma: \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is said to be positively homogeneous if $\Sigma(ty) = t\Sigma(y)$ for all $t > 0$ and $y \in \mathbb{R}^n$. Clearly Σ is positively homogeneous if and only if its graph $\text{gph } \Sigma$ is a cone. A (closed, convex) positively homogeneous multifunction is called a (closed, convex) *process* [29, § 39].

We make use of the concept of pseudo-Lipschitzian multifunctions [2, p. 98], [31]. A multifunction $\Omega(y)$ is said to be pseudo-Lipschitzian at a point (x_0, y_0) , $x_0 \in \Omega(y_0)$, if the distance function $d_\Omega(x, y) = \text{dist}(x; \Omega(y))$ is Lipschitzian in a neighborhood of (x_0, y_0) . For an equivalent definition and thorough discussion of pseudo-Lipschitzian multifunctions see Rockafellar [31]. Here and subsequently we denote by $\text{dist}(x; S)$ the distance from a point x to a set S ,

$$\text{dist}(x; S) = \inf \{\|x - z\| : z \in S\}.$$

Finally we recall that a function (mapping) $f(y)$ is said to be directionally differentiable at y_0 if the directional derivative

$$f'(y_0; v) = \lim_{t \rightarrow 0^+} \frac{f(y_0 + tv) - f(y_0)}{t}$$

exists for all v .

2. The reduction theorem. In this section we show that under certain regularity conditions the basic problem (\mathcal{P}_y) can be replaced by an equivalent one which is more convenient for sensitivity analysis. The following regularity conditions will be required.

Assumption 1. There exist a number α and a compact set $S \subset \mathbb{R}^m$ such that $\alpha > \varphi(y_0)$ and

$$\{x \in \Omega(y) : f(x, y) \leq \alpha\} \subset S$$

for all y in a neighborhood of y_0 .

This condition is required in order to ensure some continuity properties of the optimal value function and optimal solutions. It has been employed in [30] and [32]

and goes back to Wets [38]. Assumption 1 is similar to the condition of uniform compactness of $\Omega(y)$ utilized in [10] and [11].

Assumption 2. The optimal set $M(y_0) = \{x_0\}$ is a singleton.

Notice that the uniqueness Assumption 2 does not necessarily imply that $M(y)$ is single-valued in a neighborhood of y_0 .

In order to simplify notation we assume subsequently that all inequality constraints are active at the point (x_0, y_0) , i.e., $g_i(x_0, y_0) = 0$ for $i = q+1, \dots, p$.

Assumption 3 (MF-condition). (i) The gradient vectors $\nabla_x g_i(x_0, y_0)$, $i = 1, \dots, q$, are linearly independent.

(ii) There exists a vector u such that

$$u^T \nabla_x g_i(x_0, y_0) = 0, \quad i = 1, \dots, q,$$

$$u^T \nabla_x g_i(x_0, y_0) < 0, \quad i = q+1, \dots, p.$$

Assumption 3 is the standard Mangasarian-Fromovitz constraint qualification [20] applied at the point (x_0, y_0) . It follows from Assumption 3 that the Mangasarian-Fromovitz constraint qualification is satisfied at all $(\bar{x}(y), y)$ sufficiently close to (x_0, y_0) and thus the first order (Kuhn-Tucker) necessary conditions hold: If $\bar{x} = \bar{x}(y)$ is a solution of (\mathcal{P}_y) , then there exists a vector $\bar{\lambda} = \bar{\lambda}(y) \in \mathbb{R}^p$ of Lagrange multipliers such that

$$(2.1) \quad \begin{aligned} \nabla_x L(\bar{x}, y, \bar{\lambda}) &= 0, \\ \bar{\lambda}_i g_i(\bar{x}, y) &= 0 \quad \text{and} \quad \bar{\lambda}_i \geq 0, \quad i = q+1, \dots, p. \end{aligned}$$

Here $L(x, y, \lambda)$ is the Lagrangian function associated with program (\mathcal{P}_y) ,

$$L(x, y, \lambda) = f(x, y) + \sum_{i=1}^p \lambda_i g_i(x, y).$$

We denote by $\Lambda(\bar{x}, y)$ the set of vectors $\bar{\lambda}$ satisfying the first order necessary conditions (2.1). It follows from the definition that $\Lambda(\bar{x}, y)$ is a closed convex polytope. Moreover, it is implied by the MF-condition that the set $\Lambda_0 = \Lambda(x_0, y_0)$ is nonempty and *bounded* (Gauvin [9]). Consequently Λ_0 is the convex hull of the finite set E_0 of its extreme points. For a vector $\lambda \in \mathbb{R}^p$ denote $J_+(\lambda) = \{i: \lambda_i > 0, i = q+1, \dots, p\}$, $J_0(\lambda) = \{i: \lambda_i = 0, i = q+1, \dots, p\}$ and $J(\lambda) = \{1, \dots, q\} \cup J_+(\lambda)$. It is not difficult to see that a point $\lambda \in \Lambda_0$ is an extreme point if and only if the gradient vectors $\nabla_x g_i(x_0, y_0)$, $i \in J(\lambda)$, are linearly independent.

We are prepared now to formulate a reduction theorem in which program (\mathcal{P}_y) is replaced by a more convenient one. Consider the function

$$(2.2) \quad F(x, y) = \max \{L(x, y, \lambda): \lambda \in \Lambda_0\}.$$

Since $L(x, y, \lambda)$ is linear in λ , the maximum in the right-hand side of (2.2) is attained at extreme points of Λ_0 . Consequently $F(x, y)$ is representable as the pointwise maximum of a finite family of smooth functions $F_\lambda(x, y) = L(x, y, \lambda)$,

$$(2.3) \quad F(x, y) = \max \{L(x, y, \lambda): \lambda \in E_0\}.$$

We also consider the reduced feasible set

$$(2.4) \quad \bar{\Omega}(y) = \bigcup \{\Omega_\lambda(y): \lambda \in E_0\},$$

where

$$\begin{aligned} \Omega_\lambda(y) &= \{x: g_i(x, y) = 0, i \in J(\lambda); g_i(x, y) \leq 0, i \in J_0(\lambda)\} \\ &= \Omega(y) \cap \{x: g_i(x, y) = 0, i \in J_+(\lambda)\}. \end{aligned}$$

THEOREM 2.1. *Suppose that Assumptions 1–3 hold. Then for all y in a neighborhood of y_0 ,*

$$(2.5) \quad \varphi(y) = \inf \{F(x, y) : x \in \bar{\Omega}(y)\}$$

and the optimal set $M(y)$ coincides with the set of minimizers of $F(\cdot, y)$ over $\bar{\Omega}(y)$.

We shall make use of the following continuity property of the optimal set $M(y)$, which is a variation of well-known results in stability theory of parametrized nonlinear programs (e.g. [7, § 2.2]).

LEMMA 2.1. *Suppose that Assumptions 1–3 hold and let $\bar{x}(y)$ be an optimal solution of the program (\mathcal{P}_y) . Then $\bar{x}(y)$ converges to x_0 as y tends to y_0 .*

Proof. First we observe that the multifunction Ω is closed and because of the MF-condition the set $\Omega(y)$ is nonempty for all y in a neighborhood of y_0 such that there is a point $u(y)$, $u(y) \in \Omega(y)$, converging to x_0 as $y \rightarrow y_0$. Together with the inf-compactness Assumption 1 this implies that the minimizer $\bar{x}(y)$ exists for all y near y_0 .

Suppose that $\bar{x}(y)$ is not continuous at y_0 . Then there exist a sequence $y_n \rightarrow y_0$ and a constant $\delta > 0$ such that $\|x_n - x_0\| > \delta$, where $x_n = \bar{x}(y_n)$. It follows from Assumption 1 that there is a compact set S such that $\bar{x}(y) \in S$ for all y in a neighborhood of y_0 . Therefore the sequence $\{x_n\}$ can be chosen in such a way that $\{x_n\}$ converges to a point $x^* \in S$. Since Ω is closed, $x^* \in \Omega_0 = \Omega(y_0)$. The function $f(x, y)$ is continuous and hence there exists $\varepsilon > 0$ such that

$$(2.6) \quad f(x, y_0) > f(x_0, y_0) + \varepsilon$$

for all $x \in \{v : \|v - x_0\| \geq \delta, v \in S \cap \Omega_0\}$. Now consider a sequence $u_n \in \Omega(y_n)$ converging to x_0 . It follows from the continuity of $f(x, y)$ that

$$(2.7) \quad |f(x_0, y_0) - f(u_n, y_n)| < \varepsilon/2$$

and

$$(2.8) \quad |f(x^*, y_0) - f(x_n, y_n)| < \varepsilon/2$$

for sufficiently large n . Moreover, $f(x_n, y_n) \leq f(u_n, y_n)$ and hence by (2.7) and (2.8)

$$f(x^*, y_0) < f(x_0, y_0) + \varepsilon.$$

The last inequality contradicts (2.6) and hence the proof is complete. \square

Proof of Theorem 2.1. Consider the following optimal value function:

$$\bar{\varphi}(y) = \inf \{f(x, y) : x \in \bar{\Omega}(y)\}.$$

First we show that $\varphi(y) = \bar{\varphi}(y)$ and the set $M(y)$ coincides with the set of minimizers of $f(\cdot, y)$ over $\bar{\Omega}(y)$ for all y in a neighborhood of y_0 . Clearly for every y , $\bar{\Omega}(y) \subset \Omega(y)$ and hence $\bar{\varphi}(y) \geq \varphi(y)$. By continuity of $\bar{x}(y)$ the Mangasarian–Fromovitz condition holds at $\bar{x} = \bar{x}(y)$ and hence $\Lambda(\bar{x}, y)$ is nonempty for all y near y_0 . Furthermore, a vector $\bar{\lambda}(y) \in \Lambda(\bar{x}, y)$ of Lagrange multipliers converges to Λ_0 , that is there exists $\lambda^*(y) \in \Lambda_0$ such that $\bar{\lambda}(y) - \lambda^*(y) \rightarrow 0$ as $y \rightarrow y_0$. Vector $\lambda^* = \lambda^*(y)$ is representable as a convex combination of some points from E_0 . Let $\mu = \mu(y) \in E_0$ be an extreme point which takes part in this convex combination with the largest coefficient. Since coordinates μ_i , $i = q+1, \dots, p$, of μ are nonnegative we then have that

$$\lambda_i^* \geq (1/N)\mu_i, \quad i = q+1, \dots, p,$$

where N is the number of extreme points in E_0 . Because the number of extreme points is finite it follows then by continuity that $J_+(\mu) \subset J_+(\bar{\lambda})$ for y sufficiently close to y_0 .

This implies that the inequality constraints g_i , $i \in J_+(\mu)$, are active at $(\bar{x}(y), y)$ and hence $\bar{x}(y) \in \Omega_\mu(y)$. Therefore $\varphi(y) \geq \bar{\varphi}(y)$ and $M(y) \subset \bar{\Omega}(y)$ for all y near y_0 . This completes the proof of the assertion stated above.

It remains to show that for every $x \in \bar{\Omega}(y)$, $f(x, y) = F(x, y)$. Since $\bar{\Omega}(y) \subset \Omega(y)$ we have that $g_i(x, y) = 0$, $i = 1, \dots, q$, and $g_i(x, y) \leq 0$, $i = q+1, \dots, p$, for all $x \in \bar{\Omega}(y)$. Moreover, $\lambda_i \geq 0$, $i = q+1, \dots, p$, and hence $L(x, y, \lambda) \leq f(x, y)$ for all $\lambda \in E_0$ and $x \in \bar{\Omega}(y)$. Consequently $F(x, y) \leq f(x, y)$ for all $x \in \bar{\Omega}(y)$. Now if $x \in \bar{\Omega}(y)$, then $x \in \Omega_\mu(y)$ for some $\mu \in E_0$. By the definition of $\Omega_\mu(y)$ we have that $L(x, y, \mu) = f(x, y)$ and hence $F(x, y) = f(x, y)$. \square

It follows from Theorem 2.1 that $M(y) \subset \bar{\Omega}(y)$ and hence the multifunction $\bar{\Omega}$ is nonempty valued for all y in a neighborhood of y_0 . In the remainder of this section we study local behavior of the reduced feasible set $\bar{\Omega}(y)$ near the point (x_0, y_0) . In particular it will be shown that $\bar{\Omega}$ is tangentially differentiable at (x_0, y_0) in the sense of the following definition.

DEFINITION 2.1. We say that a multifunction $\Phi: \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is tangentially differentiable at (x_0, y_0) , $x_0 \in \Phi(y_0)$, if there exists a closed process $\Sigma: \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ such that

$$(2.9) \quad \lim_{\substack{(x,y) \rightarrow (x_0,y_0) \\ (y,x) \in \text{gph } \Phi}} \frac{\text{dist}(x - x_0; \Sigma(y - y_0))}{\|(y - y_0, x - x_0)\|} = 0$$

and

$$(2.10) \quad \lim_{\substack{(x,y) \rightarrow (x_0,y_0) \\ (y-y_0, x-x_0) \in \text{gph } \Sigma}} \frac{\text{dist}(x; \Phi(y))}{\|(y - y_0, x - x_0)\|} = 0.$$

The process Σ is said to be a tangential approximation to Φ at (x_0, y_0) and is denoted $\Sigma = D\Phi(x_0, y_0)$.

The concept of tangential differentiability has been introduced and discussed in [36]. It was shown that whenever $D\Phi$ exists it is unique. Fixing $y = y_0$ in (2.9) and (2.10) we obtain that the set $\Phi_0 = \Phi(y_0)$ is approximated at x_0 by the cone $\Sigma_0 = \Sigma(0)$ [35]. Also $x_0 + \Sigma_0$ and Φ_0 are tangent at x_0 in the sense of Robinson [26, Def. 3]. For a detailed discussion and derivation of basic properties of cone and tangential approximations and relevant references the reader is referred to [26, p. 504], [22, § 4], [35] and [36], respectively.

Under the MF-condition the multifunction Ω is pseudo-Lipschitzian [31, Thm. 2.3] and tangentially differentiable at (x_0, y_0) [36]. Moreover, the tangential approximation $\Sigma = D\Omega(x_0, y_0)$ is given by the linearization

$$(2.11) \quad \Sigma(v) = \{u: \alpha_i(u, v) = 0, i = 1, \dots, q; \alpha_i(u, v) \leq 0, i = q+1, \dots, p\},$$

where

$$\alpha_i(u, v) = u^T \nabla_x g_i(x_0, y_0) + v^T \nabla_y g_i(x_0, y_0).$$

For the reduced multifunction $\bar{\Omega}$ a natural candidate for the tangential approximation is provided by the corresponding linearization

$$(2.12) \quad \bar{\Sigma}(v) = \bigcup \{\Sigma_\lambda(v): \lambda \in E_0\}$$

with

$$\Sigma_\lambda(v) = \{u: \alpha_i(u, v) = 0, i \in J(\lambda); \alpha_i(u, v) \leq 0, i \in J_0(\lambda)\}.$$

The following characterization of $\bar{\Sigma}(v)$ will be useful.

LEMMA 2.2. Suppose that the MF-condition holds. Then for all v the set $\bar{\Sigma}(v)$ is given by the set of optimal solutions of the linear program

$$(2.13) \quad \underset{u}{\text{minimize}} \quad u^T \nabla_x f(x_0, y_0) \quad \text{subject to } u \in \Sigma(v).$$

Proof. First we observe that because of the MF-condition the set E_0 is nonempty. Consider a point v and let $\bar{u} \in \Sigma_\lambda(v)$ for some $\lambda \in E_0$. Because of the first order necessary conditions (2.1), applied at the point (x_0, y_0) ,

$$\nabla_x f(x_0, y_0) + \sum_{i \in J(\lambda)} \lambda_i \nabla_x g_i(x_0, y_0) = 0$$

and $\lambda_i \geq 0$, $i = q+1, \dots, p$. By the definition of $\Sigma_\lambda(v)$, $\alpha_i(\bar{u}, v) = 0$ for all $i \in J(\lambda)$. It then follows from the optimality conditions for linear programs that \bar{u} is an optimal solution of the program (2.13). Conversely, similar to the proof of Theorem 2.1 it can be shown that if \bar{u} is an optimal solution of (2.13), then $\bar{u} \in \Sigma_\lambda(v)$ for some $\lambda \in E_0$. \square

Lemma 2.2 implies that for all v the set $\bar{\Sigma}(v)$ is a closed convex polytope. Moreover, perturbations of the linear program (2.13) are determined by the right-hand side changes corresponding to $v^T \nabla_y g_i(x_0, y_0)$, $i = 1, \dots, p$. It follows then that the multifunction $\bar{\Sigma}(v)$ is Lipschitzian (Mangasarian and Shiau [21]) and hence is pseudo-Lipschitzian at $(0, 0)$.

LEMMA 2.3. Suppose that the MF-condition holds. Then the multifunction $\bar{\Omega}$ is tangentially differentiable at (x_0, y_0) and $D\bar{\Omega}(x_0, y_0) = \bar{\Sigma}$.

Proof. Without loss of generality we can assume that $(x_0, y_0) = (0, 0)$. First we show that condition (2.9), applied to the multifunctions $\bar{\Omega}$ and $\bar{\Sigma}$, is satisfied. Suppose that (2.9) does not hold. Then there exists a sequence $(x_n, y_n) \rightarrow (0, 0)$, $x_n \in \bar{\Omega}(y_n)$, and $\varepsilon > 0$ such that

$$\text{dist}(x_n; \bar{\Sigma}(y_n)) > \varepsilon t_n,$$

where $t_n = \|(x_n, y_n)\|$. Since the set E_0 is finite we can assume that there is $\lambda \in E_0$ such that $x_n \in \Omega_\lambda(y_n)$ for all n . By the standard argument of compactness we can assume that the sequence $\{t_n^{-1}(x_n, y_n)\}$ converges to a vector (u, v) . Since $\bar{\Sigma}$ is pseudo-Lipschitzian the distance function $d_{\bar{\Sigma}}(x, y) = \text{dist}(x; \bar{\Sigma}(y))$ is Lipschitzian and hence continuous. It follows that

$$\text{dist}(t_n^{-1}x_n; \bar{\Sigma}(t_n^{-1}y_n)) > \varepsilon$$

and then by the continuity of the distance function $\text{dist}(u; \bar{\Sigma}(v)) > 0$. Therefore u does not belong to the set $\bar{\Sigma}(v)$. On the other hand

$$g_i(x_n, y_n) = \alpha_i(x_n, y_n) + o(t_n)$$

and hence $\alpha_i(u, v) = 0$ for $i \in J(\lambda)$ and $\alpha_i(u, v) \leq 0$ for $i \in J_0(\lambda)$. Consequently $u \in \Sigma_\lambda(v)$, a contradiction.

Now consider condition (2.10). Because of the MF-condition (i), for a given y sufficiently close to $y_0 = 0$ the equality constraints define a smooth manifold

$$\theta(y) = \{x: g_i(x, y) = 0, i = 1, \dots, q\}$$

near zero. It follows from the MF-condition (ii) that $\Omega(y)$ has a nonempty interior relative to this manifold in a neighborhood of $x_0 = 0$. The multifunction Ω is tangentially differentiable and hence for every $(y, x) \in \text{gph } \bar{\Sigma}$ in a neighborhood of $(0, 0)$, there exists a vector $(y, u) \in \text{gph } \Omega$ such that $\|x - u\| = o(\|(x, y)\|)$. Consider the index set

$$I(u, y) = \{i: g_i(u, y) = 0, i = q+1, \dots, p\}$$

and the normal cone to $\Omega(y)$ at u , which is polar to the contingent cone of $\Omega(y)$ at u . This normal cone is generated by the gradient vectors $\nabla_x g_i(u, y)$, $i \in I(u, y)$, with nonnegative coefficients and $\nabla_x g_i(u, y)$, $i = 1, \dots, q$, with unrestricted coefficients. By the MF-condition the point u can be chosen on the boundary of $\Omega(y)$ relative to the manifold $\theta(y)$ such that the contingent cone of $\Omega(y)$ at u tends to the contingent cone of $\Sigma(y)$ at x as y tends to $y_0 = 0$. That is, the distance from $-\nabla_x f(x_0, y_0)$ to the normal cone (to $\Omega(y)$ at u) tends to zero, i.e., there are multipliers $\lambda_i^* = \lambda_i^*(u, y)$ such that

$$\nabla_x f(x_0, y_0) + \sum_{i=1}^q \lambda_i^* \nabla_x g_i(u, y) + \sum_{i \in I(u, y)} \lambda_i^* \nabla_x g_i(u, y)$$

tends to zero as $(u, y) \rightarrow (0, 0)$. By the first order necessary conditions this implies that there is $\mu = \mu(u, y) \in E_0$ such that $J_+(\mu) \subset J_+(\lambda^*)$ for (u, y) sufficiently close to $(0, 0)$. Then $J_+(\mu) \subset I(u, y)$ and hence $u \in \Omega_\mu(y)$. Consequently $u \in \bar{\Omega}(y)$ and this completes the proof of (2.10). \square

Now let us consider the polyhedral convex cone

$$(2.14) \quad C = \{u: u^T \nabla_x g_i(x_0, y_0) = 0, i = 1, \dots, q; \\ u^T \nabla_x g_i(x_0, y_0) \leq 0, i = q+1, \dots, p; u^T \nabla_x f(x_0, y_0) \leq 0\}.$$

The cone C is called critical and is useful in deriving second order optimality conditions for program (\mathcal{P}_{y_0}) . The result of the following lemma clarifies a relation between C and the multifunction $\bar{\Sigma}$ (see [30, (1.13), (1.14)]).

LEMMA 2.4. *Suppose that the MF-condition holds. Then $C = \Sigma_\lambda(0)$ for all $\lambda \in E_0$.*

Proof. Because of the MF-condition the set E_0 is nonempty. Consider $u \in C$ and $\lambda \in E_0$. Then

$$0 \geq u^T \nabla_x f(x_0, y_0) = -u^T \left(\sum_{i \in J(\lambda)} \lambda_i \nabla_x g_i(x_0, y_0) \right) = - \sum_{i \in J_+(\lambda)} \lambda_i u^T \nabla_x g_i(x_0, y_0).$$

Since λ_i is positive and $-u^T \nabla_x g_i(x_0, y_0)$ is nonnegative for all $i \in J_+(\lambda)$, it follows that $u^T \nabla_x g_i(x_0, y_0) = 0$, $i \in J_+(\lambda)$, and hence $u \in \Sigma_\lambda(0)$. Conversely, if $\lambda \in E_0$ and $u \in \Sigma_\lambda(0)$, then $u^T \nabla_x f(x_0, y_0) = 0$ and hence $u \in C$. \square

It follows from Lemma 2.4 that $C = \bar{\Sigma}(0)$. Moreover, the characterization of $\bar{\Sigma}(v)$ given in Lemma 2.2 implies that for all v , C is the recession cone of $\bar{\Sigma}(v)$ (see [29, pp. 61, 62]).

3. Pointwise Lipschitz continuity of optimal solutions. In this section we study Lipschitz continuity of an optimal solution $\bar{x}(y) \in M(y)$ as y tends to y_0 . For that purpose we utilize second order sufficient conditions corresponding to the program (\mathcal{P}_{y_0}) . Consider the index set

$$\mathcal{J} = \bigcup_{\lambda \in E_0} \{i: \lambda_i \neq 0, i = 1, \dots, p\}$$

associated with nonzero Lagrange multipliers. It will be assumed that the functions f and g_i , $i \in \mathcal{J}$, are C^2 while the functions g_i , $i \in \{1, \dots, p\} \setminus \mathcal{J}$, are C^1 smooth. This implies that the Hessian matrices $\nabla_{xx}^2 L(x_0, y_0, \lambda)$ do exist for all $\lambda \in E_0$. Then, under the MF-condition, a feasible point x_0 satisfying the first order necessary conditions is an isolated local solution of program (\mathcal{P}_{y_0}) if

$$(3.1) \quad \max_{\lambda \in \Lambda_0} u^T \nabla_{xx}^2 L(x_0, y_0, \lambda) u > 0$$

for all $u \in C$, $u \neq 0$, where C is the critical cone defined in (2.14), [5], [13], [15]. The corresponding second order necessary conditions are obtained if the strict inequality

sign “>” in (3.1) is replaced by the sign “ \geq ”. We shall refer to (3.1) as the weak second order sufficient conditions.

Unfortunately the weak second order sufficient conditions do not guarantee pointwise Lipschitzian behavior of $\bar{x}(y)$ (see an example in [10, p. 308]). Therefore a stronger form of second order conditions is required. For a given vector v consider the following set of Lagrange multipliers:

$$\Lambda_0^*(v) = \{\lambda \in \Lambda_0: v^T \nabla_y L(x_0, y_0, \lambda) = \max_{\lambda \in \Lambda_0} v^T \nabla_y L(x_0, y_0, \lambda)\}.$$

The set $\Lambda_0^*(v)$ is a face of the convex polytope Λ_0 and hence is the convex hull of the corresponding set $E_0^*(v)$, $E_0^*(v) \subset E_0$, of its extreme points.

Assumption 4 (strong second order sufficient conditions). For every v the inequality

$$(3.2) \quad \max_{\lambda \in \Lambda_0^*(v)} u^T \nabla_{xx}^2 L(x_0, y_0, \lambda) u > 0$$

holds for all $u \in C$, $u \neq 0$.

It may be noted that the sets Λ_0 and $\Lambda_0^*(v)$ in (3.1) and (3.2) can be replaced by the corresponding sets of extreme points E_0 and $E_0^*(v)$, respectively.

Without further discussion we shall make use of the following principle. If $f_1(y), \dots, f_k(y)$ and $\varphi_1(y), \dots, \varphi_k(y)$ are continuous functions such that

$$f_i(y) = \varphi_i(y) + o(\|y\|^2), \quad i = 1, \dots, k,$$

then

$$\max \{f_i(y): i = 1, \dots, k\} = \max \{\varphi_i(y): i = 1, \dots, k\} + o(\|y\|^2).$$

The Hessian matrices $\nabla_{xx}^2 L(x_0, y_0, \lambda)$, $\nabla_{xy}^2 L(x_0, y_0, \lambda)$ and $\nabla_{yy}^2 L(x_0, y_0, \lambda)$ will be denoted by H_{xx}^λ , H_{xy}^λ and H_{yy}^λ , respectively. We write ξ_λ for the corresponding quadratic function

$$(3.3) \quad \xi_\lambda(u, v) = \frac{1}{2} u^T H_{xx}^\lambda u + u^T H_{xy}^\lambda v + \frac{1}{2} v^T H_{yy}^\lambda v.$$

THEOREM 3.1. Suppose that Assumptions 1–4 hold. Then there exists a positive constant K such that

$$(3.4) \quad \|\bar{x}(y) - x_0\| \leq K \|y - y_0\|$$

for all y in a neighborhood of y_0 .

Proof. Without loss of generality we can assume that $(x_0, y_0) = (0, 0)$ and $f(x_0, y_0) = 0$. Suppose that (3.4) is false. Then there exists a sequence $\{y_n\}$ converging to $y_0 = 0$ such that for $x_n = \bar{x}(y_n)$ and $t_n = \|x_n\|$,

$$(3.5) \quad \lim_{n \rightarrow \infty} t_n \|y_n\|^{-1} = \infty.$$

It follows from Lemma 2.1 that $t_n \rightarrow 0$. By the argument of compactness we can assume that $t_n^{-1} x_n$ tends to a vector \bar{u} . Of course $\|\bar{u}\| = 1$ and hence $\bar{u} \neq 0$. From Theorem 2.1 we have that $x_n \in \bar{\Omega}(y_n)$ for sufficiently large n . Furthermore, Lemma 2.3 implies that there exists a sequence $\{u_n\}$, with $u_n \in \bar{\Sigma}(y_n)$, such that

$$\|x_n - u_n\| = o(\|(x_n, y_n)\|).$$

It follows that $t_n^{-1} u_n$ tends to \bar{u} . Since $\text{gph } \bar{\Sigma}$ is a cone

$$t_n^{-1}(y_n, u_n) \in \text{gph } \bar{\Sigma}$$

and by (3.5), $t_n^{-1}(y_n, u_n) \rightarrow (0, \bar{u})$. Moreover, $\text{gph } \bar{\Sigma}$ is a closed set and hence $(0, \bar{u}) \in \text{gph } \bar{\Sigma}$. Consequently by Lemma 2.4, $\bar{u} \in C$.

Now expression (2.5) of Theorem 2.1 implies that $\varphi(y_n) = F(x_n, y_n)$. Then by employing second order Taylor expansions of functions $F_\lambda(x, y) = L(x, y, \lambda)$ we obtain

$$\varphi(y_n) = \max_{\lambda \in E_0} \{y_n^T \nabla_y L(x_0, y_0, \lambda) + \xi_\lambda(x_n, y_n) + o(\|(x_n, y_n)\|^2)\}.$$

Notice that because of the first order necessary conditions the gradients $\nabla_x L(x_0, y_0, \lambda)$, $\lambda \in E_0$, vanish. It follows from (3.5) that $y_n^T H_{yy}^\lambda y_n$ and $x_n^T H_{xy}^\lambda y_n$, $\lambda \in E_0$, are $o(t_n^2)$. Consequently

$$\varphi(y_n) = \max_{\lambda \in E_0} \{y_n^T \nabla_y L(x_0, y_0, \lambda) + \frac{1}{2} t_n^2 \bar{u}^T H_{xx}^\lambda \bar{u}\} + o(t_n^2)$$

and then

$$(3.6) \quad \varphi(y_n) \geq \max_{\lambda \in E_n} \{y_n^T \nabla_y L(x_0, y_0, \lambda) + \frac{1}{2} t_n^2 \bar{u}^T H_{xx}^\lambda \bar{u}\} + o(t_n^2),$$

where $E_n = E_0^*(y_n)$. Moreover, by the definition for every $\lambda \in E_n$

$$y_n \nabla_y L(x_0, y_0, \lambda) = \max_{\lambda \in E_0} y_n^T \nabla_y L(x_0, y_0, \lambda).$$

Consequently it follows from (3.6) that

$$(3.7) \quad \varphi(y_n) \geq \max_{\lambda \in E_0} y_n^T \nabla_y L(x_0, y_0, \lambda) + \frac{1}{2} t_n^2 \max_{\lambda \in E_n} \bar{u}^T H_{xx}^\lambda \bar{u} + o(t_n^2).$$

Now since the set E_0 is finite and hence the number of different sets $E_0^*(v)$ is finite, we may assume that there is a vector \bar{v} such that $E_n = E_0^*(\bar{v})$ for all n . By the strong second order sufficient conditions (Assumption 4) the number

$$\varepsilon = \frac{1}{2} \max \{\bar{u}^T H_{xx}^\lambda \bar{u} : \lambda \in E_0^*(\bar{v})\}$$

is positive. Then (3.7) implies

$$(3.8) \quad \varphi(y_n) \geq \max \{y_n^T \nabla_y L(x_0, y_0, \lambda) : \lambda \in E_0\} + t_n^2 \varepsilon + o(t_n^2).$$

On the other hand it follows, from Lemma 2.3 and pseudo-Lipschitz continuity of $\bar{\Sigma}$, that there exist $\delta > 0$ and a sequence $w_n \in \bar{\Omega}(y_n)$ such that $\|w_n\| \leq \delta \|y_n\|$. Therefore

$$\begin{aligned} \varphi(y_n) &\leq F(w_n, y_n) = \max_{\lambda \in E_0} \{y_n^T \nabla_y L(x_0, y_0, \lambda) + \xi_\lambda(w_n, y_n) + o(\|(w_n, y_n)\|^2)\} \\ &\leq \max_{\lambda \in E_0} \{y_n^T \nabla_y L(x_0, y_0, \lambda) + c \|y_n\|^2\} \\ &= \max \{y_n^T \nabla_y L(x_0, y_0, \lambda) : \lambda \in E_0\} + c \|y_n\|^2 \end{aligned}$$

for some positive constant c . Because of (3.5) this implies that

$$\varphi(y_n) \leq \max \{y_n^T \nabla_y L(x_0, y_0, \lambda) : \lambda \in E_0\} + o(t_n^2),$$

a contradiction to (3.8). \square

We say that $\bar{x}(y)$ is pointwise Lipschitz continuous at y_0 along a given direction \bar{v} if there are positive constants K and ε such that

$$(3.9) \quad \|\bar{x}(y_0 + t\bar{v}) - x_0\| \leq Kt$$

for all $t \in [0, \varepsilon)$. It follows from the proof of Theorem 3.1 that (3.9) is ensured by the second order condition (3.2) applied to $v = \bar{v}$.

Pointwise Lipschitz continuity (3.4) of optimal solutions was proved by Robinson [27, § 4] under stronger second order sufficient conditions than those of our Assumption 4.

Finally it is worthwhile to note that Assumption 4 is reduced to the weak second order sufficient conditions in the following particular cases.

(i) The constraint functions g_i , $i = 1, \dots, p$, are independent of y and hence the feasible set $\Omega(y) = \Omega_0$ is constant.

(ii) The constraint functions $g_i(x, y)$ are linear in x and hence $\nabla_{xx}^2 L(x_0, y_0, \lambda) = \nabla_{xx}^2 f(x_0, y_0)$ for all λ .

(iii) The Lagrange multipliers set $\Lambda_0 = \{\lambda_0\}$ is a singleton.

4. Second order analysis of the optimal value function. In this section we study second order differential properties of the optimal value function $\varphi(y)$. Consider the function

$$(4.1) \quad \beta(u, v) = \max \{v^T \nabla_y L(x_0, y_0, \lambda) + \xi_\lambda(u, v) : \lambda \in E_0\}$$

where ξ_λ is the quadratic function defined in (3.3). We have that $F(x_0, y_0) + \beta(x - x_0, y - y_0)$ gives a second order approximation of the function $F(x, y)$ at the point (x_0, y_0) . The main result of this section is given now in the following theorem.

THEOREM 4.1. *Suppose that Assumptions 1-4 hold. Then*

$$(4.2) \quad \varphi(y_0 + v) - \varphi(y_0) = \inf \{\beta(u, v) : u \in \bar{\Sigma}(v)\} + o(\|v\|^2).$$

Proof. Without loss of generality we can assume that $(x_0, y_0) = (0, 0)$ and $\varphi(y_0) = 0$. Consider the following optimal value function:

$$\psi(y) = \inf \{\beta(x, y) : x \in \bar{\Omega}(y)\}.$$

First we show that

$$(4.3) \quad \varphi(y) = \psi(y) + o(\|y\|^2).$$

By the reduction Theorem 2.1 we have that $\varphi(y) = F(\bar{x}(y), y)$ and hence

$$\varphi(y) = \beta(\bar{x}(y), y) + o(\|y\|^2 + \|\bar{x}(y)\|^2).$$

Together with (3.4) of Theorem 3.1 this implies that

$$\varphi(y) = \beta(\bar{x}(y), y) + o(\|y\|^2).$$

By the definition of ψ , $\psi(y) \leq \beta(\bar{x}(y), y)$ and hence

$$(4.4) \quad \varphi(y) \geq \psi(y) + o(\|y\|^2).$$

Let $\bar{u}(y)$ be a minimizer of $\beta(\cdot, y)$ over $\bar{\Omega}(y)$. It follows from Theorem 3.1 that $\bar{u}(y)$ is pointwise Lipschitz continuous at $y_0 = 0$. Then the other inequality

$$(4.5) \quad \psi(y) \geq \varphi(y) + o(\|y\|^2)$$

can be proved in a way similar to the proof of (4.4). Inequalities (4.4) and (4.5) imply (4.3).

It remains to show that

$$(4.6) \quad \psi(y) = \inf \{\beta(x, y) : x \in \bar{\Sigma}(y)\} + o(\|y\|^2).$$

From Lemma 2.3 we have that $\bar{\Sigma} = D\bar{\Omega}(x_0, y_0)$. By the definition of tangential approximations (condition (2.9)) this implies that there exists $u^*(y) \in \bar{\Sigma}(y)$ such that

$$\|\bar{u}(y) - u^*(y)\| = o(\|(\bar{u}(y), y)\|).$$

Moreover, because of the pointwise Lipschitz continuity of $\bar{u}(y)$ we have

$$\|\bar{u}(y) - u^*(y)\| = o(\|y\|).$$

Then

$$\begin{aligned}\psi(y) &= \beta(\bar{u}(y), y) = \max \{y^T \nabla_y L(x_0, y_0, \lambda) + \xi_\lambda(\bar{u}(y), y) : \lambda \in E_0\} \\ &= \max \{y^T \nabla_y L(x_0, y_0, \lambda) + \xi_\lambda(u^*(y), y) : \lambda \in E_0\} + o(\|y\|^2) \\ &= \beta(u^*(y), y) + o(\|y\|^2).\end{aligned}$$

Since $\beta(u^*(y), y)$ is greater than or equal to the infimum on the right-hand side of (4.6) we obtain that

$$\psi(y) \geq \inf \{\beta(x, y) : x \in \bar{\Sigma}(y)\} + o(\|y\|^2).$$

The other inequality can be proved in a similar way by using condition (2.10) applied to the tangential approximation $\bar{\Sigma}$ of $\bar{\Omega}$. \square

The quadratic approximation result (4.2) of Theorem 4.1 allows us to calculate various forms of second order directional derivatives of the optimal value function. As an example we calculate a curved second order directional derivative of $\varphi(y)$ in the sense of Ben-Tal and Zowe [3].

It follows from Theorem 4.1 that under Assumptions 1–4, the optimal value function is directionally differentiable at y_0 and

$$(4.7) \quad \varphi'(y_0; v) = \max \{v^T \nabla_y L(x_0, y_0, \lambda) : \lambda \in E_0\}.$$

Consider a continuous curve

$$(4.8) \quad y(t) = y_0 + tv + t^2 w + t^2 \varepsilon(t), \quad t > 0,$$

with $v, w \in \mathbb{R}^n$ and $\varepsilon(t) \rightarrow 0$ as $t \rightarrow 0^+$. We have that for a fixed $\lambda \in E_0$,

$$\begin{aligned}L(x_0 + u, y(t), \lambda) &= L(x_0, y_0, \lambda) + tv^T \nabla_y L(x_0, y_0, \lambda) \\ &\quad + t^2 w^T \nabla_y L(x_0, y_0, \lambda) + \xi_\lambda(u, tv) + o(t^2 + \|u\|^2).\end{aligned}$$

Furthermore, under the MF-condition the tangential approximation to the multifunction $\bar{\Omega}(y(t))$ is given by $\bar{\Sigma}(tv)$. Then Theorem 4.1 implies that

$$\varphi(y(t)) - \varphi(y_0) = \inf \{\eta(u, t) : u \in \bar{\Sigma}(tv)\} + o(t^2),$$

where

$$\eta(u, t) = \max_{\lambda \in E_0} \{tv^T \nabla_y L(x_0, y_0, \lambda) + t^2 w^T \nabla_y L(x_0, y_0, \lambda) + \xi_\lambda(u, tv)\}.$$

Let $\bar{u}(t)$ be a minimizer of $\eta(\cdot, t)$ over $\bar{\Sigma}(tv)$. By Theorem 3.1, $\|\bar{u}(t)\|/t$ is bounded for all positive t sufficiently close to zero and hence

$$t^2 w^T \nabla_y L(x_0, y_0, \lambda) + \xi_\lambda(\bar{u}(t), tv) = o(t).$$

Therefore the set E_0 in the definition of $\eta(u, t)$ can be replaced by the set of extreme points $E_0^*(v)$, which corresponds to the maximal value of $v^T \nabla_y L(x_0, y_0, \lambda)$, $\lambda \in E_0$. By (4.7) this maximal value is equal to $\varphi'(y_0; v)$. Consequently, we obtain

$$(4.9) \quad \varphi(y(t)) - \varphi(y_0) = t\varphi'(y_0; v) + \inf \{\zeta_{tv, t^2 w}(u) : u \in \bar{\Sigma}(tv)\} + o(t^2),$$

where

$$(4.10) \quad \zeta_{v, w}(u) = \max \{w^T \nabla_y L(x_0, y_0, \lambda) + \xi_\lambda(u, v) : \lambda \in E_0^*(v)\}.$$

We have that $\bar{\Sigma}(tv) = t\bar{\Sigma}(v)$ and

$$t^{-2} \zeta_{tv, t^2 w}(u) = \zeta_{v, w}(t^{-1}u).$$

It then follows from (4.9) that the second order directional derivative

$$\varphi''(y_0; v, w) = \lim_{t \rightarrow 0^+} \frac{\varphi(y(t)) - \varphi(y_0) - t\varphi'(y_0; v)}{t^2}$$

of Ben-Tal and Zowe [3] exists and can be calculated as follows.

THEOREM 4.2. *Suppose that Assumptions 1–4 hold. Then $\varphi''(y_0; v, w)$ exists and is equal to the minimum of $\zeta_{v,w}(\cdot)$ over the set $\bar{\Sigma}(v)$.*

Notice that it will be sufficient in Theorem 4.2 to assume second order conditions (3.2) (Assumption 4) only for the vector v which appears in the definition (4.8) of the curve $y(t)$. Since C is the recession cone of $\bar{\Sigma}(v)$, these conditions imply that the minimum of $\zeta_{v,w}(\cdot)$ over $\bar{\Sigma}(v)$ exists and the corresponding set of minimizers is nonempty and bounded.

Without strong second order sufficient conditions the following inequality holds.

THEOREM 4.3. *Let $x_0 \in M(y_0)$ and suppose that the MF-condition holds at (x_0, y_0) . Then*

$$(4.11) \quad \limsup_{t \rightarrow 0^+} \frac{\varphi(y(t)) - \varphi(y_0) - t\delta(y_0; v)}{t^2} \leq Z(v, w),$$

where

$$\delta(y_0; v) = \max \{v^T \nabla_y L(x_0, y_0, \lambda) : \lambda \in \Lambda_0\}$$

and

$$Z(v, w) = \inf \{\zeta_{v,w}(u) : u \in \bar{\Sigma}(v)\}.$$

Proof. For a given $\alpha > 0$ consider the function

$$f_\alpha(x, y) = f(x, y) + \alpha \|x - x_0\|^2$$

and the corresponding problem of minimizing $f_\alpha(x, y)$ over the feasible set $\Omega(y)$. It will be enough to show that for all positive α , (4.11) holds with f replaced by f_α (see [33, p. 211] for details). We have that x_0 is the unique minimizer of $f_\alpha(\cdot, y_0)$ over $\Omega(y_0)$ and Assumption 1, applied to f_α , is satisfied. Therefore without loss of generality we can assume that our regularity conditions (Assumptions 1–3) hold, and hence the reduction theorem 2.1 is applicable. Consider now a vector $u \in \bar{\Sigma}(v)$ and choose $u' = u'(t) \rightarrow u$ such that $x_0 + tu' \in \bar{\Omega}(y(t))$ as $t \rightarrow 0^+$. Then we obtain from Theorem 2.1 that

$$\begin{aligned} \varphi(y(t)) &\leq \max \{L(x_0 + tu', y(t), \lambda) : \lambda \in E_0\} \\ &\leq \max \{L(x_0 + tu', y(t), \lambda) : \lambda \in E_0^*(v)\} \\ &= \varphi(y_0) + t\delta(y_0; v) + t^2\zeta_{v,w}(u') + o(t^2). \end{aligned}$$

Since the function $\zeta_{v,w}(\cdot)$ is continuous, it follows that \limsup in the left-hand side of (4.11) is less than or equal to $\zeta_{v,w}(u)$. Vector u is an arbitrary point of $\bar{\Sigma}(v)$ and hence (4.11) follows. \square

Notice that unless Assumptions 1–4 are satisfied, $\delta(y_0; v)$ defined in Theorem 4.3 is not necessarily the directional derivative $\varphi'(y_0; v)$ of the optimal value function. In some cases the infimum $Z(v, w)$ can be $-\infty$. Then the optimal solution $\bar{x}(y)$ cannot be pointwise Lipschitzian at y_0 .

COROLLARY 4.1. *Suppose that Assumptions 1–3 of Theorem 4.3 hold and $Z(v, w) = -\infty$. Then*

$$(4.12) \quad \lim_{t \rightarrow 0^+} t^{-1} \|\bar{x}(y(t)) - x_0\| = \infty.$$

Proof. Without loss of generality we can assume that $\bar{x}(y(t)) \rightarrow x_0$ as $t \rightarrow 0^+$. Suppose that (4.12) is false and hence there exists $t_n \rightarrow 0^+$ such that $t_n^{-1} \|x_n - x_0\|$ is bounded, $x_n = \bar{x}(y_n)$ and $y_n = y(t_n)$. Then by Theorem 2.1

$$\begin{aligned} \varphi(y_n) &= \max \{L(x_n, y_n, \lambda) : \lambda \in E_0\} \\ &= \varphi(y_0) + t_n \delta(y_0; v) + \max_{\lambda \in E_0^B(v)} \{t_n^2 w^T \nabla_y L(x_0, y_0, \lambda) + \xi_\lambda(x_n - x_0, y_n - y_0)\} + o(t_n^2). \end{aligned}$$

It follows that

$$(\varphi(y_n) - \varphi(y_0) - t_n \delta(y_0; v)) / t_n^2$$

is bounded, a contradiction to (4.11). \square

In particular $Z(v, w)$ is $-\infty$ if there exists a vector $\bar{u} \in C$ such that

$$\max_{\lambda \in \Lambda_0^B(v)} \bar{u}^T \nabla_{xx}^2 L(x_0, y_0, \lambda) \bar{u} < 0.$$

Therefore, in order to ensure pointwise Lipschitz continuity of the optimal solution, the strong second order sufficient conditions of Assumption 4 are the weakest possible.

Now let us consider some particular cases of Theorem 4.1. If the constraint functions g_i , $i = 1, \dots, p$, are independent of y , then $\bar{\Sigma}(v)$ coincides with the critical cone C and

$$\beta(u, v) = v^T \nabla_y f(x_0, y_0) + \Xi(u, v),$$

where

$$\Xi(u, v) = \max \{\xi_\lambda(u, v) : \lambda \in E_0\}.$$

COROLLARY 4.2. *Suppose that functions g_i , $i = 1, \dots, p$, are independent of y and Assumptions 1–4 hold. Then*

$$\varphi(y_0 + v) - \varphi(y_0) = v^T \nabla_y f(x_0, y_0) + \kappa(v) + o(\|v\|^2),$$

where

$$\kappa(v) = \inf \{\Xi(u, v) : u \in C\}.$$

If the functions g_i are affine, then $\xi_\lambda(u, v) = \xi(u, v)$ for all λ with

$$\xi(u, v) = \frac{1}{2} u^T \nabla_{xx}^2 f(x_0, y_0) u + u^T \nabla_{xy}^2 f(x_0, y_0) v + \frac{1}{2} v^T \nabla_{yy}^2 f(x_0, y_0) v.$$

COROLLARY 4.3. *Suppose that g_i , $i = 1, \dots, p$, are affine functions and Assumptions 1–4 hold. Then*

$$\varphi(y_0 + v) - \varphi(y_0) = \varphi'(y_0; v) + \kappa(v) + o(\|v\|^2),$$

where $\varphi'(y_0; v)$ is the directional derivative given by (4.7) and

$$\kappa(v) = \inf \{\xi(u, v) : u \in \bar{\Sigma}(v)\}.$$

Notice that Assumption 4 in Corollaries 4.2 and 4.3 is reduced to the weak second order sufficient conditions. In the case $\Lambda_0 = \{\lambda_0\}$ is a singleton the MF-condition follows [18], [34] and we obtain the next corollary.

COROLLARY 4.4. *Suppose that Λ_0 is a singleton, Assumptions 1 and 2 hold and the weak second order sufficient conditions are satisfied. Then*

$$\varphi(y_0 + v) - \varphi(y_0) = v^T \nabla_y L(x_0, y_0, \lambda_0) + \kappa(v) + o(\|v\|^2),$$

where

$$\kappa(v) = \inf \{\xi_{\lambda_0}(u, v) : u \in \Sigma_{\lambda_0}(v)\}.$$

The second order conditions employed in Corollary 4.4 are weaker than a strong second order condition utilized in [34, Assumption 8]. Therefore the result of Corollary 4.4 is stronger than a similar result in [34, Thm. 4.1].

In the situations of Corollaries 4.2–4.4 the first term in the second order approximation of φ is the directional derivative $\varphi'(y_0; v)$. The corresponding function $\kappa(v)$ is positively homogeneous of degree 2 and, since $\bar{\Sigma}$ is pseudo-Lipschitzian, is continuous. Consequently $\kappa(v)$ gives a second order directional derivative of φ in the following sense:

$$\lim_{\substack{t \rightarrow 0^+ \\ \bar{v} \rightarrow v}} \frac{\varphi(y_0 + t\bar{v}) - \varphi(y_0) - t\varphi'(y_0; \bar{v})}{t^2} = \kappa(v).$$

5. Directional differentiability of optimal solutions. In this section we study differentiability properties of an optimal solution $\bar{x}(y) \in M(y)$. Consider the curve $y(t)$ defined in (4.8) and the function $\zeta_{v,w}$ given in (4.10). Denote by $\Pi(v, w)$ the set of minimizers of $\zeta_{v,w}(\cdot)$ over $\bar{\Sigma}(v)$. As has been mentioned earlier, the second order condition (3.2) ensures that the set $\Pi(v, w)$ is nonempty and compact. The result of the following theorem shows that if $\Pi(v, w) = \{z\}$ is a singleton, then z gives the directional derivative of $\bar{x}(y)$ along the curve $y = y(t)$.

THEOREM 5.1. *Suppose that Assumptions 1–4 hold. Then*

$$(5.1) \quad \lim_{t \rightarrow 0^+} \frac{\text{dist}(\bar{x}(y(t)) - x_0; t\Pi(v, w))}{t} = 0.$$

Proof. We assume that $(x_0, y_0) = (0, 0)$ and $\varphi(y_0) = 0$. For $v = 0$ the set $\Pi(0, w)$ is $\{0\}$ and, by Theorem 3.1, $\bar{x}(y(t))$ is $o(t)$. Thus in this case (5.1) trivially holds. Therefore we assume subsequently that $v \neq 0$. Suppose that (5.1) is false and hence there exist a sequence $t_n \rightarrow 0^+$ and $\varepsilon > 0$ such that

$$(5.2) \quad \text{dist}(x_n; t_n \Pi(v, w)) \geq \varepsilon t_n,$$

where $x_n = \bar{x}(y_n)$ and $y_n = y(t_n)$. Theorem 3.1 implies that the sequence $\{\|y_n\|^{-1}x_n\}$ and hence $\{t_n^{-1}x_n\}$ are bounded. Therefore we can assume that $\{t_n^{-1}x_n\}$ converges to a vector \bar{u} . Since the distance function $d_{\Pi}(\cdot) = \text{dist}(\cdot; \Pi(v, w))$ is continuous it follows from (5.2) that

$$\text{dist}(\bar{u}; \Pi(v, w)) \geq \varepsilon$$

and thus \bar{u} does not belong to the set $\Pi(v, w)$. Since $x_n \in \bar{\Omega}(y_n)$ and $\bar{\Sigma}(tv)$ is tangent to $\bar{\Omega}(y(t))$, there exists a sequence $\{u_n\}$, with $u_n \in \bar{\Sigma}(t_nv)$, such that $\|x_n - u_n\| = o(t_n)$. It follows that $t_n^{-1}u_n$ tends to \bar{u} . Furthermore, $\text{gph } \bar{\Sigma}$ is a cone and hence $t_n^{-1}(t_nv, u_n) \in \text{gph } \bar{\Sigma}$. Since $\bar{\Sigma}$ is closed we obtain that $(v, \bar{u}) \in \text{gph } \bar{\Sigma}$, i.e., $\bar{u} \in \bar{\Sigma}(v)$.

On the other hand,

$$\begin{aligned} \varphi(y_n) &= F(x_n, y_n) = \max \{y_n^T \nabla_y L(x_0, y_0, \lambda) + \xi_\lambda(x_n, y_n) : \lambda \in E_0\} + o(t_n^2) \\ &= \max \{t_n v^T \nabla_y L(x_0, y_0, \lambda) + t_n^2 w^T \nabla_y L(x_0, y_0, \lambda) + t_n^2 \xi_\lambda(\bar{u}, v) : \lambda \in E_0\} + o(t_n^2) \\ &= t_n \varphi'(y_0; v) + t_n^2 \zeta_{v,w}(\bar{u}) + o(t_n^2). \end{aligned}$$

This implies that $\varphi''(y_0; v, w) = \zeta_{v,w}(\bar{u})$. It then follows from Theorem 4.2 that

$$\zeta_{v,w}(\bar{u}) = \inf \{\zeta_{v,w}(u) : u \in \bar{\Sigma}(v)\}$$

and we already have shown that $\bar{u} \in \bar{\Sigma}(v)$. Consequently $\bar{u} \in \Pi(v, w)$, a contradiction. \square

Consider the case of $w = 0$ and let the set $\Pi(v, 0) = \{z\}$ be a singleton. Then it follows from (5.1) that $\bar{x}(y)$ is directionally differentiable at $y = y_0$ in the direction v

and the corresponding directional derivative is given by $\bar{x}'(y_0; v) = z$. It is interesting to note that since the set $E_0^*(v)$ may change for various values of v , the optimal set $\Pi(v, 0)$ can depend discontinuously on v even if $\Pi(v, 0)$ is a singleton for all v . Therefore it may happen that $\bar{x}(y)$ is directionally differentiable at $y = y_0$, but the directional derivative $\bar{x}'(y_0; v)$ is *not continuous* as a function of v . This then implies that $\bar{x}(y)$ is not locally Lipschitz near y_0 . An example of this type is given in Robinson [26, p. 219]. We consider below a simplified version of that example.

Example. Consider the following program:

$$\begin{aligned} &\text{minimize} && \frac{1}{2}(x_1 - 1)^2 + \frac{1}{2}x_2^2 \\ &\text{subject to} && x_1 \leq 0, \\ &&& x_1 + y_1x_2 + y_2 \leq 0 \end{aligned}$$

depending on the parameter vector $y = (y_1, y_2)$. For $y_0 = (0, 0)$ it has a unique solution $x_0 = (0, 0)$. Assumptions 1-4 are satisfied with the set E_0 consisting of two points $(1, 0)$ and $(0, 1)$. Consider directions $v = (1, 0)$, $w = (0, \alpha)$ and the associated curve $y(t) = (t, \alpha t^2)$. We have that $E_0^*(v) = E_0$, $\bar{\Sigma}(v) = \{u: u_1 = 0\}$ and

$$\zeta_{v,w}(u) = \max \left\{ \frac{1}{2}u_1^2 + \frac{1}{2}u_2^2; \alpha + u_2 + \frac{1}{2}u_1^2 + \frac{1}{2}u_2^2 \right\}.$$

It follows that $\Pi(v, w) = \{(0, 0)\}$ if $\alpha \leq 0$, $\Pi(v, w) = \{(0, -\alpha)\}$ if $0 < \alpha \leq 1$ and $\Pi(v, w) = \{(0, -1)\}$ if $1 < \alpha$, which gives the directional derivative of $\bar{x}(y)$ along the curve $y(t)$ for various values of α . For $y_1 \geq 0$, straightforward calculations give $\bar{x}(y) = (0, 0)$ if $y_2 \leq 0$, $\bar{x}(y) = (0, -y_1^{-1}y_2)$ if $0 < y_2 \leq y_1^2$ and

$$(5.3) \quad \bar{x}(y) = (1 + y_1^2)^{-1}(y_1^2 - y_2, -y_1 - y_1y_2)$$

for $y_1^2 \leq y_2$.

Now consider $y(t) = tv$ with $v = v(\gamma) = (1, \gamma)$, $\gamma > 0$. Then $E_0^*(v)$ consists of one point $(0, 1)$ and $\bar{\Sigma}(v) = \{u: u_1 = -\gamma\}$. The set $\Pi(v, 0)$ consists of the minimizer of the function $\frac{1}{2}u_1^2 + \frac{1}{2}u_2^2 + u_2$ over the set $\bar{\Sigma}(v)$. This minimizer is $(-\gamma, -1)$ and hence $\bar{x}'(y_0; v) = (-\gamma, -1)$. Straightforward calculations, based on (5.3), confirm this result. As $\gamma \rightarrow 0^+$ this directional derivative tends to $(0, -1)$ while the corresponding directional derivative for $\gamma = 0$ is $(0, 0)$. A reason for such discontinuous behavior of the directional derivative is that $E_0^*(v)$ contains two points for $\gamma = 0$ while $E_0^*(v)$ is a singleton for $\gamma > 0$.

6. Differentiability of metric projections. In this section we study differentiability properties of metric projections P_Ω at a point $y_0 \in \mathbb{R}^n \setminus \Omega(y_0)$. For the sake of simplicity we consider the case where $\Omega(y)$ is defined by inequality constraints only:

$$(6.1) \quad \Omega(y) = \{x: g_i(x, y) \leq 0, i = 1, \dots, p\},$$

with g_i being C^2 functions and $x, y \in \mathbb{R}^n$. We also suppose that the functions $g_i(\cdot, y_0)$, $i = 1, \dots, p$, are convex and hence the set $\Omega_0 = \Omega(y_0)$ is convex. Now sensitivity analysis of the previous section can be applied straightforwardly if the objective function is taken to be $f(x, y) = \|y - x\|$, where $\|\cdot\|$ is a given norm in \mathbb{R}^n .

It will be assumed that the norm $\|\cdot\|$ is strictly convex and C^2 smooth in a neighborhood of the point $y_0 - x_0$. We write $\nabla\|y\|$ and $\nabla^2\|y\|$ for the gradient vector and the Hessian matrix of the norm $\|\cdot\|$ at a point y . It will be assumed that the matrix

$$U = \frac{1}{2}\nabla^2\|y_0 - x_0\|$$

is positive definite. Because of the convexity assumption there is a unique minimizer x_0 of the objective function $f(\cdot, y_0)$ over Ω_0 and hence our Assumption 2 is satisfied.

By definition, $P_\Omega(y_0) = x_0$. Moreover, here the MF-condition is equivalent to the Slater condition; that is, there exists a point u such that $g_i(u, y_0) < 0$ for $i = 1, \dots, p$.

The associated Lagrangian function is

$$L(x, y, \lambda) = \|y - x\| + \sum_{i=1}^p \lambda_i g_i(x, y).$$

Under the Slater condition the first order necessary conditions hold at x_0 (we suppose subsequently that all constraints are active at (x_0, y_0) , i.e., $g_i(x_0, y_0) = 0$, $i = 1, \dots, p$): The set Λ_0 of nonnegative vectors λ satisfying the equation

$$\nabla_x L(x_0, y_0, \lambda) = -\nabla \|y_0 - x_0\| + \sum_{i=1}^p \lambda_i \nabla_x g_i(x_0, y_0) = 0$$

is nonempty. Furthermore, Λ_0 is a convex compact polytope and hence is the convex hull of the set E_0 of its extreme points. We retain essentially the same notation and terminology as in the previous sections applied to the present situation. In particular the set $E_0^*(v)$ is defined as the set of Lagrange multipliers $\lambda \in E_0$ maximizing $v^T \nabla_y L(x_0, y_0, \lambda)$ and we consider the function

$$(6.2) \quad \pi(u, v) = \max \{ \xi_\lambda(u, v) : \lambda \in E_0^*(v) \}.$$

For all v the function $\pi(\cdot, v)$ is the pointwise maximum of strictly convex (quadratic) functions and hence is strictly convex. Consequently $\pi(\cdot, v)$ has a unique minimizer over the convex set $\bar{\Sigma}(v)$. We denote this minimizer by $\Pi(v)$. Finally we note that here Assumptions 1 and 4 hold automatically. Therefore we obtain from Theorem 5.1 the following result.

THEOREM 6.1. *Suppose that the Slater condition holds. Then P_Ω is directionally differentiable at y_0 and its directional derivative $P'_\Omega(y_0; v)$ is given by $\Pi(v)$.*

As we have mentioned earlier, the mapping $\Pi(v)$ and hence the directional derivative $P'_\Omega(y_0; v)$ are not necessarily continuous in v (see the example in § 5).

Now we consider some particular cases of Theorem 6.1 parallel to the situations of Corollaries 4.2–4.4. First suppose that the functions g_i , $i = 1, \dots, p$, are independent of y and hence $\Omega(y) = \Omega_0$ is constant. Then

$$(6.3) \quad \pi(u, v) = (v - u)^T U(v - u) + \mu(u),$$

where

$$\mu(u) = \frac{1}{2} \max \left\{ u^T \left(\sum_{i=1}^p \lambda_i \nabla^2 g_i(x_0) \right) u : \lambda \in E_0 \right\}.$$

COROLLARY 6.1. *Suppose that functions g_i , $i = 1, \dots, p$, are independent of y and the Slater condition holds. Then P_Ω is directionally differentiable at y_0 and $P'_\Omega(y_0; v)$ is equal to the minimizer of the function $\pi(\cdot, v)$ defined in (6.3), over the critical cone C .*

In the situation of Corollary 6.1 the metric projection P_Ω is differentiable at y_0 in the usual sense if and only if $P'_\Omega(y_0; v)$ is linear in v . (Strictly speaking, linearity of $P'_\Omega(y_0; v)$ implies Gâteaux differentiability only. However, here P_Ω is Lipschitz continuous (e.g., [39]) and hence Fréchet differentiability follows.) A sufficient condition for such linearity is that the function $\mu(u)$ is quadratic and C is a linear space. It is not difficult to see that the critical cone C becomes a linear space if and only if $\bigcup \{J_+(\lambda) : \lambda \in E_0\} = \{1, \dots, p\}$. When Λ_0 is a singleton this means that all Lagrange multipliers are positive, i.e., there is strict complementary slackness. Moreover, if Λ_0 is a singleton, then $\mu(u)$ is quadratic and the strict complementary slackness becomes a necessary and sufficient condition for differentiability of P_Ω . In the case where the

gradient vectors of the constraint functions are linearly independent this result is due to Malanowski [19].

Now suppose that the functions g_i , $i = 1, \dots, p$, are affine. Then

$$(6.4) \quad \pi(u, v) = (v - u)^T U(v - u)$$

and we obtain the following result.

COROLLARY 6.2. *Suppose that g_i , $i = 1, \dots, p$, are affine functions and the Slater condition holds. Then P_Ω is directionally differentiable at y_0 and $P'_\Omega(y_0; v)$ is equal to the minimizer of the quadratic function given in (6.4) over the set $\bar{\Sigma}(v)$.*

If in addition the norm $\|\cdot\|$ is Euclidean, then $P'_\Omega(y_0; v)$ is given by the metric projection of v onto the set $\bar{\Sigma}(v)$. When the (affine) functions g_i , $i = 1, \dots, p$, are independent of y , and hence $\bar{\Sigma}(v) = C$ for all v , this result is due to Haraux [12].

COROLLARY 6.3. *Suppose that $\Lambda_0 = \{\lambda_0\}$ is a singleton. Then P_Ω is directionally differentiable at y_0 and $P'_\Omega(y_0; v)$ is equal to the minimizer of the quadratic function $\xi_{\lambda_0}(\cdot, v)$ over $\Sigma_{\lambda_0}(v)$.*

Notice that in the cases of Corollaries 6.1–6.3 the directional derivative $P'_\Omega(y_0; v)$ is continuous in v .

REFERENCES

- [1] T. J. ABATZOGLOU, *The minimum norm projection on C^2 -manifolds in R^n* , Trans. Amer. Math. Soc., 243 (1978), pp. 115–122.
- [2] J.-P. AUBIN, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.
- [3] A. BEN-TAL AND J. ZOWE, *A unified theory of first and second-order conditions for extremum problems in topological vector spaces*, Math. Programming Stud., 19 (1982), pp. 39–76.
- [4] J. H. BIGELOW AND N. Z. SHAPIRO, *Implicit function theorems for mathematical programming and for systems of inequalities*, Math. Programming, 6 (1974), pp. 141–156.
- [5] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.
- [6] A. V. Fiacco, *Sensitivity analysis for nonlinear programming using penalty methods*, Math. Programming, 10 (1976), pp. 287–311.
- [7] ———, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, New York, 1983.
- [8] S. FITZPATRICK AND R. R. PHELPS, *Differentiability of the metric projection in Hilbert space*, Trans. Amer. Math. Soc., 270 (1982), pp. 483–501.
- [9] J. GAUVIN, *A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming*, Math. Programming, 12 (1977), pp. 136–138.
- [10] J. GAUVIN AND J. W. TOLLE, *Differential stability in nonlinear programming*, this Journal, 15 (1977), pp. 294–311.
- [11] J. GAUVIN AND F. DUBEAU, *Differential properties of the marginal function in mathematical programming*, Math. Programming Stud., 19 (1982), pp. 101–119.
- [12] A. HARAUX, *How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities*, J. Math. Soc. Japan, 29 (1977), pp. 615–631.
- [13] M. R. HESTENES, *Optimization Theory—The Finite Dimensional Case*, John Wiley, New York, 1975.
- [14] R. B. HOLMES, *Smoothness of certain metric projections on Hilbert space*, Trans. Amer. Math. Soc., 184 (1973), pp. 87–100.
- [15] A. D. IOFFE, *Necessary and sufficient conditions for a local minimum. 3: Second order conditions and augmented duality*, this Journal, 17 (1979), pp. 266–288.
- [16] K. JITTORNTUM, *Solution point differentiability without strict complementarity in nonlinear programming*, Math. Programming Stud., 21 (1984), pp. 127–138.
- [17] J. KRUSKAL, *Two convex counterexamples: a discontinuous envelope function and a non-differentiable nearest point mapping*, Proc. Amer. Math. Soc., 23 (1969), pp. 697–703.
- [18] J. KYPARISIS, *On uniqueness of Kuhn–Tucker multipliers in non-linear programming*, Math. Programming, 32 (1985), pp. 242–246.

- [19] K. MALANOWSKI, *Differentiability with respect to parameters of solutions to convex programming problems*, Math. Programming, 33 (1985), pp. 352–361.
- [20] O. L. MANGASARIAN AND S. FROMOVITZ, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 7 (1967), pp. 37–47.
- [21] O. L. MANGASARIAN AND T. H. SHIAU, *Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems*, this Journal, 25 (1987), pp. 583–595.
- [22] J.-P. PENOT, *On the existence of Lagrange multipliers in nonlinear programming in Banach spaces*, Lecture Notes in Control and Information Sciences 30, Springer, Berlin, 1981, pp. 89–104.
- [23] R. R. PHELPS, *Metric projections and the gradient projection method in Banach spaces*, this Journal, 23 (1985), pp. 973–977.
- [24] ———, *The gradient projection method using Curry's steplength*, this Journal, 24 (1986), pp. 692–699.
- [25] S. M. ROBINSON, *Perturbed Kuhn–Tucker points and rates of convergence for a class of nonlinear programming algorithms*, Math. Programming, 7 (1974), pp. 1–16.
- [26] ———, *Stability theory for systems of inequalities, Part II: Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
- [27] ———, *Generalized equations and their solutions, Part II: Applications to nonlinear programming*, Math. Programming Stud., 19 (1982), pp. 200–221.
- [28] ———, *Local structure of feasible sets in nonlinear programming, Part III: Stability and sensitivity*, Math. Programming Stud., to appear.
- [29] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.
- [30] ———, *Directional differentiability of the optimal value function in a nonlinear programming problem*, Math. Programming Stud., 21 (1984), pp. 213–226.
- [31] ———, *Lipschitzian Properties of Multifunctions*, Nonlinear Anal. Theory Methods Appl., 9 (1985), pp. 867–885.
- [32] ———, *Extensions of Subgradient Calculus with Applications to Optimization*, Nonlinear Analysis, Theory, Methods and Applications, 9 (1985), pp. 665–698.
- [33] A. SHAPIRO, *Second-order derivatives of extremal-value functions and optimality conditions for semi-infinite programs*, Math. Oper. Res., 10 (1985), pp. 207–219.
- [34] ———, *Second order sensitivity analysis and asymptotic theory of parametrized nonlinear programs*, Math. Programming, 33 (1985), pp. 280–299.
- [35] ———, *On differentiability of metric projections in \mathbb{R}^n , 1: Boundary case*, Proc. Amer. Math. Soc., 99 (1987), pp. 123–128.
- [36] ———, *Directional differentiability of metric projections onto moving sets at boundary points*, J. Math. Anal. Appl., to appear.
- [37] J. E. SPINGARN, *Fixed and variable constraints in sensitivity analysis*, this Journal, 18 (1980), pp. 297–310.
- [38] R. J.-B. WETS, *On inf-compact mathematical programs*, Fifth Conf. on Optimization Techniques, Part I, Lecture Notes Comput. Sci., 3, Springer, Berlin, 1973, pp. 426–436.
- [39] E. H. ZARANTONELLO, *Projections on convex sets in Hilbert space and spectral theory*, in Contributions to Nonlinear Functional Analysis, Academic Press, New York, 1971, pp. 237–424.

A CONTROLLER DEGREE BOUND FOR \mathcal{H}^∞ -OPTIMAL CONTROL PROBLEMS OF THE SECOND KIND*

D. J. N. LIMEBEER† AND G. D. HALIKIAS†

Abstract. This paper is a continuation of our work on \mathcal{H}^∞ -optimal control problems which may be embedded in the linear fractional configuration of Fig. 1. In two previous articles [19], [20], a controller degree bound was established for problems in which both $P_{12}(s)$ and $P_{21}(s)$ are square (problems of the first kind). If the McMillan degree of $P(s)$ is n , it was shown that there exist \mathcal{H}^∞ -optimal controllers with McMillan degree no greater than $n - 1$.

Here we switch our attention to problems of the second kind. That is, we allow $P_{12}(s)$ to have more rows than columns (with $P_{21}(s)$ square), or alternatively, we allow $P_{21}(s)$ to have more columns than rows (with $P_{12}(s)$ square). Our main result shows that the degree bound derived previously for problems of the first kind carries over to problems of the second kind without change. In addition to the controller degree bound, our analysis suggests a number of modifications which are easily made to currently available computer programs [7], [26]. Test calculations (for problems of the second kind) show that these improvements result in a marked reduction in computation time and also enhance the numerical robustness of the software.

Key words. \mathcal{H}^∞ -optimal control, approximation theory, cancellations, degree bound, Nehari's theorem

AMS (MOS) subject classification. 93C35

1. Introduction. The generalized regulator in Fig. 1 has been adopted as the standard configuration on which \mathcal{H}^∞ -optimal control studies are based. By appropriately choosing the four partitions of $P(s)$, most design examples of engineering interest may be embedded in this diagram. Early \mathcal{H}^∞ studies were special in the sense that $P_{12}(s)$ and $P_{21}(s)$ could be chosen square. Examples of such problems (which we call problems of the first kind) are the optimal sensitivity problem [5], [11], [12], [25], [30], [31], and the robust stabilization problem [15], [16]. Problems of the first kind have now been fully analysed and a controller degree bound has also been found [19], [20]. Although this class of problems admits a particularly elegant and simple solution, it is too special for most practical engineering problems.

If we allow one of the off-diagonal blocks of $P(s)$ to be nonsquare, the range of problems which we may study becomes considerably larger. We call such problems, problems of the second kind. A popular example of which is the so-called mixed-sensitivity problem [8], [9], [10], [13], [27]. As one would expect, this enlarged class of problems is more difficult to analyse as well as being computationally more demanding. In this paper we carry out a detailed analysis of these problems and prove that

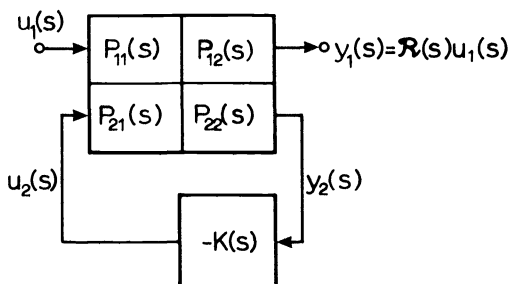


FIG. 1. Generalised regulator configuration.

* Received by the editors February 17, 1987; accepted for publication (in revised form) August 12, 1987.

† Department of Electrical Engineering, Imperial College, Exhibition Road, London, U.K.

the transition from problems of the first kind to problems of the second kind is free in terms of the controller state dimension. We believe that this is a most surprising and encouraging development. Despite advances in microprocessor technology, the potentially high McMillan degree of \mathcal{H}^∞ -optimal controllers is an issue which worries engineers. The implementation of high order controllers is not necessarily straightforward. Computation time constraints and finite precision effects are the obvious bugbears.

In a conference paper [21] we derived a bound of $2n - 1$ (where $n = \deg(P)$) for the degree of a class of \mathcal{H}^∞ -optimal controllers for problems of the second kind. I. Postlethwaite and his colleagues inform us that several of their computer examples suggest the existence of controllers requiring only $n - 1$ states. Subsequent to their comments, we traced this discrepancy (of n states) to a weak step in our original (unpublished) proof. This oversight has been rectified, and the present paper contains a cancellation analysis of the state-space algorithm described in [7], [26]. This work has not only lead to the degree bound, but it has also suggested a number of improvements to our current software [26].

A recent paper by Ball and Cohen [4] addresses the central model matching problem from a geometric viewpoint. Although their approach is still essentially iterative, it may lead to better computer algorithms and a shorter proof of the controller degree bound. This work may also be helpful in the case of problems of the third kind.

Section 2 contains the notation, a problem description and a brief review of the parametrization and optimization theory. In § 3 we use Riccati equation balancing techniques to derive minimal realizations for the transfer functions associated with the model-matching problem [7], [26]. Section 4 contains the cancellation analysis, the degree bound derivation and some computer time trials. The conclusions are in § 5. All the detailed calculations are contained in appendices at the end of the paper.

2. Notation and background theory.

2.1. Notation.

$\mathbb{R}, \mathbb{R}_+, \mathbb{C}$	real, nonnegative and complex numbers;
$\mathbb{R}(s)$	field of rational functions in s with real coefficients;
$\mathbb{F}^{m \times l}$	set of $m \times l$ matrices with elements in \mathbb{F} ($=\mathbb{R}, \mathbb{C}, \mathbb{R}(s)$ etc.);
$\mathbb{C}_+, \bar{\mathbb{C}}_+$	open (respectively, closed) right half-plane;
$\mathbb{C}_-, \bar{\mathbb{C}}_-$	open (respectively, closed) left half-plane;
$\lambda(A), \lambda_{\max}(A)$	eigenvalue of a square matrix A , largest eigenvalue of A ;
A^*	complex conjugate transpose of $A \in \mathbb{C}^{m \times l}$ (transpose if $A \in \mathbb{R}^{m \times l}$)!
$A \geq 0, A > 0$	A is positive semidefinite (respectively, positive definite);
$A \leq 0, A < 0$	A is negative semidefinite (respectively, negative definite);
\mathcal{RL}^∞	space of matrices in $\mathbb{R}(s)^{m \times l}$ which have no poles on the $j\omega$ axis (including the point at ∞);
$\ \cdot\ _\infty$	\mathcal{L}^∞ -norm of matrices in \mathcal{RL}^∞ ;
$\mathcal{RH}_+, \mathcal{RH}_-$	subspaces of \mathcal{RL}^∞ ; matrices which have no poles in $\bar{\mathbb{C}}_+$ (respectively, $\bar{\mathbb{C}}_-$);
Γ_G	Hankel operator associated with $G(s) \in \mathcal{RH}_+^\infty$;
$\sigma_i(G(s))$	i th Hankel singular value of $G(s)$ (i.e., of Γ_G) in decreasing order of magnitude;
$\ G(s)\ _H$	$= \sigma_1(G(s))$, the Hankel norm of $G(s)$;
$\operatorname{Re}(s), \bar{s}, s $	the real part, complex conjugate and modulus of $s \in \mathbb{C}$;
$G^*(s)$	$= G(-\bar{s})^*$, the para-Hermitian conjugate of $G(s)$;
$\Rightarrow, \Leftarrow, \Leftrightarrow$	implies, is implied by, if and only if.

Associated with a transfer function matrix $G(s) \in \mathbb{R}(s)^{m \times l}$ of McMillan degree n is a state-space realisation

$$(2.1) \quad G(s) = D + C(sI - A)^{-1}B$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times l}$, $C \in \mathbb{R}^{m \times n}$ and $D \in \mathbb{R}^{m \times l}$. We will use the alternative notation $G(s) \triangleq (A, B, C, D)$ or

$$(2.2) \quad G(s) \stackrel{s}{=} \begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

In the notation above, we have $G^*(s) \stackrel{s}{=} (-A^*, C^*, -B^*, D^*)$ and in the case that D is nonsingular, we also have $G^{-1}(s) \stackrel{s}{=} (A - BD^{-1}C, BD^{-1}, -D^{-1}C, D^{-1})$. If $G^{-1}(s) = G^*(s)$, then $G(s)$ is all-pass. $G(s)$ is called asymptotically stable if it has no poles in $\bar{\mathbb{C}}_+$.

If $G(s) \triangleq (A, B, C, D)$, the system matrix of the given realisation is defined as [24]:

$$\begin{bmatrix} sI - A & -B \\ C & D \end{bmatrix}$$

and the system zeros are defined to be the points at which the system matrix loses normal rank. In the case when D is nonsingular, the system zeros are also given by $\lambda(A - BD^{-1}C)$. The input decoupling zeros (uncontrollable modes) are points at which $[sI - A | B]$ loses rank. The output decoupling zeros (unobservable modes) are the points at which $[sI - A^* | C^*]$ loses rank. In the sequel, the term “zero” refers to “system zero” unless stated otherwise. Obviously, {input decoupling zeros} \cup {output decoupling zeros} are a subset of both $\lambda(A)$ and the set of system zeros. The realisation (A, B, C, D) is minimal if it has no input/output decoupling zeros. A sufficient condition for this is that all system zeros are distinct from $\lambda(A)$.

If $G_1(s) \triangleq (A_1, B_1, C_1, D_1)$ and $G_2(s) \triangleq (A_2, B_2, C_2, D_2)$ then the cascade system $G_1 G_2(s)$ has a realisation given by

$$\begin{bmatrix} A_1 & B_1 \\ C_1 & D_1 \end{bmatrix} * \begin{bmatrix} A_2 & B_2 \\ C_2 & D_2 \end{bmatrix} = \left[\begin{array}{cc|c} A_1 & B_1 C_2 & B_1 D_2 \\ 0 & A_2 & B_2 \\ \hline C_1 & D_1 C_2 & D_1 D_2 \end{array} \right]$$

where we have taken the “multiplication” of two realisations to mean cascading the two systems. This is not to be confused with ordinary matrix multiplication. The context will always make the distinction between these two possible interpretations clear.

If a basis change T is introduced into the state space of $G(s)$, we will take this to mean $G(s) \stackrel{s}{=} (TAT^{-1}, TB, CT^{-1}, D)$. The McMillan degree of $G(s)$ will be written as $\deg(G)$ and the set of poles (zeros) of $G(s)$ will be denoted {poles of G } ({zeros of G }).

Let $P(s)$ be a partitioned matrix with a state space realisation given by

$$(2.3) \quad P(s) = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}(s) \stackrel{s}{=} \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{21} \end{array} \right];$$

then

$$(2.4) \quad P_{ij}(s) = C_i(sI - A)^{-1}B_j + D_{ij}$$

is a state-space realisation of $P_{ij}(s)$. A linear fractional transformation for the partitioned matrix P and a matrix K is defined as

$$F_l(P, K) = P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21}$$

where K is of dimension $l \times m$ if P_{22} has dimension $m \times l$.

When proving the main theorem of this paper we will require a standard result describing the properties of the solutions of algebraic Riccati equations of the form

$$(2.5) \quad A^*P + PA + PBB^*P + C^*C = 0.$$

The Hamiltonian matrix associated with (2.5) is

$$(2.6) \quad H = \begin{bmatrix} A & BB^* \\ -C^*C & -A^* \end{bmatrix}.$$

LEMMA 2.1. (i) If (A, B) is stabilisable and H is free of imaginary axis eigenvalues, there exists a unique stabilising solution $P = P^* \geq 0$ to (2.5).

(ii) If (A, C) is observable, every solution P to (2.5) is nonsingular.

Proof. The proof of the first part is essentially due to Kučera [17], although Doyle noted that changing the sign of quadratic term in (2.5) did not invalidate Kučera's proof [7].

To prove the second part, we suppose for contradiction that P is singular. This supposition means that there exists $v \neq 0$ such that $Pv = 0$. Now $v^*(2.5)v \Rightarrow Cv = 0$ and $(2.5)v \Rightarrow PAv = 0$. If $v' := Av$, then $v'^*(2.5)v' \Rightarrow Cv' = CAv = 0$ and $(2.5)v' \Rightarrow PA^2v = 0$. Continuing in this way gives $Cv = 0$, $CAv = 0, \dots, CA^{n-1}v = 0$ or $v^*[C^*|A^*C^*|\dots|A^{(n-1)*}C^*] = 0$ which contradicts the assumed observability of (A, C) . Since these arguments apply to any solution, every solution P to (2.5) is nonsingular. \square

2.2. Problem description. The aim of our work is to analyse the cancellation phenomena which occur in the general class of \mathcal{H}^∞ design problems characterised by the assumptions that $P_{21}(s)$ is square while $P_{12}(s)$ has more rows than columns. We also assume that D_{21} is nonsingular, D_{12} has full column rank and that $\text{Re}(\lambda(A - B_1D_{21}^{-1}C_2)) \neq 0$ and $\text{Re}(\lambda(A - B_2(D_{12}^*D_{12})^{-1}D_{12}C_1)) \neq 0$. If we consider $\mathcal{R}^T(s)$, it is easy to see that an equivalent characterisation is given by the assumptions that $P_{12}(s)$ is square while $P_{21}(s)$ has more columns than rows. Any problem fitting either of these alternative descriptions will be called a problem of the second kind. Our analysis will show that certain cancellations are a direct consequence of \mathcal{H}^∞ optimality.

The weighted sensitivity problem [5], [9], [10], [12], [13], [25], [30], [31] is given by

$$(2.7) \quad \inf_{K \in \Xi} \|W_2(I + GK)^{-1}W_1\|_\infty = \inf_{K \in \Xi} \|\mathcal{R}_s(s)\|_\infty$$

where Ξ is the set of stabilizing compensators. By choosing

$$(2.8) \quad P_s(s) = \begin{bmatrix} W_2W_1 & W_2G \\ W_1 & G \end{bmatrix}(s)$$

we may embed this problem in the generalized regulator configuration in Fig. 1. Since

$$(2.9) \quad \mathcal{R}(s) = F_l(P, -K) = P_{11} - P_{12}K(I + P_{22}K)^{-1}P_{21}$$

direct comparison with (2.9) reveals that

$$(2.10) \quad \mathcal{R}_s(s) = F_l(P_s(s), -K(s)).$$

In the case that $G(s)$ is square, this problem falls into the class of problems of the first kind which have been analysed elsewhere [19], [20]. If, on the other hand, $G(s)$ has more outputs than inputs and $W_2(s)$ is square, $P_{12}(s) = W_2G(s)$ will have more rows than columns giving rise to a problem of the second kind.

Another problem which has received attention is the mixed sensitivity problem [8], [9], [10], [13], [27]. In this case we seek

$$(2.11) \quad \inf_{K \in \Xi} \left\| \begin{bmatrix} W_2 G K (I + G K)^{-1} W_1 \\ W_3 (I + G K)^{-1} W_1 \end{bmatrix} \right\|_{\infty} = \inf_{K \in \Xi} \|\mathcal{R}_{ms}(s)\|_{\infty}.$$

Setting

$$(2.12) \quad P_{ms}(s) = \left[\begin{array}{c|c} 0 & W_2 G \\ \hline W_3 W_1 & W_3 G \\ \hline W_1 & G \end{array} \right](s)$$

we have that

$$(2.13) \quad \mathcal{R}_{ms}(s) = F_l(P_{ms}(s), -K(s))$$

which is also an example of a problem of the second kind.

Several other problems of the second kind may be found in the literature. See for example [9], [13] and the numerous references therein. Rather than analyse these problems individually, we have chosen to identify the common characteristics shared by all \mathcal{H}^{∞} control problems of the second kind.

2.3. Review of \mathcal{H}^{∞} optimisation theory. In this section we will briefly mention the \mathcal{H}^{∞} theory which is required in the later analysis. In the next subsection we summarize the Youla parametrisation [6], [29] which is used to characterise the class of all stabilising compensators and the corresponding closed-loop transfer functions $\mathcal{R}(s)$ in Fig. 1. Following that, the closed loop transfer functions which have minimum \mathcal{L}^{∞} norm are identified.

2.3.1. Parametrization of all stabilising controllers. Let $P(s)$ in Fig. 1 be given by

$$(2.14) \quad P(s) = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}(s) \stackrel{s}{=} \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ \hline C_2 & I & D_{22} \end{array} \right]$$

and suppose that (A, B_2, C_2) is stabilisable and detectable. We remind the reader that there is no loss of generality in assuming that $D_{21} = I$ and that D_{12} is part of an orthogonal matrix [20], [26]. It is always possible to achieve this by a constant rescaling of the problem. Under these assumptions there exist unique stabilising positive semi-definite solutions to the algebraic Riccati equations [7], [26]

$$(2.15) \quad \mathfrak{X}(A - B_2 D_{12}^* C_1) + (A - B_2 D_{12}^* C_1)^* \mathfrak{X} - \mathfrak{X} B_2 B_2^* \mathfrak{X} + C_1^* D_{\perp} D_{\perp}^* C_1 = 0$$

and

$$(2.16) \quad \mathfrak{Y}(A - B_1 C_2)^* + (A - B_1 C_2) \mathfrak{Y} - \mathfrak{Y} C_2^* C_2 \mathfrak{Y} = 0.$$

D_{\perp} has been chosen to make the augmented matrix $[D_{12} | D_{\perp}]$ orthogonal [7], [26]. Associated with these two Riccati equations we have the stabilizing matrices F and H given by [7], [20], [26],

$$(2.17) \quad F = D_{12}^* C_1 + B_2^* \mathfrak{X}$$

and

$$(2.18) \quad H = B_1 + \mathfrak{Y} C_2^*.$$

The Youla parametrization theory allows us to write [6], [7], [23], [26], [29]

$$(2.19) \quad \mathcal{R}(s) = [T_{11} - T_{12}XT_{21}](s)$$

and

$$(2.20) \quad K(s) = F_i(K_0(s), X(s))$$

in which

$$(2.21) \quad K_0(s) = \left[\begin{array}{c|cc} A - B_2F - HC_2 + HD_{22}F & -H & B_2 - HD_{22} \\ \hline -F & 0 & I \\ C_2 - D_{22}F & I & D_{22} \end{array} \right]$$

and

$$(2.22) \quad \left[\begin{array}{ccc} T_{11} & T_{12} & T_{\perp} \\ T_{21} & 0 & 0 \end{array} \right] \stackrel{s}{=} \left[\begin{array}{cc|ccc} A - B_2F & B_2F & B_1 & B_2 & -\mathfrak{X}^* C_1^* D_{\perp} \\ \hline 0 & A - HC_2 & -\mathfrak{Y} C_2^* & 0 & 0 \\ C_1 - D_{12}F & D_{12}F & D_{11} & D_{12} & D_{\perp} \\ \hline 0 & C_2 & I & 0 & 0 \end{array} \right].$$

\mathfrak{X}^* is the Moore–Penrose generalized inverse of \mathfrak{X} . With the particular choice of the matrices F and H given in (2.17) and (2.18), T_{21} and $[T_{12} | T_{\perp}]$ are inner [7], [26]. We call $T_{\perp}(s)$ an inner or all-pass extension of $T_{12}(s)$.

2.3.2. The γ -iteration. In this section we will outline an algorithm [7], [8], [9], [10], [13], [26], [27] which reduces the problem of finding an upper bound for

$$(2.23) \quad \inf_{X(s) \in \mathcal{RH}^\infty} \|(T_{11} - T_{12}XT)(s)\|_\infty = \gamma_{\text{opt}}$$

to the Nehari problem [22] of identifying those matrices in \mathcal{RH}_+^∞ which are closest (in the sense of the \mathcal{L}^∞ -norm to a given point (matrix) in \mathcal{RL}^∞ . The approach will be to generate a sequence $\gamma_i = \|R_i(s)\|_\infty = \|T_{11} - T_{12}X_iT_{21}\|_\infty$ which converges on γ_{opt} (from above). In applications, the calculation of this sequence is terminated when an upper bound α for γ_{opt} has been found such that $(\alpha - \gamma_{\text{opt}}) < \varepsilon$ for some sufficiently small $\varepsilon > 0$.

It may be seen from (2.19) that

$$(2.24) \quad \mathcal{R}(s) = \left(T_{11} - [T_{12} | T_{\perp}] \begin{bmatrix} X \\ 0 \end{bmatrix} T_{21} \right)(s).$$

Since $[T_{12} | T_{\perp}](s)$ and $T_{21}(s)$ are inner and thus norm-preserving, we can write

$$(2.25) \quad \begin{aligned} \|\mathcal{R}(s)\|_\infty &= \left\| \begin{bmatrix} (T_{12}^* T_{11} T_{21}^* - X(s)) \\ T_{\perp}^* T_{11} T_{21}^*(s) \end{bmatrix} \right\|_\infty \\ &= \left\| \begin{bmatrix} (R_1 - X)(s) \\ R_2(s) \end{bmatrix} \right\|_\infty \end{aligned}$$

and a routine direct calculation from (2.22) shows that

$$(2.26) \quad \begin{aligned} R(s) &= \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}^{(s)} \\ &\stackrel{s}{=} \left[\begin{array}{cc|c} -(A - B_2F)^* & -\{(C_1 - D_{12}F)^* D_{11} + \mathfrak{X} B_1\} C_2 \mathfrak{Y} & (C_1 - D_{12}F)^* D_{11} + \mathfrak{X} B_1 \\ \hline 0 & -(A - HC_2)^* & -C_2^* \\ -B_2^* & F \mathfrak{Y} - D_{12}^* D_{11} C_2 \mathfrak{Y} & D_{12}^* D_{11} \\ \hline D_{\perp}^* C_1 \mathfrak{X}^* & -D_{\perp}^* D_{11} C_2 \mathfrak{Y} & D_{\perp}^* D_{11} \end{array} \right] \end{aligned}$$

$\Rightarrow R(s)$ is completely unstable.

We also observe that if

$$(2.27) \quad \|\mathcal{R}(s)\|_{\infty} \leq \gamma$$

then

$$(2.28) \quad (2.25) \Rightarrow (R_1 - X)^*(s)(R_1 - X)(s) \leq \gamma^2 I - R_2^*(s)R_2(s) = M^*(s)M(s)$$

where $M(s)$ is a stable and minimum-phase spectral factor of $\gamma^2 I - R_2^*(s)R_2(s)$ (the existence of which requires $\gamma \geq \|R_2(s)\|_{\infty}$). Continuing, we deduce from (2.28) that

$$(2.29a) \quad \|R_1 M^{-1}(s) - X M^{-1}(s)\|_{\infty} = \mu \leq 1.$$

Decomposing $R_1 M^{-1}(s)$ into stable and unstable parts gives

$$(2.29b) \quad \|[R_1 M^{-1}(s)]_- - \tilde{X}(s)\|_{\infty} = \mu \leq 1$$

in which

$$(2.30) \quad \tilde{X}(s) = [R_1 M^{-1}(s)]_+ + X M^{-1}(s).$$

Rearranging leads to

$$(2.31) \quad X(s) = (\tilde{X}(s) - [R_1 M^{-1}(s)]_+) M(s).$$

We observe that (2.29) is a Nehari or optimal approximation problem [14], [22].

The specific algorithm for the γ -iteration is essentially a binary search procedure which is described elsewhere [7], [9], [10], [13], [26], [27].

2.3.3. Characterization of all solutions to the Nehari approximation problem. The purpose of the γ -iteration, which was mentioned in the last subsection, was to reduce the problem in (2.23) to a Nehari problem. That is, the problem of finding all $X(s)$ s which achieve

$$(2.32) \quad \inf_{X \in \mathcal{RH}_+^{\infty}} \|H(s) - X^*(s)\|_{\infty}$$

or else which satisfy

$$(2.33) \quad \|H(s) - X^*(s)\|_{\infty} \leq \rho > \|H(s)\|_H.$$

Glover [14] has shown that all the solutions to these problems may be characterised in terms of a balanced realisation of $H(s)$. In [14], this characterisation is in terms of a linear fractional transformation which contains a free matrix contraction. In [20], we give a different version of these results which characterize all the solutions in terms of a bounded real type condition. We will use this characterisation to establish the main theorem in § 4. We refer the reader to Theorem 2.1 and Corollary 2.2, together with Remarks 2.1–2.4 in [20].

3. Balancing the Riccati equations. Our subsequent analysis is greatly simplified if the Riccati equations (2.15) and (2.16) have a diagonal balanced structure. As we have already established [20], this procedure allows us to remove the right half-plane zeros of the square off-diagonal block of $P(s)$ in the early stages of the analysis. A similar dimension deflation corresponding to the nonsquare off-diagonal block is also possible. In addition, the balanced Riccati equations lead directly to a minimal realisation for

$$\begin{bmatrix} T_{11} & T_{12} & T_{\perp} \\ T_{21} & 0 & 0 \end{bmatrix}(s)$$

(review (2.22)).

We begin by considering any basis change T in the state-space of $P(s)$ in (2.14). In the new basis, the realization of $P(s)$ becomes

$$(3.1) \quad P(s) = \begin{bmatrix} TAT^{-1} & TB_1 & TB_2 \\ C_1T^{-1} & D_{11} & D_{12} \\ C_2T^{-1} & D_{21} & D_{22} \end{bmatrix}$$

and the algebraic Riccati equation (2.15) becomes

$$(3.2a) \quad \begin{aligned} & \mathcal{X}(TAT^{-1} - TB_2D_{12}^*C_1T^{-1}) + (TAT^{-1} - TB_2D_{12}^*C_1T^{-1})^*\mathcal{X} \\ & - \mathcal{X}TB_2B_2^*T^*\mathcal{X} + T^{-*}C_1^*D_{11}^*C_1T^{-1} = 0 \end{aligned}$$

where T^{-*} denotes $(T^{-1})^*$. After pre-multiplication by T^* and post-multiplication by T we get

$$(3.2b) \quad \begin{aligned} & T^*\mathcal{X}T(A - B_2D_{12}^*C_1) + (A - B_2D_{12}^*C_1)^*T^*\mathcal{X}T \\ & - T^*\mathcal{X}TB_2B_2^*T^*\mathcal{X}T + C_1^*D_{11}^*C_1 = 0 \end{aligned}$$

which shows that \mathcal{X} undergoes the congruence transformation

$$(3.3) \quad \mathcal{X} \rightarrow T^{-*}\mathcal{X}T^{-1}.$$

Similarly, in the new basis, (2.16) becomes [20]:

$$(3.4) \quad T^{-1}\mathcal{Y}T^{-*}(A - B_1C_2)^* + (A - B_1C_2)T^{-1}\mathcal{Y}T^{-*} - T^{-1}\mathcal{Y}T^{-*}C_2^*C_2T^{-1}\mathcal{Y}T^{-*} = 0$$

and hence

$$(3.5) \quad \mathcal{Y} \rightarrow T\mathcal{Y}T^*.$$

Formulae (3.3) and (3.5), together with $\mathcal{X} = \mathcal{X}^* \geq 0$ and $\mathcal{Y} = \mathcal{Y}^* \geq 0$, show that the construction in [14, App. B] may be used to select a T so that

$$(3.6) \quad \mathcal{X} = \begin{bmatrix} \tilde{\Sigma}_1 & & & \\ & \tilde{\Sigma}_2 & & \\ & & 0 & \\ & & & 0 \end{bmatrix} = \begin{bmatrix} \Sigma_1 & \\ & 0 \end{bmatrix}$$

and

$$(3.7) \quad \mathcal{Y} = \begin{bmatrix} \tilde{\Sigma}_1 & & & \\ & 0 & & \\ & & \tilde{\Sigma}_3 & \\ & & & 0 \end{bmatrix}.$$

It is convenient to introduce a permutation matrix J so that

$$(3.8) \quad J\mathcal{Y}J^* = \begin{bmatrix} \tilde{\Sigma}_1 & & & \\ & \tilde{\Sigma}_3 & & \\ & & 0 & \\ & & & 0 \end{bmatrix} = \begin{bmatrix} \Sigma_2 & \\ & 0 \end{bmatrix}.$$

Clearly, $\Sigma_1 > 0$ and $\Sigma_2 > 0$.

In the rest of the analysis we will assume that the realization in (2.14) has been put into a basis corresponding to balanced Riccati equations. We continue by defining

$$(3.9) \quad M := J(A - B_1C_2)J^*,$$

$$(3.10) \quad Z := A - B_2D_{12}^*C_1$$

and introduce the partitioning

$$(3.11) \quad C_1 = [C_{11} | C_{12}],$$

$$(3.12) \quad [B_1 | B_2] = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

which is consistent with that in (3.6). We will also make use of the partitioning

$$(3.13) \quad C_2 J^* = [\hat{C}_{21} | \hat{C}_{22}]$$

and

$$(3.14) \quad FJ^* = [\hat{F}_1 | \hat{F}_2]$$

which is consistent with that in (3.8). Allowing (3.6) to induce a partitioning on Z and substituting into (2.15) gives

$$(3.15) \quad \begin{aligned} & \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} + \begin{bmatrix} Z_{11}^* & Z_{21}^* \\ Z_{12}^* & Z_{22}^* \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \\ & - \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} B_{12} \\ B_{22} \end{bmatrix} [B_{12}^* & B_{22}^*] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \\ & + \begin{bmatrix} C_{11}^* \\ C_{12}^* \end{bmatrix} D_\perp D_\perp^* [C_{11} C_{12}] = 0. \end{aligned}$$

From the (2, 2) block of (3.15) we get

$$(3.16) \quad C_{12}^* D_\perp D_\perp^* C_{12} = 0 \Rightarrow D_\perp^* C_{12} = 0 \Rightarrow C_{11}^* D_\perp D_\perp^* C_{12} = 0.$$

Consequently, from the (1, 2) block of (3.15) we obtain

$$(3.17) \quad \Sigma_1 Z_{12} = 0 \Rightarrow Z_{12} = 0 \quad (\text{since } \Sigma_1 > 0).$$

Finally, the (1, 1) block gives

$$(3.18) \quad \Sigma_1 Z_{11} + Z_{11}^* \Sigma_1 - \Sigma_1 B_{12} B_{12}^* \Sigma_1 + C_{11}^* D_\perp D_\perp^* C_{11} = 0$$

which is a deflated Riccati equation with a positive definite solution Σ_1 . Since

$$(3.19) \quad A - B_2 F = Z - B_2 B_2^* \mathfrak{X}$$

we get from (3.17) that

$$(3.20) \quad \begin{aligned} A - B_2 F &= \begin{bmatrix} Z_{11} & 0 \\ Z_{21} & Z_{22} \end{bmatrix} - \begin{bmatrix} B_{12} \\ B_{22} \end{bmatrix} [B_{12}^* & B_{22}^*] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} Z_{11} - B_{12} B_{12}^* \Sigma_1 & 0 \\ Z_{21} - B_{22} B_{12}^* \Sigma_1 & Z_{22} \end{bmatrix}. \end{aligned}$$

Since $(A - B_2 F)$ is asymptotically stable, Z_{22} is also. We conclude also that each eigenvalue of Z_{22} corresponds to a stable mode of $A - B_2 D_{12}^* C_1$ which is undetectable through $D_\perp^* C_1$.

A similar procedure has already been applied to (2.16) [20] to obtain

$$(3.21) \quad J(A - HC_2)J^* = \begin{bmatrix} -\Sigma_2 M_{11}^* \Sigma_2^{-1} & M_{12} - \Sigma_2 \hat{C}_{21}^* \hat{C}_{22} \\ 0 & M_{22} \end{bmatrix}$$

in which $\{\lambda(M_{11})\} = \{\text{right half-plane zeros of } P_{21}\}$.

Making use of (3.12), (3.14), (3.16), (3.20), (3.21) and (2.15), we can rewrite (2.22) as

$$(3.22) \quad \begin{bmatrix} T_{11} & T_{12} & T_{\perp} \\ T_{21} & 0 & 0 \end{bmatrix} \stackrel{s}{=} \left[\begin{array}{c|ccc} Z_{11} - B_{12}B_{12}^*\Sigma_1 & B_{12}\hat{F}_1 & B_{11} & B_{12} & -\Sigma_1^{-1}C_{11}^*D_{\perp} \\ 0 & -\Sigma_2 M_{11}^*\Sigma_2^{-1} & -\Sigma_2 \hat{C}_{21}^* & 0 & 0 \\ \hline D_{\perp}D_{\perp}^*C_{11} - D_{12}B_{12}^*\Sigma_1 & D_{12}\hat{F}_1 & D_{11} & D_{12} & D_{\perp} \\ 0 & \hat{C}_{21} & I & 0 & 0 \end{array} \right].$$

We also note that the change of basis

$$(3.23) \quad T = \begin{bmatrix} \Sigma_1^{1/2} & 0 \\ 0 & \Sigma_2^{-1/2} \end{bmatrix}$$

will balance the realisation in (3.22). From the balanced version of (3.22) we obtain

$$(3.24) \quad T_{21}(s) \stackrel{s}{=} \begin{bmatrix} -\Sigma_2^{1/2}M_{11}\Sigma_2^{-1/2} & -\Sigma_2^{1/2}\hat{C}_{21}^* \\ \hat{C}_{21}\Sigma_2^{-1/2} & I \end{bmatrix}$$

which has the identity as both its controllability and observability gramians. We conclude, therefore, that (3.24) is both a minimal and a balanced realisation of $T_{21}(s)$. We also deduce that

$$(3.25) \quad [T_{12} | T_{\perp}] \stackrel{s}{=} \left[\begin{array}{c|c} \Sigma_1^{1/2}(Z_{11} - B_{12}B_{12}^*\Sigma_1)\Sigma_1^{-1/2} & \Sigma_1^{1/2}B_{12} \quad -\Sigma_1^{-1/2}C_{11}^*D_{\perp} \\ \hline D_{\perp}D_{\perp}^*C_{11}\Sigma_1^{-1/2} - D_{12}B_{12}^*\Sigma_1^{1/2} & D_{12} \quad D_{\perp} \end{array} \right]$$

is balanced with controllability and observability gramians the identity; this realisation is thus also minimal.

We conclude this section by pointing out that the realisation in (3.22) is also minimal. This may be established by proving that all the system zeros in (3.22) lie in the open right half-plane and consequently cannot cancel any of the poles of this realisation which lie in the left half-plane. An almost identical argument may be found in § 3 in [20].

As one would expect, replacing the realisation (2.22) with (3.22) allows the realisation in (2.26) to be reduced to

$$(3.26) \quad R(s) \stackrel{s}{=} \left[\begin{array}{c|c} -(Z_{11} - B_{12}B_{12}^*\Sigma_1)^* & A_R(1, 2) \\ 0 & M_{11} \\ \hline -B_{12}^* & D_{12}^*D_{11}\hat{C}_{21} - \hat{F}_1 \\ D_{\perp}^*C_{11}\Sigma_1^{-1} & D_{\perp}^*D_{11}\hat{C}_{21} \end{array} \middle| \begin{array}{c} B_R(1, 1) \\ \Sigma_2\hat{C}_{21}^* \\ \hline D_{12}^*D_{11} \\ D_{\perp}^*D_{11} \end{array} \right]$$

where

$$A_R(1, 2) = C_{11}^*D_{\perp}D_{\perp}^*D_{11}\hat{C}_{21} + \Sigma_1 B_{11}\hat{C}_{21} - \Sigma_1 B_{12}D_{12}^*D_{11}\hat{C}_{21}$$

and

$$B_R(1, 1) = C_{11}^*D_{\perp}D_{\perp}^*D_{11} - \Sigma_1 B_{12}D_{12}^*D_{11} + \Sigma_1 B_1.$$

The realisation in (3.26) need not be minimal. The results of the analysis of this section are summarised in the following lemma.

LEMMA 3.1.

- (i) (a) *The number of zeros of $P_{21}(s)$ in $\mathbb{C}_+ = \text{rank } (\mathcal{Y})$*
 (b) *The number of stable modes of $A - B_2 D_{12}^* C_1$ which are undetectable through $D_{12}^* C_1 = \text{rank defect } (\mathcal{X})$*
- (ii) *The realization (3.22) is minimal with McMillan degree = $\text{rank } (\mathcal{X}) + \text{rank } (\mathcal{Y})$*
- (iii) *The realization (3.24) is minimal and $\deg(T_{21}) = \text{rank } (\mathcal{Y})$*
- (iv) *The realization (3.25) is minimal and $\deg([T_{12} | T_{\perp}]) = \text{rank } (\mathcal{X})$*
- (v) $\deg(R) \leq \text{rank } (\mathcal{X}) + \text{rank } (\mathcal{Y})$ (see 3.26). \square

4. Main results. In this section we combine the results we have already obtained with a new theorem to obtain the controller degree bound for all problems of the second kind. We will be treating both the optimal and suboptimal cases.

Suppose $n = \deg(P)$, $t = \deg(\mathcal{R})$ and let $c = (\text{number of cancellations which occur between } P(s) \text{ and } K(s) \text{ as a result of closing the feedback loop in Fig. 1})$. Then

$$(4.1) \quad t = n + \deg(K) - c,$$

that is,

$$(4.2) \quad \deg(K) \leq t_b + c_b - n$$

where t_b and c_b are upper bounds on t and c , respectively. We will derive the upper bound t_b in § 4.1 while c_b will be found in § 4.2. These results will be combined in § 4.3 to give our main theorem.

4.1. An upper bound for the McMillan degree of all closed loop systems of the second kind. Our derivation of the bound t_b for the degree of the closed loop requires several steps. Before stating and proving the main theorem of this section we will briefly sketch the route we intend to take:

(a) The reader will recall from Lemma 3.1(ii) that

$$(4.3) \quad \deg \begin{bmatrix} T_{11} & T_{12} & T_{\perp} \\ T_{21} & 0 & 0 \end{bmatrix} = \text{rank } (\mathcal{X}) + \text{rank } (\mathcal{Y}).$$

(b) If T_{11} and T_{21} are all-pass right coprime, and T_{11} and $[T_{12} | T_{\perp}]$ are all-pass left coprime, we have shown that [20, Thm. 4.1]

$$(4.4) \quad \deg \begin{bmatrix} R_1 \\ R_2 \end{bmatrix} = \deg \begin{bmatrix} T_{12}^* T_{11} T_{21}^* \\ T_{\perp}^* T_{11} T_{21}^* \end{bmatrix} = \text{rank } (\mathcal{X}) + \text{rank } (\mathcal{Y}).$$

T_{11} and T_{21} will be called all-pass right coprime if in

$$(4.5) \quad \begin{bmatrix} T_{11} \\ T_{21} \end{bmatrix}(s) = \begin{bmatrix} \tilde{T}_{11} \\ \tilde{T}_{21} \end{bmatrix}(s) A(s)$$

all all-pass common right divisors $A(s)$ of $T_{11}(s)$ and $T_{12}(s)$ are constant orthogonal matrices. All-pass left coprimeness is defined in a similar way.

(c) We will assume throughout that the all-pass coprimeness condition is satisfied. If this is not the case, there is always a factorisation

$$(4.6) \quad \begin{bmatrix} T_{11} & T_{12} & T_{\perp} \\ T_{21} & 0 & 0 \end{bmatrix}(s) = \begin{bmatrix} A_l(s) & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \tilde{T}_{11} & \tilde{T}_{12} & \tilde{T}_{\perp} \\ \tilde{T}_{21} & 0 & 0 \end{bmatrix} \begin{bmatrix} A_r(s) & 0 \\ 0 & I \end{bmatrix}$$

in which \tilde{T}_{11} and \tilde{T}_{21} are all-pass right coprime, and \tilde{T}_{11} and $[\tilde{T}_{12} | \tilde{T}_{\perp}]$ are all-pass left coprime [20, Thm. 4.1]. For reasons which are almost identical to those given in [20], the existence or nonexistence of these all-pass common factors makes no difference to the bound t_b that we seek.

(d) Next, it will be proved that

$$(4.7) \quad \deg ([R_1 M^{-1}]_-) \leq \text{rank } (\mathcal{X}) + \text{rank } (\mathcal{Y}).$$

We remind the reader that the definition of $M(s)$ may be found in (2.28). In the hypothesis of Theorem 4.1 we will assume that (4.7) is met with equality since this assumption simplifies one of our later calculations.

(e) Section 4.1(d), together with the work of Glover [14], ensures that

$$(4.8a) \quad \deg (\tilde{X}) \leq \text{rank } (\mathcal{X}) + \text{rank } (\mathcal{Y}) + \deg (U) - 1$$

in the optimal case, and

$$(4.8b) \quad \deg (\tilde{X}) \leq \text{rank } (\mathcal{X}) + \text{rank } (\mathcal{Y}) + \deg (U)$$

in the suboptimal case. In the above $U(s) \in \mathcal{RH}^\infty$ is a free matrix contraction to be chosen by the designer and $\tilde{X}(s)$ is defined in (2.29b).

(f) We will prove that

$$(4.9) \quad \deg (X) \leq \deg (\tilde{X})$$

and that

$$(4.10) \quad \deg (\mathcal{R}) \leq \deg (\tilde{X}).$$

$X(s)$ and $\tilde{X}(s)$ are related in (2.31). Consequently,

$$(4.11a) \quad t_b = \text{rank } (\mathcal{X}) + \text{rank } (\mathcal{Y}) + \deg (U) - 1$$

in the optimal case, or

$$(4.11b) \quad t_b = \text{rank } (\mathcal{X}) + \text{rank } (\mathcal{Y}) + \deg (U)$$

in the suboptimal case. In the first instance the reader may wish to skip to §§ 4.2–4.4. In this way we may get an initial overview without getting swamped in the details surrounding the proofs of claims (d)–(f) above; these details are considerable.

The general form of the state space model for

$$\begin{bmatrix} T_{11} & T_{12} & T_{1\perp} \\ T_{21} & 0 & 0 \end{bmatrix}(s)$$

(in (3.22)) is the basis of the hypothesis for our next theorem.

THEOREM 4.1.

Let

$$(4.12) \quad \begin{bmatrix} T_{11} & T_{12} & T_{1\perp} \\ T_{21} & 0 & 0 \end{bmatrix}(s) \stackrel{s}{=} \left[\begin{array}{ccc|cc} A_{11} & A_{12} & B_{11} & B_{12} & B_{13} \\ 0 & A_{22} & B_{21} & 0 & 0 \\ \hline C_{11} & C_{12} & D_{11} & D_{12} & D_{13} \\ 0 & C_{22} & I & 0 & 0 \end{array} \right]$$

be asymptotically stable and suppose also that

$$(i) \quad T_{21}(s) \stackrel{s}{=} \begin{bmatrix} A_{22} & B_{21} \\ C_{22} & I \end{bmatrix}$$

is all-pass, minimal and balanced;

$$(ii) \quad [T_{12} | T_{1\perp}](s) \stackrel{s}{=} \left[\begin{array}{c|cc} A_{11} & B_{12} & B_{13} \\ \hline C_{11} & D_{12} & D_{13} \end{array} \right]$$

is all-pass, minimal and balanced;

(iii) $T_{11}(s)$ and $T_{21}(s)$ are all-pass right coprime;

(iv) $T_{11}(s)$ and $[T_{12} | T_{12}^\perp]$ are all-pass left coprime;

$$(4.13) \quad (v) \quad D_{13}^* C_{12} = 0;$$

$$(4.14) \quad (vi) \quad A_{12} + C_{11}^* C_{12} = 0.$$

Then

$$(4.15) \quad (a) \quad \left[\begin{array}{c} R_1 \\ R_2 \end{array} \right] \stackrel{s}{=} \left[\begin{array}{cc|c} -A_{11}^* & -C_{11}^* D_{11} B_{21}^* - B_{11} B_{21}^* & C_{11}^* D_{11} + B_{11} \\ 0 & -A_{22}^* & C_{22}^* \\ \hline -B_{12}^* & -D_{12}^* (D_{11} B_{21}^* + C_{12}) & D_{12}^* D_{11} \\ -B_{13}^* & -D_{13}^* D_{11} B_{21}^* & D_{13}^* D_{11} \end{array} \right]$$

$$\stackrel{s}{=} \left[\begin{array}{c|c} -\bar{A}^* & \bar{C}^* \\ \hline -\bar{B}_1^* & \bar{D}_1^* \\ -\bar{B}_2^* & \bar{D}_2^* \end{array} \right]$$

is a minimal balanced realisation.

$$(4.16) \quad (b) \quad (R_1 M^{-1})_- \stackrel{s}{=} \left[\begin{array}{cc} -\bar{A}^* & F^* E^{1/2} \\ -\bar{B}_1^* & 0 \end{array} \right]$$

where the matrices E and F are defined in the proof; see (4.41) and (4.45), respectively.

$$(c) \quad \text{If } (R_1 M^{-1})_- \stackrel{s}{=} \left[\begin{array}{cc} -\bar{A}^* & F^* E^{1/2} \\ -\bar{B}_1^* & 0 \end{array} \right]$$

is minimal and

$$\tilde{X}(s) := \left[\begin{array}{cc} \hat{A} & \hat{B} \\ \hat{C} & \hat{D} \end{array} \right]$$

is chosen so that

$$(4.17) \quad [(R_1 M^{-1})_- + \tilde{X}](s) \stackrel{s}{=} \left[\begin{array}{cc|c} -\bar{A}^* & 0 & F^* E^{1/2} \\ 0 & \hat{A} & \hat{B} \\ \hline -\bar{B}_1^* & \hat{C} & \hat{D} \end{array} \right] = \left[\begin{array}{cc} \tilde{A} & \tilde{B} \\ \tilde{C} & \tilde{D} \end{array} \right]$$

satisfies the bounded real-type equations

$$(4.18) \quad \left[\begin{array}{cc} -(\tilde{A}P + P\tilde{A}^* + \tilde{B}\tilde{B}^*) & -(\tilde{B}\tilde{D}^* + P\tilde{C}^*) \\ -(\tilde{D}\tilde{B}^* + \tilde{C}P) & I - \tilde{D}\tilde{D}^* \end{array} \right] = \left[\begin{array}{c} L \\ W \end{array} \right] [L^* \quad W^*]$$

and their duals

$$(4.19) \quad \left[\begin{array}{cc} -(\tilde{A}^*Q + Q\tilde{A} + \tilde{C}^*\tilde{C}) & -(\tilde{C}^*\tilde{D} + Q\tilde{B}) \\ -(\tilde{D}^*\tilde{C} + \tilde{B}^*Q) & I - \tilde{D}^*\tilde{D} \end{array} \right] = \left[\begin{array}{c} L_d^* \\ W_d^* \end{array} \right] [L_d \quad W_d]$$

in which

$$(4.20) \quad (i) \quad QP = I$$

and

$$(4.21) \quad (ii) \quad L^* = [0 \quad L_{21}^*]; \quad L_d = [0 \quad L_{21d}]$$

where the partitioning in (4.21) is conformal with that in (4.17).

Then

$$(4.22) \quad (1) \quad \{\text{poles of } X\} \subseteq \{\text{poles of } \tilde{X}\},$$

$$(4.23) \quad (2) \quad \{\text{poles of } T_{11} + T_{12}XT_{21}\} \subseteq \{\text{poles of } \tilde{X}\}.$$

Explicit formulae for $X(s)$ and $[T_{11} + T_{12}XT_{21}](s)$ are given in (4.59) and (4.71) below.

Remark 4.1. The validity of the bounds in (4.11) is proven by applying Theorem 4.1 to a balanced version of (3.22). In this regard we should note the following:

(a) The general form of the realisations in (3.22) and (4.12) is the same.

(b) T_{21} and $[T_{12}|T_\perp]$ are both inner and their realizations are minimal and balanced.

(c) Equations (4.13) and (4.14) are easily seen to be satisfied after balancing (3.22). This will also be true should it be necessary to extract all-pass common factors.

(d) Theorem 2.1 and Cor. 2.2 in [20] ensure that the error systems corresponding to any Nehari or suboptimal extension of $[R_1M^{-1}]_-$ will satisfy the bounded real-type equations (4.18) and (4.19).

(e) Theorem 4.1 and Lemma 3.1(ii) thus establish the validity of (4.7)–(4.11).

Remark 4.2. In the case that the final value of $\gamma > \gamma_{\text{opt}}$, the cancellation phenomena predicted by Theorem 4.1 can only be guaranteed if $\tilde{X}(s)$ is a suboptimal extension of $(R_1M^{-1})_-(s)$ corresponding to an error system with an infinity norm of one.

Proof. In the interests of clarity, we have relegated long calculations and the treatment of certain technical details to a sequence of appendices. The appendices and the main body of the proof will share common notation.

The assumed properties of the realisation of $T_{21}(s)$ enforces

$$(4.24) \quad A_{22} + A_{22}^* + B_{21}B_{21}^* = 0$$

and

$$(4.25) \quad C_{22} = -B_{21}^*.$$

Similarly, the realisation of $[T_{12}|T_{12}^\perp]$ must satisfy

$$(4.26) \quad A_{11} + A_{11}^* + B_{12}B_{12}^* + B_{13}B_{13}^* = 0,$$

$$(4.27) \quad A_{11} + A_{11}^* + C_{11}C_{11}^* = 0,$$

$$(4.28) \quad D_{12}^*C_{11} + B_{12}^* = 0,$$

$$(4.29) \quad D_{13}^*C_{11} + B_{13}^* = 0,$$

$$(4.30) \quad D_{12}B_{12}^* + D_{13}B_{13}^* + C_{11} = 0,$$

$$(4.31) \quad [D_{12}|D_{13}][D_{12}|D_{13}]^* = I.$$

The reader may wish to consult Glover [14, Thm. 5.1] for a state-space characterisation of all-pass matrices.

From (4.12) and (4.25) we get

$$\begin{aligned} T_{11}T_{21}^* &\stackrel{s}{=} \left[\begin{array}{cc|c} A_{11} & A_{12} & B_{11} \\ 0 & A_{22} & B_{21} \\ \hline C_{11} & C_{12} & D_{11} \end{array} \right] * \left[\begin{array}{cc} -A_{22}^* & B_{21} \\ B_{21}^* & I \end{array} \right] \\ &\stackrel{s}{=} \left[\begin{array}{ccc|c} A_{11} & A_{12} & B_{12}B_{21}^* & B_{11} \\ 0 & A_{22} & B_{21}B_{21}^* & B_{21} \\ 0 & 0 & -A_{22}^* & B_{21} \\ \hline C_{11} & C_{12} & D_{11}B_{21}^* & D_{11} \end{array} \right]. \end{aligned}$$

Introducing the change of basis

$$T = \begin{bmatrix} I & 0 & 0 \\ 0 & I & -I \\ 0 & 0 & I \end{bmatrix}$$

and making use of (4.24) gives

$$(4.32) \quad \begin{aligned} T_{11} T_{21}^* &\stackrel{s}{=} \left[\begin{array}{ccc|c} A_{11} & A_{12} & A_{12} + B_{11} B_{21}^* & B_{11} \\ 0 & A_{22} & 0 & 0 \\ 0 & 0 & -A_{22}^* & B_{21} \\ \hline C_{11} & C_{12} & D_{11} B_{21}^* + C_{12} & D_{11} \end{array} \right] \\ &\stackrel{s}{=} \left[\begin{array}{cc|c} A_{11} & A_{12} + B_{11} B_{21}^* & B_{11} \\ 0 & -A_{22}^* & B_{21} \\ \hline C_{11} & D_{11} B_{21}^* + C_{12} & D_{11} \end{array} \right]. \end{aligned}$$

Next,

$$\begin{aligned} \begin{bmatrix} R_1 \\ R_2 \end{bmatrix} (s) &= \begin{bmatrix} T_{12}^* \\ T_{12}^* \end{bmatrix} T_{11} T_{21}^* (s) \stackrel{s}{=} \left[\begin{array}{c|c} -A_{11}^* & C_{11}^* \\ -B_{12}^* & D_{12}^* \\ -B_{13}^* & D_{13}^* \end{array} \right] * \left[\begin{array}{cc|c} A_{11} & A_{12} + B_{11} B_{21}^* & B_{11} \\ 0 & -A_{22}^* & B_{21} \\ \hline C_{11} & D_{11} B_{21}^* + C_{12} & D_{11} \end{array} \right] \\ &\stackrel{s}{=} \left[\begin{array}{ccc|c} -A_{11}^* & C_{11}^* C_{11} & C_{11}^* (D_{11} B_{21}^* + C_{12}) & C_{11}^* D_{11} \\ 0 & A_{11} & A_{12} + B_{11} B_{21}^* & B_{11} \\ 0 & 0 & -A_{22}^* & B_{21} \\ \hline -B_{12}^* & D_{12}^* C_{11} & D_{12}^* (D_{11} B_{21}^* + C_{12}) & D_{12}^* D_{11} \\ -B_{13}^* & D_{13}^* C_{11} & D_{13}^* D_{11} B_{21}^* & D_{13}^* D_{11} \end{array} \right] \end{aligned}$$

by (4.13). Introducing the change of basis

$$T = \begin{bmatrix} I & I & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}$$

and invoking (4.14), (4.27)–(4.29) gives

$$(4.33) \quad \begin{aligned} \begin{bmatrix} R_1 \\ R_2 \end{bmatrix} (s) &\stackrel{s}{=} \left[\begin{array}{cc|c} -A_{11}^* & -C_{11}^* D_{11} B_{21}^* - B_{11} B_{21}^* & C_{11}^* D_{11} + B_{11} \\ 0 & -A_{22}^* & -B_{21} \\ \hline -B_{12}^* & -D_{12}^* (D_{11} B_{21}^* + C_{12}) & D_{12}^* D_{11} \\ -B_{13}^* & -D_{13}^* D_{11} B_{21}^* & D_{13}^* D_{11} \end{array} \right] \\ &\stackrel{s}{=} \left[\begin{array}{c|c} -\bar{A}^* & \bar{C}^* \\ \hline -\bar{B}_1^* & \bar{D}_1^* \\ -\bar{B}_2^* & \bar{D}_2^* \end{array} \right] \end{aligned}$$

which is the same as (4.15). The minimality of this realisation follows from assumptions (i)–(iv) together with Limebeer and Hung [20, Thm. 4.1]. This completes the proof of (a).

We begin the proof of the remainder of the theorem by writing down the equations describing the spectral factorization of $(\gamma^2 I - R_2^* R_2)(s)$. Clearly,

$$(4.34) \quad (\gamma^2 I - R_2^* R_2)(s) \stackrel{s}{=} \left[\begin{array}{cc|c} \bar{A} & \bar{B}_2 \bar{B}_2^* & \bar{B}_2 \bar{D}_2^* \\ 0 & -\bar{A}^* & -\bar{C}^* \\ \hline -\bar{C} & -\bar{D}_2 \bar{B}_2^* & \gamma^2 I - \bar{D}_2 \bar{D}_2^* \end{array} \right].$$

Our first step is to carry out the decomposition

$$(4.35) \quad (\gamma^2 I - R_2^* R_2)(s) = Z(s) + Z^*(s)$$

in which $Z(s)$ is positive real. Since \bar{A} is asymptotically stable, there exists a unique $\Theta = \Theta^* \geq 0$ which satisfies the Lyapunov equation

$$(4.36) \quad \bar{A}\Theta + \Theta\bar{A}^* + \bar{B}_2 \bar{B}_2^* = 0.$$

For our later convenience we will also introduce the partitioning

$$(4.37) \quad \Theta = \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{12}^* & \Theta_{22} \end{bmatrix}$$

which is conformable with that of \bar{A} in (4.33).

The change of basis

$$(4.38) \quad T = \begin{bmatrix} I & -\Theta \\ 0 & I \end{bmatrix}$$

in the state space of (4.34) gives

$$(4.39) \quad Z(s) \stackrel{s}{=} [\bar{A}, N, -\bar{C}, \frac{1}{2}E]$$

in which

$$(4.40) \quad N = \bar{B}_2 \bar{D}_2^* + \Theta \bar{C}^* = \begin{bmatrix} N_1 \\ N_2 \end{bmatrix}$$

where the partitioning is also induced by that in (4.33). Finally,

$$(4.41) \quad E = \gamma^2 I - \bar{D}_2 \bar{D}_2^*.$$

The spectral factor $M(s)$ may now be expressed in terms of the unique positive definite solution of the Riccati equation [1]

$$(4.42) \quad Y(\bar{A} + NE^{-1}\bar{C}) + (\bar{A} + NE^{-1}\bar{C})^* Y + YNE^{-1}N^* Y + \bar{C}^* E^{-1} \bar{C} = 0.$$

The next lemma shows that the required solution to (4.42) always exists.

LEMMA A. The Riccati equation (4.42) always has a unique positive definite stabilizing solution.

Proof. See Appendix A.

As with (4.37), the solution to (4.42) has a partitioning induced by (4.33):

$$(4.43) \quad Y = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{12}^* & Y_{22} \end{bmatrix}.$$

Since Y is stabilising,

$$(4.44) \quad M(s) \stackrel{s}{=} [\bar{A}, N, -E^{1/2}F, E^{1/2}]$$

in which

$$(4.45) \quad F = E^{-1}(\bar{C} + N^* Y) = [F_1 | F_2]$$

is a minimum phase spectral factor (it is easy to check that $M^*(s)M(s) = Z(s) + Z^*(s)$). $\gamma > \|R_2\|_\infty$ ensures the positive definiteness of E and therefore that

$$(4.46) \quad M^{-1}(s) \stackrel{s}{=} [\bar{A} + NF, NE^{-1/2}, F, E^{-1/2}]$$

is proper.

We will now carry out the calculations which lead to state space realisations of $R_1 M^{-1}(s)$, $[R_1 M^{-1}(s)]_+$, and $[R_1 M^{-1}(s)]_-$.

$$(4.47) \quad R_1 M^{-1}(s) \stackrel{s}{=} \begin{bmatrix} -\bar{A}^* & \bar{C}^* \\ -\bar{B}_1^* & \bar{D}_1^* \end{bmatrix}^* \begin{bmatrix} \bar{A} + NF & NE^{-1/2} \\ F & E^{-1/2} \end{bmatrix} \\ = \begin{bmatrix} -\bar{A}^* & \bar{C}^* F & \bar{C}^* E^{-1/2} \\ 0 & \bar{A} + NF & NE^{-1/2} \\ -\bar{B}_1^* & \bar{D}_1^* F & \bar{D}_1^* E^{-1/2} \end{bmatrix}.$$

The change of basis

$$T = \begin{bmatrix} I & Y \\ 0 & I \end{bmatrix}$$

in the state space of (4.47) together with (4.42) and (4.45) yields

$$(4.48) \quad R_1 M^{-1}(s) \stackrel{s}{=} \begin{bmatrix} -\bar{A}^* & 0 & F^* E^{1/2} \\ 0 & \bar{A} + NF & NE^{-1/2} \\ -\bar{B}_1^* & \bar{D}_1^* F + \bar{B}_1^* Y & \bar{D}_1^* E^{-1/2} \end{bmatrix}.$$

In (4.48) we note that $-\bar{A}^*$ is completely unstable while $\bar{A} + NF$ is asymptotically stable. Thus

$$(4.49) \quad R_1 M^{-1}(s)_- \stackrel{s}{=} [-\bar{A}^*, F^* E^{1/2}, -\bar{B}_1^*, 0]$$

and

$$(4.50) \quad [R_1 M^{-1}(s)]_+ \stackrel{s}{=} [\bar{A} + NF, NE^{-1/2}, \bar{D}_1^* F + \bar{B}_1^* Y, \bar{D}_1^* E^{-1/2}].$$

The remainder of the proof is based on detailed manipulations requiring various partitions of the bounded real-type equations (4.18) and (4.19). The (1, 1) and (2, 1) blocks of (4.18) may be written out in full as

$$(4.51) \quad \begin{bmatrix} -A_{11}^* & -(C_{11}^* D_{11} + B_{11}) B_{21}^* & 0 \\ 0 & -A_{22}^* & 0 \\ 0 & 0 & \hat{A} \end{bmatrix} \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{12}^* & P_{22} & P_{23} \\ P_{13}^* & P_{23}^* & P_{33} \end{bmatrix} \\ + \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{12}^* & P_{22} & P_{23} \\ P_{13}^* & P_{23}^* & P_{33} \end{bmatrix} \begin{bmatrix} -A_{11} & 0 & 0 \\ -B_{21}(D_{11}^* C_{11} + B_{11}^*) & -A_{22} & 0 \\ 0 & 0 & \hat{A}^* \end{bmatrix} \\ + \begin{bmatrix} F_1^* E^{1/2} \\ F_2^* E^{1/2} \\ \hat{B} \end{bmatrix} [E^{1/2} F_1 \quad E^{1/2} F_2 \quad \hat{B}^*] + \begin{bmatrix} 0 \\ 0 \\ L_{21} \end{bmatrix} [0 \quad 0 \quad L_{21}^*] = 0$$

and

$$(4.52) \quad \hat{D}[E^{1/2}F_1 \quad E^{1/2}F_2 \quad \hat{B}^*] + [-\bar{B}_{11}^* \quad -\bar{B}_{21}^* \quad \hat{C}] \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{12}^* & P_{22} & P_{23} \\ P_{13}^* & P_{23}^* & P_{33} \end{bmatrix} + W[0 \quad 0 \quad L_{21}^*] = 0.$$

In the same way the (1, 1) and (2, 1) blocks of (4.19) may be written out as

$$(4.53) \quad \begin{bmatrix} -A_{11} & 0 & 0 \\ -B_{21}(D_{11}^*C_{11} + B_{11}^*) & -A_{22} & 0 \\ 0 & 0 & \hat{A}^* \end{bmatrix} \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{12}^* & Q_{22} & Q_{23} \\ Q_{13}^* & Q_{23}^* & Q_{33} \end{bmatrix} + \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{12}^* & Q_{22} & Q_{23} \\ Q_{13}^* & Q_{23}^* & Q_{33} \end{bmatrix} \begin{bmatrix} -A_{11}^* & -(C_{11}^*D_{11} + B_{11})B_{21}^* & 0 \\ 0 & -A_{22}^* & 0 \\ 0 & 0 & \hat{A} \end{bmatrix} + \begin{bmatrix} -\bar{B}_{11} \\ -\bar{B}_{21} \\ \hat{C}^* \end{bmatrix} [-\bar{B}_{11}^* - \bar{B}_{21}^* \quad \hat{C}] + \begin{bmatrix} 0 \\ 0 \\ L_{21d}^* \end{bmatrix} [0 \quad 0 \quad L_{21d}] = 0$$

and

$$(4.54) \quad \hat{D}^*[-\bar{B}_{11}^* | -\bar{B}_{21}^* | \hat{C}] + [E^{1/2}F_1 | E^{1/2}F_2 | \hat{B}^*] \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{12}^* & Q_{22} & Q_{23} \\ Q_{13}^* & Q_{23}^* & Q_{33} \end{bmatrix} + W_d^*[0 \quad 0 \quad L_{21d}] = 0.$$

An easy rearrangement shows that (4.42) may be written in the alternative form

$$(4.55) \quad Y\bar{A} + \bar{A}^*Y + F^*EF = 0$$

which together with (4.51) yields

$$(4.56) \quad \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{12}^* & Y_{22} \end{bmatrix} = - \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^* & P_{22} \end{bmatrix}.$$

The (1, 1) block of (4.53) together with (4.26), (4.36) and (4.37) yields

$$(4.57) \quad I + Q_{11} = \Theta_{11}.$$

Next, we have

$$\begin{aligned} X(s) &= (\tilde{X} - (R_1 M^{-1})_+) M(s) \\ &\stackrel{s}{=} \left[\begin{array}{c|c} \hat{A} & 0 \\ 0 & \bar{A} + NF \\ \hline \hat{C} & \bar{D}_1^* F + \bar{B}_1^* Y \end{array} \middle| \begin{array}{c} \hat{B} \\ -NE^{-1/2} \\ \hline -\bar{D}_1^* E^{-1/2} + \hat{D} \end{array} \right] * \begin{bmatrix} \bar{A} & N \\ -E^{1/2}F & E^{1/2} \end{bmatrix} \\ &\stackrel{s}{=} \left[\begin{array}{c|c|c} \hat{A} & 0 & -\hat{B}E^{1/2}F \\ 0 & \bar{A} + NF & NF \\ 0 & 0 & \bar{A} \\ \hline \hat{C} & \bar{D}_1^* F + \bar{B}_1^* Y & (\bar{D}_1^* E^{-1/2} - \hat{D})E^{1/2}F \end{array} \middle| \begin{array}{c} \hat{B}E^{1/2} \\ -N \\ N \\ \hline \hat{D}E^{1/2} - \bar{D}_1^* \end{array} \right] \end{aligned}$$

which after the change of basis

$$T = \begin{bmatrix} I & 0 & 0 \\ 0 & I & I \\ 0 & 0 & I \end{bmatrix}$$

becomes

$$(4.58) \quad X(s) \stackrel{s}{=} \left[\begin{array}{c|c} \hat{A} & -\hat{B}E^{1/2}F \\ 0 & \bar{A} \\ \hline \hat{C} & -\hat{D}E^{1/2}F - \bar{B}_1^*Y \end{array} \middle| \begin{array}{c} \hat{B}E^{1/2} \\ N \\ \hline \hat{D}E^{1/2} - \bar{D}_1^* \end{array} \right].$$

The change of basis

$$T = \begin{bmatrix} I & P_{13}^* & P_{23}^* \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix},$$

together with the (3, 1) and (3, 2) blocks of (4.51), and the (1, 1) and (1, 2) blocks of (4.52), gives

$$(4.59) \quad X(s) \stackrel{s}{=} \left[\begin{array}{c|c} \hat{A} & \hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2 \\ \hline \hat{C} & \hat{D}E^{1/2} - \bar{D}_1^* \end{array} \right]$$

thereby proving (4.22).

The last part of the proof is concerned with showing that (4.23) is true. We will establish this by a chain of intricate manipulations of the state-space realization of $T_{12}XY_{21}(s)$. From (4.12) and (4.59) we have

$$(4.60) \quad T_{12}XT_{21}(s) \stackrel{s}{=} \left[\begin{array}{c|c} \frac{A_{11}}{C_{11}} & \frac{B_{12}}{D_{12}} \end{array} \right] * \left[\begin{array}{c|c} \hat{A} & \hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2 \\ \hline \hat{C} & \hat{D}E^{1/2} - \bar{D}_1^* \end{array} \right] * \left[\begin{array}{c|c} \frac{A_{22}}{C_{22}} & \frac{B_{21}}{I} \end{array} \right]$$

$$\stackrel{s}{=} \left[\begin{array}{ccc|c} A_{11} & B_{12}(\hat{D}E^{1/2} - \bar{D}_1^*)C_{22} & B_{12}\hat{C} & B_{12}(\hat{D}E^{1/2} - \bar{D}_1^*) \\ 0 & A_{22} & 0 & B_{21} \\ 0 & (\hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2)C_{22} & \hat{A} & \hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2 \\ \hline C_{11} & D_{12}(\hat{D}E^{1/2} - \bar{D}_1^*)C_{22} & D_{12}\hat{C} & D_{12}(\hat{D}E^{1/2} - \bar{D}_1^*) \end{array} \right].$$

The first change of basis in the state space of (4.60) is given by

$$(4.61) \quad T = \begin{bmatrix} I & 0 & -Q_{13} \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}.$$

This together with the (1, 3) block of (4.53) gives

(4.62)

$$T_{12}XT_{21}(s) \stackrel{s}{=} \left[\begin{array}{ccc|c} A_{11} & B_{12}(\hat{D}E^{1/2} - \bar{D}_1^*)C_{22} - Q_{13}(\hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2)C_{22} & 0 & \Pi \\ 0 & A_{22} & 0 & B_{21} \\ 0 & (\hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2)C_{22} & \hat{A} & \hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2 \\ \hline C_{11} & D_{12}(\hat{D}E^{1/2} - \bar{D}_1^*)C_{22} & D_{12}\hat{C} + C_{11}Q_{13} & D_{12}(\hat{D}E^{1/2} - \bar{D}_1^*) \end{array} \right].$$

Our next lemma gives a simplified expression for Π in (4.62).

LEMMA B.

$$\Pi = (\Theta_{12} - Q_{12})B_{21} - B_{11}.$$

Proof. See Appendix B.

A second change of basis

$$(4.63) \quad T = \begin{bmatrix} I & Q_{12} - \Theta_{12} & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}$$

gives

$$T_{12}XT_{21}(s) \stackrel{s}{=} \left[\begin{array}{ccc|c} A_{11} & \Sigma & 0 & -B_{11} \\ 0 & A_{22} & 0 & B_{21} \\ 0 & (\hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2)C_{22} & \hat{A} & \hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2 \\ \hline C_{11} & D_{12}(\hat{D}E^{1/2} - \hat{D}_1^*)C_{22} - C_{11}(Q_{12} - \Theta_{12}) & D_{12}\hat{C} + C_{11}Q_{13} & D_{12}(\hat{D}E^{1/2} - \hat{D}_1^*) \end{array} \right].$$

LEMMA C.

$$\Sigma = -A_{12}.$$

Proof. See Appendix C.

Making use of Lemma C yields

$$(4.64) \quad T_{12}XT_{21}(s) \stackrel{s}{=} \left[\begin{array}{ccc|c} A_{11} & A_{12} & 0 & -B_{11} \\ 0 & A_{22} & 0 & -B_{21} \\ 0 & -(\hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2)C_{22} & \hat{A} & \hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2 \\ \hline C_{11} & -D_{12}(\hat{D}E^{1/2} - \hat{D}_1^*)C_{22} + C_{11}(Q_{12} - \Theta_{12}) & D_{12}\hat{C} + C_{11}Q_{13} & D_{12}(\hat{D}E^{1/2} - \hat{D}_1^*) \end{array} \right].$$

Before continuing further, we need to link Θ and Y (the solutions to (4.36) and (4.42), respectively). This connection is provided next.

LEMMA D. (a) *There exists a unique "largest" solution Z_0 to the Riccati equation*

$$(4.65) \quad \begin{aligned} &\{A_{11} + B_{13}D_{13}^*D_{11}E^{-1}(D_{11}^*C_{11} + B_{11}^*)\}Z + Z\{A_{11} + B_{13}D_{13}^*D_{11}E^{-1}(D_{11}^*C_{11} + B_{11}^*)\}^* \\ &+ Z(C_{11}^*D_{11} + B_{11})E^{-1}(D_{11}^*C_{11} + B_{11}^*)Z \\ &+ B_{13}[I + D_{13}^*D_{11}E^{-1}D_{11}^*D_{13}]B_{13} = 0. \end{aligned}$$

Further,

$$(4.66) \quad H_0 = (B_{13}D_{13}^*D_{11} + Z_0(C_{11}^*D_{11} + B_{11}))E^{-1}$$

is a destabilizing output injection for $[A_{11}, (D_{11}^*C_{11} + B_{11}^*)]$. In other words, $\text{Re}[\lambda(A_{11} + H_0(D_{11}^*C_{11} + B_{11}^*))] > 0$.

(b) *The stabilizing solution to (4.42) is related to Z_0 and Θ by*

$$(4.67) \quad Y^{-1} + \Theta = \begin{bmatrix} Z_0 & 0 \\ 0 & \gamma^2 I \end{bmatrix}$$

where the partitioning is induced by that in (4.37) and (4.43).

Proof. See Appendix D.

For convenience we introduce the notation

$$(4.68) \quad Y^{-1} = \begin{bmatrix} \hat{Y}_{11} & \hat{Y}_{12} \\ \hat{Y}_{12}^* & \hat{Y}_{22} \end{bmatrix}$$

and the last transformation we require is

$$(4.69) \quad T = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & P_{13}^* \hat{Y}_{12} + P_{23}^* \hat{Y}_{22} & I \end{bmatrix}.$$

Applying this to (4.64) gives

$$(4.70) \quad T_{12} X T_{21} \stackrel{s}{=} \left[\begin{array}{ccc|c} A_{11} & A_{12} & 0 & -B_{11} \\ 0 & A_{22} & 0 & -B_{21} \\ 0 & \Phi & \hat{A} & \Delta \\ \hline C_{11} & \Psi & D_{12} \hat{C} + C_{11} Q_{13} & D_{12}(\hat{D}E^{1/2} - \bar{D}_1^*) \end{array} \right].$$

LEMMA E.

$$\Psi = C_{12},$$

$$\Phi = 0,$$

$$\Delta = \hat{B}E^{1/2} + P_{13}^* B_{13} D_{13}^* D_{11} + P_{23}^* C_{22}^* E + (P_{13}^* \Theta_{11} + P_{23}^* \Theta_{12}^*)(C_{11}^* D_{11} + B_{11}).$$

Proof. See Appendix E.

Invoking Lemma E gives

$$(4.71) \quad (T_{11} + T_{12} X T_{21})(s) \stackrel{s}{=} \begin{bmatrix} \hat{A} & \Delta \\ D_{12} \hat{C} + C_{11} Q_{13} & D_{12}(\hat{D}E^{1/2} - \bar{D}_1^*) \end{bmatrix}$$

which verifies (4.23) and concludes our proof. \square

4.2. The bound c_b . We will establish c_b by counting those points (including multiplicities) at which cancellations may occur in $F_i(P(s), -K(s))$ in the case that $K(s)$ is stabilizing.

Theorem 4.2 below shows that every uncontrollable mode in $F_i(P(s), -K(s))$ is due to a cancellation at a zero of $P_{21}(s)$. In the case that $K(s)$ is (internally) stabilizing, the number of uncontrollable modes is bounded above by the number of zeros of $P_{21}(s)$ in \mathbb{C}_- (counting multiplicities). In the same way, the number of unobservable modes is bounded above by the number of zeros of $P_{12}(s)$ in \mathbb{C}_- (counting multiplicities).

THEOREM 4.2 [2], [19], [20]. Let

$$(4.72) \quad \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \stackrel{s}{=} \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{array} \right]$$

in which $P_{12}(s) \in \mathbb{R}^{p_1 \times m_2}(s)$ with $p_1 \geq m_2$ and $P_{21}(s) \in \mathbb{R}^{p_2 \times m_1}(s)$ with $m_1 \geq p_2$. Suppose also that

$$(4.73) \quad K(s) \stackrel{s}{=} \begin{bmatrix} \hat{A} & \hat{B} \\ \hat{C} & \hat{D} \end{bmatrix}$$

is a minimal realization and that the well-posedness condition $\det(I - D_{22}\hat{D}) \neq 0$ is satisfied. Then, in the closed loop of Fig. 1,

(a) Every unobservable mode (from y_1) is a Smith zero of

$$(4.74) \quad \begin{bmatrix} sI - A & B_2 \\ C_1 & D_{12} \end{bmatrix}.$$

(b) Every uncontrollable mode (from u_1) is a Smith zero of

$$(4.75) \quad \begin{bmatrix} sI - A & B_1 \\ C_2 & D_{21} \end{bmatrix}.$$

Proof. See [19], [20].

Theorem 4.2 thus allows us to write

$$c_b = \{\text{number of zeros of } P_{12} \text{ in } \mathbb{C}_-\} + \{\text{number of zeros of } P_{21} \text{ in } \mathbb{C}_-\}$$

and, as a result of Lemma 3.1(i) and Lemma 4.3 below, this becomes

$$(4.76) \quad \begin{aligned} c_b &= \{n - \text{rank}(\mathcal{X})\} + \{n - \text{rank}(\mathcal{Y})\} \\ &= 2n - \text{rank}(\mathcal{X}) - \text{rank}(\mathcal{Y}). \end{aligned}$$

Lemma 4.3 provides a link between the Smith zeros of $P_{12} \stackrel{s}{=} (A, B_2, C_1, D_{12})$ and the modes of $(A - B_2 D_{12}^* C_1)$ which are undetectable through $D_{\perp}^* C_1$.

LEMMA 4.3. Suppose that

$$(4.77) \quad P(s) = \begin{bmatrix} sI - A & -B_2 \\ C_1 & D_{12} \end{bmatrix}$$

is a polynomial matrix of dimension $(n+p) \times (n+m)$ with $p > m$. If D_{12} is part of an orthogonal matrix and D_{\perp} is its orthogonal completion, then every Smith zero of $P(s)$ is an unobservable mode of $[A - B_2 D_{12}^* C_1, D_{\perp}^* C_1]$ and vice versa.

Proof. If s_0 is a Smith zero of $P(s)$, there exists a vector $[w^* | v^*] \neq 0$ such that

$$(4.78) \quad \begin{bmatrix} s_0 I - A & -B_2 \\ C_1 & D_{12} \end{bmatrix} \begin{bmatrix} w \\ v \end{bmatrix} = 0.$$

We note also that $w \neq 0$ since if this were not the case, we would have $D_{12}v = 0$ which is impossible. From (4.78) we get

$$(4.79) \quad (s_0 I - A)w - B_2 v = 0$$

and

$$(4.80) \quad C_1 w + [D_{12} | D_{\perp}] \begin{bmatrix} v \\ 0 \end{bmatrix} = 0.$$

Multiplying (4.80) on the left by $[D_{12} | D_{\perp}]^*$ gives

$$(4.81a) \quad D_{12}^* C_1 w + v = 0,$$

$$(4.81b) \quad D_{\perp}^* C_1 w = 0.$$

Substituting (4.81a) into (4.79) and combining the result with (4.81b) yields

$$(4.82) \quad \begin{bmatrix} s_0 I - A + B_2 D_{12}^* C_1 \\ D_{\perp}^* C_1 \end{bmatrix} w = 0$$

which completes the proof in one direction.

If (4.82) is satisfied, we may write

$$(4.83) \quad (s_0 I - A)w - B_2 v = 0$$

where

$$(4.84) \quad v := -D_{12}^* C_1 w.$$

Combining (4.84) with the (2, 1) block of (4.82) gives

$$\begin{bmatrix} D_{12}^* \\ D_1^* \end{bmatrix} C_1 w + \begin{bmatrix} v \\ 0 \end{bmatrix} = 0$$

and hence also

$$(4.85) \quad C_1 w + D_{12} v = 0.$$

Finally, we note that (4.85) combined with (4.83) gives (4.78) thereby establishing the result. \square

4.3. The controller degree bound. The main theorem is proved by substituting (4.11) and (4.76) into (4.2).

THEOREM 4.4. For any \mathcal{H}^∞ -optimal control problem of the second kind, every \mathcal{H}^∞ -optimal controller satisfies

$$(4.86) \quad (1) \quad \deg(K) \leq n + \deg(U) - 1$$

and every suboptimal controller ($\|\mathcal{R}(s)\|_\infty > \gamma_{\text{opt}}$) satisfies

$$(4.87) \quad (2) \quad \deg(K) \leq n + \deg(U).$$

In (4.86) and (4.87), $U(s) \in \mathcal{RH}_-^\infty$ is an arbitrary matrix contraction of specified dimensions, which may be chosen constant (or even zero). \square

4.4. \mathcal{H}^∞ -optimal control problems of the third kind. The $n-1$ degree bound has now been proved in the case of problems of the first and second kind. With this background it is natural to ask: "Does this bound carry over to problems of the third kind?" Problems of the third kind are characterized by the assumption that $P_{12}(s)$ in (2.14) has more rows than columns while $P_{21}(s)$ has more columns than rows. Several small computer examples ($n = 1, 2, 3, 4, 5$) indicate that the answer to this question is indeed "yes." In the case of larger problems, finite precision effects make cancellation phenomena increasingly difficult to detect. On the basis of these experimental observations, we offer the following conjecture.

CONJECTURE. For any \mathcal{H}^∞ -optimal control problem of the third kind, every \mathcal{H}^∞ -optimal controller satisfies

$$(i) \quad \deg(K) \leq n + \deg(U) - 1$$

and every suboptimal controller ($\|\mathcal{R}(s)\|_\infty > \gamma_{\text{opt}}$) satisfies

$$(ii) \quad \deg(K) \leq n + \deg(U).$$

As before, $U(s) \in \mathcal{RH}_-^\infty$ is an arbitrary matrix contraction which may be chosen constant.

4.5. Computation time trials. In this subsection we present quantitative data which substantiate our claims regarding the importance of removing cancellation phenomena from \mathcal{H}^∞ computer software. We performed a number of test computations on both the original software, and an improved program which takes into account many of the cancellation phenomena predicted by the results in this paper. Both programs were run on a VAX 750 computer under UNIX. The timing data was obtained using the UNIX routine dtime. In every case we allowed the γ -iteration to run until the solution was almost optimal: The computation exited from the iterative loop when μ (see again (2.29)) was in the interval $1 - 0.5 \times 10^{-5} \leq \mu \leq 1$.

TABLE 4.1

Number of states	Modified program	Original program
1	4.517 s	18.23 s
2	11.617 s	67.95 s
3	53.417 s	406.98 s
4	102.47 s	442.32 s
6	141.43 s	692.98 s
8	215.77 s	1611.42 s
14	1128.48 s	25273.70 s
17	1703.47 s	terminated after 12 hours of processor time

The results given in Table 4.1 show the state dimension of $P(s)$, the execution time of the original program and the execution time of the improved program (both in seconds). Apart from a marked reduction in computation time, the improved program demonstrated improved robustness properties.

5. Conclusions. The purpose of this paper has been to generalise the analysis in [20] to problems of the second kind. If in Fig. 1 $\deg(P) = n$, we have shown that any \mathcal{H}^∞ -optimal control problem of the second kind has an associated controller which requires no more than $n - 1$ states. In the case that a suboptimal value of $\gamma (> \gamma_{\text{opt}})$ is chosen, there is a continuum of controllers with $\deg(K) \leq n$. These results are stated formally in Theorem 4.4.

Our experience has been that the solution of “large” (big n) \mathcal{H}^∞ control problems is time consuming, especially when several iterations corresponding to various weight selections are required. Further, long calculations of this type are susceptible to severe numerical difficulties. The work in this paper has shown that there is considerable scope for reducing these problems by using cancellation theory to remove state inflation effects from computer code. Although the level of benefit varies from problem to problem, a cpu calculation time reduction of between five and ten times is easy to achieve. The modified code is also considerably more robust from a numerical point of view.

Appendix A.

Proof of Lemma A. Since \bar{A} is asymptotically stable, $[\bar{A}, N]$ is stabilisable. $\gamma > \|R_2(s)\|_\infty \Rightarrow \gamma^2 I - \bar{D}_2 \bar{D}_2^* = E$ is positive definite and

$$[sI - \bar{A} | N] \begin{bmatrix} I & 0 \\ -E^{-1} \bar{C} & E^{-1/2} \end{bmatrix} = [sI - \bar{A} - NE^{-1} \bar{C} | NE^{-1/2}]$$

ensures that $[\bar{A} + NE^{-1} \bar{C}, NE^{-1/2}]$ is also stabilisable.

The Hamiltonian matrix associated with (4.42) is

$$H = \begin{bmatrix} \bar{A} + NE^{-1} \bar{C} & NE^{-1} N^* \\ -\bar{C}^* E^{-1} \bar{C} & -(\bar{A} + NE^{-1} \bar{C})^* \end{bmatrix}$$

and we note that

$$(A.1) \quad \{\lambda(H)\} \subseteq \{\text{Smith-McMillan zeros of } (\gamma^2 I - R_2^* R_2)(s)\} \cup \{\lambda_i(\bar{A})\} \cup \{\lambda_i(-\bar{A})\}.$$

Since $\gamma > \|R_2(s)\|_\infty$ no Smith-McMillan zero of $(\gamma^2 I - R_2^* R_2)(s)$ lies on the imaginary axis. This together with the asymptotic stability of \bar{A} ensures that H is free of imaginary axis eigenvalues. Finally, the stabilisability of $[\bar{A} + NE^{-1} \bar{C}, NE^{-1/2}]$ and Lemma 1(i)

ensure the existence of a unique stabilising solution $Y = Y^* \geq 0$ to (4.42). Part (a) in the theorem statement ensures that $[\bar{A}, E^{-1/2}\bar{C}]$ is observable and Lemma 1(ii) $\Rightarrow Y$ is nonsingular or else that the unique stabilising solution satisfies $Y = Y^* > 0$. \square

Appendix B.

Proof of Lemma B. The change of basis (4.61) in the state space of (4.60) gives

$$\Pi = B_{12}(\hat{D}E^{1/2} - \bar{D}_1^*) - Q_{13}(\hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2).$$

Making use of the (1, 1) block of (4.54) yields

$$\Pi = Q_{11}F_1^*E + Q_{12}F_2^*E - B_{12}D_{12}^*D_{11} - Q_{13}P_{13}^*N_1 - Q_{13}P_{23}^*N_2.$$

From the (1, 1) and (1, 2) blocks of (4.20), and from (4.56), we get

$$\Pi = -N_1 + Q_{11}(C_{11}^*D_{11} + B_{11}) + Q_{12}C_{22}^* - B_{12}D_{12}^*D_{11}.$$

Equation (4.57) gives

$$\Pi = -C_{11}^*D_{11} - B_{11} + (Q_{12} - \Theta_{12})C_{22}^* - B_{13}D_{13}^*D_{11} - B_{12}D_{12}^*D_{11}.$$

Finally, from (4.30) we get

$$\Pi = -B_{11} + (Q_{12} - \Theta_{12})C_{22}^*.$$

\square

Appendix C.

Proof of Lemma C. After the coordinate change (4.61) in (4.60) we get

$$\begin{aligned} \Sigma &= (Q_{12} - \Theta_{12})A_{22} - A_{11}(Q_{12} - \Theta_{12}) + B_{12}(\hat{D}E^{1/2} - \bar{D}_1^*)C_{22} \\ &\quad - Q_{13}(\hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2)C_{22}. \end{aligned}$$

Making use of (4.15) and the (1, 1) block of (4.54) gives

$$\begin{aligned} \Sigma &= (Q_{12} - \Theta_{12})A_{22} - A_{11}(Q_{12} - \Theta_{12}) - B_{12}D_{12}^*D_{11}C_{22} \\ &\quad - Q_{13}(P_{13}^*N_1 + P_{23}^*N_2)C_{22} + (Q_{11}F_1^*E + Q_{12}F_2^*E)C_{22}. \end{aligned}$$

Substituting from the (1, 1) and (1, 2) blocks of (4.20), (4.40), (4.45), (4.24), (4.25) and (4.15) yields

$$\begin{aligned} \Sigma &= (Q_{12} - \Theta_{12})A_{22} - A_{11}(Q_{12} - \Theta_{12}) - (B_{13}D_{13}^*D_{11} + \Theta_{11}(C_{11}^*D_{11} + B_{11}) + \Theta_{12}C_{22}^*)C_{22} \\ &\quad + Q_{11}(B_{11} + C_{11}^*D_{11})C_{22} - B_{12}D_{12}^*D_{11}C_{22} + Q_{12}(-A_{22} - A_{22}^*). \end{aligned}$$

Invoking (4.24), (4.25), (4.30) and (4.57) gives

$$\Sigma = A_{11}(\Theta_{12} - Q_{12}) - (Q_{12} - \Theta_{12})A_{22}^* - B_{11}C_{22}.$$

Finally, the (1, 2) block of (4.36), the (1, 2) block of (4.53) and (4.25) lead to the required result

$$\Sigma = -A_{12}.$$

\square

Appendix D.

Proof of Lemma D. A minor variant of Lemma 2.1(i) shows that the existence of the destabilising solution Z_0 is established by proving

- (i) $[A_{11}^*, C_{11}^*D_{11} + B_{11}]$ is controllable;
- (ii) The Hamiltonian matrix corresponding to (4.65) is free of imaginary axis eigenvalues.

The fact that Z_0 is the largest solution (i.e., $Z_0 - Z \geq 0$ for all other solutions) is only of peripheral interest and consequently will not be proved here. In fact, all that is needed is a minor modification of an argument in [28, Lemma 3].

We begin by showing that (i) is true. We know from part (a) of Theorem 4.1 and (4.25) that

$$\left\{ \begin{bmatrix} -A_{11}^* & (C_{11}^* D_{11} + B_{11}) C_{22} \\ 0 & -A_{22}^* \end{bmatrix}, \begin{bmatrix} C_{11}^* D_{11} + B_{11} \\ C_{22}^* \end{bmatrix} \right\}$$

is controllable. Next, we suppose for contradiction that (i) is not satisfied. That is, there exists a vector $w \neq 0$ such that

$$w^* A_{11}^* = \lambda w^*,$$

$$w^* (C_{11}^* D_{11} + B_{11}) = 0.$$

From this we have

$$(D.1) \quad [w^* | 0] \begin{bmatrix} -A_{11}^* & (C_{11}^* D_{11} + B_{11}) C_{22} \\ 0 & -A_{22}^* \end{bmatrix} = [-\lambda w^* | 0]$$

and

$$(D.2) \quad [w^* | 0] \begin{bmatrix} C_{11}^* D_{11} + B_{11} \\ C_{22}^* \end{bmatrix} = [0 | 0]$$

which contradicts part (a) of Theorem 4.1 (which is already proved). This contradiction establishes (i).

As we will now show, (ii) is in fact a consequence of Lemma A. Direct substitution into the Hamiltonian in Lemma A gives

$$(D.3) \quad H = \begin{bmatrix} A_{11} + N_1 E^{-1} (D_{11}^* C_{11} + B_{11}^*) & N_1 E^{-1} C_{22} \\ (-C_{22}^* + N_2 E^{-1}) (D_{11}^* C_{11} + B_{11}^*) & A_{22} + N_2 E^{-1} C_{22} \\ -(C_{11}^* D_{11} + B_{11}) E^{-1} (D_{11}^* C_{11} + B_{11}^*) & -(C_{11}^* D_{11} + B_{11}) E^{-1} C_{22} \\ -C_{22}^* E^{-1} (D_{11}^* C_{11} + B_{11}^*) & -C_{22}^* E^{-1} C_{22} \end{bmatrix} \\ + \begin{bmatrix} N_1 E^{-1} N_1^* & N_1 E^{-1} N_2^* \\ N_2 E^{-1} N_1^* & N_2 E^{-1} N_2^* \\ -A_{11}^* - (C_{11}^* D_{11} + B_{11}) E^{-1} N_1^* & (B_{11} + C_{11}^* D_{11}) (E^{-1} N_2^* - C_{22}) \\ -C_{22}^* E^{-1} N_1^* & -A_{22}^* - C_{22}^* E^{-1} N_2^* \end{bmatrix}.$$

A tortuous but routine computation based on (4.13), (4.24), (2.25), (4.33), (4.40) and (4.41) shows that

$$(D.4) \quad -THT^{-1} = \begin{bmatrix} A_{11}^* + (C_{11}^* D_{11} + B_{11}) E^{-1} D_{11}^* D_{13} B_{13}^* & -B_{13} D_{13}^* [I + D_{11} E^{-1} D_{11}^*] D_{13} B_{13}^* & 0 \\ 0 & C_{22}^* E^{-1} D_{11}^* D_{13} B_{13}^* & 0 \\ (C_{11}^* D_{11} + B_{11}) E^{-1} (D_{11}^* C_{11} + B_{11}^*) & (C_{11}^* D_{11} + B_{11}) E^{-1} C_{22} & 0 \\ -A_{11} - B_{13} D_{13}^* D_{11} E^{-1} (D_{11}^* C_{11} + B_{11}^*) & -B_{13} D_{13}^* D_{11} E^{-1} C_{22} & 0 \\ 0 & A_{22}^* & 0 \\ C_{22}^* E^{-1} (D_{11}^* C_{11} + B_{11}^*) & C_{22}^* E^{-1} C_{22} & -A_{22} \end{bmatrix}$$

where

$$(D.5) \quad T = \begin{bmatrix} 0 & 0 & I & 0 \\ I & 0 & \Theta_{11} & \Theta_{12} \\ 0 & I & \Theta_{12}^* & \Theta_{22} - \gamma^2 I \\ 0 & 0 & 0 & I \end{bmatrix}.$$

Since H in (D.3) is free of imaginary axis eigenvalues (Lemma A), so too is the matrix in (D.4). Suppose

$$(D.6) \quad S = \begin{bmatrix} (1, 1) & (1, 2) \\ (2, 1) & (2, 2) \end{bmatrix}$$

where (i, j) represents the (i, j) th block of (D.4). Then it is easy to see that

- (a) S is the Hamiltonian associated with (4.65);
- (b) $\lambda(S) \subset \lambda(H) \Rightarrow$ (ii).

This concludes the proof of the existence part of (a).

Let us suppose that the orthogonal matrix W transforms S into an ordered upper Schur form [18], specifically

$$(D.7) \quad S \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}$$

in which T_{11} is completely unstable and T_{22} is asymptotically stable. Substituting from (D.4) into (D.6) gives

$$(D.8) \quad \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} W_{11} \\ W_{21} \end{bmatrix} = \begin{bmatrix} W_{11} \\ W_{21} \end{bmatrix} [T_{11}]$$

where

$$\begin{aligned} S_{11} &= \{A_{11} + B_{13}D_{13}^*D_{11}E^{-1}(D_{11}^*C_{11} + B_{11})\}^*, \\ S_{12} &= (C_{11}^*D_{11} + B_{11})E^{-1}(D_{11}^*C_{11} + B_{11}^*), \\ S_{21} &= -B_{13}(I + D_{13}^*D_{11}E^{-1}D_{11}^*D_{13})B_{13}^*, \\ S_{22} &= -S_{11}^*. \end{aligned}$$

and the Riccati equation solution is [18]

$$(D.9) \quad Z_0 = W_{21}W_{11}^{-1}.$$

Conjugating the $(1, 1)$ block of (D.8) and multiplying on the left by W_{11}^{-*} gives

$$(D.10) \quad A_{11} + \{(B_{13}D_{13}^*D_{11} + Z_0(C_{11}^*D_{11} + B_{11}))E^{-1}\}(D_{11}^*C_{11} + B_{11}) = W_{11}^{-*}T_{11}^*W_{11}^*$$

which proves that H_0 in (4.66) is a destabilising output injection. Thus (a) is proved.

The (b) part of the lemma will be established in two steps. First, we will prove that *any* solution to (4.65) will generate a solution to (4.42) via (4.67). Following that, we show that the largest (destabilising) solution to (4.65) generates the stabilising solution to (4.42).

From (4.55) we get

$$(D.11) \quad Y^{-1}\bar{A}^* + \bar{A}Y^{-1} + Y^{-1}F^*EFY^{-1} = 0;$$

adding this to (4.36) gives

$$(D.12) \quad (Y^{-1} + \Theta)\bar{A}^* + \bar{A}(Y^{-1} + \Theta) + Y^{-1}F^*EFY^{-1} + \bar{B}_2\bar{B}_2^* = 0.$$

Substituting from (4.67) and writing (D.12) out in full gives

$$\begin{aligned}
 & \begin{bmatrix} Z & 0 \\ 0 & \gamma^2 I \end{bmatrix} \begin{bmatrix} A_{11}^* & \bar{A}_{12}^* \\ 0 & A_{22}^* \end{bmatrix} + \begin{bmatrix} A_{11} & 0 \\ \bar{A}_{12} & A_{22} \end{bmatrix} \begin{bmatrix} Z & 0 \\ 0 & \gamma^2 I \end{bmatrix} \\
 & + \begin{bmatrix} Z - \Theta_{11} & -\Theta_{12} \\ -\Theta_{12}^* & \gamma^2 I - \Theta_{22} \end{bmatrix} \begin{bmatrix} C_{11}^* D_{11} + B_{11} \\ C_{22}^* \end{bmatrix} E^{-1} [D_{11}^* C_{11} + B_{11}^* | C_{22}] \\
 & + \begin{bmatrix} Z - \Theta_{11} & -\Theta_{12} \\ -\Theta_{12}^* & \gamma^2 I - \Theta_{22} \end{bmatrix} \begin{bmatrix} N_1 \\ N_2 \end{bmatrix} E^{-1} [D_{11}^* C_{11} | C_{22}] \\
 (D.13) \quad & + \begin{bmatrix} Z - \Theta_{11} & -\Theta_{12} \\ -\Theta_{12}^* & \gamma^2 I - \Theta_{22} \end{bmatrix} + \begin{bmatrix} B_{13} \\ B_{21} D_{11}^* D_{13} \end{bmatrix} [B_{13}^* | D_{13}^* D_{11} B_{21}^*] \\
 & + \begin{bmatrix} Z - \Theta_{11} & -\Theta_{12} \\ -\Theta_{12}^* & \gamma^2 I - \Theta_{22} \end{bmatrix} \begin{bmatrix} C_{11}^* D_{11} + B_{11} \\ C_{22}^* \end{bmatrix} E^{-1} [N_1 | N_2] \\
 & + \begin{bmatrix} N_1 \\ N_2 \end{bmatrix} E^{-1} [N_1^* \quad N_2^*] = 0
 \end{aligned}$$

in which

$$(D.14) \quad \bar{A}_{12} = B_{21}(B_{11}^* + D_{11}^* C_{11}).$$

Equations (4.40) and (4.25) allow the (2, 2) block (denoted (2, 2)) of (D.13) to be written out in full as

$$\begin{aligned}
 (2, 2) &= \gamma^2 A_{22} + \gamma^2 A_{22}^* + [B_{21} E + N_2] E^{-1} [E B_{21}^* + N_2^*] - N_2 [B_{21}^* + E^{-1} N_2^*] \\
 &\quad + N_2 E^{-1} N_2^* - [B_{21} + N_2 E^{-1}] N_2^* + B_{21} (\gamma^2 I - E) B_{21}^*.
 \end{aligned}$$

After we cancel terms this becomes

$$\begin{aligned}
 (D.15) \quad (2, 2) &= \gamma^2 (A_{22} + A_{22}^* + B_{21} B_{21}^*) \\
 &= 0 \quad \text{by (4.24).}
 \end{aligned}$$

The definitions of N and E in (4.40) and (4.41) allow the (1, 2) block of (D.13) (denoted (1, 2)) to be written out in full as

$$\begin{aligned}
 (1, 2) &= Z \bar{A}_{12}^* + [Z (C_{11}^* D_{11} + B_{11}) + B_{13} D_{13}^* D_{11} - N_1] E^{-1} [-E B_{21}^* - N_2^*] \\
 &\quad - N_1 [B_{21}^* + E^{-1} N_2^*] + [Z (C_{11}^* D_{11} + B_{11}) + B_{13} D_{13}^* D_{11} - N_1] E^{-1} N_2^* \\
 &\quad + N_1 E^{-1} N_2^* + B_{13} D_{13}^* D_{11} B_{21}^*.
 \end{aligned}$$

After we cancel terms this becomes

$$\begin{aligned}
 (D.16) \quad (1, 2) &= Z (\bar{A}_{12}^* - (C_{11}^* D_{11} + B_{11}) B_{21}^*) \\
 &= 0 \quad \text{by (D.14).}
 \end{aligned}$$

Clearly, (2, 1) = 0 follows by symmetry.

As with (D.15) and (D.16), it is easy to show that the (1, 1) block of (D.13) is zero provided Z is a solution of (4.65). This verification only requires the definitions of N_1 and N_2 in (4.40).

We now begin a sequence of arguments which prove that the largest solution Z_0 to (4.65) generates the stabilizing solution to (4.42). Since H_0 in (4.66) is a destabilizing output injection, and since A_{22} is stable, the matrix

$$(D.17) \quad \begin{bmatrix} A_{11}^* + (C_{11}^* D_{11} + B_{11}) H_0^* & 0 \\ C_{22}^* E^{-1} [D_{11}^* D_{13} B_{13}^* + (D_{11}^* C_{11} + B_{11}^*) Z_0] & -A_{22} \end{bmatrix}$$

is completely unstable,

$$(D.18) \Rightarrow \begin{bmatrix} A_{11}^* + (C_{11}^* D_{11} + B_{11}) H_0^* & \bar{A}_{12}^* + (C_{11}^* D_{11} + B_{11}) C_{22} \\ C_{22}^* E^{-1} [D_{11}^* D_{13} B_{13}^* + (D_{11}^* C_{11} + B_{11}^*) Z_0] & A_{22}^* + C_{22}^* C_{22} \end{bmatrix}$$

is completely unstable by (D.14), (4.24) and (4.25), and

$$(D.19) \Rightarrow \begin{bmatrix} A_{11}^* & \bar{A}_{12}^* \\ 0 & A_{22}^* \end{bmatrix} + \begin{bmatrix} C_{11}^* D_{11} + B_{11} \\ C_{22}^* \end{bmatrix} E^{-1} [D_{11}^* D_{13} B_{13}^* + (D_{11}^* C_{11} + B_{11}^*) Z_0 | EC_{22}]$$

is completely unstable by (4.66). Using the notation

$$(D.20) \quad Y^{-1} = \begin{bmatrix} \hat{Y}_{11} & \hat{Y}_{12} \\ \hat{Y}_{12}^* & \hat{Y}_{22} \end{bmatrix}$$

together with (4.67) and (4.40) establishes that

$$(D.21) \quad (D.19) \Rightarrow \begin{bmatrix} A_{11}^* & \bar{A}_{12}^* \\ 0 & A_{22}^* \end{bmatrix} + \begin{bmatrix} C_{11}^* D_{11} + B_{11} \\ C_{22}^* \end{bmatrix} E^{-1} [\mathbb{A} | \mathbb{B}]$$

where

$$\begin{aligned} \mathbb{A} &= N_1^* + (D_{11}^* C_{11} + B_{11}^*) \hat{Y}_{11} + C_{22} \hat{Y}_{12}^*, \\ \mathbb{B} &= N_2^* + (D_{11}^* C_{11} + B_{11}^*) \hat{Y}_{12} + C_{22} \hat{Y}_{22} \end{aligned}$$

is completely unstable. Substituting from

$$(D.22) \quad (4.15) \Rightarrow \bar{A}^* + \bar{C}^* E^{-1} (N^* + \bar{C}^* \hat{Y})$$

is completely unstable. From (4.42) we get

$$(D.23) \quad \bar{A}^* + \bar{C}^* E^{-1} (N^* + \bar{C}^* \hat{Y}) = -Y \{ \bar{A} + NE^{-1} \bar{C} + NE^{-1} N^* Y \} \hat{Y},$$

whence by (4.45)

$$(D.24) \quad (D.22) \Rightarrow \bar{A} + NE^{-1} (\bar{C} + N^* Y) = \bar{A} + NF$$

is completely stable. This thus proves that Z_0 generates the required solution to (4.42). \square

Appendix E.

Proof of Lemma E. We begin by proving that

$$(E.1) \quad \begin{aligned} \Psi &= C_{11} (Q_{12} - \Theta_{12}) - D_{12} (\hat{D} E^{1/2} - D_{12}^* D_{11}) C_{22} \\ &\quad - (D_{12} \hat{C} + C_{11} Q_{13}) (P_{13}^* \hat{Y}_{12} + P_{23}^* \hat{Y}_{22}) = C_{12}. \end{aligned}$$

From the (1, 1) and (1, 2) blocks of (4.52), and from (4.56) and (4.15), we get

$$(E.2) \quad -\hat{C} P_{13}^* = \hat{D} E^{1/2} F_1 + B_{12}^* Y_{11} + D_{12}^* (D_{11} B_{21}^* + C_{12}) Y_{12}^*,$$

$$(E.3) \quad -\hat{C} P_{23}^* = \hat{D} E^{1/2} F_2 + B_{12}^* Y_{12} + D_{12}^* (D_{11} B_{21}^* + C_{12}) Y_{22}.$$

We will also use the three equations from (see (4.68))

$$(E.4) \quad \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{12}^* & Y_{22} \end{bmatrix} \begin{bmatrix} \hat{Y}_{11} & \hat{Y}_{12} \\ \hat{Y}_{12}^* & \hat{Y}_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Substituting (E.2), (E.3) and (E.4) into (E.1) gives

$$(E.5) \quad \begin{aligned} \Psi = & C_{11}(Q_{12} - \Theta_{12}) - D_{12}\hat{D}E^{1/2}C_{22} + D_{12}\hat{D}E^{1/2}F_1\hat{Y}_{12} + D_{12}D_{12}^*C_{12} \\ & - C_{11}Q_{13}P_{13}^*\hat{Y}_{12} + D_{12}\hat{D}E^{1/2}F_2\hat{Y}_{22} - C_{11}Q_{13}P_{23}^*\hat{Y}_{22}. \end{aligned}$$

Using (4.13), (4.31), (4.45), (4.25) and (E.4) gives

$$(E.6) \quad \begin{aligned} \Psi = & C_{11}(Q_{12} - \Theta_{12} - Q_{13}P_{13}^*\hat{Y}_{12} - Q_{13}P_{23}^*\hat{Y}_{22}) + C_{12} \\ & + D_{12}\hat{D}E^{-1/2}[(D_{11}^*C_{11} + B_{11}^*)\hat{Y}_{12} + C_{22}\hat{P}_{22} + N_2^* + EC_{22}]. \end{aligned}$$

Using (4.13), (4.25), (4.67) and (4.41) we get from (E.6) that

$$\Psi = C_{11}(Q_{12} - \Theta_{12} - Q_{13}P_{13}^*\hat{Y}_{12} - Q_{13}P_{23}^*\hat{Y}_{22}) + C_{12}.$$

By (4.20) and

$$(E.7) \quad \hat{Y}_{22} = (Y_{22} - Y_{12}^*Y_{11}^{-1}Y_{12})^{-1} \quad (Y > 0 \Rightarrow Y_{11} > 0 \Rightarrow Y_{11}^{-1} \text{ exists}),$$

$$(E.8) \quad \hat{Y}_{12} = Y_{11}^{-1}Y_{12}(Y_{22} - Y_{12}^*Y_{11}^{-1}Y_{12})^{-1}$$

gives (E.1) as required.

We now establish that

$$(E.9) \quad \begin{aligned} \Phi = & (P_{13}^*\hat{Y}_{12} + P_{23}^*\hat{Y}_{22})A_{22} - \hat{A}(P_{13}^*\hat{Y}_{12} + P_{23}^*\hat{Y}_{22}) \\ & - (\hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2)C_{22} = 0. \end{aligned}$$

Substituting the (3, 1) and (3, 2) blocks of (4.51), (4.45) and (E.4) gives

$$(E.10) \quad \begin{aligned} \Phi = & \hat{B}E^{-1/2}\{D_{11}^*C_{11}\hat{Y}_{12} + B_{11}^*\hat{Y}_{12} + N_2^* + C_{22}\hat{Y}_{22} - EC_{22}\} \\ & - P_{23}^*\{\bar{A}_{12}\hat{Y}_{12} + A_{22}\hat{Y}_{22} + N_2C_{22} - \hat{Y}_{22}A_{22}\} \\ & - P_{13}^*\{A_{11}\hat{Y}_{12} + N_1C_{22} - \hat{Y}_{12}A_{22}\}. \end{aligned}$$

By (4.40), (4.41) and (4.67) we get

$$(E.11) \quad \Phi = -P_{23}^*\{\bar{A}_{12}\hat{Y}_{12} + A_{22}\hat{Y}_{22} + N_2C_{22} - \hat{Y}_{22}A_{22}\} - P_{13}^*\{A_{11}\hat{Y}_{12} + N_1C_{22} - \hat{Y}_{12}A_{22}\}.$$

Making use of (D.14), (4.13), (4.24), (4.40), (4.67), and the (2, 2) block of (4.36) gives

$$(E.12) \quad \Phi = -P_{13}^*\{A_{11}\hat{Y}_{12} + B_{13}D_{13}^*D_{11}C_{22} + \Theta_{11}(C_{11}^*D_{11} + B_{11})C_{22} - \Theta_{12}A_{22}^*\}.$$

From the (1, 2) block of (4.36) we obtain

$$(E.13) \quad A_{11}\Theta_{12} - \Theta_{11}(C_{11}^*D_{11} + B_{11})C_{22} + \Theta_{12}A_{22}^* = B_{13}D_{13}^*D_{11}C_{22}$$

and this together with the (1, 2) block of (4.67) gives

$$(E.14) \quad \Phi = 0$$

as required.

Finally, we have from (4.64), (4.69) and (4.70) that

$$(E.15) \quad \Delta = \hat{B}E^{1/2} + P_{13}^*N_1 + P_{23}^*N_2 - (P_{13}^*\hat{Y}_{12} + P_{23}^*\hat{Y}_{22})B_{21}.$$

From (4.40) this becomes

$$\begin{aligned} \Delta = & \hat{B}E^{1/2} + P_{13}^*\{B_{13}D_{13}^*D_{11} + \Theta_{11}(C_{11}^*D_{11} + B_{11}) + \Theta_{12}C_{22}^* - \hat{Y}_{12}B_{21}\} \\ & + P_{23}^*\{B_{21}D_{11}^*D_{13}D_{13}^*D_{11} + \Theta_{12}^*(C_{11}^*D_{11}B_{11}) + \Theta_{22}C_{22}^* - \hat{Y}_{22}B_{21}\}. \end{aligned}$$

Using (4.25), (4.41) and (4.67) to cancel terms gives

$$\begin{aligned}\Delta &= \hat{B}E^{1/2} + P_{13}^*(B_{13}D_{13}^*D_{11} + \Theta_{11}(C_{11}^*D_{11} + B_{11})) + P_{23}^*(C_{22}^*E + \Theta_{12}^*(C_{11}^*D_{11} + B_{11})) \\ &= \hat{B}E^{1/2} + P_{13}^*B_{13}D_{13}^*D_{11} + P_{23}^*C_{22}^*E + (P_{13}^*\Theta_{11} + P_{23}^*\Theta_{12}^*)(C_{11}^*D_{11} + B_{11})\end{aligned}$$

as required. \square

6. Acknowledgments. It is our pleasure to thank E. Kasenally and H. Wang for their assistance with the computer programming and for running numerous examples.

REFERENCES

- [1] B. D. O. ANDERSON, *An algebraic solution to the spectral factorisation problem*, IEEE Trans. Automat. Control, AC-12 (1967), pp. 410-414.
- [2] B. D. O. ANDERSON AND A. LINNEMANN, *Control of decentralized systems with distributed controller complexity*, Proc. 24th IEEE Conf. on Decision and Control, 1985, pp. 1468-1472.
- [3] B. D. O. ANDERSON AND S. VONGPANITLERD, *Network analysis and synthesis. A modern approach*, Prentice-Hall, Englewood Cliffs, NJ, 1972.
- [4] J. A. BALL AND N. COHEN, *Sensitivity minimization in \mathcal{H}^∞ norm: Parametrization of all suboptimal solutions*, Internal report, Dept. of Math., Virginia Polytechnic Inst. and State Univ., Blacksburg, VA, 1987.
- [5] B. C. CHANG AND B. PEARSON, *Optimal disturbance rejection in linear multivariable systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 880-887.
- [6] C. A. DESOER, R. W. LIU, J. MURRAY AND R. SAEKS, *Feedback system design: The fractional representation approach to analysis and synthesis*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 399-412.
- [7] J. C. DOYLE, *Advances in multivariable control*, Office of Naval Research-Honeywell Workshop, 1984.
- [8] Y. K. FOO AND I. POSTLETHWAITE, *An \mathcal{H}^∞ -minimax approach to the design of robust control systems*, Systems Control Lett., 5 (1984), pp. 81-82.
- [9] B. A. FRANCIS, *A course in \mathcal{H}^∞ control theory*, Lecture Notes in Control and Information Sci., Springer-Verlag, Berlin-New York, 1987.
- [10] ———, *Notes on \mathcal{H}^∞ -optimal linear feedback systems*, Lecture Notes, Linköping Univ., Linköping, Sweden, 1983.
- [11] B. A. FRANCIS AND G. ZAMES, *On optimal sensitivity theory for SISO feedback systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 9-16.
- [12] B. A. FRANCIS, J. W. HELTON AND G. ZAMES, *\mathcal{H}^∞ -optimal feedback controllers for linear multivariable systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 888-900.
- [13] B. A. FRANCIS AND J. C. DOYLE, *Linear control theory with an \mathcal{H}^∞ optimality criterion*, this Journal, 25 (1987), pp. 815-844.
- [14] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their \mathcal{L}^∞ -error bounds*, Internat. J. Control, 39 (1984), pp. 1115-1193.
- [15] ———, *Robust stabilization of linear multivariable systems: relations to approximation*, Internat. J. Control, 43 (1986), pp. 741-766.
- [16] H. KIMURA, *Robust stabilization for a class of transfer functions*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 788-793.
- [17] V. KUČERA, *A contribution to matrix quadratic equations*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 344-347.
- [18] A. J. LAUB, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 913-921.
- [19] D. J. N. LIMEBEER AND B. D. O. ANDERSON, *An interpolation theory approach to \mathcal{H}^∞ controller degree bounds*, Linear Algebra Appl., to appear.
- [20] D. J. N. LIMEBEER AND Y. S. HUNG, *An analysis of the pole-zero cancellations in \mathcal{H}^∞ optimal control problems of the first kind*, this Journal, 25 (1987), pp. 1457-1493.
- [21] D. J. N. LIMEBEER AND G. D. HALIKIAS, *A controller degree bound for \mathcal{H}^∞ -optimal control problems of the second kind*, IEEE Conf. on Decision and Control, Athens, Greece, 1986.
- [22] Z. NEHARI, *On bounded bilinear forms*, Ann. of Math., 65 (1984), pp. 153-162.
- [23] C. N. NETT, C. A. JACOBSON AND M. J. BALAS, *A connection between state-space and doubly coprime fractional representations*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 831-832.
- [24] H. H. ROSENBRCK, *State-Space and Multivariable Theory*, Nelson, Nashville, TN, 1970.

- [25] M. G. SAFONOV AND M. S. VERMA, \mathcal{L}^∞ -sensitivity optimization and Hankel approximation, IEEE Trans. Automat. Control, AC-30 (1985), pp. 278–280.
- [26] M. G. SAFONOV, E. A. JONCKHEERE, M. VERMA AND D. J. N. LIMBEER, Synthesis of positive real multivariable feedback systems, Internat. J. Control, 45 (1987), pp. 817–842.
- [27] M. VERMA AND E. JONCKHEERE, \mathcal{L}^∞ -compensation with mixed sensitivity as a broadband matching problem, Systems Control Lett., 4 (1984), pp. 125–130.
- [28] J. C. WILLEMS, Least squares stationary optimal control and the algebraic Riccati equation, IEEE Trans. Automat. Control, AC-16 (1971), pp. 621–634.
- [29] D. C. YOULA, H. JABR AND J. J. BONGIORNO, Modern Wiener–Hopf design of optimal controllers, Part II: The multivariable case, IEEE Trans. Automat. Control, AC-21 (1976), pp. 319–338.
- [30] G. ZAMES AND B. A. FRANCIS, Feedback, minimax sensitivity, and optimal robustness, IEEE Trans. Automat. Control, AC-28 (1983), pp. 585–600.
- [31] G. ZAMES, Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses, IEEE Trans. Automat. Control, AC-26 (1981), pp. 301–320.

AN ASYMPTOTIC EXPANSION FOR AN OPTIMAL RELAXATION OSCILLATOR*

CHIEN-HSIUNG CHUANG†, JASON L. SPEYER†, AND JOHN V. BREAKWELL‡

Abstract. For certain reduced-order optimization problems where assumed fast dynamics are neglected, *chattering* optimal solutions occur. The chattering optimal solution is represented by some of the variables alternating between distinctively different values at an infinite rate. For a simple and somewhat transparent periodic optimal control problem, the neglected dynamics are included by an asymptotic expansion about the chattering solution. The periodic chattering arc is approached as a weighting parameter, associated with the control penalty in the performance index, goes toward zero. This weighting parameter is used as the expansion parameter to form an asymptotic expansion about the chattering arc. In particular, two time scales are used in the expansions. A time scale proportional to the period is used to transform the problem to one similar to that of a *relaxation oscillator* where the problem is characterized by slow, almost equilibrium motions connected by fast, jump type transitions. The asymptotic expansion is divided into two parts, an outer part at the time scale of the period and an inner part characterized by an even faster time scale which captures the fast transitions. These two solutions are matched together to obtain the resulting asymptotic solution in which the performance index and the optimal period are obtained up to third order, and the states and the control are obtained up to second order. Comparison with the exact solution shows extremely good agreement.

Key words. periodic optimal control problems, relaxation oscillation, chattering solution, asymptotic expansion, singular perturbation

AMS(MOS) subject classification. 41A60

1. Introduction. For many models of physical systems, optimization leads to a *chattering* optimal solution where the control alternates between distinctly different values at an infinite rate. For this class of optimization problems, the hodograph of the velocity space is not convex. Problems in this class are given for a chemical batch process of Horn and Lin [1] and for aircraft cruise [2]–[4]. This phenomenon seems to occur because the model simplification assumed that certain dynamics are fast and can be neglected. Motivated by the physical problems and analysis given in [1]–[5], the objective of this work is to show how these neglected dynamics can be included by an asymptotic expansion about the chattering solution. In particular, a simple and somewhat transparent optimal control problem that induces a periodic optimal control is selected [6]. Asymptotic expansions are obtained for this problem about the minimizing static optimal solution [7]. Here, the chattering phenomenon is approached as a weighting parameter associated with the control penalty in the performance index goes toward zero. Therefore, this weighting parameter is used as the expansion parameter to form an asymptotic expansion about the chattering arc.

Periodic solutions of singularly perturbed differential equations are called *relaxation oscillations*. The most prominent example of a relaxation oscillator is the singularly perturbed Van der Pol equation in which the order of the differential equation is two and the frequency of the oscillation remains finite for all values of the perturbation parameter [8]–[10]. This and other interesting examples can be found in [11]. In this paper, we investigate a fourth-order, singularly perturbed, differential equation representing the first-order necessary conditions for a periodic optimal control problem,

* Received by the editors August 18, 1986; accepted for publication (in revised form) August 18, 1987.

† Department of Aerospace Engineering and Engineering Mechanics, University of Texas, Austin, Texas 78712. This research was supported by National Science Foundation grant ECS-8413475.

‡ Department of Aeronautics and Astronautics, Stanford University, Stanford, California 94305.

where the periodic optimal solution converges to a minimizing chattering arc as a perturbation parameter goes to zero. This example is considered to extend the perspective of relaxation oscillations to the infinite frequency case. By scaling the optimal control problem with a fast time scale, say τ_1 , the resulting first-order necessary conditions have properties very similar to that of a relaxation oscillator. The essential feature characterizing both problems is the slow, almost equilibrium motion connected by a fast or jump type motion. Therefore, the asymptotic expansion is divided into two parts, an outer part characterized by the time scale τ_1 and an inner part characterized by an even faster time scale τ_2 . These two solutions are matched together, and a resulting asymptotic solution is given for the performance index, the optimal period, the state variables and the Lagrange multipliers.

This paper is organized as follows: In § 2, a two-dimensional periodic optimal control problem is formulated and necessary conditions for its optimal solution are stated. In § 3 the optimization problem is transformed with respect to two time scales: a time scale proportional to the period (called the outer region) and a time scale associated with the fast motion of the relaxation oscillation (called the inner region). In § 4 an expansion in the outer region is developed, whereas in § 5, an inner expansion is developed. The matching of the outer expansion with the inner expansion, a minimization of a performance index with respect to an optimal period, and comparison of the asymptotic expansion with the numerical solution are presented in § 6. Finally, conclusions are given in § 7.

2. Problem statement and conditions for optimality. A two-dimensional periodic optimal control problem which was first proposed by Speyer and Evans [6] is formulated in this section. First-order necessary conditions which supply differential equations and boundary conditions for states and Lagrange multipliers, and second-order necessary conditions, are reviewed. An optimality condition utilizing a frequency method is also discussed.

The object of this periodic optimal control problem is to find a period T , a scalar control function $u(\cdot)$, and an initial state vector $x^T(0) = [x_1(0), x_2(0)]$ that will minimize the performance index

$$(2.1) \quad J = \frac{1}{T} \int_0^T \left(\frac{x_1^2}{2} + \frac{x_2^4}{4} - \frac{x_2^2}{2} + \frac{bu^2}{2} \right) dt$$

subject to a dynamic constraint of a second-order differential equation

$$(2.2) \quad \dot{x}_1 = x_2, \quad \dot{x}_2 = u$$

and periodic boundary conditions

$$(2.3) \quad x_1(0) = x_1(T), \quad x_2(0) = x_2(T),$$

where the dot operation denotes derivatives with respect to time and b is a weighting parameter for the control. The periodic optimal control problem above has been conceptually considered as the simplest example of a periodic optimal control problem [6]. Since a nonconvex cost seems to induce periodic optimal paths, a negative quadratic term is included in the performance index (2.1). A quartic term is also included in (2.1) so that the states of the optimal solutions are bounded. The weighting parameter b determines whether the optimal solution is a periodic or a static path. This will be discussed in the latter part of this section.

The problem might physically represent a sailboat attempting to maximize its average velocity into the direction of the wind. Suppose y is the distance in the direction of the prevailing wind, and x_1, x_2 in (2.2) represent the lateral position and velocity.

The cost function (2.1) may be rewritten as

$$(2.4) \quad J = -\frac{y(T)}{T} + \frac{1}{T} \int_0^T \left(\frac{x_1^2}{2} + \frac{bu^2}{2} \right) dt,$$

where the average velocity $-y(T)/T$ is to be minimized subject to an integral penalty on the lateral position $\int_0^T (x_1^2/2) dt/T$ associated with the tack of the sailboat, and the cost of tacking $\int_0^T (bu^2/2) dt/T$. The longitudinal velocity $\dot{y}(t)$ is assumed to be a function of the lateral velocity as $\dot{y} = -x_2^4/4 + x_2^2/2$.

The first-order necessary conditions [6] which are used to obtain a two-point boundary-value problem for the states and Lagrange multipliers are presented as follows. A variational Hamiltonian function is defined in terms of (2.1) and (2.2) as

$$(2.5) \quad H = \frac{x_1^2}{2} + \frac{x_2^4}{4} - \frac{x_2^2}{2} + \frac{bu^2}{2} + \hat{\lambda}_1 x_2 + \hat{\lambda}_2 u,$$

where $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are Lagrange multipliers associated with the dynamics of the system (2.2). The first-order necessary conditions produce

$$(2.6) \quad \dot{x}_1 = x_2, \quad \dot{x}_2 = -\frac{\hat{\lambda}_2}{b}, \quad \dot{\hat{\lambda}}_1 = -x_1, \quad \dot{\hat{\lambda}}_2 = -x_2^3 + x_2 - \hat{\lambda}_1,$$

with boundary conditions

$$(2.7) \quad x_1(0) = x_1(T), \quad x_2(0) = x_2(T), \quad \hat{\lambda}_1(0) = \hat{\lambda}_1(T), \quad \hat{\lambda}_2(0) = \hat{\lambda}_2(T),$$

$$(2.8) \quad H = J^o,$$

where the superscript o indicates an optimal value. The optimal control $u = -\hat{\lambda}_2/b$, solved by using $H_u = 0$, has been eliminated in obtaining (2.6).

The transversality condition (2.8) is a special boundary condition which exists only for periodic optimal control problems [1]. Since the period optimizes the performance index (2.1), the derivative of J with respect to T must vanish; that is,

$$(2.9) \quad \frac{\partial J}{\partial T} = \frac{1}{T} [-J^o + H(T)] = 0.$$

Furthermore, for an autonomous Hamiltonian system the Hamiltonian function H is a constant of the motion, i.e., $H(T) = H(t) = H$, and this implies (2.8). Finally, for the second-order necessary conditions of this problem, if $b > 0$, it follows that $H_{uu} > 0$, and hence the Legendre-Clebsch condition is satisfied in its strong form.

A frequency test [12], [13] is used to determine the range of the parameter b where the steady-state optimal control solutions are not minimizing. According to the frequency test, a steady-state solution which satisfies the first-order necessary conditions is a locally minimizing solution if a function $\pi(\omega) = 1/\omega^4 - 1/\omega^2 + b$ is nonnegative for all values of the frequency ω . It is easy to show that $\pi(\omega) \geq -\frac{1}{4} + b \geq 0$ for local optimality. Therefore, for $b \geq \frac{1}{4}$, $\pi(\omega)$ is nonnegative for all ω , and thus the steady-state solution is locally minimizing. For $0 < b < \frac{1}{4}$ there exists ω such that $\pi(\omega)$ is negative and the steady-state solution is not minimizing. This implies that periodic optimal paths which are locally minimizing can be obtained for $0 < b < \frac{1}{4}$. For the case $b = (1 - \bar{\epsilon}^2)/4$, where $\bar{\epsilon}$ is a small expansion parameter, a regular perturbation scheme is developed by using the Lindstedt-Poincaré expansion method [7]. The steady-state solution for $b = \frac{1}{4}$ is used as a reference path to expand the solution. However, in this paper the expansion will be obtained for a small positive b . Therefore, a chattering optimal solution with an infinite frequency, where H_{uu} is identically zero, is used as a reference solution for the asymptotic expansion. This will be illustrated in the following sections.

3. Time scales and associated optimal control formulations. The chattering solution of the system specified by (2.1)–(2.3) as b goes to zero is presented in this section. Then, some useful symmetry properties are presented. Finally, two time scales are suggested and the associated optimal control formulations are presented. The relationship between the first-order necessary conditions for these two optimal problems is also given in this section.

The chattering solution for $b=0$ is demonstrated as follows. The performance index J in (2.1) can be rewritten in the form

$$(3.1) \quad J = \frac{1}{T} \int_0^T \left[\frac{x_1^2}{2} + \frac{(x_2^2 - 1)^2}{4} + \frac{bu^2}{2} \right] dt - \frac{1}{4}.$$

From this, it is clear that the minimum solution is $J = -\frac{1}{4}$, and the states which give this minimum solution are $x_1 = 0$ and $x_2 = \pm 1$. However, this solution requires a control which is neither piecewise continuous nor bounded. Conceptually, if x_2 jumps back and forth between $+1$ and -1 in an infinite rate, then the magnitude of x_1 remains zero, and x_2 stays at an absolute value of one. It is about this seemingly complex arc that an asymptotic expansion will be produced.

To simplify the problem, the existence of symmetric planes in the surface of initial conditions is used. These symmetries are found by examining the Lagrangian, which is being extremized:

$$(3.2) \quad \begin{aligned} L &= H - \hat{\lambda}^T \dot{x} \\ &= \frac{x_1^2}{2} + \frac{x_2^4}{4} - \frac{x_2^2}{2} + \frac{bu^2}{2} + \hat{\lambda}_1 x_2 + \hat{\lambda}_2 u - \hat{\lambda}_1 \dot{x}_1 - \hat{\lambda}_2 \dot{x}_2 \end{aligned}$$

to determine the symmetries in x_1 , x_2 , $\hat{\lambda}_1$, $\hat{\lambda}_2$ and u that keep L invariant. The symmetry used for all the expansions produced here is

$$(3.3) \quad \begin{aligned} x_1(t) &= -x_1(-t), & x_2(t) &= x_2(-t), & \hat{\lambda}_1(t) &= \hat{\lambda}_1(-t), & \hat{\lambda}_2(t) &= -\hat{\lambda}_2(-t), \\ u(t) &= -u(-t). \end{aligned}$$

Since x_1 is odd and x_2 is even, then $\dot{x}_1(t)$ is even and \dot{x}_2 is odd. Introduction of these symmetries into (3.2) shows that L remains invariant. Note that $u = -\hat{\lambda}_2/b$, so that if λ_2 is even, so is u , as specified in (3.3). A second symmetry, however, not investigated here, is found by reversing the symmetry in (3.3) by taking x_1 and λ_2 and u as even functions, and x_2 and $\hat{\lambda}_1$ as odd functions. Again (3.2) is left invariant. However, both symmetric paths are only different in phase. The condition that the extremal path be closed is replaced by the stronger conditions that the initial values of $[\hat{\lambda}_1, \hat{\lambda}_2]$ be orthogonal to the initial values of $[x_1, x_2]^T$. As a consequence of this restriction, nonsymmetric periodic paths which can exist are not determined.

Two time scales can be used to expand the solution; one time scale approximates the magnitude of the period such that x_1 and the part of x_2 before its jump can be expanded, and another faster time scale is used to expand the jump. The time domain defined by the first time scale is called an outer region and that defined by the second time scale is called an inner region.

The outer time scale and outer variables are defined as

$$(3.4) \quad \tau_1 = \frac{t}{b^m}, \quad x_1^* = \frac{x_1}{b^m}, \quad v = b^{1/2}u, \quad \tau_{1f} = \frac{T}{4b^m},$$

where m is to be chosen. Since x_2 is of order one and $\dot{x}_1 = x_2$ in the limit as $b \rightarrow 0$, then x_1^* and τ_1 are scaled by the same power of b such that $dx_1^*/d\tau_1 = x_2$ remains of order

one. The control variable is scaled by $b^{1/2}$ so that effect of the small parameter on the control appears in only one place, the differential equation (see (3.6)) rather than the cost (see (3.5)). Finally, note that τ_{1f} is only a fourth of the scaled period T/b^m since the solution is periodic and symmetric such that each quarter provides equal contributions to the cost. The periodic optimal control problem defined by (2.1)–(2.3) can be rewritten in terms of the new variables in (3.4):

$$(3.5) \quad J = \frac{1}{\tau_{1f}} \int_0^{\tau_{1f}} \left(b^{2m} \frac{x_1^{*2}}{2} + \frac{x_2^4}{4} - \frac{x_2^2}{2} + \frac{v^2}{2} \right) d\tau_1,$$

$$(3.6) \quad \frac{dx_1^*}{d\tau_1} = x_2, \quad b^{1/2-m} \frac{dx_2}{d\tau_1} = v,$$

$$(3.7) \quad x_1^*(0) = 0, \quad x_2(\tau_{1f}) = 0.$$

If $m = \frac{1}{6}$ is chosen, then only one expansion parameter $\varepsilon \triangleq b^{1/3}$ will appear in the expansion since the powers of b in (3.5) and (3.6) are both equal to $\frac{1}{3}$. The boundary conditions (3.7) are different from (2.3), since τ_{1f} is only one-fourth of the original period, and by symmetry x_1 must be an even function of $(\tau_1 - \tau_{1f})$ and x_2 must be an odd function of $(\tau_1 - \tau_{1f})$.

Since ε multiplies the derivative $dx_2/d\tau_1$ in (3.6), this problem is seen as a singular perturbation problem. Therefore, to determine the motion in the “boundary layer” a faster time scale is required; that is,

$$(3.8) \quad \tau_2 = \frac{\tau_1}{\varepsilon} = \frac{t}{b^{1/2}}.$$

If we change the time variable from τ_1 to τ_2 , the system defined by (3.5)–(3.7) becomes

$$(3.9) \quad J = \frac{1}{\tau_{2f}} \int_0^{\tau_{2f}} \left(\varepsilon \frac{x_1^{*2}}{2} + \frac{x_2^4}{4} - \frac{x_2^2}{2} + \frac{v^2}{2} \right) d\tau_2,$$

$$(3.10) \quad \frac{dx_1^*}{d\tau_2} = \varepsilon x_2, \quad \frac{dx_2}{d\tau_2} = v,$$

$$(3.11) \quad x_1^*(0) = 0, \quad x_2(\tau_{2f}) = 0,$$

where $\varepsilon = b^{1/3}$ is substituted into (3.5) and (3.6) to obtain (3.9) and (3.10). The Hamiltonian function H for (3.9) and (3.10) is

$$(3.12) \quad H^{(i)} = \varepsilon \frac{x_1^{*(i)2}}{2} + \frac{x_2^{(i)4}}{4} - \frac{x_2^{(i)2}}{2} + \frac{v^{(i)2}}{2} + \varepsilon \lambda_1^{(i)} x_2^{(i)} + \lambda_2^{(i)} v^{(i)},$$

and the two-point boundary-value problem related to (3.9)–(3.11) is

$$(3.13) \quad \frac{dx_1^{*(i)}}{d\tau_2} = \varepsilon x_2^{(i)}, \quad \frac{dx_2^{(i)}}{d\tau_2} = -\lambda_2^{(i)}, \quad \frac{d\lambda_1^{(i)}}{d\tau_2} = -\varepsilon x_1^{*(i)}, \quad \frac{d\lambda_2^{(i)}}{d\tau_2} = -x_2^{(i)3} + x_2^{(i)} - \varepsilon \lambda_1^{(i)},$$

$$(3.14) \quad x_2^{(i)}(\tau_{2f}) = 0, \quad \lambda_1^{(i)}(\tau_{2f}) = 0,$$

where the superscript (i) denotes the inner variable and the use of an equality $v^{(i)} = -\lambda_2^{(i)}$ to obtain (3.13). Note that the boundary conditions $x_1^*(0) = 0$ and $\lambda_2(0) = 0$ will not be used to obtain the inner expansion since the left-hand boundary conditions are not required for the inner region.

By changing the independent variable from τ_2 to τ_1 in (3.12) and (3.13), we obtain the Hamiltonian function and the first-order equations for the outer expansion:

$$(3.15) \quad H^{(o)} = \varepsilon \frac{x_1^{*(o)2}}{2} + \frac{x_2^{(o)4}}{4} - \frac{x_2^{(o)2}}{2} + \frac{v^{(o)2}}{2} + \varepsilon \lambda_1^{(o)} x_2^{(o)} + \lambda_2^{(o)} v^{(o)},$$

$$(3.16) \quad \frac{dx_1^{*(o)}}{d\tau_1} = x_2^{(o)}, \quad \varepsilon \frac{dx_2^{(o)}}{d\tau_1} = -\lambda_2^{(o)}, \quad \frac{d\lambda_1^{(o)}}{d\tau_1} = -x_1^{*(o)},$$

$$\varepsilon \frac{d\lambda_2^{(o)}}{d\tau_1} = -x_2^{(o)3} + x_2^{(o)} - \varepsilon \lambda_1^{(o)},$$

$$(3.17) \quad x_1^{*(o)}(0) = 0, \quad \lambda_2^{(o)}(0) = 0,$$

where the superscript (o) denotes the outer variable. Note that the boundary conditions $x_2(\tau_{1f}) = 0$ and $\lambda_1(\tau_{1f}) = 0$ which are not presented in (3.17) will not be used to obtain the outer expansion since the outer expansion is not valid near τ_{1f} . Note also that the definitions for the state variables and Lagrange multipliers are the same for both regions.

4. Asymptotic expansion in the outer region. The series solution for the system specified by (3.15)–(3.17) in the outer region is developed as follows. Let the states $x_1^{*(o)}$ and $x_2^{(o)}$ and the Lagrange multipliers $\lambda_1^{(o)}$ and $\lambda_2^{(o)}$ be expanded in ascending powers of ε :

$$(4.1) \quad x_1^{*(o)}(\tau_1, \varepsilon) = x_{10}^{*(o)}(\tau_1) + \varepsilon x_{11}^{*(o)}(\tau_1) + \varepsilon^2 x_{12}^{*(o)}(\tau_1) + \cdots,$$

$$(4.2) \quad x_2^{(o)}(\tau_1, \varepsilon) = x_{20}^{(o)}(\tau_1) + \varepsilon x_{21}^{(o)}(\tau_1) + \varepsilon^2 x_{22}^{(o)}(\tau_1) + \cdots,$$

$$(4.3) \quad \lambda_1^{(o)}(\tau_1, \varepsilon) = \lambda_{10}^{(o)}(\tau_1) + \varepsilon \lambda_{11}^{(o)}(\tau_1) + \varepsilon^2 \lambda_{12}^{(o)}(\tau_1) + \cdots,$$

$$(4.4) \quad \lambda_2^{(o)}(\tau_1, \varepsilon) = \lambda_{20}^{(o)}(\tau_1) + \varepsilon \lambda_{21}^{(o)}(\tau_1) + \varepsilon^2 \lambda_{22}^{(o)}(\tau_1) + \cdots.$$

By substituting (4.1)–(4.4) into (3.16), sets of differential equations for the parameters in the series are obtained by equating like powers of ε on either sides of (3.16). For the order of ε^0 , the following equations are obtained:

$$(4.5) \quad \frac{dx_{10}^{*(o)}}{d\tau_1} = x_{20}^{(o)}, \quad 0 = \lambda_{20}^{(o)}, \quad \frac{d\lambda_{10}^{(o)}}{d\tau_1} = -x_{10}^{*(o)}, \quad 0 = -x_{20}^{(o)3} + x_{20}^{(o)},$$

which determine the zeroth-order solutions

$$(4.6) \quad x_{10}^{*(o)} = \tau_1, \quad \lambda_{20}^{(o)} = 0, \quad \lambda_{10}^{(o)} = -\frac{1}{2}\tau_1^2 + c_1, \quad x_{20}^{(o)} = 1,$$

where c_1 is an arbitrary constant. The boundary condition $x_{10}^{*(o)}(0) = 0$ is used to eliminate an integration constant in (4.6), and the integration constant c_1 will be determined during the matching. The Hamiltonian function in ascending powers of ε is

$$(4.7) \quad H^{(o)} = H_0^{(o)} + \varepsilon H_1^{(o)} + \varepsilon^2 H_2^{(o)} + \cdots,$$

where $H_i^{(o)}$, $i = 0, 1, 2, \dots$, can be determined in terms of the series parameters in (4.1)–(4.4) by substituting (4.1)–(4.4) into (3.15) and by using $v^{(o)} = -\lambda_2^{(o)}$. The resulting $H_0^{(o)}$ is

$$(4.8) \quad H_0^{(o)} = \frac{1}{4}x_{20}^{(o)4} - \frac{1}{2}x_{20}^{(o)2} - \frac{1}{2}\lambda_{20}^{(o)2} = -\frac{1}{4},$$

where the zeroth-order solutions (4.6) are used to determine the value of $H_0^{(o)}$.

If this procedure is extended to higher orders, then the differential equations for the order ε^1 are

$$(4.9) \quad \begin{aligned} \frac{dx_{11}^{*(o)}}{d\tau_1} &= x_{21}^{(o)}, \quad \frac{dx_{20}^{(o)}}{d\tau_1} = -\lambda_{21}^{(o)}, \quad \frac{d\lambda_{11}^{(o)}}{d\tau_1} = -x_{11}^{*(o)}, \\ \frac{d\lambda_{20}^{(o)}}{d\tau_1} &= -3x_{20}^{(o)2} x_{21}^{(o)} + x_{21}^{(o)} - \lambda_{10}^{(o)}. \end{aligned}$$

It follows that the first-order solutions are determined by using the above equations as

$$(4.10) \quad \begin{aligned} x_{11}^{*(o)} &= \frac{1}{12}\tau_1^3 - \frac{1}{2}c_1\tau_1, \quad \lambda_{21}^{(o)} = 0, \quad \lambda_{11}^{(o)} = -\frac{1}{48}\tau_1^4 + \frac{1}{4}c_1\tau_1^2 + c_2, \\ x_{21}^{(o)} &= \frac{1}{4}\tau_1^2 - \frac{1}{2}c_1, \end{aligned}$$

where c_2 is an integration constant and $x_{11}^{*(o)}(0) = 0$ is used to eliminate the integration constant for $x_{11}^{*(o)}$. The Hamiltonian function $H_1^{(o)}$ is also obtained by using the above solutions

$$(4.11) \quad H_1^{(o)} = \frac{1}{2}x_{10}^{*(o)2} + x_{20}^{(o)3} x_{21}^{(o)} - x_{20}^{(o)} x_{21}^{(o)} - \lambda_{20}^{(o)} \lambda_{21}^{(o)} + \lambda_{10}^{(o)} x_{20}^{(o)} = c_1.$$

To second order in ε , the differential equations are

$$(4.12) \quad \begin{aligned} \frac{dx_{12}^{*(o)}}{d\tau_1} &= x_{22}^{(o)}, \quad \frac{dx_{21}^{(o)}}{d\tau_1} = -\lambda_{22}^{(o)}, \quad \frac{d\lambda_{12}^{(o)}}{d\tau_1} = -x_{12}^{*(o)}, \\ \frac{d\lambda_{21}^{(o)}}{d\tau_1} &= -3x_{20}^{(o)2} x_{22}^{(o)} - 3x_{20}^{(o)} x_{21}^{(o)2} + x_{22}^{(o)} - \lambda_{11}^{(o)}, \end{aligned}$$

and the second-order solutions are determined to be

$$(4.13) \quad \begin{aligned} x_{12}^{*(o)} &= -\frac{1}{60}\tau_1^5 + \frac{1}{12}c_1\tau_1^3 + (-\frac{3}{8}c_1^2 - \frac{1}{2}c_2)\tau_1 \\ \lambda_{22}^{(o)} &= -\frac{1}{2}\tau_1, \\ \lambda_{12}^{(o)} &= \frac{1}{360}\tau_1^6 - \frac{1}{48}c_1\tau_1^4 - \frac{1}{2}(-\frac{3}{8}c_1^2 - \frac{1}{2}c_2)\tau_1^2 + c_3, \\ x_{22}^{(o)} &= -\frac{1}{12}\tau_1^4 + \frac{1}{4}c_1\tau_1^2 + (-\frac{3}{8}c_1^2 - \frac{1}{2}c_2), \end{aligned}$$

where c_3 is an arbitrary integration constant for $\lambda_{12}^{(o)}$ and where $x_{12}^{*(o)}(0) = 0$ is used to eliminate an integration constant for $x_{12}^{*(o)}$. It follows that we obtain the Hamiltonian function $H_2^{(o)}$:

$$(4.14) \quad \begin{aligned} H_2^{(o)} &= x_{10}^{*(o)} x_{11}^{*(o)} + \frac{3}{2}x_{20}^{(o)2} x_{21}^{(o)2} + x_{20}^{(o)3} x_{22}^{(o)} - \frac{1}{2}x_{21}^{(o)2} - x_{20}^{(o)} x_{22}^{(o)} - \frac{1}{2}\lambda_{21}^{(o)2} \\ &\quad - \lambda_{20}^{(o)} \lambda_{22}^{(o)} + \lambda_{10}^{(o)} x_{21}^{(o)} + \lambda_{11}^{(o)} x_{20}^{(o)} \\ &= c_2 - \frac{1}{4}c_1^2. \end{aligned}$$

In terms of the results of (4.6), (4.10) and (4.13), we obtain the states and Lagrange multipliers for the outer region:

$$(4.15) \quad x_1^{*(o)} = \tau_1 + \varepsilon(\frac{1}{12}\tau_1^3 - \frac{1}{2}c_1\tau_1) + \varepsilon^2[-\frac{1}{60}\tau_1^5 + \frac{1}{12}c_1\tau_1^3 + (-\frac{3}{8}c_1^2 - \frac{1}{2}c_2)\tau_1],$$

$$(4.16) \quad x_2^{(o)} = 1 + \varepsilon(\frac{1}{4}\tau_1^2 - \frac{1}{2}c_1) + \varepsilon^2[-\frac{1}{12}\tau_1^4 + \frac{1}{4}c_1\tau_1^2 + (-\frac{3}{8}c_1^2 - \frac{1}{2}c_2)],$$

$$(4.17) \quad \begin{aligned} \lambda_1^{(o)} &= -\frac{1}{2}\tau_1^2 + c_1 + \varepsilon(-\frac{1}{48}\tau_1^4 + \frac{1}{4}c_1\tau_1^2 + c_2) \\ &\quad + \varepsilon^2[\frac{1}{360}\tau_1^6 - \frac{1}{48}c_1\tau_1^4 - \frac{1}{2}(-\frac{3}{8}c_1^2 - \frac{1}{2}c_2)\tau_1^2 + c_3], \end{aligned}$$

$$(4.18) \quad \lambda_2^{(o)} = \varepsilon^2(-\frac{1}{2}\tau_1).$$

The Hamiltonian function is also written by combining the results of (4.8), (4.11) and (4.14) as

$$(4.19) \quad H^{(o)} = -\frac{1}{4} + \varepsilon(c_1) + \varepsilon^2(c_2 - \frac{1}{4}c_1^2).$$

Note that the unsolved constants c_1 , c_2 , and c_3 will be determined by the matching of the outer solution with the inner solution in § 6.

5. Asymptotic expansion in the inner region. The expansion in the inner region for the system specified by (3.12), (3.13) and (3.14) is developed in this section. Let the states $x_1^{*(i)}$ and $x_2^{(i)}$ and the Lagrange multipliers $\lambda_1^{(i)}$ and $\lambda_2^{(i)}$ be expanded in ascending powers of ε as those in (4.1)–(4.4), except that all the coefficients are now functions of τ_2 instead of τ_1 . By substituting these series into (3.13) and by equating like powers of ε on either sides of (3.13), differential equations for the unknown coefficients can be obtained for different orders of ε .

For the order of ε^0 , the differential equations for the zeroth-order solutions are

$$(5.1) \quad \frac{dx_{10}^{*(i)}}{d\tau_2} = 0, \quad \frac{dx_{20}^{(i)}}{d\tau_2} = -\lambda_{20}^{(i)}, \quad \frac{d\lambda_{10}^{(i)}}{d\tau_2} = 0, \quad \frac{d\lambda_{20}^{(i)}}{d\tau_2} = -x_{20}^{(i)3} + x_{20}^{(i)}.$$

A second-order differential equation for $x_{20}^{(i)}$ can be derived from (5.1):

$$(5.2) \quad \frac{d^2 x_{20}^{(i)}}{d\tau_2^2} = x_{20}^{(i)3} - x_{20}^{(i)}.$$

By integrating (5.2) with respect to $x_{20}^{(i)}$, we obtain a first-order differential equation, which might be considered an energy equation:

$$(5.3) \quad \frac{1}{2} \left(\frac{dx_{20}^{(i)}}{d\tau_2} \right)^2 = \frac{1}{4} x_{20}^{(i)4} - \frac{1}{2} x_{20}^{(i)2} + c,$$

where c is an integration constant which will be determined as follows. If $\lambda_{20}^{(i)} = -dx_{20}^{(i)}/d\tau_2$ is substituted into (5.3) and the expression (4.8) for the zeroth-order Hamiltonian function in the outer region is used here for the inner region, the constant c is found to be equal to $-H_0^{(i)}$; that is,

$$(5.4) \quad c = -\left(\frac{1}{4}x_{20}^{(i)4} - \frac{1}{2}x_{20}^{(i)2} - \frac{1}{2}\lambda_{20}^{(i)2}\right) = -H_0^{(i)} = -H_0^{(o)} = \frac{1}{4},$$

where the fact that the zeroth-order Hamiltonian functions are equal is used to obtain $\frac{1}{4}$. It follows that

$$(5.5) \quad \frac{dx_{20}^{(i)}}{d\tau_2} = \pm\sqrt{2}(x_{20}^{(i)2} - 1),$$

where the positive sign will be used to solve for $x_{20}^{(i)}$, since $x_{20}^{(i)}$ decreases from one to zero in the inner region and that gives $dx_{20}^{(i)}/d\tau_2 \leq 0$ and $x_{20}^{(i)} \leq 1$. Equation (5.5) is integrated again and an integration constant is chosen to fix the boundary conditions for $x_{20}^{(i)}$; that is, $x_{20}^{(i)}(\tau_{2f}) = 0$, so the solution for $x_{20}^{(i)}$ is

$$(5.6) \quad x_{20}^{(i)} = \tanh\left(\frac{\tau_{2f} - \tau_2}{\sqrt{2}}\right).$$

Therefore, the solutions for $x_{10}^{*(i)}$, $\lambda_{10}^{(i)}$ and $\lambda_{20}^{(i)}$ are obtained by substituting (5.6) into (5.1):

$$(5.7) \quad x_{10}^{*(i)} = k_1, \quad \lambda_{20}^{(i)} = \frac{1}{\sqrt{2} \cosh^2((\tau_{2f} - \tau_2)/\sqrt{2})}, \quad \lambda_{10}^{(i)} = 0,$$

where the boundary condition $\lambda_{10}^{(i)}(\tau_{2f}) = 0$ is used to determine the integration constant for $\lambda_{10}^{(i)}$ and where k_1 is an undetermined integration constant.

It is convenient to define a new time variable:

$$(5.8) \quad s = \frac{\tau_{2f} - \tau_2}{\sqrt{2}}$$

for solving the high-order inner solution since the zeroth-order solutions (5.6) and (5.7) both include a term of $(\tau_{2f} - \tau_2)/\sqrt{2}$. For the order of ε^1 , the differential equations for the first-order solutions in terms of the new time variable s are

$$(5.9) \quad \begin{aligned} \frac{dx_{11}^{*(i)}}{ds} &= -\sqrt{2} x_{20}^{(i)}, & \frac{dx_{21}^{(i)}}{ds} &= -\sqrt{2} \lambda_{21}^{(i)}, & \frac{d\lambda_{11}^{(i)}}{ds} &= \sqrt{2} x_{10}^{*(i)}, \\ \frac{d\lambda_{21}^{(i)}}{ds} &= -\sqrt{2}(-3x_{20}^{(i)2} x_{21}^{(i)} + x_{21}^{(i)} - \lambda_{10}^{(i)}). \end{aligned}$$

A second-order differential equation for $x_{21}^{(i)}$ is obtained by using (5.9):

$$(5.10) \quad \frac{d^2 x_{21}^{(i)}}{ds^2} + 2(1 - 3x_{20}^{(i)2})x_{21}^{(i)} = 0,$$

where the solution $\lambda_{10}^{(i)} = 0$ is used to obtain the above equation and where $x_{20}^{(i)} = \tanh s$. The solution to the time-varying equation (5.10) is an important step in determining the inner expansion.

It can be seen from the expansion procedure that the homogeneous solutions for all the high-order solutions of $x_{21}^{(i)}$, $\bar{x}_{2n}^{(i)}$, where $n = 1, 2, 3, \dots$, satisfy

$$(5.11) \quad \frac{d^2 \bar{x}_{2n}^{(i)}}{ds^2} + 2(1 - 3x_{20}^{(i)2})\bar{x}_{2n}^{(i)} = 0.$$

The solution to (5.11) or (5.10) is developed as follows. If the solution $x_{20}^{(i)} = \tanh s$ is substituted into (5.11) and the coefficient of $\bar{x}_{2n}^{(i)}$, $2(1 - 3 \tanh^2 s)$, is expanded in a serial form of s , then (5.11) becomes

$$(5.12) \quad \frac{d^2 \bar{x}_{2n}^{(i)}}{ds^2} - 4 \left[1 + 6 \sum_{i=1}^{\infty} (-1)^i i e^{-2is} \right] \bar{x}_{2n}^{(i)} = 0.$$

Let the solution for $\bar{x}_{2n}^{(i)}$ be assumed as

$$(5.13) \quad \bar{x}_{2n}^{(i)} = \sum_{i=1}^{\infty} a_i e^{-2is}.$$

By substituting (5.13) into (5.12) and equating the coefficients of each order of e^{-2is} to zero, the relationships $a_i = (-1)^{i-1} i a_1$, for $i = 1, 2, 3, \dots$ are obtained. Therefore, the solution for $\bar{x}_{2n}^{(i)}$ becomes

$$(5.14) \quad \bar{x}_{2n}^{(i)} = a_1 \sum_{i=1}^{\infty} (-1)^{i-1} i e^{-2is} = A_1 \left(\frac{1}{\cosh^2 s} \right),$$

where $A_1 = a_1/2$.

It follows that the solution to (5.10) can be found by using (5.14) and applying the method of undetermined coefficients. Let $x_{21}^{(i)}$ in (5.14) be

$$(5.15) \quad x_{21}^{(i)} = A_1(s) \left(\frac{1}{\cosh^2 s} \right),$$

where $A_1(s)$ is a function of s instead of a constant as in (5.14). By substituting (5.15) into (5.10), we obtain the differential equation determining $dA_1(s)/ds$:

$$(5.16) \quad \frac{1}{\cosh^2 s} \frac{d}{ds} \left(\frac{dA_1(s)}{ds} \right) + 2 \frac{d}{ds} \left(\frac{1}{\cosh^2 s} \right) \left(\frac{dA_1(s)}{ds} \right) = 0.$$

Hence, it follows that

$$(5.17) \quad \frac{dA_1(s)}{ds} = B_1 \cosh^4 s,$$

where B_1 is a constant. An integration of (5.17) gives

$$(5.18) \quad A_1(s) = B_1 \left(\frac{3}{8}s + \frac{1}{4} \sinh 2s + \frac{1}{32} \sinh 4s \right) + B_2,$$

where B_2 is a constant. Therefore, the solution to (5.10) becomes

$$(5.19) \quad x_{21}^{(i)} = B_1 \left[\frac{\frac{3}{8}s + \frac{1}{4} \sinh 2s + \frac{1}{32} \sinh 4s}{\cosh^2 s} \right] + B_2 \left(\frac{1}{\cosh^2 s} \right).$$

The solution (5.19) can also be verified by a direct differentiation of $x_{21}^{(i)}$ in (5.10).

It will be shown that the constants B_1 and B_2 must be equal to zero by the following statements. Since $x_{21}^{(i)}$ is an odd function of $(\tau_2 - \tau_{2f})$, that is, $x_{21}^{(i)}(s) = -x_{21}^{(i)}(-s)$, it follows that B_2 must be equal to zero to satisfy this relation. The solution of $x_{21}^{(i)}$ appears in the series form of $x_2^{(i)}$ as $\varepsilon x_{21}^{(i)}$, and the limit of $\lim_{\varepsilon \rightarrow 0} \varepsilon x_{21}^{(i)}$ must be bounded to ensure that the expansion for $x_2^{(i)}$ is an asymptotic series expansion. Since for large s the coefficient of B_1 goes as e^{2s} , then $\lim_{\varepsilon \rightarrow 0} \varepsilon \left[\frac{3}{8}s + \frac{1}{4} \sinh 2s + \frac{1}{32} \sinh 4s \right] / \cosh^2 s$ goes to infinity. It is concluded that B_1 must be equal to zero to bound $\lim_{\varepsilon \rightarrow 0} \varepsilon x_{21}^{(i)}$. Therefore, the solution, $x_{21}^{(i)}$, is equal to zero and higher-order solutions, $x_{2n}^{(i)}$, $n = 1, 2, 3, \dots$, may exist only if a forcing term appears in (5.11).

The first-order solutions for (5.9) are now obtained:

$$(5.20) \quad x_{11}^{*(i)} = -\sqrt{2} \ln \cosh s + k_2, \quad \lambda_{21}^{(i)} = 0, \quad \lambda_{11}^{(i)} = \sqrt{2} k_1 s, \quad x_{21}^{(i)} = 0,$$

where k_1 and k_2 are unknown constants to be determined. The Hamiltonian function $H_1^{(i)}$ for the inner solution is found by substituting the solutions of (5.6), (5.7) and (5.20) into (4.11):

$$(5.21) \quad H_1^{(i)} = \frac{1}{2} k_1^2.$$

For the order of ε^2 , differential equations for the second-order solutions in the inner region are

$$(5.22) \quad \begin{aligned} \frac{dx_{12}^{*(i)}}{ds} &= -\sqrt{2} x_{21}^{(i)}, & \frac{dx_{22}^{(i)}}{ds} &= \sqrt{2} \lambda_{22}^{(i)}, & \frac{d\lambda_{12}^{(i)}}{ds} &= \sqrt{2} x_{11}^{*(i)}, \\ \frac{d\lambda_{22}^{(i)}}{ds} &= -\sqrt{2} (-3x_{20}^{(i)} x_{21}^{(i)2} - 3x_{20}^{(i)2} x_{22}^{(i)} + x_{22}^{(i)} - \lambda_{11}^{(i)}). \end{aligned}$$

By differentiating $d\lambda_{22}^{(i)}/ds$ with respect to s , we derive a second-order differential equation for $x_{22}^{(i)}$ from (5.22):

$$(5.23) \quad \frac{d^2 x_{22}^{(i)}}{ds^2} + 2(1 - 3x_{20}^{(i)2}) x_{22}^{(i)} = 2\lambda_{11}^{(i)},$$

where $2\lambda_{11}^{(i)}$ is the forcing term for the above differential equation. By substituting $x_{20}^{(i)}$ and $\lambda_{11}^{(i)}$ from (5.6) and (5.20) into (5.23), the solution to (5.23) is obtained as follows.

Let the solution to (5.23) be

$$(5.24) \quad x_{22}^{(i)} = A_2(s) \left(\frac{1}{\cosh^2 s} \right).$$

The differential equation for $dA_2(s)/ds$ is obtained by substituting (5.24) into (5.23):

$$(5.25) \quad \frac{1}{\cosh^2 s} \frac{d}{ds} \left(\frac{dA_2(s)}{ds} \right) + 2 \frac{d}{ds} \left(\frac{1}{\cosh^2 s} \right) \left(\frac{dA_2(s)}{ds} \right) = 2\sqrt{2} k_1 s.$$

The solution for $dA_2(s)/ds$ is obtained by solving (5.25) as

$$(5.26) \quad \frac{dA_2(s)}{ds} = \int_{\infty}^s \frac{2\sqrt{2} k_1 s_1}{\cosh^2 s_1} ds_1 \cosh^4 s + D_1 \cosh^4 s,$$

where the lower limit of the integral, ∞ , is chosen such that the solution $x_{22}^{(i)}$ will not grow exponentially and all the exponentially growing terms are included in the constant D_1 . Therefore, the solution for $x_{22}^{(i)}$ is

$$(5.27) \quad x_{22}^{(i)} = \frac{2\sqrt{2} k_1}{\cosh^2 s} \int_0^s \int_{\infty}^{s_2} \frac{s_1}{\cosh^2 s_1} ds_1 \cosh^4 s_2 ds_2 \\ + D_1 \left[\frac{\frac{3}{8}s + \frac{1}{4} \sinh 2s + \frac{1}{32} \sinh 4s}{\cosh^2 s} \right] + D_2 \left(\frac{1}{\cosh^2 s} \right),$$

where D_2 is a constant. The constants D_1 and D_2 must be equal to zero for the same reasons as B_1 and B_2 are zero in (5.19); that is, D_2 is equal to zero to satisfy $x_{22}^{(i)}(s) = -x_{22}^{(i)}(-s)$, and D_1 is equal to zero to bound $\lim_{\varepsilon \rightarrow 0} \varepsilon x_{22}^{(i)}$ since for large s the coefficient of D_1 grows as e^{2s} . Note that for $D_1 = D_2 = 0$, $x_{22}^{(i)}$ behaves as $-k_1 s / \sqrt{2}$ for large s , and $dx_{22}^{(i)}/ds(s=0) \neq 0$ and $x_{22}^{(i)}(s=0) = 0$.

It follows that the second-order solutions for (5.22) are

$$(5.28) \quad x_{12}^{*(i)} = k_3, \\ \lambda_{22}^{(i)} = -\frac{4k_1 \tanh s}{\cosh^2 s} \int_0^s \int_{\infty}^{s_2} \frac{s_1}{\cosh^2 s_1} ds_1 \cosh^4 s_2 ds_2 \\ + 2k_1 \cosh^2 s \int_{\infty}^s \frac{s_1}{\cosh^2 s_1} ds_1, \\ \lambda_{12}^{(i)} = -2 \int_0^s \ln \cosh s_1 ds_1 + \sqrt{2} k_2 s, \\ x_{22}^{(i)} = \frac{2\sqrt{2} k_1}{\cosh^2 s} \int_0^s \int_{\infty}^{s_2} \frac{s_1}{\cosh^2 s_1} ds_1 \cosh^4 s_2 ds_2.$$

The Hamiltonian function $H_2^{(i)}$ is also obtained by using the inner solution and the expression (4.14):

$$(5.29) \quad H_2^{(i)} = k_1(k_2 + \sqrt{2} \ln 2).$$

Therefore, the inner solution can be written in terms of the results in this section as

$$(5.30) \quad x_1^{*(i)} = k_1 + \varepsilon(-\sqrt{2} \ln \cosh s + k_2) + \varepsilon^2(k_3),$$

$$(5.31) \quad x_2^{(i)} = \tanh s + \varepsilon^2 \left[\frac{2\sqrt{2} k_1}{\cosh^2 s} \int_0^s \int_{\infty}^{s_2} \frac{s_1}{\cosh^2 s_1} ds_1 \cosh^4 s_2 ds_2 \right],$$

$$(5.32) \quad \lambda_1^{(i)} = \varepsilon(\sqrt{2} k_1 s) + \varepsilon^2 \left[-2 \int_0^s \ln \cosh s_1 ds_1 + \sqrt{2} k_2 s \right],$$

$$(5.33) \quad \lambda_2^{(i)} = \frac{1}{\sqrt{2} \cosh^2 s} + \varepsilon^2 \left[-\frac{4k_1 \tanh s}{\cosh^2 s} \int_0^s \int_{-\infty}^{s_2} \frac{s_1}{\cosh^2 s_1} ds_1 \cosh^4 s_2 ds_2 \right. \\ \left. + 2k_1 \cosh^2 s \int_{-\infty}^s \frac{s_1}{\cosh^2 s_1} ds_1 \right].$$

The Hamiltonian function $H^{(i)}$ in ascending powers of ε for the inner region is

$$(5.34) \quad H^{(i)} = -\frac{1}{4} + \varepsilon(\frac{1}{2}k_1^2) + \varepsilon^2[k_1(k_2 + \sqrt{2} \ln 2)].$$

Note that the constants k_1 , k_2 and k_3 in (5.30)–(5.34) will be determined by a matching procedure in the next section.

6. Matching of the outer expansion with the inner expansion and the determination of the optimal period. The matching of the outer solution with the inner solution is developed in this section by using a principle proposed by Van Dyke [14]: the m -term inner expansion of the n -term outer solutions is equal to the n -term outer expansion of the m -term inner solution. After the matching is completed, a minimization of the performance index with respect to the period is presented.

The outer expansion in (4.15)–(4.18) and the inner expansion in (5.30)–(5.33) will be used in the following matching. The integrals in (5.31), (5.32) and (5.33) cannot be solved in closed forms. However, this fact does not obstruct the matching procedure since only the limits of those integrals are required for the matching. A three-term inner expansion of the three-term outer solution is obtained by replacing τ_1 in (4.15)–(4.18) with $\tau_1 = \tau_{1f} - \varepsilon(\tau_{2f} - \tau_2)$ and by using a Taylor series expansion of the resulting equation

$$(6.1) \quad x_1^{*(oi)} = \tau_{1f} + \varepsilon[-(\tau_{2f} - \tau_2) + \frac{1}{12}\tau_{1f}^3 - \frac{1}{2}c_1\tau_{1f}] \\ + \varepsilon^2[-(\frac{1}{4}\tau_{1f}^2 - \frac{1}{2}c_1)(\tau_{2f} - \tau_2) - \frac{1}{60}\tau_{1f}^5 + \frac{1}{12}c_1\tau_{1f}^3 + (-\frac{3}{8}c_1^2 - \frac{1}{2}c_2)\tau_{1f}],$$

$$(6.2) \quad x_2^{(oi)} = 1 + \varepsilon[\frac{1}{4}\tau_{1f}^2 - \frac{1}{2}c_1] \\ + \varepsilon^2[-\frac{1}{2}\tau_{1f}(\tau_{2f} - \tau_2) - \frac{1}{12}\tau_{1f}^4 + \frac{1}{4}c_1\tau_{1f}^2 + (-\frac{3}{8}c_1^2 - \frac{1}{2}c_2)],$$

$$(6.3) \quad \lambda_1^{(oi)} = -\frac{1}{2}\tau_{1f}^2 + c_1 + \varepsilon[\tau_{1f}(\tau_{2f} - \tau_2) - \frac{1}{48}\tau_{1f}^4 + \frac{1}{4}c_1\tau_{1f}^2 + c_2] \\ + \varepsilon^2[-\frac{1}{2}(\tau_{2f} - \tau_2)^2 - (-\frac{1}{12}\tau_{1f}^3 + \frac{1}{2}c_1\tau_{1f})(\tau_{2f} - \tau_2) + \frac{1}{360}\tau_{1f}^6 \\ - \frac{1}{48}c_1\tau_{1f}^4 - \frac{1}{2}(-\frac{3}{8}c_1^2 - \frac{1}{2}c_2)\tau_{1f}^2 + c_3],$$

$$(6.4) \quad \lambda_2^{(oi)} = \varepsilon^2(-\frac{1}{2}\tau_{1f}),$$

where the superscript (oi) denotes the inner expansion of the outer variables.

The three-term outer expansion of the three-term inner solution is determined by letting $s = (\tau_{1f} - \tau_1)/(\sqrt{2} \varepsilon)$ and then by expanding the solutions of (5.30)–(5.33) in a Taylor series expansion of ε with $(\tau_{1f} - \tau_1)$ fixed; that is, when $\varepsilon \rightarrow 0$ then $s \rightarrow \infty$. The integral term $\varepsilon^2(-2 \int_0^s \ln \cosh s_1 ds_1)$ in (5.32) is expanded as follows. If the integrand is expanded in a Taylor series of exponential functions for small ε , the integral becomes

$$(6.5) \quad \varepsilon^2 \left(-2 \int_0^s \ln \cosh s_1 ds_1 \right) = \varepsilon^2 \left[-\frac{(\tau_{1f} - \tau_1)^2}{\sqrt{2} \varepsilon^2} - 2 \sum_{n=1}^{\infty} \frac{(-1)^n (\tau_{1f} - \tau_1)}{n} \frac{1}{2\varepsilon} \right. \\ \left. - \sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} \exp \left(-2n \left(\frac{\tau_{1f} - \tau_1}{\sqrt{2} \varepsilon} \right) \right) + \sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} \right] \\ = -\frac{1}{2}(\tau_{1f} - \tau_1)^2 + \varepsilon[\sqrt{2} \ln 2(\tau_{1f} - \tau_1)] + \varepsilon^2 \left(-\frac{\pi^2}{12} \right),$$

where the results of $\exp(-2n((\tau_{1f} - \tau_1)/\sqrt{2}\epsilon)) \rightarrow 0$, and $\sum_{n=1}^{\infty} (-1)^n/n = -\ln 2$, and $\sum_{n=1}^{\infty} (-1)^n/n^2 = -\pi^2/12$ are used to obtain the final form of the above equation. Therefore, by using the above results, we obtain the three-term outer expansion of the three-term inner solution:

$$(6.6) \quad x_1^{*(io)} = k_1 - (\tau_{1f} - \tau_1) + \epsilon(k_2 + \sqrt{2} \ln 2) + \epsilon^2(k_3),$$

$$(6.7) \quad x_2^{(io)} = 1 + \epsilon[-\frac{1}{2}k_1(\tau_{1f} - \tau_1)],$$

$$(6.8) \quad \lambda_1^{(io)} = k_1(\tau_{1f} - \tau_1) - \frac{1}{2}(\tau_{1f} - \tau_1)^2 + \epsilon[(k_2 + \sqrt{2} \ln 2)(\tau_{1f} - \tau_1)] + \epsilon^2\left(-\frac{\pi^2}{12}\right),$$

$$(6.9) \quad \lambda_2^{(io)} = \epsilon^2(-\frac{1}{2}k_1),$$

where the superscript (io) indicates the outer expansion of the inner variables. Note that the substitution of $s = (\tau_{1f} - \tau_1)/\sqrt{2}\epsilon$ is necessary to calculate the order of ϵ in the above Taylor series expansion by fixing $(\tau_{1f} - \tau_1)$. However, in order to match (6.6)–(6.9) with (6.1)–(6.4), $(\tau_{1f} - \tau_1)$ in (6.6)–(6.9) is replaced by $\epsilon(\tau_{2f} - \tau_2)$ such that

$$(6.10) \quad x_1^{*(io)} = k_1 + \epsilon[-(\tau_{2f} - \tau_2) + (k_2 + \sqrt{2} \ln 2)] + \epsilon^2(k_3),$$

$$(6.11) \quad x_2^{(io)} = 1 + \epsilon^2[-\frac{1}{2}k_1(\tau_{2f} - \tau_2)],$$

$$(6.12) \quad \lambda_1^{(io)} = \epsilon[k_1(\tau_{2f} - \tau_2)] + \epsilon^2\left[-\frac{1}{2}(\tau_{2f} - \tau_2)^2 + (k_2 + \sqrt{2} \ln 2)(\tau_{2f} - \tau_2) - \frac{\pi^2}{12}\right],$$

$$(6.13) \quad \lambda_2^{(io)} = \epsilon^2(-\frac{1}{2}k_1).$$

The unknown constants c_1, c_2, c_3, k_1, k_2 and k_3 are determined by equating (6.10)–(6.13) with (6.1)–(6.4). Comparing (6.13) with (6.4) for λ_2 gives

$$(6.14) \quad k_1 = \tau_{1f}.$$

By comparing (6.11) with (6.2) for x_2 , the constants c_1 and c_2 are obtained:

$$(6.15) \quad c_1 = \frac{1}{2}\tau_{1f}^2, \quad c_2 = -\frac{5}{48}\tau_{1f}^4.$$

For a comparison of (6.10) and (6.1), k_2 and k_3 are

$$(6.16) \quad k_2 = -\sqrt{2} \ln 2 - \frac{1}{6}\tau_{1f}^3, \quad k_3 = -\frac{1}{60}\tau_{1f}^5,$$

and it follows that c_3 is obtained by a comparison of (6.12) and (6.3) as

$$(6.17) \quad c_3 = -\frac{19}{1440}\tau_{1f}^6 - \frac{\pi^2}{12}.$$

Finally, the outer solution is written as

$$(6.18) \quad x_1^{*(o)} = \tau_1 + \epsilon\left(\frac{1}{12}\tau_1^3 - \frac{1}{4}\tau_{1f}^2\tau_1\right) + \epsilon^2\left(-\frac{1}{60}\tau_1^5 + \frac{1}{24}\tau_{1f}^2\tau_1^3 - \frac{1}{24}\tau_{1f}^4\tau_1\right),$$

$$(6.19) \quad x_2^{(o)} = 1 + \epsilon\left(\frac{1}{4}\tau_1^2 - \frac{1}{4}\tau_{1f}^2\right) + \epsilon^2\left(-\frac{1}{12}\tau_1^4 + \frac{1}{8}\tau_{1f}^2\tau_1^2 - \frac{1}{24}\tau_{1f}^4\right),$$

$$(6.20) \quad \lambda_1^{(o)} = -\frac{1}{2}\tau_1^2 + \frac{1}{2}\tau_{1f}^2 + \epsilon\left(-\frac{1}{48}\tau_1^4 + \frac{1}{8}\tau_{1f}^2\tau_1^2 - \frac{5}{48}\tau_{1f}^4\right) \\ + \epsilon^2\left(\frac{1}{360}\tau_1^6 - \frac{1}{96}\tau_{1f}^2\tau_1^4 + \frac{1}{48}\tau_{1f}^4\tau_1^2 - \frac{19}{1440}\tau_{1f}^6 - \frac{\pi^2}{12}\right),$$

$$(6.21) \quad \lambda_2^{(o)} = \epsilon^2\left(-\frac{1}{2}\tau_1\right),$$

and the inner solution is given by

$$(6.22) \quad x_1^{*(i)} = \tau_{1f} + \varepsilon \left(-\sqrt{2} \ln \cosh s - \sqrt{2} \ln 2 - \frac{1}{6} \tau_{1f}^3 \right) + \varepsilon^2 \left(-\frac{1}{60} \tau_{1f}^5 \right),$$

$$(6.23) \quad x_2^{(i)} = \tanh s + \varepsilon^2 \left[\frac{2\sqrt{2} \tau_{1f}}{\cosh^2 s} \int_0^s \int_\infty^{s_2} \frac{s_1}{\cosh^2 s_1} ds_1 \cosh^4 s_2 ds_2 \right],$$

$$(6.24) \quad \lambda_1^{(i)} = \varepsilon(\sqrt{2} \tau_{1f} s) + \varepsilon^2 \left[-2 \int_0^s \ln \cosh s_1 ds_1 - 2 \ln 2s - \frac{\sqrt{2}}{6} \tau_{1f}^3 s \right],$$

$$(6.25) \quad \lambda_2^{(i)} = \frac{1}{\sqrt{2} \cosh^2 s} + \varepsilon^2 \left[\frac{4\tau_{1f} - \tanh s}{\cosh^2 s} \int_0^s \int_\infty^{s_2} \frac{s_1}{\cosh^2 s_1} ds_1 \cosh^4 s_2 ds_2 \right. \\ \left. + 2\tau_{1f} \cosh^2 s \int_\infty^s \frac{s_1}{\cosh^2 s_1} ds_1 \right].$$

Note that s is defined in (5.8) and τ_{1f} is the period, which will be determined later in this section. The scaled controls v in the outer and inner regions are the negative functions of (6.21) and (6.25). It can be seen from (6.21) and (6.25) that the scaled control in the outer region is almost zero for a small ε , and the scaled control in the inner region reaches a value of $-1/\sqrt{2}$. Note that the control u can be obtained by using $u = v/b^{1/2} = v/\varepsilon^{3/2}$. Therefore, u is of order ε in the outer region and of order $1/\varepsilon$ in the inner region.

The constants in (6.14)–(6.17) are substituted into (4.19) and (5.34) to determine the Hamiltonian function, and the result of the substitution shows that the two expressions in the outer region and inner region are identically equal to

$$(6.26) \quad H = -\frac{1}{4} + \varepsilon(\frac{1}{2}\tau_{1f}^2) + \varepsilon^2(-\frac{1}{6}\tau_{1f}^4).$$

An extension of the Hamiltonian function up to the order of ε^2 can be made by using the fact that H_3 is a function of the states and the Lagrange multipliers up to the order of ε^2 only; that is,

$$(6.27) \quad H_3 = -\frac{1}{360} \tau_{1f}^6 - \frac{\pi^2}{12}.$$

Therefore, the Hamiltonian function becomes

$$(6.28) \quad H = -\frac{1}{4} + \varepsilon\left(\frac{1}{2}\tau_{1f}^2\right) + \varepsilon^2\left(-\frac{1}{6}\tau_{1f}^4\right) + \varepsilon^3\left(-\frac{1}{360}\tau_{1f}^6 - \frac{\pi^2}{12}\right).$$

The optimal period can be determined as follows. The performance index J is evaluated in the two regions, the outer region and the inner region. After J is expressed in terms of the undetermined τ_{1f} , J is minimized with respect to τ_{1f} to determine the optimal value of τ_{1f} . The performance index J in (3.5) is rewritten by changing the independent variable from τ_1 to s in the inner region

$$(6.29) \quad J = \frac{1}{\tau_{1f}} \left[\int_0^{\tau_{1m}} \left(\varepsilon \frac{x_1^{*(o)2}}{2} - \frac{x_2^{(o)2}}{2} + \frac{x_2^{(o)4}}{4} + \frac{\lambda_2^{(o)2}}{2} \right) d\tau_1 \right. \\ \left. + \sqrt{2} \varepsilon \int_0^{s_m} \left(\varepsilon \frac{x_1^{*(i)2}}{2} - \frac{x_2^{(i)2}}{2} + \frac{x_2^{(i)4}}{4} + \frac{\lambda_2^{(i)2}}{2} \right) ds \right],$$

where $s_m = (\tau_{1f} - \tau_{1m})/(\sqrt{2} \varepsilon)$ and $\tau_{1m} \approx \tau_{1f}$, and where $v^{(o)} = -\lambda_2^{(o)}$ and $v^{(i)} = -\lambda_2^{(i)}$ are used to eliminate $v^{(o)}$ and $v^{(i)}$.

For convenience, the performance index J is written in a serial form of ε

$$(6.30) \quad \begin{aligned} J &= \frac{1}{\tau_{1_f}} (J^{(o)} + J^{(i)}) \\ &= \frac{1}{\tau_{1_f}} (J_0^{(o)} + \varepsilon J_1^{(o)} + \varepsilon^2 J_2^{(o)} + \varepsilon^3 J_3^{(o)} + \cdots + J_0^{(i)} + \varepsilon J_1^{(i)} + \varepsilon^2 J_2^{(i)} + \varepsilon^3 J_3^{(i)} + \cdots), \end{aligned}$$

where the superscripts (o) and (i) denote the cost for the outer region and inner region. By substituting (6.18)–(6.21) into (6.30), $J_0^{(o)}$, $J_1^{(o)}$, and $J_2^{(o)}$ are obtained as

$$(6.31) \quad J_0^{(o)} = -\frac{1}{4}\tau_{1_m}, \quad J_1^{(o)} = \frac{1}{6}\tau_{1_m}^3, \quad J_2^{(o)} = \frac{7}{240}\tau_{1_m}^5 - \frac{1}{8}\tau_{1_f}^2\tau_{1_m}^3 + \frac{1}{16}\tau_{1_f}^4\tau_{1_m}.$$

The term $J_3^{(o)}$ can be evaluated by using the results of the expansion of $x_1^{*(o)}$, $x_2^{(o)}$, and $\lambda_2^{(o)}$ up to the order ε^2 ; this can be seen from the following expression for $J_3^{(o)}$:

$$(6.32) \quad \begin{aligned} J_3^{(o)} &= \int_0^{\tau_{1_m}} \left[\frac{1}{2} x_{11}^{*(o)2} + x_{10}^{*(o)} x_{12}^{*(o)} + x_{20}^{(o)} x_{21}^{(o)3} \right. \\ &\quad \left. + (3x_{20}^{(o)2} - 1)x_{21}^{(o)} x_{22}^{(o)} + (x_{20}^{(o)3} - x_{20}^{(o)})x_{23}^{(o)} \right] d\tau_1 \\ &= -\frac{113}{20160}\tau_{1_m}^7 + \frac{1}{64}\tau_{1_f}^2\tau_{1_m}^5 - \frac{1}{64}\tau_{1_f}^4\tau_{1_m}^3 + \frac{1}{196}\tau_{1_f}^6\tau_{1_m}, \end{aligned}$$

where $x_{20}^{(o)} = 1$ is used to eliminate the term $(x_{20}^{(o)3} - x_{20}^{(o)})x_{23}^{(o)}$. Thus, $x_{23}^{(o)}$ does not appear in the expression for $J_3^{(o)}$.

For the inner region, the performance index is integrated with respect to s , and then $s = (\tau_{1_f} - \tau_{1_m})/(\sqrt{2}\varepsilon)$ is substituted into the resulting equation. In (6.29) the integral for the inner region is multiplied by ε . Therefore, the states $x_1^{*(i)}$, $x_2^{(i)}$ and Lagrange multiplier $\lambda_2^{(i)}$ are expanded to the order of ε^2 , and the integral for the inner region in (6.29) can be evaluated up to the order of ε^3 . A substitution of the serial forms of the states and Lagrange multiplier into (6.29) gives an expression for the cost in the inner region:

$$(6.33) \quad \begin{aligned} J^{(i)} &= \sqrt{2} \int_0^{s_m} \left[\varepsilon \left(-\frac{x_{20}^{(i)2}}{2} + \frac{x_{20}^{(i)4}}{4} + \frac{\lambda_{20}^{(i)2}}{2} \right) + \varepsilon^2 \left(\frac{x_{10}^{*(i)2}}{2} \right) \right. \\ &\quad \left. + \varepsilon^3 [x_{10}^{*(i)} x_{11}^{*(i)} + (x_{20}^{(i)3} - x_{20}^{(i)})x_{22}^{(i)} + \lambda_{20}^{(i)} \lambda_{22}^{(i)}] \right] ds, \end{aligned}$$

where $J^{(i)}$ is simplified by using $x_{21}^{(i)} = 0$, and $\lambda_{21}^{(i)} = 0$. The closed form of $J^{(i)}$ seems to be impossible to obtain since the integration in (6.33) involves a nonintegrable term $x_{22}^{(i)}$. However, it will be shown by the following manipulations that the term $x_{22}^{(i)}$ is eliminated. After substituting the solutions (5.30)–(5.33) into (6.33), the inner cost becomes

$$(6.34) \quad \begin{aligned} J^{(i)} &= \sqrt{2} \int_0^{s_m} \left[\varepsilon \left[-\frac{1}{4} + \frac{(1 - \tanh^2 s)^2}{2} \right] + \varepsilon^2 \left(\frac{\tau_{1_f}^2}{2} \right) + \varepsilon^3 \right. \\ &\quad \left. \cdot \left[\tau_{1_f} \left(-\sqrt{2} \ln \cosh s - \sqrt{2} \ln 2 - \frac{1}{6}\tau_{1_f}^3 \right) + \frac{1}{2} \frac{d}{ds} [(1 - \tanh^2 s)x_{22}^{(i)}] \right] \right] ds. \end{aligned}$$

Therefore, the integration of $J^{(i)}$ gives

$$\begin{aligned}
 J_0^{(i)} &= -\frac{1}{4}(\tau_{1_f} - \tau_{1_m}), \\
 J_1^{(i)} &= \frac{1}{\sqrt{2}} \tanh\left(\frac{\tau_{1_f} - \tau_{1_m}}{\sqrt{2}\epsilon}\right) \\
 &\quad - \frac{1}{3\sqrt{2}} \tanh^3\left(\frac{\tau_{1_f} - \tau_{1_m}}{\sqrt{2}\epsilon}\right) + \frac{1}{2}\tau_{1_f}^2(\tau_{1_f} - \tau_{1_m}) - \frac{1}{2}\tau_{1_f}(\tau_{1_f} - \tau_{1_m})^2, \\
 J_2^{(i)} &= \left(-\frac{1}{6}\tau_{1_f}^4\right)(\tau_{1_f} - \tau_{1_m}), \\
 J_3^{(i)} &= -\frac{\pi^2}{12}\tau_{1_f} - \sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} \exp\left(-2n\left(\frac{\tau_{1_f} - \tau_{1_m}}{\sqrt{2}\epsilon}\right)\right)\tau_{1_f} \\
 &\quad + \frac{\sqrt{2}}{2} \left[\left(1 - \tanh^2\left(\frac{\tau_{1_f} - \tau_{1_m}}{\sqrt{2}\epsilon}\right)\right) x_{22}^{(i)}(s_m) \right],
 \end{aligned}
 \tag{6.35}$$

where the result in (6.5) is used to obtain $-\pi^2/12$ and where $x_{22}^{(i)}(s=0) = 0$ comes from the boundary condition $x_2^{(i)}(s=0) = 0$.

It is noted that τ_{1_m} is the intersection point of the outer expansion and inner expansion and if τ_{1_m} is chosen as $\tau_{1_m} = \tau_{1_f} - (\epsilon\tau_{1_f})^{1/p}$, where $p > 1$, then the following approximations are true:

$$(6.36) \quad \tau_{1_m} \approx \tau_{1_f}, \quad (\tau_{1_f} - \tau_{1_m}) \approx 0, \quad \tanh^n\left(\frac{\tau_{1_f} - \tau_{1_m}}{\sqrt{2}\epsilon}\right) \approx 1, \quad \exp\left(-2n\left(\frac{\tau_{1_f} - \tau_{1_m}}{\sqrt{2}\epsilon}\right)\right) \approx 0,$$

where $n = 1, 2, 3, \dots$. Hence, by using the approximations of (6.36) and the results of (6.31), (6.32) and (6.35), the performance index J is given by

$$(6.37) \quad J = -\frac{1}{4} + \epsilon\left(\frac{\sqrt{2}}{3\tau_{1_f}} + \frac{1}{6}\tau_{1_f}^2\right) + \epsilon^2\left(-\frac{1}{30}\tau_{1_f}^4\right) + \epsilon^3\left(-\frac{\pi^2}{12} - \frac{1}{2520}\tau_{1_f}^6\right).$$

Since the performance index J in (6.37) is minimized with respect to τ_{1_f} , the first derivative of J with respect to τ_{1_f} must be zero; that is,

$$(6.38) \quad \frac{\partial J}{\partial \tau_{1_f}} = \epsilon\left(-\frac{\sqrt{2}}{3\tau_{1_f}^2} + \frac{1}{3}\tau_{1_f}\right) + \epsilon^2\left(-\frac{2}{15}\tau_{1_f}^3\right) + \epsilon^3\left(-\frac{1}{420}\tau_{1_f}^5\right) = 0.$$

To obtain τ_{1_f} , τ_{1_f} is assumed as a serial form of ϵ

$$(6.39) \quad \tau_{1_f} = \tau_{1_f0} + \epsilon\tau_{1_f1} + \epsilon^2\tau_{1_f2} + \dots$$

By substituting (6.39) into (6.38), an expression of ascending powers of ϵ is obtained, and it follows that the coefficients in (6.39) are determined by setting every coefficient of (6.38) for each order of ϵ zero:

$$(6.40) \quad \tau_{1_f} = 2^{1/6} + \epsilon \frac{2 \cdot 2^{3/6}}{15} + \epsilon^2 \frac{463 \cdot 2^{5/6}}{6300}.$$

Therefore, the performance index is obtained by substituting the expression for τ_{1_f} in (6.40) into (6.37):

$$(6.41) \quad J = -\frac{1}{4} + \epsilon \frac{1}{2^{2/3}} - \epsilon^2 \frac{2^{2/3}}{30} - \epsilon^2 \left(\frac{\pi^2}{12} + \frac{117}{6300} \right).$$

In terms of the original system parameter b and period T , (6.40) and (6.41) are rewritten as

$$(6.42) \quad T = 4b^{1/6} \left(2^{1/6} + b^{1/3} \frac{2 \cdot 2^{3/6}}{15} + b^{2/3} \frac{463 \cdot 2^{5/6}}{6300} \right),$$

$$(6.43) \quad J = -\frac{1}{4} + b^{1/3} \frac{1}{2^{2/3}} - b^{2/3} \frac{2^{2/3}}{30} - b \left(\frac{\pi^2}{12} + \frac{117}{6300} \right).$$

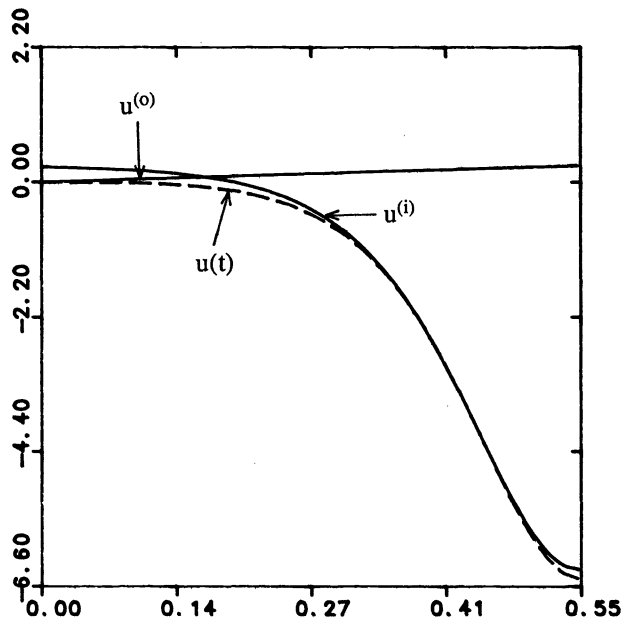
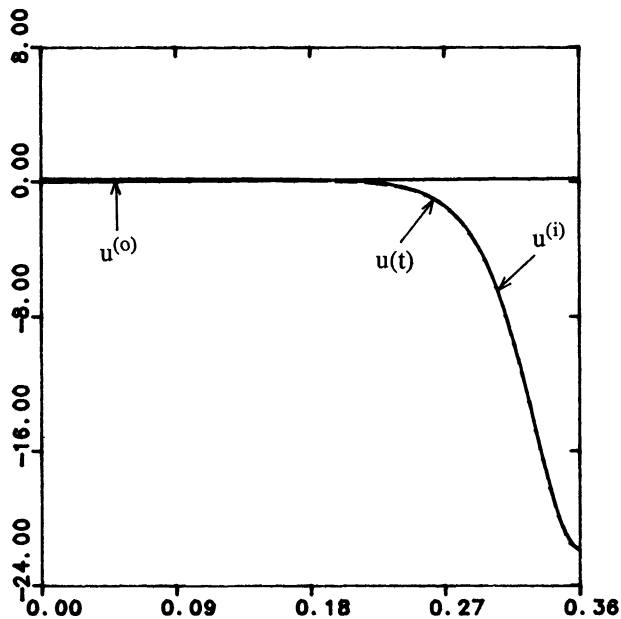
The transversality condition (2.8), which was not used for the expansion of τ_{1r} , is verified by substituting τ_{1r} in (6.40) back into (6.28) and comparing the resulting equation with (6.43). This ensures that both the outer and inner expansions are correct.

The asymptotic expansion is compared to the numerical optimal path. Due to the assumed symmetry, the numerical optimal path is obtained by minimizing the average cost over one-fourth of the period subject to the boundary conditions $x_1(0) = 0$, $\lambda_2(0) = 0$, $x_2(T/4) = 0$, $\lambda_1(T/4) = 0$, and $H = J^o$. The numerical algorithm of [4] is used to find the optimal solution. First, all the boundary conditions are satisfied except $H = J^o$, and then, a shooting method is used to search the family solutions, which are indexed by the period, until all the boundary conditions are satisfied. The comparison presented in Table 1 shows good agreement of the exact and approximate solutions for several values of the expansion parameter. The outer, inner, and numerical solutions for the control u are presented in Fig. 1 for $b = 0.01$ and in Fig. 2 for $b = 0.001$. For $b = 0.001$, the asymptotic expansion is almost identical to the numerical solution indicating that the matching procedure works well in this case.

7. Conclusions. An asymptotic expansion for an optimal relaxation oscillator has been obtained in terms of a small parameter b of the system by proper scalings of the outer and inner variables. The expressions for the states, Lagrange multipliers, control, optimal period and performance index have also been derived. Since the system is autonomous, the Hamiltonian function is a constant and equal in both regions, and its related transversality condition $H = J^o$ is shown to be satisfied. The performance index is equal to $-\frac{1}{4}$, and the period is equal to zero for the chattering solution of the

TABLE 1
Comparative results.

b	ε		Series solution	Numerical solution
0.01	0.215443	$x_2(0)$	0.922761	0.926674
		$\lambda_1(0)$	0.129654	0.135114
		$u(T/4)$	-6.31876	-6.48800
		$T/4$	0.542678	0.545912
		J	-0.125146	-0.123484
0.005	0.170998	$x_2(0)$	0.940427	0.943486
		$\lambda_1(0)$	0.104998	0.105603
		$u(T/4)$	-9.33586	-9.45965
		$T/4$	0.479076	0.480376
		J	-0.148031	-0.147349
0.001	0.1	$x_2(0)$	0.966649	0.967318
		$\lambda_1(0)$	0.0626525	0.0627276
		$u(T/4)$	-21.8598	-21.9192
		$T/4$	0.361331	0.361497
		J	-0.188374	-0.188289

FIG. 1. Inner, outer and numerical solutions for $b = 0.01$.FIG. 2. Inner, outer and numerical solutions for $b = 0.001$.

problem in which the system parameter b is made to be zero. For nonzero b , the performance index of the optimal solution is greater than $-\frac{1}{4}$, and the period is no longer zero. Comparison of the expansion of the cost and period agree extremely well with the numerical solution.

The results of this paper demonstrate a procedure for expanding an optimal solution about the high-frequency chattering solution obtained from a reduced-order

system. An outer expansion is obtained by scaling time to the order of the optimal period, and an inner expansion is obtained when the fast variable approaches the switch point by using a scaling faster than that used in the outer expansion. A careful study of the chattering solution and formulation here may give clues as to how to scale the outer and inner variables for more complex problems.

REFERENCES

- [1] F. J. M. HORN AND R. C. LIN, *Periodic processes: a variational approach*, *Indust. Engrg. Chem. Process Design Development*, 6 (1967), pp. 21–30.
- [2] J. L. SPEYER, *Nonoptimality of the steady-state cruise for aircraft*, *J. Aircraft*, 14 (1976), pp. 1604–1610.
- [3] E. G. GILBERT AND M. G. PARSONS, *Periodic control and the optimality of aircraft cruise*, *J. Aircraft*, 13 (1976), pp. 828–830.
- [4] C.-H. CHUANG AND J. L. SPEYER, *Periodic optimal hypersonic scramjet cruise*, *Optimal Control Appl. Methods*, 8 (1987), pp. 231–242.
- [5] S. C. HOULIHAN, E. M. CLIFF AND H. J. KELLEY, *Study of chattering cruise*, *J. Aircraft*, 19 (1982), pp. 119–124.
- [6] J. L. SPEYER AND R. T. EVANS, *A second variation theory for optimal periodic processes*, *IEEE Trans. Automat. Control*, AC-29 (1984), pp. 138–148.
- [7] R. T. EVANS, J. L. SPEYER AND C.-H. CHUANG, *Solution of a periodic optimal control problem by asymptotic series*, *J. Optim. Theory Appl.*, 52 (1987), pp. 343–364.
- [8] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, John Wiley, New York, 1985.
- [9] A. H. NAYFEH, *Introduction to Perturbation Techniques*, John Wiley, New York, 1981.
- [10] E. F. MISCENKO AND N. KH. ROZOV, *Differential Equations with Small Parameters and Relaxation Oscillations*, Plenum Press, New York, 1980.
- [11] J. GRASMAN, *Asymptotic Methods for Relaxation Oscillations and Applications*, Springer-Verlag, New York, 1987.
- [12] D. S. BERNSTEIN AND E. G. GILBERT, *Optimal periodic control: the π test revisited*, *IEEE Trans. Automat. Control* (1980), pp. 673–684.
- [13] S. BITTANTI, G. FRONZA AND G. GUARDABASSI, *Periodic control: a frequency domain approach*, *IEEE Trans. Automat. Control*, AC-18 (1973), pp. 33–38.
- [14] M. VAN DYKE, *Perturbation Methods in Fluid Mechanics*, Academic Press, New York, 1964.

NOT ALL FEEDBACK STABILIZED HYPERBOLIC SYSTEMS ARE ROBUST WITH RESPECT TO SMALL TIME DELAYS IN THEIR FEEDBACKS*

RICHARD DATKO†

Abstract. We present two examples of hyperbolic partial differential equations which are stabilized by boundary feedback controls and then destabilized by small delays in these controls. We show that in a general case, when the controls are distributed, stabilized hyperbolic systems possess nontrivial periodic solutions if small time delays are introduced into their feedbacks. We also indicate by means of an example that the general case of this phenomenon is harder to demonstrate for boundary control problems.

Key words. stabilization, hyperbolic, delays

AMS(MOS) subject classifications. 93D15, 93C20, 35L35, 35R10

Introduction. The purpose of this paper is to indicate that certain hyperbolic partial differential equations involving Neumann type boundary conditions which stabilize the system are not robust with respect to arbitrarily small delays in the boundary conditions. These boundary conditions are of feedback type in that they depend on certain time derivatives of the solution evaluated on a part of the boundary. The lack of robustness with respect to time delays means that when arbitrarily small delays are introduced into their feedbacks the resulting system has periodic or even exponentially increasing solutions. A previous paper [3] demonstrated this phenomenon for the one-dimensional wave equation. In this paper, in §§ 1 and 2, we supply further examples of this "anomaly." In § 1 we consider a stabilized Euler beam equation and in § 2 the two-dimensional wave equation on a square.

The question naturally arises as to whether this behavior is able to be demonstrated for general classes of hyperbolic systems. The answer to this is probably yes. As yet we have not been able to accomplish this generalization. However in § 3 we do indicate, by means of an example, where the main difficulty seems to be in such an extension. We also show in this section that for distributed systems in a Hilbert space H of the form

$$(0.1) \quad \ddot{x}(t) + B\dot{x}(t) + Ax(t) = 0,$$

where A is an unbounded positive definite self-adjoint linear operator on H and B is a linear symmetric operator on H such that $A^{-1/2}BA^{-1/2}$ is compact, an arbitrarily small delay ε in the first derivative of (0.1) which leads to the system

$$(0.2) \quad \ddot{x}(t) + B\dot{x}(t - \varepsilon) + Ax(t) = 0$$

results in nontrivial periodic solutions of (0.2).

Thus for distributed control systems of the form

$$(0.3) \quad \ddot{x}(t) + Ax(t) = Cu(t),$$

where C is a linear mapping from a Hilbert space $H_1 \rightarrow H$, the control $u(t) = -C^*\dot{x}(t)$ (C^* is the adjoint mapping), which in some sense stabilizes (0.3), is not robust with respect to small time delays if $A^{-1/2}CC^*A^{-1/2}$ is compact. Hence for distributed controls of the form (0.3) there is a general theory concerning destabilization with respect to small delays.

* Received by the editors May 12, 1986; accepted for publication (in revised form) July 17, 1987.

† Mathematical Reviews, 416 Fourth Street, Ann Arbor, Michigan; Georgetown University, Department of Mathematics, Washington, D.C.

1. Consider the system

$$(1.1) \quad u_{tt} + u_{xxxx} = 0, \quad 0 < x < 1, \quad t > 0,$$

$$(1.2) \quad u(0, t) = u_x(0, t) = u_{xxx}(1, t) = 0,$$

$$(1.3) \quad u_{xx}(1, t) = -u_{xt}(1, t - \varepsilon) \quad \text{where } \varepsilon \geq 0 \text{ is fixed.}$$

The initial values of u and u_t are

$$(1.4) \quad u(x, t) = \varphi(x, t),$$

$$(1.5) \quad u_t(x, t) = \psi(x, t),$$

for $0 \leq x \leq 1$, $-\varepsilon \leq t \leq 0$, and for simplicity we assume that φ and ψ are infinitely differentiable in (x, t) .

The Laplace transform of (1.1) with the initial conditions (1.4) and (1.5) satisfies the two-parameter family of ordinary differential equations in x

$$(1.6) \quad \omega^2 U(x, \omega, \varepsilon) + \frac{d^4 U}{dx^4}(x, \omega, \varepsilon) = \omega \varphi(x, 0) + \psi(x, 0).$$

If we let

$$(1.7) \quad \sqrt{\omega} e^{(\pi/4)i} = \tau, \quad \omega = -i\tau^2$$

and apply the first two conditions in (1.2) we can write the solutions of (1.6) in the form

$$(1.8) \quad \begin{aligned} U(x, \tau, \varepsilon) &= A(\cosh \tau x - \cos \tau x) + B(\sinh \tau x - \sin \tau x) \\ &+ \frac{1}{2\tau^3} \int_0^x [\sinh \tau(x - \sigma) - \sin \tau(x - \sigma)][-i\tau^2 \varphi(\sigma) + \psi(\sigma)] d\sigma \\ &= A(\cosh \tau x - \cos \tau x) + B(\sinh \tau x - \sin \tau x) + U_p(x, \tau). \end{aligned}$$

The following statement is obvious but will be needed below so we state it as a proposition.

PROPOSITION 1.1. (i) *The expression $U_p(x, \tau)$ (implicitly defined in (1.8)) is an entire function in τ .*

(ii) *The term $(\cosh \tau x - \cos \tau x)$ in (1.8) has a zero of order two at $\tau = 0$.*

(iii) *The term $(\sinh \tau x - \sin \tau x)$ in (1.8) has a zero of order three at $\tau = 0$.*

To obtain the coefficients A and B in (1.8) we apply the last boundary condition in (1.2) and the boundary condition (1.3) via their Laplace transforms. The first condition leads to the equation

$$(1.9) \quad \begin{aligned} \tau^3 [A(\sinh \tau - \sin \tau) + B(\cosh \tau + \cos \tau)] \\ = -\frac{1}{2} \int_0^1 (\cosh \tau(1 - \sigma) + \cos \tau(1 - \sigma))(-i\tau^2 \varphi(\sigma) + \psi(\sigma)) d\sigma. \end{aligned}$$

The second condition results in the equation

$$(1.10) \quad \begin{aligned} \tau^2 [A(\cosh \tau + \cos \tau) + B(\sinh \tau + \sin \tau)] \\ \cdot \frac{1}{2\tau} \int_0^1 (\sinh \tau(1 - \sigma) + \sin \tau(1 - \sigma))(-i\tau^2 \varphi(\sigma) + \psi(\sigma)) d\sigma \\ = i\tau^3 e^{i\varepsilon\tau^2} [A(\sinh \tau + \sin \tau) + B(\cosh \tau - \cos \tau)] \\ + i\tau^2 \frac{dU_p}{dx}(1, \tau) + \varphi_x(1, -\varepsilon) + i\tau^2 e^{i\varepsilon\tau^2} \int_{-\varepsilon}^0 \varphi_x(\sigma, 1) e^{i\tau^2\sigma} d\sigma. \end{aligned}$$

Dividing (1.9) by τ^3 and (1.10) by τ^2 and rearranging terms we obtain the following equations for the solution of A and B :

$$(1.11) \quad A(\sinh \tau - \sin \tau) + B(\cosh \tau + \cos \tau) = \frac{1}{\tau^3} f_1(\tau),$$

$$(1.12) \quad A[(\cosh \tau + \cos \tau) - i\tau e^{i\epsilon\tau^2}(\sinh \tau + \sin \tau)] \\ + B[(\sinh \tau + \sin \tau) - i\tau e^{i\epsilon\tau^2}(\cosh \tau - \cos \tau)] = g(\tau) + \frac{1}{\tau^2} f_2(\tau),$$

where f_1, f_2 and g are entire functions of τ . The matrix equation associated with (1.11) and (1.12) has the structure

$$(1.13) \quad \begin{bmatrix} (\sinh \tau - \sin \tau) & (\cosh \tau + \cos \tau) \\ (\cosh \tau + \cos \tau) + \tau q_1(\tau) & (\sinh \tau + \sin \tau) + \tau q_2(\tau) \end{bmatrix} \begin{pmatrix} A \\ B \end{pmatrix} \\ = \begin{pmatrix} \frac{1}{\tau^3} f_1(\tau) \\ g(\tau) + \frac{1}{\tau^2} f_2(\tau) \end{pmatrix} = R(\tau) \begin{pmatrix} A \\ B \end{pmatrix},$$

where $q_1(\tau)$ and $q_2(\tau)$ are entire functions of τ . Notice that, at $\tau=0$, $\det R(0) \neq 0$. The inversion of (1.13) leads to the following equations for A and B :

$$(1.14) \quad A(\tau) = \frac{(\sinh \tau + \sin \tau + \tau q_2(\tau)) f_1(\tau)}{\det R(\tau) \tau^3} - \frac{(\cosh \tau + \cos \tau)(g(\tau) + f_2(\tau)/\tau^2)}{\det R(\tau)},$$

$$(1.15) \quad B(\tau) = \frac{-(\cosh \tau + \cos \tau + \tau q_1(\tau)) f_1(\tau)}{\det R(\tau) \tau^3} + \frac{(\sinh \tau - \sin \tau) \left(g(\tau) + \frac{f_2(\tau)}{\tau^2} \right)}{\det R(\tau)}.$$

Thus (1.14) and (1.15) show that $A(\tau)$ has a pole of order at most two and $B(\tau)$ a pole of order at most three at $\tau=0$. Hence by Proposition 1.1 it follows that $U(x, \tau, \epsilon)$ has no poles at $\tau=0$, and we can state the following proposition.

PROPOSITION 1.2. *The poles of $U(x, \tau, \epsilon)$ are determined by the zeros of $\det R(\tau)$, i.e., by*

$$(1.16) \quad F(\tau, \epsilon) = -\frac{1}{2} \det R(\tau) = (1 + \cosh \tau \cos \tau) - i\tau e^{i\epsilon\tau^2}(\sinh \tau \cos \tau + \cosh \tau \sin \tau) = 0.$$

Proof. By the discussion preceding the statement of Proposition 1.2 we know that $\tau=0$ is not a pole of $U(x, \tau, \epsilon)$ and by Proposition 1.1 that $U_p(x, \tau)$ is entire. Thus the only poles of (1.8), i.e., of $U(x, \tau, \epsilon)$, that can occur are in the coefficients $A(\tau)(\cosh \tau x - \cos \tau x)$ and $B(\tau)(\sinh \tau x - \sin \tau x)$. By (1.14) and (1.15) these occur where $\det R(\tau)=0$, which proves the proposition.

Chen et al. have proven in [1] that when $\epsilon=0$ the system (1.1)–(1.5) is uniformly exponentially stable. We shall show, using elementary techniques, that for $\epsilon=0$, system (1.1)–(1.5) has at most a finite number of poles and that these, if they exist, must be in the left half plane $\operatorname{Re} \omega < 0$.

We first consider the energy functional associated with (1.1)–(1.5) when $\epsilon=0$:

$$(1.17) \quad E(t) = \frac{1}{2} \int_0^1 [u_t^2(x, t) + u_{xx}^2(x, t)] dx.$$

Along trajectories of (1.1)–(1.5) an easy computation shows that

$$(1.18) \quad \frac{dE}{dt}(t) = -u_{xt}^2(1, t) \leq 0.$$

Thus if a pole of $U(x, \omega, 0)$ occurs in $\operatorname{Re} \omega > 0$, the energy function associated with the real part of the inverse Laplace transform of that solution would have to increase exponentially, which because of (1.18) is absurd. Hence the eigenvalues of (1.1)–(1.5) or equivalently the poles of $U(x, \omega, 0)$ must be in $\operatorname{Re} \omega \leq 0$.

PROPOSITION 1.3. *The poles of $U(x, \omega, 0)$ are at most finite in number and if they exist lie in $\operatorname{Re} \omega < 0$.*

Proof. The proof is in two parts. (i) If $\omega = i\omega_0$, ω_0 real, is a pole of $U(x, \omega, 0)$, then by (1.7) the corresponding value of τ is either real or imaginary, depending on whether ω_0 is negative or positive ($\omega_0 = 0$ is clearly impossible). But since the poles of $U(x, \omega, 0)$ must satisfy (1.16) this cannot occur.

(ii) To prove there can exist at most a finite number of poles in $\operatorname{Re} \omega = 0$ we assume the contrary, that is, there are an infinite number of poles in $\operatorname{Re} \omega < 0$ of the form

$$(1.19) \quad \omega = -2\alpha + i\beta, \quad \tau = a + ib,$$

where α, β, a and b are real. (We omit using sequence notation for economy of expression.) Since $\tau^2 = i\omega$, (1.19) reduces to the two real equations

$$(1.20) \quad b^2 - a^2 = \beta, \quad \alpha = -ab.$$

Clearly if $|\omega| \rightarrow \infty$, then $|\tau| \rightarrow \infty$, and, from (1.20), we see that since $\alpha > 0$, a and b must be of opposite sign. Also observe that, when $\tau = a + ib$ and $\varepsilon = 0$, (1.16) can be rewritten as

$$(1.21) \quad 1 + \frac{1}{4}(e^{a+ib} + e^{-a-ib})(e^{ia-b} + e^{-ia+b}) + \frac{b-ia}{4}(e^{a+ib} - e^{-a-ib})(e^{ia-b} + e^{-ia+b}) \\ + \frac{b-ia}{4i}(e^{a+ib} + e^{-a-ib})(e^{ia-b} - e^{-ia+b}) = 0.$$

Since the left-hand side of (1.21) is an entire function in $\tau = a + ib$, it follows that the zeros of (1.21) can have no finite point of accumulation in the complex τ plane. Hence if an infinite number of zeros of (1.21) exist, one of the following three cases must hold for some infinite sequence of these zeros (we again omit the sequence notation for convenience of expression):

- (a) $|a| \rightarrow \infty, |b| \rightarrow 0, |b| \rightarrow \infty, |a| \rightarrow 0,$
- (b) $|a| \rightarrow \infty, \lim |b| \rightarrow b_0 > 0, |b| \rightarrow \infty, \lim |a| \rightarrow a_0 > 0,$
- (c) $|a| \rightarrow \infty, |b| \rightarrow \infty.$

We shall now show each case is impossible and hence that (1.21) can have at most a finite number of zeros in the τ plane which is equivalent to saying that for $\varepsilon = 0$, $U(x, \omega, 0)$ has at most a finite number of poles.

Proof that case (a) is impossible. Assume $a \rightarrow \infty$ and $b \rightarrow 0$. Then (1.21) may be written as follows:

$$(1.22) \quad 1 + \frac{e^{a+ib}}{4}(1 + e^{-2a}q_1(a, b))(e^{ia-b} + e^{-ia+b}) \\ + \frac{(e^{a+ib})}{4}(b-ia)(1 - e^{-2a}q_1(a, b))(e^{ia-b} + e^{-ia+b}) \\ + \frac{e^{a+ib}}{4i}(b-ia)(1 + e^{-2a}q_1(a, b))(e^{ia-b} - e^{-ia+b}) = 0,$$

where $|q_1(a, b)| < q_0 < \infty$ for all a, b . Multiplying both sides of (1.22) by $e^{-(a+ib)}$ and observing that $b \rightarrow 0$, we conclude that the following expression must tend to zero as $a \rightarrow \infty$.

(1.23)

$$\frac{1}{4}(e^{ia} + e^{-ia}) + \left(\frac{-ia}{4}\right)(e^{ia} + e^{-ia}) - \frac{ia}{4}(e^{ia} - e^{-ia}) = \frac{1}{2} \cos a - \frac{ia}{2}(\cos a + \sin a) \rightarrow 0.$$

But this is impossible since it would imply that $\cos a \rightarrow 0$ and $\cos a + \sin a \rightarrow 0$. The case when $a \rightarrow -\infty$ and $b \rightarrow 0$ is similarly proved. In this case we factor out e^{-a-ib} in (1.21).

When $b \rightarrow \infty$ and $|a| \rightarrow 0$, we factor out e^{-ia+b} in (1.21) to obtain

$$\begin{aligned} (1.24) \quad & e^{ia-b} + \frac{1}{4}(e^{a+ib} + e^{-a-ib})(e^{-2b}q_2(a, b) + 1) \\ & + \frac{b-ia}{4}(e^{a+ib} - e^{-a-ib})(e^{-2b}q_2(a, b) + 1) \\ & + \frac{b-ia}{4i}(e^{a+ib} + e^{-a-ib})(e^{-2b}q_2(a, b) - 1) = 0, \end{aligned}$$

where $|q_2(a, b)| < q_0 < \infty$ for all a, b . Since $a \rightarrow 0$ this implies that

$$\frac{1}{2} \cos b + \frac{ib}{2}(\sin b + \cos b) \rightarrow 0$$

as $b \rightarrow \infty$, which is impossible.

The cases where $a \rightarrow -\infty$, $b \rightarrow 0$ or $b \rightarrow -\infty$, $a \rightarrow 0$ can be treated similarly. Thus case (a) is not possible.

Proof that case (b) is impossible. Assume $b \rightarrow \infty$ and $a \rightarrow a_0 \neq 0$, a constant. Then factoring out e^{-ia+b} in (1.21) we obtain (1.24). Dividing (1.24) by $(b-ia)$ we conclude that

$$\begin{aligned} & \frac{1}{4}(e^{a+ib} - e^{-a-ib}) - \frac{1}{4i}(e^{a+ib} + e^{-a-ib}) \\ & = \frac{e^a}{4}[(\cos b - \sin b) + i(\sin b + \cos b)] \\ & \quad + \frac{e^{-a}}{4}[(-\cos b + \sin b) + i(\sin b + \cos b)] \\ & = \frac{(e^a - e^{-a})}{4}(\cos b - \sin b) + i \frac{(e^a + e^{-a})}{4}(\cos b + \sin b) \rightarrow 0. \end{aligned}$$

But, unless $e^a \rightarrow 1$, this is impossible. Thus, since case (a) does not hold, neither does case (b) for $b \rightarrow \infty$, $a \rightarrow a_0$. The other three possibilities $b \rightarrow -\infty$; $a \rightarrow a_0 \neq 0$, $a \rightarrow \pm\infty$; $b \rightarrow b_0 \neq 0$ are treated in the same manner and in each situation a reduction to case (a) results which has already been shown to be impossible.

Proof that case (c) is impossible. We assume $a \rightarrow \infty$ and $-b \rightarrow \infty$. (The other three cases are similar, as will be seen by the treatment.) Dividing (1.21) by $e^{a+ib}e^{ia-b}((b-ia)/4)$ we arrive at the contradiction

$$(1.25) \quad \frac{1}{4}(1-i) \rightarrow 0 \quad \text{as } a^2 + b^2 \rightarrow \infty.$$

This concludes the proof of the proposition.

Remark 1.1. Proposition 1.3 may not seem as unusual as it first appears if we recall that the system (1.1), (1.4), (1.5) with (1.2) and (1.3) replaced by

$$(1.2)' \quad u(0, t) = u_x(0, t) = u_{xx}(1, t) = 0,$$

$$(1.3)' \quad u_{xxx}(1, t) = \varphi(t)$$

(φ —the control) is controllable between two arbitrary finite energy states in an arbitrarily short time for some $\varphi \in L^2[0, T]$, T -arbitrary (see, e.g., [7, § 4]).

PROPOSITION 1.4. *There exist $\omega = -i\omega_n$, $\omega_n \rightarrow \infty$ and $\varepsilon_n > 0$, $\varepsilon_n \rightarrow 0$, which satisfy (1.16) when ω is substituted for τ via the relations (1.7).*

Proof. Let $n = 1, 2, \dots$, and consider the interval $I_n = (2n\pi - \pi/2, 2n\pi - \pi/4)$. Choose ε_n such that $\tau^2 \varepsilon_n = \pi/2$. Then the left-hand side of (1.16) is

$$(1.26) \quad F\left(\tau, \frac{\pi}{2\tau^2}\right) = (1 + \cosh \tau \cos \tau) + \tau(\sinh \tau \cos \tau + \cosh \tau \sin \tau).$$

Notice that the value of (1.26) at the right-hand end point of I_n satisfies the condition

$$1 + \frac{\sqrt{2}}{2} \cosh\left(2n\pi - \frac{\pi}{4}\right) + \left(2n\pi - \frac{\pi}{4}\right) \frac{\sqrt{2}}{2} \left(\sinh\left(2n\pi - \frac{\pi}{4}\right) - \cosh\left(2n\pi - \frac{\pi}{4}\right)\right) > 0$$

for n sufficiently large, whereas the left-hand endpoint of (1.26) satisfies

$$1 - \left(2n\pi - \frac{\pi}{2}\right) \cosh\left(2n\pi - \frac{\pi}{2}\right) < 0$$

for all n . Thus for n sufficiently large we can find a $\tau_n \in I_n$ for which (1.26) is zero, and when $\varepsilon_n = \pi/2\tau_n^2$ and $\tau = \tau_n$, (1.16) is satisfied. The corresponding value of ω is from (1.7) $\omega_n = -i\tau_n^2$ which proves the proposition.

We proved in Proposition 1.4 that there exists

$$(1.27) \quad \tau_n^2 = -i\omega_n \quad \text{and} \quad \varepsilon_n = \frac{\pi}{2\tau_n^2} \quad \text{with} \quad \tau_n \rightarrow \infty$$

such that (1.16) is satisfied for these values of $-i\omega_n$ and ε_n . This is equivalent to stating that for these values of ε_n the spectrum of (1.1)–(1.5) has points on the imaginary axis. We shall now show that there exist $\varepsilon_n \rightarrow 0$ such that the spectrum of (1.1)–(1.5) has points in the right half plane. First notice that (1.16) is actually a function of ω and ε , because of the relations (1.7), which implicitly define the functions $\omega(\varepsilon)$ or $\varepsilon(\omega)$. We seek to compute their derivatives at the points which satisfy Proposition 1.4. The partial derivatives of F are:

$$\begin{aligned} \frac{\partial F}{\partial \omega} &= \frac{\partial F}{\partial \tau} \frac{\partial \tau}{\partial \omega} = \frac{i}{2\tau} \frac{\partial F}{\partial \tau} \\ &= \frac{i}{2\tau} [(\sinh \tau \cos \tau - \cosh \tau \sin \tau) - ie^{i\varepsilon\tau^2}(\sinh \tau \cos \tau + \cosh \tau \sin \tau) \\ &\quad + 2\tau^2 \varepsilon e^{i\varepsilon\tau^2}(\sinh \tau \cos \tau + \cosh \tau \sin \tau) - 2i\tau e^{i\varepsilon\tau^2} \cosh \tau \cos \tau], \\ \frac{\partial F}{\partial \varepsilon} &= \tau^3 e^{i\varepsilon\tau^2}(\sinh \tau \cos \tau + \cosh \tau \sin \tau). \end{aligned}$$

Evaluating $\partial F/\partial \omega$ and $\partial F/\partial \varepsilon$ at the points obtained in Proposition 1.4 and noting that $e^{i\varepsilon_n \tau_n^2} = i$ we obtain the following expression:

$$(1.28) \quad \left. \frac{\partial F}{\partial \omega} / \frac{\partial F}{\partial \varepsilon} \right|_{\substack{\varepsilon = \varepsilon_n \\ \omega = -i\omega_n}} = \frac{1}{2\tau_n^4} \frac{\sinh \tau_n \cos \tau_n - \cosh \tau_n \sin \tau_n + 2\tau_n(\cosh \tau_n \cos \tau_n)}{\sinh \tau_n \cos \tau_n + \cosh \tau_n \sin \tau_n} + iq(\tau_n),$$

where $q(\tau_n)$ is real. The denominator in the first term on the right in (1.28) cannot be zero for large values of n since, because (1.16) must also be satisfied, this would imply that $|\cos \tau_n| \cong 1/\sqrt{2}$ and $1 + \cosh \tau_n \cos n = 0$ which is impossible for n sufficiently large. Similarly the numerator of the first term on the right in (1.28) cannot be zero, since this would imply $\cos \tau_n \cong 0$, which is impossible if (1.16) is satisfied.

Since (1.28) has the form $a + ib$, where $a \neq 0$, and since $-1/(a + ib) = -(a - ib)/(a^2 + b^2)$ this implies that

$$(1.29) \quad \operatorname{Re} \left(\frac{d\omega}{d\varepsilon} \right) \bigg|_{\varepsilon = \varepsilon_n, \omega = -i\omega_n^2} \neq 0.$$

Hence for ε "near" ε_n (1.16) has solutions in $\operatorname{Re} \omega > 0$. Thus we can state the following result.

THEOREM 1.1. *The system (1.1)–(1.5) has for $\varepsilon = 0$ at most a finite point spectrum which must lie in $\operatorname{Re} \omega < 0$ if it exists. However there exist values $\varepsilon_n \rightarrow 0$ for which the resulting system has its spectrum in $\operatorname{Re} \omega > 0$.*

Proof. The proof follows from Proposition 1.2 which essentially states that the spectrum of (1.1)–(1.5) is determined by (1.16), Proposition 1.3 and (1.28), which states that for some sufficiently small values of $\varepsilon \rightarrow 0$, (1.16) has solutions in $\operatorname{Re} \omega > 0$.

2. Consider the two-dimensional wave equation described by

$$(2.1) \quad u_{tt}(x, y, t) = u_{xx}(x, y, t) + u_{yy}(x, y, t), \quad 0 < x < 1, \quad 0 < y < 1, \quad t > 0,$$

$$(2.2) \quad u(0, y, t) = u(x, 0, t) = u(1, y, t) = 0,$$

$$(2.3) \quad u_y(x, 1, t) = -u_t(x, 1, t - \varepsilon), \quad t \geq \varepsilon, \quad \text{where } \varepsilon \geq 0 \text{ is a fixed constant,}$$

$$(2.4) \quad u(x, y, 0) = f(x, y), \quad u_t(x, y, 0) = g(x, y).$$

Condition (2.3) may be viewed as a boundary feedback control. In fact we shall show (Proposition 2.5 below) that for $\varepsilon = 0$ it places the point spectrum of (2.1)–(2.4) in the left half of the complex plane.

Remark 2.1. We shall not be unduly concerned with the space of initial conditions (2.4) except to assume that they represent functions which are real valued and analytic in the variables x and y , i.e., we want to ensure well-posedness of the problem at hand in terms of L_2 representations of the solutions with regard to the variables x and y .

DEFINITION 2.1. For a given f, g and ε we shall denote the solution (2.1)–(2.4) by $S(f, g, \varepsilon)$.

Notation. Since frequent reference will be made to complex functions whose arguments depend on the term $(s^2 + n^2\pi^2)^{1/2}$ (s is complex and n is a natural number), we let

$$(2.5) \quad (s^2 + n^2\pi^2)^{1/2} = \varphi(s, n).$$

The Laplace transform with respect to the t variable of $S(f, g, \varepsilon)$ for a fixed $(x, y) \in (0, 1) \times (0, 1)$ will be denoted by

$$(2.6) \quad U(x, y, s, \varepsilon) = \int_0^\infty e^{-st} u(x, y, t) dt.$$

For f, g and n fixed we denote by $r_n(\sigma, f, g, s)$ the function

$$(2.7) \quad r_n(\sigma, f, g, s) = \int_0^1 \sin n\pi\tau (sf(\tau, \sigma) + g(\tau, \sigma)) d\tau.$$

DEFINITION 2.2. By the spectrum of (2.1)–(2.4) we mean the set of points in the complex plane for which the Laplace transform of some nontrivial solution has a pole.

The following propositions will be used to prove the main result.

PROPOSITION 2.1. An energy function for (2.1)–(2.4) in the case $\varepsilon = 0$ is

$$(2.8) \quad E(t) = \int_0^1 \int_0^1 [u_x^2(x, y, t) + u_y^2(x, y, t) + u_t^2(x, y, t)] dx dy$$

and its derivative along solutions is

$$(2.9) \quad \frac{dE}{dt}(t) = - \int_0^1 u_t^2(x, 1, t) dt.$$

Proof. The proof is by direct calculation using conditions (2.2) and (2.3).

PROPOSITION 2.2. For f, g fixed the Laplace transform of $S(f, g, \varepsilon)$ is given by the expression

$$(2.10) \quad \begin{aligned} U(x, y, s) &= \sum_{n=1}^{\infty} \frac{\sin n\pi x}{\sinh \varphi(s, n)} \left[A_n(s) \sinh \varphi(s, n) y \right. \\ &\quad + \int_0^y \frac{(\sinh \varphi(s, n)(1-y))(\sinh \varphi(s, n)\sigma) r_n(\sigma, f, g, s)}{\varphi(s, n)} d\sigma \\ &\quad \left. + \int_y^1 \frac{(\sinh \varphi(s, n)y)(\sinh \varphi(s, n)(1-\sigma)) r_n(\sigma, f, g, s)}{\varphi(s, n)} d\sigma \right] \\ &= \sum_{n=1}^{\infty} \frac{\sin n\pi x}{\sinh \varphi(s, n)} \left[A_n(s) \sinh \varphi(s, n) y + \int_0^1 G_n(y, \sigma, s) r_n(\sigma, f, g, s) d\sigma \right]. \end{aligned}$$

Proof. The proof is a direct application of the Laplace transform to (2.1), (2.2), (2.4). A description of the procedure may be found in [2, Ex. 3.3, p. 14].

PROPOSITION 2.3. If $\varepsilon = 0$, then $A_n(s)$ in (2.10) satisfies the equation

$$(2.11) \quad \frac{A_n(s)}{\sinh \varphi(s, n)} = \left[\frac{1}{\varphi(s, n) \cosh \varphi(s, n) + s \sinh \varphi(s, n)} \right] \cdot \left[f_n + \int_0^1 \frac{(\sinh \varphi(s, n)\sigma) r_n(\sigma, f, g, s)}{\sinh \varphi(s, n)} d\sigma \right],$$

where f_n is the n th coefficient of the sine expansion

$$(2.12) \quad f(x, 1) = \sum_{n=1}^{\infty} f_n \sin n\pi x.$$

If $\varepsilon > 0$, then $A_n(s)$ satisfies

$$(2.13) \quad \frac{A_n(s)}{\sinh \varphi(s, n)} = \left[\frac{1}{\varphi(s, n) \cosh \varphi(s, n) + s e^{-s\varepsilon} \sinh \varphi(s, n)} \right] \cdot \left[b_n(s) + \int_0^1 \frac{(\sinh \varphi(s, n)\sigma) r_n(\sigma, f, g)}{\sinh \varphi(s, n)} d\sigma \right],$$

where $b_n(s)$ is the entire function of the n th coefficient of the sine expansion

$$e^{-s\varepsilon}f(x, 1, 0) - \int_{-\varepsilon}^0 e^{-s(\sigma+\varepsilon)}u_t(x, l, \sigma)d\sigma = \sum_{n=1}^{\infty} b_n(s) \sin n\pi x.$$

Proof. The proof is to observe that the Laplace transform of (2.3) for $\varepsilon = 0$ is

$$(2.14) \quad U_y(x, 1, s) = -sU(x, 1, s) + f(x, 1)$$

and for $\varepsilon > 0$ is

$$(2.15) \quad U_y(x, 1, s) = -e^{-s\varepsilon} \left[sU(x, 1, s, \varepsilon) - u(x, 1, 0) + \int_{-\varepsilon}^0 e^{-s\sigma}u_t(x, 1, \sigma)d\sigma \right].$$

Then we apply (2.14) or (2.15) to (2.10) to obtain (2.11) or (2.13), respectively.

PROPOSITION 2.4. *The poles of $U(x, y, s, 0)$ for a given f and g are determined by the zeros of the function*

$$(2.16) \quad \tau(s, n) = \cosh \varphi(s, n) + \frac{s}{\varphi(s, n)} \sinh \varphi(s, n), \quad n = 1, 2, \dots$$

Proof. When $\varepsilon = 0$ we see, using (2.11), that $U(x, y, s)$ can be written by Propositions 2.2 and 2.3 as

$$(2.17) \quad U(x, y, s) = \sum_{n=1}^{\infty} \frac{(\sin n\pi x) \sinh \varphi(s, n)y}{\varphi(s, n) \cosh \varphi(s, n) + s \sinh \varphi(s, n)} \cdot \left[f_n + \int_0^1 \frac{(\sinh \varphi(s, n))r_n(\sigma, f, g)}{\sinh \varphi(s, n)} d\sigma + \int_0^1 \frac{G_n(y, \sigma, s)r_n(\sigma, f, g, s)}{\sinh \varphi(s, n)} d\sigma \right].$$

Using l'Hôpital's rule we can exclude points of the form $s = \pm n\pi i$, $n = 1, 2, \dots$ as poles, since the numerators and denominators in (2.17) have zeros which cancel each other out at such points. Hence the only poles of (2.17) occur when

$$(2.18) \quad \varphi(s, n) \cosh \varphi(s, n) + s \sinh \varphi(s, n) = 0,$$

$s \neq \pm n\pi i$. But another way of phrasing this is to look for the zeros of

$$\tau(s, n) = \cosh \varphi(s, n) + \frac{s \sinh \varphi(s, n)}{\varphi(s, n)} = 0.$$

This proves the proposition.

PROPOSITION 2.5. *For each n the set of zeros of $\tau(s, n)$ lies in $\operatorname{Re} s < 0$ and is uniformly bounded away from the imaginary axis.*

Proof. It follows from Proposition 2.1 that $U(x, y, s, 0)$ can have no poles in $\operatorname{Re} s > 0$ since $E(t)$ is decreasing for each set of initial values (f, g) . There can also be no zeros of $\tau(s, n)$ on the imaginary axis since elementary calculations show that $|\tau(s, n)| \geq 1$ if $s = i\omega$. Hence the only zeros of $\tau(s, n)$ lie in $\operatorname{Re} s < 0$. Thus assume there is a sequence of zeros of $\tau(s, n)$ of the form $s_m = -\delta_m + i\omega_m$, where $0 < \delta_m$ and $\delta_m \rightarrow 0$ as $m \rightarrow \infty$. Clearly $|\omega_m| \rightarrow \infty$ since $|\tau(s, n)| \geq 4$ on the imaginary axis. But if $|\omega_m| \rightarrow \infty$ and $\delta_m \rightarrow 0$ then using the identities $\cosh i\omega = \cos \omega$ and $\sinh i\omega = i \sin \omega$ we can write for m sufficiently large

$$(2.19) \quad \tau(s_m, n) = \cos \sqrt{\omega_m^2 - n^2 \pi^2} + i \frac{\omega_m \sin \sqrt{\omega_m^2 - n^2 \pi^2}}{\sqrt{\omega_m^2 - n^2 \pi^2}} + \varepsilon(s_m),$$

where $|\varepsilon(s_m)| \rightarrow 0$ as $m \rightarrow \infty$. But then

$$\lim_{m \rightarrow \infty} |\tau(s_m, n)| = 1$$

which is impossible. This proves the proposition.

PROPOSITION 2.6. *If $\varepsilon > 0$, then the spectrum of (2.1)–(2.4), i.e., of $S(f, g, \varepsilon)$, is given by the zeros of*

$$(2.20) \quad \psi(s, n, \varepsilon) = \cosh \varphi(s, n) + \frac{se^{-\varepsilon s}}{\varphi(s, n)} \sinh \varphi(s, n).$$

Proof. Using (2.11) and (2.13) we can write $U(s, y, s, \varepsilon)$ in the form

$$(2.21) \quad U(x, y, s, \varepsilon) = \sum_{n=1}^{\infty} \frac{(\sin n\pi x) \sinh \varphi(s, n)y}{\varphi(s, n) \cosh \varphi(s, n) + se^{-\varepsilon s} \sinh \varphi(s, n)} \cdot \left[b_n(s) + \int_0^1 \frac{(\sinh \varphi(s, n)\sigma) r_n(\sigma, f, g) d\sigma}{\sinh \varphi(s, n)} + \frac{\int_0^1 G_n(y, \sigma, s) r_n(\sigma, f, g, s) d\sigma}{\sinh \varphi(s, n)} \right].$$

Arguing as in the proof of Proposition 2.4 we see that the poles (2.21) are determined by the zeros of

$$\cosh \varphi(s, n) + \frac{se^{-\varepsilon s}}{\varphi(s, n)} \sinh \varphi(s, n) = 0,$$

which proves the proposition.

PROPOSITION 2.7. *For any $r > 0$ there exists an $\varepsilon > 0$ such that $0 < \varepsilon < r$ and such that (2.20), i.e., $\psi(s, n, \varepsilon)$, has a zero on the imaginary axis.*

Proof. Let $\omega > n\pi$ and set $s = i\omega$. Then,

$$(2.22) \quad \psi(i\omega, n, \varepsilon) \cos \sqrt{\omega^2 - n^2\pi^2} + \frac{i\omega e^{-i\omega\varepsilon} \sin \sqrt{\omega^2 - n^2\pi^2}}{\sqrt{\omega^2 - n^2\pi^2}}.$$

If (2.22) has a zero, then

$$(2.23) \quad \frac{i\omega}{\sqrt{\omega^2 - n^2\pi^2}} \tan \sqrt{\omega^2 - n^2\pi^2} = e^{i\omega\varepsilon}$$

must be satisfied. Since the right side of (2.23) has absolute value equal to one, we shall seek a value of ω for which

$$(2.24) \quad \frac{\omega}{\sqrt{\omega^2 - n^2\pi^2}} \tan \sqrt{\omega^2 - n^2\pi^2} = 1.$$

Such a choice is possible since when $\sqrt{\omega^2 - n^2\pi^2} = 2m\pi + \pi/3$, m a positive integer, or zero, the left side of (2.24) is greater than $\sqrt{3}$ and when $\sqrt{\omega^2 - n^2\pi^2} = 2m\pi$ the left side of (2.24) is zero. Hence, by continuity, in each interval of the form

$$(2.25) \quad I = (\sqrt{4m^2\pi^2 + n^2\pi^2}, \sqrt{(2m\pi + \pi/3)^2 + n^2\pi^2})$$

there exists a zero $\hat{\omega}$ of (2.24). Set

$$(2.26) \quad \hat{\omega} = 2k\pi + \frac{\pi}{2}, \quad k = 0, 1, \dots$$

For these values of $\hat{\omega}$ and ε , (2.23) is satisfied. But then $i\hat{\omega}$ is a zero of (2.22) and

$$(2.27) \quad 0 < \varepsilon = \frac{2k\pi + \pi/2}{\hat{\omega}} < \frac{2k\pi + \pi/2}{\sqrt{4m\pi^2 + n^2\pi^2}}.$$

Since the right-hand side of (2.27) can be made arbitrarily small by choosing m large this proves the proposition.

PROPOSITION 2.8. *Given $r > 0$ there exists an ε_0 and s_0 , with $\operatorname{Re} s_0 > 0$ and $0 < \varepsilon_0 < r$, such that $\varphi(s_0, n\varepsilon_0) = 0$. That is, $U(x, y, s, \varepsilon_0)$ has a pole at s_0 where $\operatorname{Re} s_0 > 0$.*

Proof. Let $\hat{\varepsilon}$ and $i\omega$ satisfy Proposition 2.8 for a given $r > 0$. We shall compute $\partial\varphi/\partial s$ and $\partial\varphi/\partial\varepsilon$ at (ε, ω) and show that an implicitly defined function $s(\varepsilon)$ exists in a neighborhood of $\hat{\varepsilon}$ whose derivative is not zero at $\hat{\varepsilon}$ and is such that $\varphi(s(\varepsilon), n, \varepsilon) = 0$ in that neighborhood. Thus

$$(2.28) \quad \begin{aligned} \frac{\partial\psi}{\partial s} &= \frac{s}{\varphi(s, n)} \sinh \varphi(s, n) - \frac{s^2 e^{-\varepsilon s}}{(\varphi(s, n))^3} \sinh \varphi(s, n) \\ &\quad - \frac{\varepsilon s e^{-\varepsilon s}}{\varphi(s, n)} \sinh \varphi(s, n) + \frac{s^2 e^{\varepsilon s}}{(\varphi(s, n))^2} \cosh \varphi(s, n) + e^{\varepsilon s} \frac{\sinh \varphi(s, n)}{\varphi(s, n)} \end{aligned}$$

and

$$(2.29) \quad \frac{\partial\psi}{\partial\varepsilon} = -\frac{s^2 e^{-\varepsilon s} \sinh \varphi(s, n)}{\varphi(s, n)}.$$

At $s = i\omega$ and $\varepsilon = \hat{\varepsilon}$ the following relations, which are consequences of (26), the fact that $e^{i\omega\hat{\varepsilon}} = i$, $\sinh is = i \sin s$ and $\cosh is = \cos s$ hold:

$$(2.30) \quad \frac{\omega}{\sqrt{\omega^2 - n^2\pi^2}} \tan \sqrt{\omega^2 - \pi^2} = 1,$$

$$(2.31) \quad \varphi(i\omega) = i\sqrt{\omega^2 - n^2\pi^2},$$

$$(2.32) \quad i\omega\hat{\varepsilon} = \left(2k\pi + \frac{\pi}{2}\right)i, \quad k = 0, 1, \dots$$

Substituting the above into (2.28) and (2.29) we obtain

$$(2.33) \quad \begin{aligned} \frac{1}{\cosh \varphi(s, n)} \frac{\partial\psi}{\partial s} \Big|_{s=i\omega, \varepsilon=\hat{\varepsilon}} &= i + \frac{i\omega}{\omega^2 - n^2\pi^2} - \hat{\varepsilon} + \frac{i}{\omega^2 - n^2\pi^2} - \frac{i}{\omega} \\ &= -\hat{\varepsilon} + i \left(1 + \frac{\omega}{\omega^2 - n^2\pi^2} + \frac{1}{\omega^2 - n^2\pi^2} - \frac{1}{\omega}\right), \end{aligned}$$

$$(2.34) \quad \frac{1}{\cosh \varphi(s, n)} \frac{\partial\psi}{\partial\varepsilon} \Big|_{s=i\omega, \varepsilon=\hat{\varepsilon}} = i\omega.$$

Thus from (2.33) and (2.34) we deduce that there exists an implicitly defined function $s(\varepsilon)$ such that $\psi(s(\varepsilon), n, \varepsilon) = 0$ and

$$(2.35) \quad \frac{ds}{d\varepsilon} \Big|_{\varepsilon=\hat{\varepsilon}} = i\omega / \left(-\hat{\varepsilon} + i \left(1 + \frac{\omega}{\omega^2 - n^2\pi^2} + \frac{1}{\omega^2 - n^2\pi^2} - \frac{1}{\omega}\right)\right).$$

Since ω may be chosen to be arbitrarily large, the imaginary term in the denominator of (2.35) can be made to approach one. Consequently

$$(2.36) \quad \frac{ds}{d\varepsilon} \Big|_{\varepsilon=\hat{\varepsilon}} = a + ib,$$

where $a \neq 0$. Relation (2.36) implies that as ε varies on any sufficiently small interval $I(\hat{\varepsilon}) = (\hat{\varepsilon} - \delta_1, \hat{\varepsilon} + \delta_1)$, $\delta_1 > 0$, $\operatorname{Re} s(\varepsilon) > 0$ for some ε in $I(\hat{\varepsilon})$, which prove the proposition.

DEFINITION 2.3. Let $X_n = \{t(x, y): t(x, y) = (\sin n\pi x)q(y), \text{ where } q \text{ is analytic on the set } [0, 1]\}$.

By Propositions 2.4 and 2.6 any point in the spectrum of (2.1)–(2.4), where $\varepsilon \geq 0$, has its corresponding eigenfunctions in X_n for some n . Moreover if $f(x, y) = u(x, y, 0)$ and $g(x, y) = u_t(x, y, 0)$ are in any X_n for some n then, using (2.7), (2.17) and (2.21), we see that the corresponding solution of (2.1)–(2.4) for any $\varepsilon \geq 0$ is also in X_n for each fixed $t > 0$. Thus we can state the following theorem which is the main result of this section.

THEOREM 2.1. For each $t > 0$, fixed, and $\varepsilon \geq 0$ the solution of (2.1)–(2.4) with initial values at $t = 0$ in X_n is also in X_n . The eigenfunctions associated with any point of the spectrum of (2.1)–(2.4) lie in X_n for some n . For $\varepsilon = 0$ and n fixed the solutions of (2.1)–(2.4) in X_n have a spectrum which lies in $\operatorname{Re} s \leq -\delta_n$, where $\delta_n > 0$. However given any $r > 0$, there exists an ε_0 , $0 < \varepsilon_0 < r$, such that at least one point of the spectrum of the projected system lies in $\operatorname{Re} s > 0$.

Remark 2.2. It would be most satisfactory to state that for $\varepsilon = 0$ the system (2.1)–(2.4) has its spectrum uniformly bounded away from the imaginary axis. However our boundary feedback prohibits this. What Theorem 2.1 does state is that for each n fixed the projected solutions of (2.1)–(2.4) on X_n have a “jump” in the real part of their spectrum across the imaginary axis of width at least δ_n as ε changes from 0 to $\varepsilon > 0$.

3. Distributed control versus boundary control. In this section we shall indicate that when the stabilizing control for a hyperbolic system is distributed a type of destabilization phenomenon similar to that exhibited by the examples in the previous sections is easily obtained. We first prove a general result for linear hyperbolic ordinary differential equations in a Hilbert space and then compare it vis-à-vis a simple example of the wave equation with boundary feedback.

Thus let H be an infinite-dimensional real Hilbert space with inner product (\cdot, \cdot) and norm $|\cdot|$ and let H_c be the complex extension of H . We assume A is an unbounded linear self-adjoint positive definite operator on H with domain $\mathcal{D}(A)$ whose inverse A^{-1} is compact. Let $A^{1/2}$ denote the positive square root of A and let B be a linear self-adjoint operator on H such that (i) $\mathcal{D}(B) \cap \mathcal{D}(A)$ is dense in H , (ii) $A^{-1/2}BA^{-1/2}$ is compact when considered as a mapping on H and (iii) the ordinary differential equation defined on $H \times H$ by (3.1):

$$(3.1) \quad \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} -B & A^{1/2} \\ -A^{1/2} & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}, \quad t > 0,$$

has “mild” solutions on $H \times H$ and strong solutions for a dense set $\hat{\mathcal{D}}$ of initial values

$$\left\{ \left(\begin{pmatrix} x(0) \\ y(0) \end{pmatrix} \right) = \left(\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \right) \right\}$$

on $H \times H$.

If B is positive in the sense that $(Bx, x) > 0$ for $x \in \mathcal{D}(B)$, $|x| \neq 0$, then the solutions of (3.1) are nonincreasing with respect to the functional

$$(3.2) \quad V(t) = \frac{1}{2}[(x(t), x(t)) + (y(t), y(t))],$$

since strong solutions of (3.1) satisfy

$$(3.3) \quad \frac{dV}{dt} = -(Bx(t), x(t)) \leq 0.$$

Let $\varepsilon > 0$ be fixed and consider the delay system

$$(3.4i) \quad \begin{pmatrix} \dot{x}(t) \\ \dot{y}(t) \end{pmatrix} = \begin{pmatrix} 0 & A^{1/2} \\ -A^{1/2} & 0 \end{pmatrix} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} + \begin{pmatrix} -B & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x(t-\varepsilon) \\ y(t-\varepsilon) \end{pmatrix}, \quad t > \varepsilon,$$

$$(3.4ii) \quad \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} \varphi(t) \\ \psi(t) \end{pmatrix}, \quad t \in [-\varepsilon, 0].$$

THEOREM 3.1. *For any $\varepsilon_0 > 0$ there exists an $\varepsilon_1, 0 < \varepsilon_1 < \varepsilon_0$, such that (3.4) possesses a nontrivial strong periodic solution.*

The proof of Theorem 3.1 is based on the following lemma.

LEMMA 3.1. *The eigenvalue problem*

$$(3.5) \quad [\lambda^2 I - \lambda B - A]x = 0,$$

$|x| = 1$ (I is the identity mapping in H), has an infinite number of eigenvalues $\{\omega_n\}$ such that $|\omega_n| \rightarrow \infty$.

Proof. Equation (3.5) is equivalent to the equation

$$(3.6) \quad \begin{pmatrix} \lambda I - B & -A^{1/2} \\ -A^{1/2} & \lambda I \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

in $H \times H$.

If

$$(3.7) \quad \hat{A} = \begin{pmatrix} B & A^{1/2} \\ A^{1/2} & 0 \end{pmatrix},$$

then

$$(3.8) \quad \hat{A}^{-1} \begin{pmatrix} 0 & A^{-1/2} \\ A^{-1/2} & -A^{-1/2}BA^{-1/2} \end{pmatrix}.$$

By hypothesis \hat{A}^{-1} is compact and self-adjoint in $H \times H$; hence there exists an infinite number of solutions of (3.6), i.e., eigenvalues of \hat{A} satisfying the conclusion of the lemma (see e.g. [4, VII, 4.1]).

Proof of Theorem 3.1. We seek $x_0 \in H$, $\omega \in \mathbf{R}^1$ and $\varepsilon > 0$, with $\varepsilon < \varepsilon_0$, such that $e^{-i\omega\varepsilon} = -i$ and $x_0 e^{i\omega t}$ is a solution in H_ε of (3.4).

Such a triple must satisfy

$$(3.9) \quad [-\omega^2 I + i\omega e^{-i\varepsilon\omega} B + A]x_0 e^{i\omega t} = [-\omega^2 I + \omega B + A]x_0 e^{i\omega t} = 0.$$

By Lemma 3.1, (3.9) has an unbounded sequence $\{\omega_n\}$ of real eigenvalues. Thus for each n we select $\varepsilon_n = \pi/2\omega_n$ if $\omega_n > 0$ and $\varepsilon_n = -(3/2)(\pi/\omega_n)$ if $\omega_n < 0$. In either case for each n (3.9) is satisfied for some $x_n \in H$, $|x_n| = 1$.

Clearly, $\varepsilon_n \rightarrow 0$ as $|\omega_n| \rightarrow \infty$ and

$$(3.10) \quad x(t) = (\cos \omega_n t)x_0$$

satisfies (3.4) for $\varepsilon = \varepsilon_n$, which proves the theorem.

Remark 3.1. If $H = \mathbf{R}^1$ and $B > 0$ (3.4) is uniformly asymptotically stable for $0 \leq \varepsilon \leq \varepsilon_0$, where

$$\varepsilon_0 = \min \left[\frac{\pi}{B + \sqrt{B^2 + 4A}}, \frac{3\pi}{B - \sqrt{B^2 + 4A}} \right],$$

and has a periodic solution for $\varepsilon = \varepsilon_0$ (see e.g. [5]). Thus, if we let $A \rightarrow \infty$ and $B/A \rightarrow 0$, we see that $\varepsilon_0 \rightarrow 0$.

Hence, if the abstract system (3.4) can be reduced to an infinite number of uncoupled harmonic oscillators of the form

$$(3.11) \quad \ddot{x}_n + B_n \dot{x}_n(t - \varepsilon) + A_n x_n(t) = 0, \quad n = 1, 2, \dots,$$

where $A_n \rightarrow \infty$ and $B_n/A_n \rightarrow 0$, there can exist no minimum interval $[0, \varepsilon_0)$ for which the overall system is asymptotically stable. Clearly system (3.11) is a special system (3.4), where $H = l_2$.

Remark 3.2. Theorem 3.1 has the following obvious application to distributed control stabilization problems. Suppose we are given the abstract problem

$$(3.12) \quad \ddot{x}(t) + Ax(t) = Cu(t), \quad t > 0, \quad lx(0) = x_0, \quad \dot{x}(0) = x_0,$$

where $x(t) \in H$ (a real Hilbert space), A is a closed unbounded linear self-adjoint positive definite operator on H with a dense domain and $C: H_1 \rightarrow H$ is a closed linear mapping from a real Hilbert space H_1 into H such that C^* , the adjoint of C , and C satisfy the conditions $\mathcal{D}(CC^*) \cap \mathcal{D}(A)$ is dense in H , CC^* is closed and

$$(3.13) \quad A^{-1/2}CC^*A^{-1/2} = K$$

is a compact mapping.

If $u(t)$ in (3.12) is defined by

$$(3.14) \quad -u(t) = -C^*\dot{x}(t)$$

and the resulting system

$$(3.15) \quad \ddot{x}(t) + CC^*\dot{x}(t) + Ax(t) = 0$$

is asymptotically stable in some sense, then the conditions of Theorem 3.1 hold and we can always find an $\varepsilon_1 > 0$ sufficiently small such that

$$(3.16) \quad \ddot{x}(t) + CC^*\dot{x}(t - \varepsilon) + Ax(t) = 0$$

has a nontrivial periodic solution.

We might suppose that Remark 3.2 could be extended to hyperbolic systems which are stabilized by boundary feedback and such may be the case. This would then partially obviate the necessity of examples such as the ones given in §§ 1 and 2. However the next example, which is a special case of one given in [3], indicates that even when analogous formulations of type (3.12) are given to boundary control problems the extension is nontrivial. The reason is that the operator C in (3.12) is often either a distributional operator or (3.12) may be equivalent to variational problems of the type considered in Lions' book [6, Chap. 4, §§ 7.1-7.2], which lead to operators CC^* which are not closed in H .

Example 3.1. Consider the one-dimensional wave equation

$$(3.17) \quad u_{tt}(x, t) = u_{xx}(x, t), \quad 0 < x < 1, \quad t > 0,$$

with initial conditions

$$(3.18) \quad u(x, 0) = \tau(x), \quad u_t(x, 0) = \gamma(x)$$

and boundary conditions

$$(3.19) \quad u_x(1, t) = v(t), \quad u(0, t) = 0.$$

It is easy to show that all solutions of (3.17)-(3.19) (with reasonable initial data) are zero for $t > 2$ if $v(t)$ in (3.19) satisfies

$$(3.20) \quad v(t) = -u_t(1, t).$$

Moreover we can always find $\varepsilon > 0$, arbitrarily small, such that if (3.19) is replaced by

$$(3.21) \quad u_x(1, t) = -u_t(1, t - \varepsilon), \quad u(0, t) = 0$$

the resulting system is unstable (see e.g. [3]). What we wish to do is frame (3.17)–(3.19) in a form similar to (3.12).

Using the formal procedure described in Lions [6, pp. 342–345] we define for $T > 0$ the family of functions

$$\Phi = \{\varphi: \varphi \in L^2[(0, T); H^1(0, 1)]; \varphi_t \in L^2[(0, T) \times (0, 1)]; \varphi_{tt} - \varphi_{xx} \in L^2[(0, T) \times (0, 1)]; \\ \varphi_x(1, t) = 0, \varphi(0, t) = 0, \varphi(x, T) - \varphi_t(x, T) = 0\},$$

where $H^1(0, 1)$ and L^2 denote the usual Sobolev and square integrable spaces, respectively.

We can easily show that the solutions of (3.17)–(3.19) satisfy the equation

$$(3.22) \quad \int_0^T \int_0^1 u(\varphi_{tt} - \varphi_{xx}) dx dt = - \int_0^1 \tau(x) \varphi_t(x, 0) dx + \int_0^1 v(x) \varphi(x, 0) dx \\ + \int_0^T v(t) \varphi(1, t) dt$$

for all $\varphi \in \Phi$.

We express the solutions of (3.17)–(3.19) in the form

$$(3.23) \quad u(x, t) = \sum_{n=0}^{\infty} \sqrt{2} \left(\sin \left(n\pi + \frac{\pi}{2} \right) x \right) g_n(t),$$

where g_n are to be determined using (3.22). (Note that $\sqrt{2} \sin(n\pi + \pi/2)x$, $n = 0, 1, \dots$, satisfy the problem $w_{xx}'' = \lambda_n w''$, $w''(0) = 0$, $w''(1) = 0$, $\int_0^1 w''(x) w^j(x) dx = \delta_{n,j}$, where $\delta_{n,j}$ is the Kronecker delta.)

For each $n = 0, 1, \dots$, we consider $\varphi \in \Phi$ of the form

$$(3.24) \quad \varphi(x, t) = \sqrt{2} \left(\sin \left(n\pi + \frac{\pi}{2} \right) x \right) \psi(t), \quad \psi(T) = \dot{\psi}(T) = 0.$$

Substituting (3.23) and (3.24) into (3.22) and integrating by parts with respect to t on the left-hand side we obtain the infinite system of equations

$$(3.25) \quad \int_0^T \left[\ddot{g}_n + \sqrt{2}(-1)^{n+1}v + \left(n\pi + \frac{\pi}{2} \right)^2 g_n \right] \psi dt + \dot{g}_n(0)\psi(0) - g_n(0)\dot{\psi}(0) = 0.$$

Thus the formal solution of the original problem assumes the Hilbert space formulation of (3.12) (with one exception) when $H = l_2$. To see this observe that the operator corresponding to A in (3.12) is

$$(3.26) \quad A = \text{diag} \left[\left(\frac{\pi}{2} \right)^2, \left(\pi + \frac{\pi}{2} \right)^2, \dots, \left(n\pi + \frac{\pi}{2} \right)^2, \dots \right]^T.$$

However the C operator obtained from (3.25) is

$$(3.27) \quad C = \sqrt{2}(1, -1, 1, \dots, (-1)^n, \dots)^T,$$

which unfortunately is not a vector in l_2 . Moreover CC^* is not a closed operator in l_2 even though it is symmetric. In fact CC^* is the infinite square matrix

$$(3.28) \quad CC^* = (c_{ij}) \quad \text{where } c_{ij} = 2(-1)^{i+j}, \quad i, j = 0, 1, \dots.$$

It can be easily seen that the domain of CC^* in l_2 consists of all sequences $a = (a_0, a_1, \dots)$ such that

$$(3.29) \quad \sum_{j=0}^{\infty} (-1)^j a_j = 0.$$

Moreover for such an a

$$(3.30) \quad CC^*a = 0.$$

It is also easy to show that the domain of CC^* is dense in l_2 .

Interestingly enough for A defined by (3.26) and CC^* by (3.28),

$$(3.31) \quad A^{-1/2}CC^*A^{-1/2} = Q = (q_{ij}),$$

where

$$(3.32) \quad q_{ij} = \frac{2(-1)^{i+j}}{(i\pi + \pi/2)(j\pi + \pi/2)}, \quad i, j = 0, 1, \dots,$$

is a compact operator in l_2 , since if a in l_2 satisfies

$$(3.33) \quad a = (a_0, -a_1, \dots, (-1)^n a_j, \dots),$$

then

$$A^{-1/2}CC^*A^{-1/2}a = \left(2 \sum_{j=0}^{\infty} \frac{a_j}{(j\pi + \pi/2)} \right) \left(\frac{1}{\pi/2}, \dots, \frac{(-1)^n}{k\pi + \pi/2}, \dots \right)^T.$$

From (3.25) the explicit Hilbert space formulation of the control problem is

$$(3.34i) \quad \ddot{g}_n + \left(n\pi + \frac{\pi}{2} \right)^2 g_n = \sqrt{2}(-1)^{n+1}v,$$

$$(3.34ii) \quad g_n(0) = x_n^0, \quad \dot{g}_n(0) = y_n^0, \quad n = 0, 1, \dots$$

If we consider the homogeneous system

$$(3.35) \quad \ddot{G}_n + \left(n\pi + \frac{\pi}{2} \right)^2 G_n = 0, \quad n = 0, 1, \dots,$$

with initial conditions (3.34ii) we can show that for

$$(3.36i) \quad v(t) = \sum_{j=0}^{\infty} \frac{(-1)^j}{\sqrt{2}} \dot{G}_j(t) \quad \text{if } 0 \leq t \leq 2,$$

and

$$(3.36ii) \quad v(t) = 0 \quad \text{if } t > 2,$$

the control problem (3.34) satisfies the conditions

$$g_n(t) = \dot{g}_n(t) = 0 \quad \text{for } t \geq 2, \quad i = 0, 1, \dots$$

(See the Appendix for a proof.)

Thus this example has some of the flavor of a distributed control problem, but not enough to handle the relatively simple theory given in Remark 3.2.

Appendix. For each n the explicit solution of (3.34) when $t \geq 2$ is

$$(4.1) \quad g_n(t) = x_n^0 \cos\left(n\pi + \frac{\pi}{2}\right)t + \frac{y_n^0}{n\pi + \pi/2} \sin\left(n\pi + \frac{\pi}{2}\right)t + \frac{(-1)^{n+1}}{n\pi + \pi/2} \int_0^2 \sin\left(n\pi + \frac{\pi}{2}\right)(t - \sigma) \\ \cdot \left[\sum_{j=0}^{\infty} (-1)^j \left(-\left(j\pi + \frac{\pi}{2}\right) x_j^0 \sin\left(j\pi + \frac{\pi}{2}\right)\sigma y_j^0 \cos\left(j\pi + \frac{\pi}{2}\right)\sigma \right) \right] d\sigma.$$

Since

$$(i) \quad \sin\left(n\pi + \frac{\pi}{2}\right)(t - \sigma) = \sin\left(n\pi + \frac{\pi}{2}\right)t \cos\left(n\pi + \frac{\pi}{2}\right)\sigma \\ - \cos\left(n\pi + \frac{\pi}{2}\right)t \sin\left(n\pi + \frac{\pi}{2}\right)\sigma, \\ (ii) \quad \int_0^2 \sin\left(\left(n\pi + \frac{\pi}{2}\right)\sigma\right) \cos\left(\left(j\pi + \frac{\pi}{2}\right)\sigma\right) d\sigma = 0, \\ (iii) \quad \int_0^2 \sin\left(\left(n\pi + \frac{\pi}{2}\right)\sigma\right) \sin\left(\left(j\pi + \frac{\pi}{2}\right)\sigma\right) dv \\ = \int_0^2 \cos\left(\left(n\pi + \frac{\pi}{2}\right)\sigma\right) \cos\left(\left(j\pi + \frac{\pi}{2}\right)\sigma\right) dv = \delta_{nj}$$

(δ_{nj} is the Kronecker delta) it follows by simple calculations that $g_n(t) \equiv 0$ for $t \geq 2$. A similar observation shows that $\dot{g}_n(t) \equiv 0$ for $t \geq 2$.

Acknowledgment. The author thanks the referees for their helpful suggestions, particularly in regard to § 1, where the original result was much improved due to their comments.

REFERENCES

- [1] G. CHEN, S. G. KRANTZ, D. W. MA, C. E. WAYNE AND H. H. WEST, *The Euler-Bernoulli beam equation with boundary energy dissipation*, to appear.
- [2] R. DATKO, *Applications of the finite Laplace transform to linear control problems*, this Journal, 18 (1980), pp. 1-20.
- [3] R. DATKO, J. LAGNESE AND M. P. POLIS, *An example on the effect of time delays in boundary feedback of wave equations*, this Journal, 24 (1986), pp. 152-156.
- [4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I: General Theory*, Interscience, New York, 1958.
- [5] V. B. KOLMANOVSKII AND V. R. NOSOV, *Stability of Functional Differential Equations*, Academic Press, New York, 1986.
- [6] J. L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.
- [7] D. L. RUSSELL, *Mathematical Models for the Elastic Beam and Their Control-Theoretic Implications: Semigroups, Theory and Applications*, Vol. II, Longmans, London, 1986.

ON THE MATHEMATICAL MODEL FOR LINEAR ELASTIC SYSTEMS WITH ANALYTIC DAMPING*

FALUN HUANG†

Abstract. In the present paper we investigate linear elastic systems with damping $\ddot{y} + B\dot{y} + Ay = 0$ in Hilbert spaces, where A is a positive definite unbounded linear operator and B is a closed linear operator related in various ways to A^α ($\frac{1}{2} \leq \alpha \leq 1$). We discuss the spectral property of these systems and obtain some fundamental results for the holomorphic property and the exponential stability of the semigroups associated with these systems.

Key words. elastic systems, damping, exponential stability, holomorphic property

AMS(MOS) subject classifications. 34G, 93A, 93D

1. Introduction. Let H be a Hilbert space with inner product (\cdot, \cdot) and let it be associated with norm $\|\cdot\|$. An unbounded self-adjoint linear operator A in H is called positive definite if $(Ax, x) \geq \lambda_1 \|x\|^2$, for all $x \in D(A)$, the domain of A , where $\lambda_1 > 0$ is a constant number. Let A^α for a real number α be the fractional power of a positive definite operator A [1], [2]. Chen and Russell [3] studied the following linear elastic systems with structure damping:

$$(1.1) \quad \begin{aligned} \ddot{y} + B\dot{y} + Ay &= 0, \\ y(0) &= y_0, \quad \dot{y}(0) = y_1, \end{aligned}$$

where \cdot means d/dt , $y, y_0, y_1 \in H$, A and B are positive definite self-adjoint unbounded linear operators in H and B is $A^{1/2}$ -bounded. Letting $x_1 = A^{1/2}y$, $x_2 = \dot{y}$, we get the equivalent first-order linear systems

$$(1.2) \quad \begin{aligned} \frac{d}{dt} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= \begin{pmatrix} 0 & A^{1/2} \\ -A^{1/2} & -B \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = L_B \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \\ x_1(0) &= A^{1/2}y_0, \quad x_2(0) = y_1. \end{aligned}$$

Chen and Russell [3] have proved that

$$L_B = \begin{pmatrix} 0 & A^{1/2} \\ -A^{1/2} & -B \end{pmatrix}$$

generates an analytic semigroup on $W = H \oplus H$, if some additional conditions are satisfied. In the same paper, they pose the following conjecture proved by Huang [4], [5]: Let $D(B) \supset D(A^{1/2})$; then either of the following conditions (1) and (2) implies that L_B generates an analytic semigroup on W :

$$(1) \quad \rho_1(A^{1/2}x, x) \leq (Bx, x) \leq \rho_2(A^{1/2}x, x) \quad \forall x \in D(A^{1/2})$$

or (not, in general, equivalent)

$$(2) \quad \rho_1(Ax, x) \leq (B^2x, x) \leq \rho_2(Ax, x) \quad \forall x \in D(A)$$

for some $\rho_1, \rho_2 > 0$ with $\rho_1 \leq \rho_2$. In addition, the semigroup e^{tL_B} generated by L_B is exponentially stable, i.e., there exist positive numbers M and σ such that $\|e^{tL_B}\| \leq Me^{-\sigma t}$ for $t \geq 0$.

* Received by the editors August 5, 1986; accepted for publication (in revised form) July 17, 1987.

† Department of Mathematics, Sichuan University, Chengdu, People's Republic of China. This author's research was supported by the National Foundation of China.

But these results do not contain the situation that $B = \rho A$, which could possibly appear in engineering applications. For this situation, Massatt [6] shows that if $B = \rho A$ with $\rho > 0$, then

$$\mathcal{A} = \begin{pmatrix} 0 & 1 \\ -A & -\rho A \end{pmatrix}$$

generates an analytic semigroup which is exponentially stable.

In the present paper we will investigate the more widely used linear elastic systems (1.1) with damping B related in various ways to A^α ($\frac{1}{2} \leq \alpha \leq 1$), so that the C_0 -semigroups associated with them are analytic and exponentially stable. Our approach agrees with those of previous studies. Moreover, we will still make use of the representation (1.2) to study the systems (1.1). This approach is new for $\alpha \in [\frac{1}{2}, 1]$ and has many advantages in engineering applications. In § 2 we describe some preliminary results. In § 3 we investigate the spectral property of the systems (1.1) associated with $\alpha \in [\frac{1}{2}, 1]$. Finally, in the last section, we obtain some new results for analytic property and exponential stability of the semigroup associated with the systems (1.1).

2. Preliminary results. Let the spectral measure E be the resolution of the identity for a self-adjoint operator A and let f be a complex Borel function defined E almost everywhere on the real axis R . Then the operator $f(A)$ is defined by

$$(2.1) \quad D(f(A)) = \left\{ x \in H; \int_R |f(t)|^2 (E(dt)x, x) < \infty \right\}$$

and

$$(2.2) \quad (f(A)x, y) = \int_R f(t) (E(dt)x, y) \quad \text{for } x \in D(f(A)) \text{ and } y \in H$$

(see [7], [8] for the A -function calculus used in this paper).

Below, we assume that A is a positive definite unbounded linear operator in H and B is a closed linear operator in H with the domain $D(B) \supset D(A^\alpha)$ for some $\alpha \in [\frac{1}{2}, 1]$. Let $\sigma(A)$ and $\rho(A)$ be the spectrum and the resolvent set of A , respectively. We will study the linear elastic system

$$(2.3) \quad \ddot{y} + B\dot{y} + Ay = 0.$$

Setting $x_1 = A^{1/2}y$, $x_2 = \dot{y}$ and $w = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (x_1, x_2)^\perp$, we obtain the first order representation of the system (2.3)

$$(2.4) \quad \dot{w} = \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} 0 & A^{1/2} \\ -A^{1/2} & -B \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = L_B w,$$

where

$$(2.5) \quad D(L_B) = D(A^{1/2}) \oplus D(A^{1/2}) \cap D(B).$$

The energy of the system (2.3)

$$(2.6) \quad E_0(t) = \frac{1}{2} [\|\dot{y}\|^2 + \|A^{1/2}y\|^2],$$

provided that the initial data belongs to $D(A^{1/2}) \oplus H$, decays exponentially if and only if \bar{L}_B generates an exponentially stable c_0 -semigroup on $W = H \oplus H$, where the bar indicates closure.

LEMMA 2.1. Let $D(B^*) \cap D(B) \supset D(A^\alpha)$ for some $\alpha \in [\frac{1}{2}, 1]$. Then we have:

- (a) $A^{-(\alpha-\tau)}BA^{-\tau} \in \mathcal{L}(H)$, for $\tau \in [0, \alpha]$ and $\|\overline{A^{-(\alpha-\tau)}BA^{-\tau}}\| \leq \|\overline{A^{-\alpha}B}\|^{\tau/\alpha} \|BA^{-\alpha}\|^{(\alpha-\tau)/\alpha}$ for $\tau \in [0, \alpha]$, where $A^{-\alpha}B$ have bounded extension on H ;
 (b) $\overline{A^{-1/2}BA^{-1/2}} \in \mathcal{L}(H)$ and $\|\overline{A^{-1/2}BA^{-1/2}}\| \leq \lambda_1^{-(1-\alpha)} \|\overline{A^{-\alpha}B}\|^{1/2\alpha} \|BA^{-\alpha}\|^{(2\alpha-1)/2\alpha}$, where $\lambda_1 = \inf\{\lambda; \lambda \in \sigma(A)\}$;
 (c) $D(A^{-1/2}BA^{1/2}) \supset D(A^\alpha)$ and $\|A^{-1/2}BA^{1/2}x\| \leq \|A^{-\alpha}B\|^{(2\alpha-1)/2\alpha} \|BA^{-\alpha}\|^{1/2\alpha} \|A^\alpha x\|$, for all $x \in D(A^\alpha)$.

Proof. For $x \in H$ and $\lambda = \tau + i\omega$ with $\omega \in \mathbb{R}$ and $\tau \in [-\alpha, 0]$, we define

$$A^\lambda x = \int_{\lambda_1}^{\infty} t^\lambda E(dt)x.$$

Clearly, $A^\lambda x$ is analytic for $(-\alpha) < \operatorname{Re} \lambda < 0$ and

$$\|A^{\pm i\omega}x\|^2 = \int_{\lambda_1}^{\infty} |t^{\pm i\omega}|^2 \|E(dt)x\|^2 = \int_{\lambda_1}^{\infty} \|E(dt)x\|^2 = \|x\|^2.$$

Now for $x \in D(A^\alpha)$ and $\lambda = \tau + i\omega$, since $BA^{-\alpha} \in \mathcal{L}(H)$, $T(\lambda)x = A^{-(\alpha+\lambda)}BA^\lambda x = A^{-(\alpha+\lambda)}BA^{-\alpha}A^\lambda x$ is analytic for $\tau \in (-\alpha, 0)$. Observing that $B^*A^{-\alpha} \in \mathcal{L}(H)$, $\overline{A^{-\alpha}B} = (B^*A^{-\alpha})^*$ is bounded on H ; thus we have

$$\begin{aligned} \|T(0+i\omega)x\| &= \|A^{-(\alpha+i\omega)}BA^{i\omega}x\| = \|A^{-i\omega}A^{-\alpha}BA^{i\omega}x\| \\ &\leq \|A^{-i\omega}\| \|\overline{A^{-\alpha}B}\| \|A^{i\omega}\| \|x\| \leq \|\overline{A^{-\alpha}B}\| \|x\| \end{aligned}$$

and

$$\|T(-\alpha+i\omega)x\| = \|A^{-i\omega}BA^{-\alpha}A^{i\omega}x\| \leq \|BA^{-\alpha}\| \|x\|.$$

By the three lines theorem [8, pp. 520], for $\tau \in [-\alpha, 0]$, we have

$$\begin{aligned} \sup_{\omega \in \mathbb{R}} \|T(\tau+i\omega)x\| &\leq \left[\sup_{\omega \in \mathbb{R}} \|T(0+i\omega)x\| \right]^{-\tau/\alpha} \left[\sup_{\omega \in \mathbb{R}} \|T(-\alpha+i\omega)x\| \right]^{(\alpha+\tau)/\alpha} \\ &\leq \|\overline{A^{-\alpha}B}\|^{-\tau/\alpha} \|BA^{-\alpha}\|^{(\alpha+\tau)/\alpha} \|x\| \quad \forall x \in D(A^\alpha). \end{aligned}$$

Especially, for $\lambda = \tau \in [-\alpha, 0]$, we have

$$\|T(\tau)x\| = \|A^{-(\alpha+\tau)}BA^\tau x\| \leq \|\overline{A^{-\alpha}B}\|^{-\tau/\alpha} \|BA^{-\alpha}\|^{(\alpha+\tau)/\alpha} \|x\| \quad \forall x \in D(A^\alpha).$$

Since $D(A^\alpha)$ is dense, (a) is proved.

From (a), we have $\overline{A^{-1/2}BA^{-1/2}} = A^{-(1-\alpha)}\overline{A^{-(\alpha-1/2)}BA^{-1/2}} \in \mathcal{L}(H)$ and $\|\overline{A^{-1/2}BA^{-1/2}}\| \leq \|A^{-(1-\alpha)}\| \|\overline{A^{-(\alpha-1/2)}BA^{-1/2}}\| \leq \lambda_1^{-(1-\alpha)} \|\overline{A^{-\alpha}B}\|^{1/2\alpha} \|BA^{-\alpha}\|^{(2\alpha-1)/2\alpha}$. This is (b). Also from (a), $\overline{A^{-1/2}BA^{-(\alpha-1/2)}} \in \mathcal{L}(H)$ and $\|\overline{A^{-1/2}BA^{-(\alpha-1/2)}}\| \leq \|\overline{A^{-\alpha}B}\|^{(2\alpha-1)/2\alpha} \|BA^{-\alpha}\|^{1/2\alpha}$. Therefore $D(A^{-1/2}BA^{1/2}) = D(A^{-1/2}BA^{-(\alpha-1/2)}A^\alpha) \supset D(A^\alpha)$, and so (c) holds. \square

THEOREM 2.2. Let $D(B^*) \cap D(B) \supset D(A^\alpha)$ for some $\alpha \in [\frac{1}{2}, 1]$; then L_B is closable and $D(\tilde{L}_B) = \{(-A^{-1/2}BA^{-1/2}x - A^{-1/2}y, A^{-1/2}x)^\perp; (x, y)^\perp \in W\}$.

Proof. Let $\omega_n = (x_n, y_n)^\perp \in D(L_B)$, $\omega_n \rightarrow \omega_0 = (x_0, y_0)^\perp$ and $L_B \omega_n = (A^{1/2}y_n, -A^{1/2}x_n - By_n)^\perp \rightarrow (u, v)^\perp$, as $n \rightarrow \infty$. Then by the closedness of $A^{1/2}$, we have $y_0 \in D(A^{1/2})$ and $A^{1/2}y_0 = u$. Moreover, from $A^{1/2}x_n + By_n \rightarrow -v$ and $x_n \rightarrow x_0$ as $n \rightarrow \infty$, we have $A^{-1/2}By_n \rightarrow -A^{-1/2}v - x_0$ as $n \rightarrow \infty$. But $A^{-1/2}By_n = A^{-1/2}BA^{-1/2}A^{1/2}y_n \rightarrow \overline{A^{-1/2}BA^{-1/2}}u$ as $n \rightarrow \infty$ by Lemma 2.1(b); thus $x_0 = -\overline{A^{-1/2}BA^{-1/2}}u - A^{-1/2}v$. Therefore, if $w_0 = 0$, then $u = 0$ and $v = 0$, so L_B is closable. If $w_0 \neq 0$, then $w_0 = (-\overline{A^{-1/2}BA^{-1/2}}u - A^{-1/2}v, A^{-1/2}u)^\perp$ and the set of all these w_0 is $D(\tilde{L}_B)$ by the definition of closure. Thus $D(\tilde{L}_B) \subset D_0 = \{(-\overline{A^{-1/2}BA^{-1/2}}x - A^{-1/2}y, A^{-1/2}x)^\perp; (x, y)^\perp \in W\}$. On the other hand, for any $(u, v)^\perp \in W$ and $(x_0, y_0) = (-\overline{A^{-1/2}BA^{-1/2}}u - A^{-1/2}v, A^{-1/2}u)$,

there exists $u_n \in D(BA^{-1/2})$ such that $u_n \rightarrow u$ and $A^{-1/2}BA^{-1/2}u_n \rightarrow \overline{A^{-1/2}BA^{-1/2}u}$ as $n \rightarrow \infty$. Thus $w_n = (x_n, y_n)^\perp \rightarrow (x_0, y_0)^\perp$, where $x_n = -A^{-1/2}BA^{-1/2}u_n - A^{-1/2}v$, $y_n = A^{-1/2}u_n$ and $L_B w_n = (u_n, v) \rightarrow (u, v)$ as $n \rightarrow \infty$. Therefore $(x_0, y_0)^\perp \in D(\bar{L}_B)$ and $D_0 \subset D(\bar{L}_B)$, and so $D(\bar{L}_B) = D_0$. \square

DEFINITION 2.3. The elastic system (2.3) is said to have the Analytic Damping Property (ADP) if and only if \bar{L}_B generates an exponentially stable analytic semigroup. In this case B is called an AD-operator for the system (2.3).

Clearly, if $\alpha = \frac{1}{2}$, then analytic damping is structural damping as in [3]. For $\alpha = 1$, it is strong damping as in [6].

3. Spectral analysis. In the present section we investigate the spectral property of the system (2.3) and the representation of the resolvent of the operator \bar{L}_B .

THEOREM 3.1. Let B be a closed linear operator with $D(B^*) \cap D(B) \supset D(A^\alpha)$ for some $\alpha \in [\frac{1}{2}, 1]$. Then:

(a) $\lambda \in \rho(\bar{L}_B)$ if and only if $\Delta(\lambda) = (\lambda^2 I + \lambda B + A)$ has inverse $\Delta^{-1}(\lambda) \in \mathcal{L}(H)$ and $\overline{\Delta^{-1}(\lambda)A^{1/2}} \in \mathcal{L}(H)$ and $\overline{\Delta^{-1}(\lambda)A^{1/2}}H \subset D(A^{1/2})$.

(b) For $\lambda \in \rho(\bar{L}_B)$, we have

$$\begin{aligned} R(\lambda; \bar{L}_B) &= (\lambda I - \bar{L}_B)^{-1} = \begin{pmatrix} \overline{A^{-1/2}(\lambda I + B)\Delta^{-1}(\lambda)A^{1/2}} & A^{1/2}\Delta^{-1}(\lambda) \\ -\overline{\Delta^{-1}(\lambda)A^{1/2}} & \lambda\Delta^{-1}(\lambda) \end{pmatrix} \\ &= \begin{pmatrix} A^{1/2}\overline{\Delta^{-1}(\lambda)(\lambda I + B)A^{-1/2}} & A^{1/2}\Delta^{-1}(\lambda) \\ -\overline{\Delta^{-1}(\lambda)A^{1/2}} & \lambda\Delta^{-1}(\lambda) \end{pmatrix}. \end{aligned} \quad (3.1)$$

(c) For $\frac{1}{2} \leq \alpha < 1$, \bar{L}_B has compact resolvent if A has, but if $\alpha = 1$, \bar{L}_B does not, in general.

Proof. Let $\lambda \in \rho(\bar{L}_B)$. Then for $(u, v) \in W = H \oplus H$ there exists a unique $(x, y)^\perp \in D(\bar{L}_B)$ such that $(\lambda - \bar{L}_B)(x, y)^\perp = (u, v)^\perp$. Thus there exist $(x_n, y_n) \in D(L_B)$ such that $(x_n, y_n) \rightarrow (x, y)$, $(\lambda - L_B)(x_n, y_n)^\perp = (u_n, v_n)^\perp$, i.e.,

$$\lambda x_n - A^{1/2}y_n = u_n, \quad (3.2)$$

$$A^{1/2}x_n + (\lambda + B)y_n = v_n \quad (3.3)$$

and $(u_n, v_n) \rightarrow (u, v)$ as $n \rightarrow \infty$. Equation (3.3) implies that $x_n = -A^{-1/2}(\lambda + B)y_n + A^{-1/2}v_n$ and from (3.2), $[A^{1/2} + \lambda A^{-1/2}(\lambda + B)]y_n = \lambda A^{-1/2}v_n - u_n$. Letting $n \rightarrow \infty$, we obtain

$$[A^{1/2} + \lambda A^{-1/2}(\lambda + B)]y = \lambda A^{-1/2}v - u. \quad (3.4)$$

Taking $v = 0$ in (3.4), we see that $[A^{1/2} + \lambda A^{-1/2}(\lambda + B)]$ is one to one from $D(A^{1/2})$ onto H , because $D(A^{1/2}) = \{y; (x, y)^\perp \in D(\bar{L}_B)\}$ by Theorem 2.2. It follows from the closed graph theorem that $[A^{1/2} + \lambda A^{-1/2}(\lambda I + B)]^{-1} = \overline{\Delta^{-1}(\lambda)A^{1/2}} \in \mathcal{L}(H)$ and its range is $D(A^{1/2})$. Therefore $A^{1/2}\overline{\Delta^{-1}(\lambda)A^{1/2}}$, $\overline{A^{-1/2}B\Delta^{-1}(\lambda)A^{1/2}} \in \mathcal{L}(H)$, and so $A^{1/2}\overline{\Delta^{-1}(\lambda)A^{1/2}}$ has bounded extension on H and $\Delta^{-1}(\lambda)$, $A^{1/2}\Delta^{-1}(\lambda) \in \mathcal{L}(H)$. Moreover, from (3.4), we have

$$y = -\overline{\Delta^{-1}(\lambda)A^{1/2}}u + \lambda\Delta^{-1}(\lambda)v \quad (3.5)$$

and

$$\begin{aligned} x &= \overline{(\lambda A^{-1/2} + A^{-1/2}B)\Delta^{-1}(\lambda)A^{1/2}}u - \overline{\lambda A^{-1/2}(\lambda + B)\Delta^{-1}(\lambda)}v + A^{-1/2}v \\ &= \overline{A^{-1/2}(\lambda + B)\Delta^{-1}(\lambda)A^{1/2}}u + A^{1/2}\Delta^{-1}(\lambda)v. \end{aligned} \quad (3.6)$$

But for $\lambda \neq 0$,

$$(3.7) \quad \begin{aligned} \overline{A^{-1/2}(\lambda+B)\Delta^{-1}(\lambda)A^{1/2}}u &= \frac{1}{\lambda} [1 - \overline{A^{1/2}\Delta^{-1}(\lambda)A^{1/2}}]u \\ &= A^{1/2}\Delta^{-1}(\lambda)(\lambda+B)A^{-1/2}u. \end{aligned}$$

Thus, from (3.5)–(3.7) we have obtained (3.1).

Conversely, let $\Delta^{-1}(\lambda) \in \mathcal{L}(H)$ and $\overline{\Delta^{-1}(\lambda)A^{1/2}} = [\overline{A^{1/2} + \lambda A^{-1/2}(\lambda+B)}]^{-1} \in \mathcal{L}(H)$ and $\overline{\Delta^{-1}(\lambda)A^{1/2}}H \subset D(A^{1/2})$. Then $A^{1/2}\Delta^{-1}(\lambda)A^{1/2} \in \mathcal{L}(H)$; thus $\overline{A^{-1/2}(\lambda+B)\Delta^{-1}(\lambda)A^{1/2}} = \overline{A^{-1/2}(\lambda+B)A^{-1/2}A^{1/2}\Delta^{-1}(\lambda)A^{1/2}} \in \mathcal{L}(H)$ by Lemma 2.1, and so the matrix operator in (3.1) belongs to $\mathcal{L}(W)$ and it is equal to $(\lambda - \bar{L}_B)^{-1}$. Thus $\lambda \in \rho(\bar{L}_B)$.

Finally, since $\Delta^{-1}(0) = A^{-1} \in \mathcal{L}(H)$ and $\overline{A^{-1/2}BA^{-1/2}} \in \mathcal{L}(H)$ by Lemma 2.1, we have $\lambda = 0 \in \rho(\bar{L}_B)$ and

$$(3.8) \quad R(0; \bar{L}_B) = \begin{pmatrix} \overline{A^{-1/2}BA^{-1/2}} & A^{-1/2} \\ -A^{-1/2} & 0 \end{pmatrix}.$$

If $\alpha \in [\frac{1}{2}, 1)$ and A is discrete, i.e., having compact resolvent, then $A^{-\beta}$ is compact for $\beta > 0$. And so, since $A^{-(\alpha-1/2)}BA^{-1/2} \in \mathcal{L}(H)$ by Lemma 2.1, $A^{-1/2}BA^{-1/2} = A^{-(1-\alpha)}(A^{-(\alpha-1/2)}BA^{-1/2})$ is compact. Also from (3.8) we get that $R(0; \bar{L}_B)$ is compact, and so $R(\lambda; \bar{L}_B)$ is compact for $\lambda \in \rho(\bar{L}_B)$. If $\alpha = 1$ and $B = \rho A$ with $\rho > 0$, then $A^{-1/2}BA^{-1/2} = \rho I$ is not compact, and so $\bar{L}_{\rho A}$ cannot have compact resolvent. \square

COROLLARY 3.2. *If $B = \rho A^\alpha$ for some $\alpha \in [\frac{1}{2}, 1]$ and $\rho > 0$, then $\lambda \in \rho(\bar{L}_B)$ if and only if $\Delta^{-1}(\lambda)$ exists and $\in \mathcal{L}(H)$.*

Proof. If $\Delta^{-1}(\lambda) \in \mathcal{L}(H)$, then $(\lambda + \rho A^\alpha)\Delta^{-1}(\lambda) \in \mathcal{L}(H)$ and it is a bounded extension of $A^{1/2}\Delta^{-1}(\lambda)(\lambda+B)A^{-1/2}$ on H . Also $\Delta^{-1}(\lambda)A^{1/2}$ has the bounded extension $A^{1/2}\Delta^{-1}(\lambda)$ on H . \square

Remark 3.3. If $D(B) \supset D(A^\alpha)$ for some $\alpha \in (-\infty, \frac{1}{2}]$, then, by the analogical method used to prove Theorem 3.1, we can prove that $\lambda \in \rho(\bar{L}_B)$ if and only if $\Delta^{-1}(\lambda) \in \mathcal{L}(H)$. In that case, $D(\bar{L}_B) = D(A^{1/2}) \oplus D(A^{1/2})$.

THEOREM 3.4. *Let $\rho > 0$, $\alpha \in (-\infty, 1]$ and*

$$(3.9) \quad \mu_\alpha^\pm(t) = \frac{1}{2}(-\rho t^\alpha \pm \sqrt{\rho^2 t^{2\alpha} - 4t}) = -2t^{1-\alpha}\rho^{-1}(1 \pm \sqrt{1 - 4\rho^{-2}t^{(1-2\alpha)}})^{-1}.$$

Then

- (a) For $\alpha \in (-\infty, 1)$, $\sigma(\bar{L}_{\rho A^\alpha}) = \{\mu_\alpha^\pm(t); t \in \sigma(A)\}$;
- (b) $\sigma(\bar{L}_{\rho A}) = \{\mu_1^\pm(t); t \in \sigma(A)\} \cup \{-\rho^{-1}\}$;
- (c) For $\alpha \in (-\infty, \frac{1}{2})$, $\arg(\mu_\alpha^\pm(t)) \rightarrow \pm\pi/2$ as $t \rightarrow \infty$.

Proof. Let $f_\lambda(t) = \lambda^2 + \lambda\rho t^\alpha + t$; then $f_\lambda(A) \supset \Delta(\lambda) = \lambda^2 I + \lambda\rho A^\alpha + A$. If $\alpha < 1$ (or $\lambda \neq -\rho^{-1}$), we have $|t| = 0 (|f_\lambda(t)|)$ as $t \rightarrow \infty$. From (2.1) it follows that $D(f_\lambda(A)) \subset D(A)$. Hence we obtain that $f_\lambda(A) = \Delta(\lambda)$, and so $\Delta^{-1}(\lambda) \in \mathcal{L}(H)$ if and only if $[f_\lambda(t)]^{-1}$ is E almost everywhere bounded on R . By Corollary 3.2 and the observation that $|\mu_\alpha^\pm(t)| \rightarrow \infty$ as $t \rightarrow \infty$, (a) holds.

Let $\alpha = 1$. Then by the same reasoning used above, we have that $\{\mu_1^\pm(t); t \in \sigma(A)\} \subset \sigma(\bar{L}_{\rho A}) \subset \{\mu_1^\pm(t); t \in \sigma(A)\} \cup \{-\rho^{-1}\}$. Furthermore, $\Delta(-\rho^{-1}) = \rho^{-2}I|_{D(A)} \not\subseteq f_{(-\rho^{-1})}(A)$, and so $\Delta^{-1}(-\rho^{-1}) \notin \mathcal{L}(H)$. Hence (b) holds. Obviously (c) holds from (3.9). \square

It is seen that $\sigma(\bar{L}_{\rho A}) \ni \mu_1^\pm(t) \rightarrow -\rho^{-1}$ as $t \rightarrow \infty$ in $\sigma(A)$, even if A is discrete. We do not know what this means in regard to engineering applications.

LEMMA 3.5. *Let ρ , α and $\mu_\alpha^\pm(t)$ be the same as in Theorem 3.4. Let $\operatorname{Re} \lambda \geq 0$ and $\Delta(\lambda) = \lambda^2 + \lambda\rho A^\alpha + A$. Then*

- (a) $\|\lambda A^\alpha \Delta^{-1}(\lambda)\| \leq \rho^{-1}$;

- (b) $\|\lambda^2 \Delta^{-1}(\lambda)\| \leq 1 + \rho^{-1} \lambda_1^{(1/2-\alpha)}$, for $\alpha \in [\frac{1}{2}, 1]$;
 (c) $\|A \Delta^{-1}(\lambda)\| \leq \max \{1, 4\rho^{-2} \lambda_1^{-(2\alpha-1)}\}$, for $\alpha \in [\frac{1}{2}, 1]$.

Proof. Observing that for $\lambda = \tau + i\omega$ with $\tau \geq 0$,

$$\begin{aligned} |f_\lambda(t)|^2 &= |\lambda^2 + \lambda \rho t^\alpha + t|^2 \\ (3.10) \quad &= \rho^2 t^{2\alpha} |\lambda|^2 + 2\rho t^\alpha \tau |\lambda|^2 + 2\rho t^{1+\alpha} \tau + 4\omega^2 \tau^2 + (\tau^2 - \omega^2 + t)^2 \\ &\geq \rho^2 t^{2\alpha} |\lambda|^2, \end{aligned}$$

we have $E\text{-sup} |\lambda t^\alpha [f_\lambda(t)]^{-1}| \leq \rho^{-1} (t \geq \lambda_1 > 0)$. Hence (a) holds from $\lambda A^\alpha \Delta^{-1}(\lambda) = \lambda A^\alpha [f_\lambda]^{-1}(A)$.

If we set $\mu^\pm = \mu_\alpha^\pm(t)$, then $f_\lambda(t) = (\lambda - \mu^+)(\lambda - \mu^-)$. If $\rho^2 t^{2\alpha} > 4t$ it is obvious that $|\lambda - \mu^\pm| > |\lambda|$ for $\text{Re } \lambda \geq 0$. If $\rho^2 t^{2\alpha} < 4t$, it is also obvious that $|\lambda - \mu^+| > |\lambda|$ when $\text{Im } \lambda < 0$, $|\lambda - \mu^-| > |\lambda|$ when $\text{Im } \lambda \geq 0$, and $|\mu^\pm| = \frac{1}{2} \rho t^\alpha |1 \pm i\sqrt{4\rho^{-2} t^{(1-2\alpha)} - 1}| = t^{1/2}$. This implies that for $\text{Im } \lambda < 0$ and $\text{Im } \lambda \geq 0$, respectively,

$$\begin{aligned} |\lambda^2 f_\lambda^{-1}(t)| &= |\lambda| \left| \frac{1}{\lambda - \mu^\pm} + \frac{\mu^\mp}{(\lambda - \mu^+)(\lambda - \mu^-)} \right| \\ &\leq 1 + \frac{|\mu^\mp|}{\text{Re } (\lambda - \mu^\mp)} \\ &= 1 + \frac{t^{1/2}}{\text{Re } \lambda + \frac{1}{2} \rho t^\alpha} \quad \text{for } \rho^2 t^{2\alpha} < 4t. \end{aligned}$$

Therefore, for $\alpha \in [\frac{1}{2}, 1]$ we obtain that

$$E\text{-sup} |\lambda^2 f_\lambda^{-1}(t)| \leq 1 + 2\rho^{-1} \lambda_1^{1/2-\alpha},$$

and so (b) holds.

Finally, we prove (c). If $\rho^2 t^{2\alpha} \geq 4t$, then from (3.10) we have

$$\begin{aligned} |f_\lambda(t)|^2 &= \rho^2 t^{2\alpha} \tau^2 + \omega^2 (\rho^2 t^{2\alpha} - 2t) + 2\rho t^\alpha \tau |\lambda|^2 + 2\rho t^{1+\alpha} \tau + |\lambda|^4 + 2\tau^2 t + t^2 \\ (3.11) \quad &\geq t^2, \end{aligned}$$

and so $|tf_\lambda^{-1}(t)| \leq 1$. If $\rho^2 t^{2\alpha} < 4t$, then

$$\begin{aligned} t^2 |f_\lambda^{-1}(t)|^2 &= t^2 |(\lambda - \mu^+)^{-1}|^2 |(\lambda - \mu^-)^{-1}|^2 \leq t^2 |\text{Re } (\lambda - \mu^+)|^{-2} |\text{Re } (\lambda - \mu^-)|^{-2} \\ &\leq t^2 (\tau + \frac{1}{2} \rho t^\alpha)^{-4} \leq 4^2 \rho^{-4} \lambda_1^{-2(2\alpha-1)}. \end{aligned}$$

Therefore (c) holds.

LEMMA 3.6. Let B be a positive definite self-adjoint linear operator with the domain $D(B) \supset D(A^\alpha)$ for some $\alpha \in [\frac{1}{2}, 1]$. Then

$$\sup_{\lambda \in \sigma(\bar{L}_B)} \text{Re } \lambda \leq -\min \left\{ \frac{\nu_1}{2}, \rho^{-1} \lambda_1^{(1-\alpha)} \right\},$$

where $\rho > 0$ such that $\|Bx\| \leq \rho \|A^\alpha x\|$, for all $x \in D(A^\alpha)$ and $\nu_1 = \inf \{\lambda; \lambda \in \sigma(B)\}$.

Proof. We need only to prove that $\Delta^{-1}(\lambda) = (\lambda^2 I + \lambda B + A)^{-1}$, $\Delta^{-1}(\lambda) A^{1/2}$ and $A^{1/2} \Delta^{-1}(\lambda) A^{1/2} \in \mathcal{L}(H)$ for $\text{Re } \lambda > -\delta_0$ by Theorem 3.1, where $\delta_0 = \min \{\frac{1}{2} \nu_1, \rho^{-1} \lambda_1^{(1-\alpha)}\}$. In order to do this, it is sufficient to prove $A^{1/2} \Delta^{-1}(\lambda) A^{1/2} \in \mathcal{L}(H)$ for $\text{Re } \lambda > -\delta_0$. Since $D(B) \supset D(A^\alpha) \supset D(A)$, we have that $D(B^{1/2}) \supset D(A^{1/2})$

[9, § 12.5], and so $B^{1/2}A^{-1/2}, \overline{A^{-1/2}B^{1/2}} \in \mathcal{L}(H)$. Let $\rho_1 = \|B^{1/2}A^{-1/2}\|$; then $\|A^{1/2}x\| \leq \rho_1^{-1}\|B^{1/2}x\|$ for $x \in D(A^{1/2})$. Thus, for $\lambda = \tau + i\omega$ with $\tau > -\delta_0$ and $x \in D(A)$, if $\omega = 0$, since $\|A^{-1/2}BA^{-1/2}\| \leq \lambda_1^{-(1-\alpha)}\rho$ by Lemma 2.1, we have

$$\begin{aligned} \operatorname{Re}(\Delta(\lambda)x, x) &= \tau^2\|x\|^2 + \|A^{1/2}x\|^2 + \tau\|B^{1/2}x\|^2 \\ (3.12) \quad &\geq \tau\|B^{1/2}x\|^2 \quad \text{or} \quad (1 - \lambda_1^{-(1-\alpha)}\rho|\tau|)\rho_1^{-2}\|B^{1/2}x\|^2 \\ &\quad (\text{accordingly as } \tau \geq 0 \text{ or } -\delta_0 < \tau < 0). \end{aligned}$$

If $\omega \neq 0$, then

$$\begin{aligned} |\operatorname{Im}(\Delta(\lambda)x, x)| &= |\omega|[\|B^{1/2}x\|^2 + 2\tau\|x\|^2] \\ (3.13) \quad &\leq |\omega|\|B^{1/2}x\|^2 \quad \text{or} \quad |\omega|(1 - 2\nu_1^{-1}|\tau|)\|B^{1/2}x\|^2 \\ &\quad (\text{accordingly as } \tau \geq 0 \text{ or } -\delta_0 < \tau < 0). \end{aligned}$$

Since $A^{1/2}\Delta^{-1}(0)A^{1/2} = I \in \mathcal{L}(H)$, and so $A^{1/2}\Delta^{-1}(0)B^{1/2}, B^{1/2}\Delta^{-1}(0)A^{1/2}, B^{1/2}\Delta^{-1}(0)B^{1/2} \in \mathcal{L}(H)$ and $\|B^{1/2}\Delta^{-1}(0)B^{1/2}\| = \|B^{1/2}A^{-1}B^{1/2}\| \leq \rho_1^2$. Therefore, for $0 < \tau \leq \gamma_1 = \frac{1}{2}[-\nu_1 + (\nu_1^2 + 4\nu_1\rho_1^{-2}\theta_0)^{1/2}]$ with $0 < \theta_0 < 1$, we have $A^{1/2}\Delta^{-1}(\tau)A^{1/2} \in \mathcal{L}(H)$ and

$$\begin{aligned} A^{1/2}\Delta^{-1}(\tau)A^{1/2} &= A^{1/2}\Delta^{-1}(0)[1 + \tau(\tau + B)\Delta^{-1}(0)]^{-1}A^{1/2} = A^{1/2}\Delta^{-1}(0)A^{1/2} \\ &\quad + A^{1/2}\Delta^{-1}(0)B^{1/2} \sum_{n=1}^{\infty} (-\tau)^n [(\tau B^{-1} + I)B^{1/2}\Delta^{-1}(0)B^{1/2}]^{(n-1)} \\ &\quad \cdot (\tau B^{-1} + I)B^{1/2}\Delta^{-1}(0)A^{1/2}, \end{aligned}$$

because $|\tau| \|(\tau B^{-1} + I)B^{1/2}\Delta^{-1}(0)B^{1/2}\| \leq \gamma_1(\gamma_1\nu_1^{-1} + 1)\rho_1^2 \leq \theta_0$ and so the series is absolutely convergent in norm. Let $\gamma_2 = \frac{1}{2}[-(\nu_1 + 2\gamma_1) + ((\nu_1 + 2\gamma_1)^2 + 4\nu_1\gamma_1\theta_0)^{1/2}]$, observing that $\|B^{1/2}\Delta^{-1}(\gamma_1)B^{1/2}\| \leq \gamma_1^{-1}$ from (3.12) and $(\tau - \gamma_1)\|[(\gamma_1 + \tau)B^{-1} + I]B^{1/2}\Delta^{-1}(\gamma_1)B^{1/2}\| \leq \gamma_2[(2\gamma_1 + \gamma_2)\nu_1^{-1} + 1]\gamma_1^{-1} \leq \theta_0$ for $\tau \in [\gamma_1, \gamma_1 + \gamma_2]$. We also obtain $A^{1/2}\Delta^{-1}(\tau)A^{1/2} \in \mathcal{L}(H)$ and

$$\begin{aligned} A^{1/2}\Delta^{-1}(\tau)A^{1/2} &= A^{1/2}\Delta^{-1}(\gamma_1)A^{1/2} + A^{1/2}\Delta^{-1}(\gamma_1)B^{1/2} \\ &\quad \cdot \sum_{n=1}^{\infty} (\gamma_1 - \tau)^n [((\gamma_1 + \tau)B^{-1} + I)B^{1/2}\Delta^{-1}(\gamma_1)B^{1/2}]^{(n-1)} \\ &\quad \cdot ((\gamma_1 + \tau)B^{-1} + I)B^{1/2}\Delta^{-1}(\gamma_1)A^{1/2} \end{aligned}$$

for $\tau \in [\gamma_1, \gamma_1 + \gamma_2]$. Repeating the process, we see that $A^{1/2}\Delta^{-1}(\tau)A^{1/2} \in \mathcal{L}(H)$ for $\tau \geq 0$. Similarly, we have from (3.12) and (3.13) that $A^{1/2}\Delta^{-1}(\lambda)A^{1/2} \in \mathcal{L}(H)$ for $\lambda = \tau \in (-\delta_0, 0)$ and $\operatorname{Re} \lambda > -\delta_0$ with $\operatorname{Im} \lambda \neq 0$, respectively. \square

4. Sufficient conditions for AD-operators. By semigroup theory [1], [2], [9], a closed linear operator A_0 densely defined in H is the infinitesimal generator of a holomorphic semigroup if and only if there exist real numbers $\tau_0 > 0$ and $M \geq 1$ such that $\lambda \in \rho(A_0)$ and $\|(\lambda - A_0)^{-1}\| \leq M|\lambda|^{-1}$ whenever $\operatorname{Re} \lambda \geq \tau_0$. In addition, for $\eta \in [\pi/2, \pi/2 + \arcsin(1/2M)]$ we have that $\{\lambda = \tau_0 + re^{i\theta}; 0 \leq r < \infty, 0 \leq |\theta| \leq \eta\} \subset \rho(A_0)$ if A_0 is such. Therefore by (a) and (c) in Theorem 3.4, for $\alpha \in [0, \frac{1}{2})$ and $B = \rho A^\alpha$, L_B cannot generate a holomorphic semigroup. That is, the elastic system (2.3) with this B does not have ADP. However for $\alpha \in [\frac{1}{2}, 1]$ we have the following theorem.

THEOREM 4.1. *Let $B = \rho A^\alpha + B_1$, where $\rho > 0$, $\alpha \in [\frac{1}{2}, 1]$ and B_1 is a closed linear operator with $D(B_1) \cap D(B_1^*) \supset D(A^\alpha)$ and $\|B_1x\| \leq \theta\rho\|A^\alpha x\|$, $\|B_1^*x\| \leq \theta\rho\|A^\alpha x\|$ for all $x \in D(A^\alpha)$ with some $\theta \in [0, 1)$. Then B is an AD-operator for the system (2.3). Moreover, we have the estimate*

$$(4.1) \quad \|e^{tL_B}\| \leq M_\sigma e^{-\sigma t} \quad \text{for } t \geq 0,$$

where $0 < \sigma < \sigma_0 = \min \{ \frac{1}{6} \rho \lambda_1^\alpha (1 - \theta^2), \frac{1}{2} \rho^{-1} \lambda_1^{(1-\alpha)} \}$ and $e^{t\bar{L}_B}$ denotes the semigroup generated by \bar{L}_B .

Proof. We first prove that \bar{L}_B generates a holomorphic semigroup on W . For $\lambda \in \rho(\bar{L}_B)$, from (3.1), we have

$$(4.2) \quad (\lambda - \bar{L}_B)^{-1} = \begin{pmatrix} A^{1/2} \bar{\Delta}^{-1}(\lambda)(\lambda + B)A^{-1/2} & A^{1/2} \Delta^{-1}(\lambda) \\ -\bar{\Delta}^{-1}(\lambda)A^{1/2} & \lambda \Delta^{-1}(\lambda) \end{pmatrix},$$

where $\Delta^{-1}(\lambda) = (\lambda^2 + \lambda B + A)^{-1}$. For $\operatorname{Re} \lambda \geq 0$ and $\lambda \neq 0$, by Lemma 3.5(a), we obtain that

$$(4.3) \quad \begin{aligned} \|A^{1/2} \Delta^{-1}(\lambda)\| &= \|A^{1/2}(\lambda^2 I + \lambda \rho A^\alpha + A + \lambda B_1)^{-1}\| \\ &= \|A^{(1/2-\alpha)} A^\alpha (\lambda^2 I + \lambda \rho A^\alpha + A)^{-1} \\ &\quad \cdot [I + \lambda B_1 A^{-\alpha} A^\alpha (\lambda^2 I + \lambda \rho A^\alpha + A)^{-1}]^{-1}\| \\ &\leq \|A^{(1/2-\alpha)}\| \|A^\alpha (\lambda^2 I + \lambda \rho A^\alpha + A)^{-1}\| \\ &\quad \cdot \|[I + \lambda B_1 A^{-\alpha} A^\alpha (\lambda^2 I + \lambda \rho A^\alpha + A)^{-1}]^{-1}\| \\ &\leq \lambda_1^{(1/2-\alpha)} \rho^{-1} (1 - \theta)^{-1} |\lambda|^{-1}. \end{aligned}$$

Similarly, we have

$$(4.4) \quad \begin{aligned} \|\Delta^{-1}(\lambda) A^{1/2}\| &= \|[I + \lambda(\lambda^2 I + \lambda \rho A^\alpha + A)^{-1} A^\alpha A^{-\alpha} B_1]^{-1} (\lambda^2 I + \lambda \rho A^\alpha + A)^{-1} A^{1/2}\| \\ &\leq \lambda_1^{(1/2-\alpha)} \rho^{-1} (1 - \theta)^{-1} |\lambda|^{-1} \quad \text{for } \operatorname{Re} \lambda \geq 0 \text{ and } \lambda \neq 0. \end{aligned}$$

Also by Lemma 3.5(b), we have

$$(4.5) \quad \begin{aligned} \|\lambda \Delta^{-1}(\lambda)\| &= \|\lambda(\lambda^2 I + \lambda \rho A^\alpha + A)^{-1} [I + \lambda B_1 A^{-\alpha} A^\alpha (\lambda^2 I + \lambda \rho A^\alpha + A)^{-1}]^{-1}\| \\ &\leq \left[1 + \frac{\rho}{2} \lambda_1^{(1/2-\alpha)} \right] \frac{1}{(1 - \theta)|\lambda|} \quad \text{for } \operatorname{Re} \lambda \geq 0 \text{ and } \lambda \neq 0. \end{aligned}$$

Now we estimate $A^{1/2} \Delta^{-1}(\lambda)(\lambda + B)A^{-1/2}$. For $\lambda \neq 0$, from (3.7), we have

$$(4.6) \quad A^{1/2} \Delta^{-1}(\lambda)(\lambda + B)A^{-1/2} = \frac{1}{\lambda} [I - A^{1/2} \Delta^{-1}(\lambda) A^{1/2}].$$

Since $\|A^{-1/2} B_1 A^{1/2} A^{-\alpha}\| \leq \theta \rho$ by Lemma 2.1(c) and $\|A(\lambda^2 I + \lambda \rho A^\alpha + A)^{-1}\| \leq 1 + 2\rho^{-1} \lambda_1^{(1-\alpha)}$ by Lemma 3.5(c), we have

$$\begin{aligned} \|A^{1/2} \Delta^{-1}(\lambda) A^{1/2}\| &= \|A^{1/2}(\lambda^2 A^{-1/2} + \lambda \rho A^{\alpha-1/2} + A^{1/2} + \lambda A^{-1/2} B_1)^{-1}\| \\ &= \|A(\lambda^2 I + \lambda \rho A^\alpha + A + \lambda A^{-1/2} B_1 A^{1/2})^{-1}\| \\ &= \|A(\lambda^2 I + \lambda \rho A^\alpha + A)^{-1} [I + \lambda(A^{-1/2} B_1 A^{1/2} A^{-\alpha}) \\ &\quad \cdot (A^\alpha (\lambda^2 I + \lambda \rho A^\alpha + A)^{-1})^{-1}]\| \\ &\leq (1 + 2\rho^{-1} \lambda_1^{(1-\alpha)})(1 - \theta)^{-1} \quad \text{for } \operatorname{Re} \lambda \geq 0. \end{aligned}$$

Thus, from (4.6), we obtain

$$(4.7) \quad \begin{aligned} \|A^{1/2} \Delta^{-1}(\lambda)(\lambda + B)A^{-1/2}\| \\ \leq [1 + (1 + 2\rho^{-1} \lambda_1^{(1-\alpha)})(1 - \theta)^{-1}] \frac{1}{|\lambda|} \quad \text{for } \operatorname{Re} \lambda \geq 0 \text{ and } \lambda \neq 0. \end{aligned}$$

Therefore, from (4.3)–(4.7), we see that $\{\lambda; \operatorname{Re} \lambda \geq 0, \lambda \neq 0\} \subset \rho(\bar{L}_B)$ and there exists $M > 0$ such that

$$\|(\lambda - \bar{L}_B)^{-1}\| \leq \frac{M}{|\lambda|} \quad \text{for } \operatorname{Re} \lambda \geq 0 \text{ and } \lambda \neq 0.$$

It is proved that \bar{L}_B generates a holomorphic semigroup on W .

Next we prove that (4.1) holds. Since $\lim_{t \rightarrow \infty} t^{-1} \log \|T(t)\| = \sup \{\operatorname{Re} \lambda; \lambda \in \sigma(A)\}$, for any holomorphic semigroup $T(t)$ with the infinitesimal generator A , we need only to prove that

$$(4.8) \quad \sup \{\operatorname{Re} \lambda; \lambda \in \sigma(\bar{L}_B)\} \leq -\min \{\frac{1}{6}\rho\lambda_1^\alpha(1-\theta^2), \frac{1}{2}\rho^{-1}\lambda_1^{(1-\alpha)}\}.$$

In fact, above, we have proved that $\lambda \in \rho(\bar{L}_B)$ for $\operatorname{Re} \lambda \geq 0$ and $\lambda \neq 0$. Also by (3.8), we have that $\lambda = 0 \in \rho(\bar{L}_B)$. Let $\lambda = \tau + i\omega$ with $\tau \in (-\sigma_0, 0)$; then

$$\begin{aligned} |\lambda^2 + \lambda\rho t^\alpha + t|^2 &= \rho^2 t^{2\alpha} |\lambda|^2 + |\lambda|^4 + 2\rho\tau t^{1+\alpha} + 2\rho\tau |\lambda|^2 t^\alpha + t(\lambda^2 + \bar{\lambda}^2) + t^2 \\ &= \theta^2 \rho^2 |\lambda|^2 t^{2\alpha} + |\lambda|^4 + t^2(1 - 2\rho|\tau|t^{\alpha-1}) + 2|\lambda|^2 t^\eta, \end{aligned}$$

where $\eta = |\lambda|^{-2}(\tau^2 - \omega^2) + \rho^2 t^{2\alpha-1}((1-\theta^2)/2 - \rho^{-1}|\tau|t^{-\alpha})$. Since $(1 - 2\rho|\tau|t^{\alpha-1})^{1/2} > (1 - 2\rho|\tau|t^{\alpha-1}) > |\lambda|^{-2}(\omega^2 - \tau^2) - \rho^2 t^{2\alpha-1}((1-\theta^2)/2 - \rho^{-1}|\tau|t^{-\alpha}) > 0$ when $\eta < 0$, it follows that $|\lambda^2 + \lambda\rho t^\alpha + t| > \theta\rho|\lambda|t^\alpha$ for $t \geq \lambda_1$ and $\tau \in (-\sigma_0, 0)$. Thus, $\|A^\alpha(\lambda^2 I + \lambda\rho A^\alpha + A)^{-1}\| \leq [\theta\rho|\lambda|]^{-1}$ and $\|\lambda A^{-1/2} B_1 A^{1/2}(\lambda^2 I + \lambda\rho A^\alpha + A)^{-1}\| \leq |\lambda| \|A^{-1/2} B_1 A^{1/2} A^{-\alpha}\| - \|A^\alpha(\lambda^2 I + \lambda\rho A^\alpha + A)^{-1}\| < 1$ by Lemma 2.1(c), and so $[I + \lambda A^{-1/2} B_1 A^{1/2}(\lambda^2 I + \lambda\rho A^\alpha + A)^{-1}]^{-1} \in \mathcal{L}(H)$. Therefore $\Delta^{-1}(\lambda) A^{1/2} = (\lambda^2 A^{-1/2} + \lambda\rho A^{\alpha-1/2} + A^{1/2} + \lambda A^{-1/2} B_1)^{-1} = A^{1/2}(\lambda^2 I + \lambda\rho A^\alpha + A)^{-1} [I + \lambda A^{-1/2} B_1 A^{1/2}(\lambda^2 I + \lambda\rho A^\alpha + A)^{-1}]^{-1} \in \mathcal{L}(H)$ and $\Delta^{-1}(\lambda) A^{1/2} H \subset A^{1/2}(\lambda^2 I + \lambda\rho A^\alpha + A)^{-1} H = (\lambda^2 A^{-1/2} + \lambda\rho A^{\alpha-1/2} + A^{1/2})^{-1} H = D(A^{1/2})$ and $\lambda \in \rho(\bar{L}_B)$ by Theorem 3.1. Thus (4.8) is proved. \square

COROLLARY 4.2. $\bar{L}_{\rho A^\alpha}$ generates an exponentially stable analytic semigroup on $W = H \oplus H$ for $\alpha \in [\frac{1}{2}, 1]$.

COROLLARY 4.3. Let B be a positive-definite self-adjoint operator with $D(B) \supset D(A^\alpha)$ for some $\alpha \in [\frac{1}{2}, 1]$ and let $BA^{-\alpha} + A^{-\alpha}B \geq \delta I$ for some $\delta > 0$. Then \bar{L}_B generates an exponentially stable holomorphic semigroup on W .

Proof. Since $D(B) \supset D(A^\alpha)$, there exists $r > 0$ such that $\|BA^{-\alpha}\| \leq r$, and so $B^2 \leq r^2 A^{2\alpha}$. Taking $\rho > \delta^{-1}r^2$, then there exists $\theta \in (0, 1)$ such that $r^2 + \rho^2(1-\theta^2) \leq \rho\delta$. Since $A^{-\alpha}B^2A^{-\alpha} + \rho^2(1-\theta^2)I \leq r^2 + \rho^2(1-\theta^2)I \leq \rho\delta I \leq \rho(BA^{-\alpha} + A^{-\alpha}B)$, we have that $B^2 + \rho^2(1-\theta^2)A^{2\alpha} \leq \rho(A^\alpha B + BA^\alpha)$, i.e., $(B - \rho A^\alpha)^2 \leq \theta^2 \rho^2 A^{2\alpha}$. Setting $B_1 = B - \rho A^\alpha$, then B_1 is self-adjoint and $\|B_1 A^{-\alpha}\| \leq \theta\rho$. Thus, by Theorem 4.1, \bar{L}_B generates an exponentially stable holomorphic semigroup on W .

THEOREM 4.4. Let $B = \rho A^\alpha + B_1$ for some $\rho > 0$ and $\alpha \in [\frac{1}{2}, 1]$ and let B_1 be a closed linear operator with $D(B_1) \cap D(B_1^*) \supset D(A^{\alpha_1})$ for some $\alpha_1 < \alpha$. Then \bar{L}_B generates a holomorphic semigroup on W . Moreover, if B_1 is positive self-adjoint, then the semigroup $e^{t\bar{L}_B}$ is exponentially stable.

Proof. If $\alpha > \frac{1}{2}$, since $D(A^{\alpha_1}) \supset D(A^\beta)$ for $\beta \in [\alpha_1, \alpha]$, we can assume that $\alpha_1 \in [\frac{1}{2}, \alpha)$. Thus, by Lemma 2.1(c), $D(A^{-1/2} B_1 A^{1/2}) \supset D(A^{\alpha_1})$, and so by Corollary 2.6.11 in [2, p. 73] there exists $C_0 > 0$ such that

$$(4.9) \quad \|B_j x\| \leq C_0(\eta^{\alpha_1/\alpha} \|x\| + \eta^{\alpha_1/\alpha-1} \|A^\alpha x\|) \quad (j=1, 2)$$

for all $\eta > 0$ and $x \in D(A^\alpha)$, where $B_2 = A^{-1/2} B_1 A^{1/2}$. Let $\Delta^{-1}(\lambda) = (\lambda^2 I + \lambda B + A)^{-1}$ and $\Delta_1^{-1}(\lambda) = (\lambda^2 I + \lambda \rho A^\alpha + A)^{-1}$, taking η_0 and $\tau_0 > 0$ such that $C_0 \eta_0^{\alpha_1/\alpha-1} < \frac{1}{4}\rho$ and

$C_0\eta_0^{\alpha_1/\alpha}\tau_0^{-1}(1+2\rho^{-1}\lambda_1^{(1/2-\alpha)})<\frac{1}{4}$. Then for $\operatorname{Re} \lambda \geq \tau_0$, by (4.9) and Lemma 3.5, we have

$$\begin{aligned} \|\mathcal{R}_2(\lambda)\| &= \|\Delta^{-1}(\lambda)A^{1/2}\| = \|A^{1/2}\Delta_1^{-1}(\lambda)[I+\lambda A^{-1/2}B_1A^{1/2}\Delta_1^{-1}(\lambda)]^{-1}\| \\ (4.10) \quad &\leq \|A^{1/2}\Delta_1^{-1}(\lambda)\|[1-|\lambda|C_0(\eta_0^{\alpha_1/\alpha}\|\Delta_1^{-1}(\lambda)\|+\eta_0^{\alpha_1/\alpha-1}\|A^\alpha\Delta_1^{-1}(\lambda)\|)]^{-1} \\ &\leq 2\lambda_1^{(1/2-\alpha)}\rho^{-1}|\lambda|^{-1}. \end{aligned}$$

By the same reasoning, from (4.9) and (4.3)-(4.7), there exists $M>0$ such that for $\operatorname{Re} \lambda \geq \tau_0$,

$$(4.11) \quad \|\mathcal{R}_j(\lambda)\| \leq M|\lambda|^{-1} \quad (j=1, 3, 4),$$

where $\mathcal{R}_1(\lambda)=A^{1/2}\Delta^{-1}(\lambda)(\lambda+B)A^{-1/2}$, $\mathcal{R}_3(\lambda)=A^{1/2}\Delta^{-1}(\lambda)$ and $\mathcal{R}_4(\lambda)=\lambda\Delta^{-1}(\lambda)$. Therefore from (4.2) it is proved that \bar{L}_B generates a holomorphic semigroup.

Writing

$$(4.12) \quad L_B = \begin{pmatrix} 0 & A^{1/2} \\ -A^{1/2} & -\rho A^{1/2} - B_1 \end{pmatrix} = L_{\rho A^{1/2}} + \begin{pmatrix} 0 & 0 \\ 0 & -B_1 \end{pmatrix} = L_{\rho A^{1/2}} + \mathcal{B}_1,$$

if $\alpha = \frac{1}{2}$, then for $\lambda \in \rho(\bar{L}_B)$ we have

$$(4.13) \quad (\lambda - \bar{L}_B)^{-1} = (\lambda - \bar{L}_{\rho A^{1/2}})^{-1} [I - \mathcal{B}_1(\lambda - \bar{L}_{\rho A^{1/2}})^{-1}]^{-1},$$

and from (3.1)

$$(4.14) \quad \mathcal{B}_1(\lambda - \bar{L}_{\rho A^{1/2}})^{-1} = \begin{pmatrix} 0 & 0 \\ B_1\Delta_1^{-1}(\lambda)A^{1/2} & -\lambda B_1\Delta_1^{-1}(\lambda) \end{pmatrix},$$

where $\Delta_1^{-1}(\lambda) = (\lambda^2 I + \lambda \rho A^{1/2} + A)^{-1}$. Thus, if we apply (4.9) for $\alpha = \frac{1}{2}$ to (4.14), there exists $\tau_0 > 0$ such that for $\operatorname{Re} \lambda \geq \tau_0$

$$(4.15) \quad \|\mathcal{B}_1(\lambda - \bar{L}_{\rho A^{1/2}})^{-1}\| \leq \frac{1}{2}.$$

While by Theorem 4.1, $\bar{L}_{\rho A^{1/2}}$ generates an exponentially stable holomorphic semigroup, there exists $M>0$ such that for $\operatorname{Re} \lambda \geq \tau_0$

$$(4.16) \quad \|(\lambda - \bar{L}_{\rho A^{1/2}})^{-1}\| \leq M|\lambda|^{-1}.$$

Therefore, from (4.13), (4.15) and (4.16), we have

$$\|(\lambda - \bar{L}_B)^{-1}\| \leq 2M|\lambda|^{-1} \quad \text{for } \operatorname{Re} \lambda \geq \tau_0.$$

This shows that \bar{L}_B generates a holomorphic semigroup on W .

Moreover, if B_1 is a positive self-adjoint operator, by Lemma 3.6 it follows that the semigroup $e^{t\bar{L}_B}$ is exponentially stable.

We note that Theorem 4.4 cannot be proved by the perturbation theory of holomorphic semigroups in [1], [2].

Example 4.5. Let $H = L^2(0, 1)$ and $A = d^4/dx^4$ with $D(A) = \{u \in H^4(0, 1); u(0) = u(1) = u''(0) = u''(1) = 0\}$. Let $B_1 = b_0(x) + b_1(x)(d/dx) + b_2(x)(d^2/dx^2) + b_3(x)(d^3/dx^3)$ with $D(B_1) = H^3(0, 1)$ and $B_0 = -(d/dx)(a(x)(d/dx)) + C(x)$ with $D(B_0) = \{u \in H^2(0, 1); u(0) = u(1) = 0\}$, where $b_j(x) \in C^j(0, 1)$ ($j=0, 1, 2, 3$), $a(x) \in C^1(0, 1)$ and ≥ 0 , $C(x) \in L^2(0, 1)$ and ≥ 0 . Then by Theorem 4.4, $\bar{L}_{(\rho A + B_1)}$ and $\bar{L}_{(\rho A + B_0)}$ generate a holomorphic semigroup on $H \oplus H$, while the semigroup $e^{t\bar{L}_{(\rho A + B_0)}}$ is exponentially stable.

Finally we remark that although A is a differential operator, this does not ensure in general that A^α is such ($\frac{1}{2} \leq \alpha < 1$). This does not impede the usefulness in computations in this paper (see the last paragraph in the § 2 of [3]). On the other hand, from Theorem 3.4, $\alpha=1$ and $\alpha=\frac{1}{2}$ are the break for the stability and the analytic property of the semigroup $e^{t\bar{L}_{\rho A^\alpha}}$, respectively. Thus, perhaps if $\frac{1}{2} < \alpha < 1$, the property of the semigroup $e^{t\bar{L}_{\rho A^\alpha}}$ should be the best for application.

REFERENCES

- [1] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, Berlin, 1966.
- [2] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, Berlin, 1983.
- [3] G. CHEN AND D. L. RUSSELL, *A mathematical model for linear elastic systems with structural damping*, Quart. Appl. Math., 39 (1981/1982), pp. 433-454.
- [4] F. L. HUANG, *On the holomorphic property of the semigroup associated with linear elastic systems with structural damping*, Acta Math. Sci. (Chinese), 55 (1985), pp. 271-277.
- [5] ———, *A problem for linear elastic systems with structural damping*, Acta Math. Sci. (Sinica), 6 (1986), pp. 107-113.
- [6] P. MASSATT, *Limiting behavior for strongly damped non-linear wave equations*, J. Differential Equations, 48 (1983), pp. 334-349.
- [7] F. RIESZ AND B. S.-NAGY, *Functional Analysis*, Ungar, New York, 1955.
- [8] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I*, Wiley-Interscience, New York, 1958.
- [9] M. A. KRASNOSELSKII, P. P. ZABREIKO, E. L. PUSTYLIK, AND P. E. SOBOLEVSKII, *Integral Operators in Space of Summable Functions*, Nauka, Moscow, 1966.

A SEPARABLE PIECEWISE LINEAR UPPER BOUND FOR STOCHASTIC LINEAR PROGRAMS*

JOHN R. BIRGE† AND STEIN W. WALLACE‡

Abstract. Stochastic linear programs require the evaluation of an integral in which the integrand is itself the value of a linear program. This integration is often approximated by discrete distributions that bound the integral from above or below. A difficulty with previous upper bounds is that they generally require a number of function evaluations that grows exponentially in the number of variables. We give a new upper bound that requires operations that only grow polynomially in the number of random variables. We show that this bound is sharp if the function is linear and give computational results to illustrate its performance.

Key words. stochastic programming, upper bounds, convex functions, integration

AMS(MOS) subject classification. 90C15

1. Introduction. Stochastic linear programs can be formulated for a variety of applications. Some examples include airline scheduling (Ferguson and Dantzig [1956]), financial planning (Kusy and Ziemba [1986]), energy modeling (Birge [1987]) and water resource planning (Prékopa and Szantai [1978]). The basic model we consider here is the stochastic linear program with recourse in the following general form:

$$\min_x \{c^T x + \mathcal{Q}(x) | Ax = b, x \geq 0\}$$

where

$$\mathcal{Q}(x) = \int Q(x, \xi, \phi) P(d\xi, d\phi)$$

and the *recourse function* is defined as

$$Q(x, \xi, \phi) = \min\{q^T y | Wy = \xi - Tx, u + \phi \geq y \geq 0\},$$

where $x \in \mathbb{R}^{n_1}$, $y \in \mathbb{R}^{n_2}$, $b \in \mathbb{R}^{m_1}$, and (ξ, ϕ) is a random vector on the probability space $(\mathbb{R}^{m_2+n_2}, \mathcal{F}, P)$ with support, $\Xi \times \Phi$. The vectors, c , q , and u , and matrices, A , W , and T are dimensioned correspondingly. The fundamental problem in stochastic programming is to evaluate the integral of \mathcal{Q} . In this paper, we describe a method for finding an upper bound on \mathcal{Q} that requires a polynomial number of operations in the number of random variables.

Previous results in bounding expressions for \mathcal{Q} are described in Birge and Wets [1986a]. The bounds are based on the convexity and positive homogeneity of Q . The first result is due to Jensen's [1906] inequality which provides a lower bound on \mathcal{Q} . The usefulness of this lower bound is that it requires an evaluation of Q at one point (the mean of the random variables) and has been found to be generally sharp in some practical examples (see, e.g., Hausch and Ziemba [1983]). Madansky [1959] provided an upper bound following Edmundson [1956] that is based on the theory of moment spaces and amounts to weighting the extreme points of the support of

* Received by the editors February 18, 1987; accepted for publication June 10, 1987.

† Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109. The work of this author was supported in part by Office of Naval Research grant N00014-86-K-0628, by Dalhousie University during a visit in the Department of Mathematics, Statistics, and Computing Science, and by the National Research Council under a Research Associateship at the Naval Postgraduate School, Monterey, California 93943.

‡ Department of Science and Technology, Chr. Michelsen Institute, N-5036 Fantoft, Norway.

the random variables. Ben-Tal and Hochman [1972] and Huang, Ziemba and Ben-Tal [1977] refined this bound for independent random variables. Dupačová [1976] formulated a bound of the same general type for dependent random variables that was extended to unbounded ranges and nonpolyhedral sets in Gassmann and Ziemba [1986]. Frauendorfer [1986] provided a sharper bound in the bounded range, dependent variable case, and Birge and Wallace [1986] gave a bound and method for refinement for special cases of dependent random variables.

The upper bounds mentioned above all have the property that they are solutions to moment problems with varying conditions. Dupačová's work on minimax solutions (Žáčková [1966]) led to these conclusions and to the use of the generalized moment problem. Ermoliev et al. [1985] provided a general programming framework for solving the general problem. It is used in Birge and Wets [1987] for bounds with piecewise linear approximations on moment constraints and in Cipra [1985] with first and second moment constraints.

The problem with each of these bounds is that they require an exponentially increasing number of function evaluations as the number of random variables increases. An alternative for this situation was given by the ray approximation procedure in Birge and Wets [1986a]. This uses the sublinearity property of the recourse function to obtain a separable function that majorizes Q . This approach is generalized in Birge and Wets [1986b]. Wallace [1987b], on the other hand, formulated a procedure that applies to problems in which the recourse function involves the solution of a network problem. Our procedure is a combination and generalization of these two basic approaches. The algorithm we give provides a separable piecewise linear function that bounds Q throughout the support of the random variables and can be easily evaluated.

Section 2 presents our basic algorithm and the separable piecewise linear upper bound (SPLU). Its properties are described in §3. Section 4 gives an illustrative small example and provides comparison with the upper bound of Edmundson and Madansky. Extensions of the basic algorithm and conclusions are given in §5.

2. The basic algorithm. We give a general method for finding an upper bound on the expected value of the value of a linear program with random right-hand sides and random upper bounds on the variables. To simplify notation and to establish general results, we consider the following system :

$$(1) \quad A_1 x = b_1 + \xi, \quad A_2 x = b_2, \quad 0 \leq x \leq c + \phi$$

where $A_1 \in \mathbb{R}^{m_1 \times n}$, $A_2 \in \mathbb{R}^{(m-m_1) \times n}$, $(A_1 | A_2)^T = A$ is the coefficient matrix, $(b_1 | b_2)^T = b$ is the fixed part of the right-hand side, c is the fixed part of the bounds on the variables, ξ is the random availability of resources and ϕ is the random part of the variable capacities, where $\phi \geq 0$. We assume that there is a positive probability that $\phi = 0$. Next define $Q(\xi, \phi)$ by

$$(2) \quad Q(\xi, \phi) = \min\{q^T x | (1)\}.$$

Finally define $\chi(\xi, \phi, d^-, d^+)$ as the set of x -vectors satisfying

$$(3) \quad A_1 x = \xi, \quad A_2 x = 0, \quad d^- \leq x \leq \phi + d^+.$$

Our goal is to find an upper bound on $Q(\xi, \phi)$, or, more precisely, on $\mathbb{E}Q(\xi, \phi)$. We do this by finding a separable piecewise linear function $U(\xi, \phi)$ defined by

$$U(\xi, \phi) = Q(\bar{\xi}, 0) + H(\phi) + \sum_{i=1}^{m_1} \begin{cases} q^T x^{i+}(\xi_i - \bar{\xi}_i) & \text{if } \xi_i \geq \bar{\xi}_i, \\ q^T x^{i-}(\bar{\xi}_i - \xi_i) & \text{if } \xi_i < \bar{\xi}_i \end{cases}$$

where $\bar{\xi}_i = E\xi_i$, and $H(\phi)$ is a piecewise linear function in ϕ .

ALGORITHM 1.

Step 0. Find $Q(\bar{\xi}, 0)$ with optimal solution x^0 , where

$$x^0(i) = \begin{cases} e_i^T B_0^{-1}(b + \bar{\xi}) & \text{if } i \text{ is basic,} \\ 0 & \text{if } i \text{ is nonbasic at lower bound,} \\ c(i) & \text{if } i \text{ is nonbasic at upper bound.} \end{cases}$$

Assume for simplicity that the first m variables are basic. Let $x^{i+} = (B_0^{-1}e_i, 0, 0, \dots, 0)$ and $x^{i-} = (-B_0^{-1}e_i, 0, 0, \dots, 0)$ where $i = 1, 2, \dots, m_1$. Let

$$\alpha^1(i) = \max - x^0(i) - \sum_{j=2}^{m_1} x^{j+}(i)y^{j+} - \sum_{j=2}^{m_1} x^{j-}(i)y^{j-}$$

subject to

$$\begin{aligned} y^{j+} - y^{j-} &= \xi_j - \bar{\xi}_j, \\ \xi_j^{\min} &\leq \xi_j \leq \xi_j^{\max}, \quad j = 2, \dots, m_1. \end{aligned}$$

Let

$$\beta^1(i) = \min - x^0(i) - \sum_{j=2}^{m_1} x^{j+}(i)y^{j+} - \sum_{j=2}^{m_1} x^{j-}(i)y^{j-} + c(i)$$

subject to

$$\begin{aligned} y^{j+} - y^{j-} &= \xi_j - \bar{\xi}_j, \\ \xi_j^{\min} &\leq \xi_j \leq \xi_j^{\max}, \quad j = 2, \dots, m_1, \end{aligned}$$

for all $i = 1, \dots, n$.

If $\alpha^1(i) > 0$ for some i or $\beta^1(i) < 0$ for some i , let $x^{i+} = x^{i-} = (0, \dots, 0)$, $\alpha^1(i) = -x^0(i)$, and $\beta^1(i) = c(i) - x^0(i)$ for all $i = 1, \dots, m_1$ and go to Step 1 with $r = 1$. Otherwise, check

$$e(i) = \max - x^{1+}(i)y^+ - x^{1-}(i)y^-$$

subject to

$$\begin{aligned} y^+ - y^- &= \xi_1 - \bar{\xi}_1, \\ \xi_1^{\min} &\leq \xi_1 \leq \xi_1^{\max} \end{aligned}$$

and

$$f(i) = \min - x^{1+}(i)y^+ - x^{1-}(i)y^-$$

subject to

$$\begin{aligned} y^+ - y^- &= \xi_1 - \bar{\xi}_1, \\ \xi_1^{\min} &\leq \xi_1 \leq \xi_1^{\max}. \end{aligned}$$

If $\alpha^1 + e \leq 0$ and $\beta^1 + f \geq 0$, then $Q(\xi, \phi)$ is linear in ξ , go to Step 4. Otherwise, let $r = 1$ and go to Step 1.

Step 1. If $\xi_r^{\max} < +\infty$, solve

$$\min \{q^T x \mid \chi[(\xi_r^{\max} - \bar{\xi}_r)e_r, 0, \alpha^r, \beta^r]\} = q^T x^{r+}(\xi_r^{\max} - \bar{\xi}_r).$$

Else (let $\beta_*^r(i) = \infty$ if $\beta^r(i) = +\infty$, $\beta_*^r(i) = 0$ otherwise) and solve

$$\min\{q^T x \mid \chi(e_r, 0, 0, \beta_*^r)\} = q^T x^{r+}.$$

If $\xi_r^{\min} > -\infty$, solve

$$\min\{q^T x \mid \chi[(\xi_r^{\min} - \bar{\xi}_r)e_r, 0, \alpha^r, \beta_*^r]\} = q^T x^{r+}(-\xi_r^{\min} + \bar{\xi}_r).$$

Else (let $\beta_*^r(i) = \infty$ if $\beta^r(i) = +\infty$, $\beta_*^r(i) = 0$ otherwise) and solve

$$\min\{q^T x \mid \chi(-e_r, 0, 0, \beta_*^r)\} = -q^T x^{r-}.$$

If Step 1 was entered with $x^{i\pm} = (0, \dots, 0)$ for all i , go to Step 2; otherwise, go to Step 4.

Step 2. For $i = 1, \dots, n$, solve

$$\alpha^{r+1}(i) = \max - x^0(i) - \sum_{j \neq r+1} x^{j+}(i)y^{j+} - \sum_{j \neq r+1} x^{j-}(i)y^{j-}$$

subject to

$$\begin{aligned} y^{j+} - y^{j-} &= \xi_j - \bar{\xi}_j, \\ \xi_j^{\min} &\leq \xi_j \leq \xi_j^{\max}, \quad j \neq r+1, \end{aligned}$$

$$\beta^{r+1}(i) = \min - x^0(i) - \sum_{j \neq r+1} x^{j+}(i)y^{j+} - \sum_{j \neq r+1} x^{j-}(i)y^{j-} + c(i)$$

subject to

$$\begin{aligned} y^{j+} - y^{j-} &= \xi_j - \bar{\xi}_j, \\ \xi_j^{\min} &\leq \xi_j \leq \xi_j^{\max}, \quad j \neq r+1. \end{aligned}$$

Step 3. If $r < m$, let $r = r + 1$ and go to Step 1. Otherwise, go to Step 4.

Step 4. Find

$$\alpha^*(i) = \max - x^0(i) - \sum_{j=1}^{m_1} x^{j+}(i)y^{j+} - \sum_{j=1}^{m_1} x^{j-}(i)y^{j-}$$

subject to

$$\begin{aligned} y^{j+} - y^{j-} &= \xi_j - \bar{\xi}_j, \\ \xi_j^{\min} &\leq \xi_j \leq \xi_j^{\max}, \quad j = 1, \dots, m_1, \end{aligned}$$

$$\beta^*(i) = \min - x^0(i) - \sum_{j=1}^{m_1} x^{j+}(i)y^{j+} - \sum_{j=1}^{m_1} x^{j-}(i)y^{j-} + c(i)$$

subject to

$$\begin{aligned} y^{j+} - y^{j-} &= \xi_j - \bar{\xi}_j, \\ \xi_j^{\min} &\leq \xi_j \leq \xi_j^{\max}, \quad j = 1, \dots, m_1, \end{aligned}$$

for $i = 1, \dots, n$.

Let $x^* = \operatorname{argmin} \{q^T x \mid \chi(0, \phi^{\max}, \alpha^*, \beta^*)\}$. Find a conformal realization of x^* (Rockafellar [1984, p.455]), so that $x^* = \sum \alpha_k x_k^*$ with $\alpha_k > 0$, such that $x^*(i) > 0 \Rightarrow x_k^*(i) \geq 0$ and $x^*(i) < 0 \Rightarrow x_k^*(i) \leq 0$, and $x^*(i) = 0 \Rightarrow x_k^*(i) = 0$. An algorithm

for finding such a realization is the “painted index algorithm” in Rockafellar [1984, p.476]. Paint all columns A_j of A such that

$$A_j \text{ is } \begin{cases} \text{white} & \text{if } x^*(j) > 0, \\ \text{black} & \text{if } x^*(j) < 0, \\ \text{red} & \text{if } x^*(j) = 0. \end{cases}$$

Let $k = 1$. Pivot until a Tucker tableau is reached in which there is a compatible column. This will always be possible in our case. Let the compatible column be A'_j , and let F be the set of indices for the basic columns in the final Tucker tableau. We now have that

$$\sum_{i \in F} A_i A'_j(i) + A_j = 0.$$

If A_j is white, let

$$x_k^*(i) = \begin{cases} A'_j(i) & \text{if } i \in F, \\ 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

If A_j is black, reverse all signs in x_k^* . (Note that the sign convention in a Tucker tableau is opposite of the convention in the standard simplex tableau.)

Let $\alpha_k = \min\{x^*(i)/x_k^*(i), x_k^*(i) \neq 0\}$, $x^*(i) = x^*(i) - \alpha_k x_k^*(i)$ and repaint every column for which $x^*(i) = 0$ red.

If $x^* \neq 0$, let $k = k + 1$ and repeat. Otherwise, go to Step 5 with the conformal realization $\sum_{k=1}^K \alpha_k x_k^*$.

Step 5. Using the cost coefficients $q^T x^{\pm}$, find $E_\xi U(\xi, \phi)$. This amounts to performing m_1 simple line integrals.

Step 6. If $x^*(i) > 0$ (so that $x_k^*(i) \geq 0$, for all k), we are using a variable $x(i)$ with random capacity $\beta^*(i) + \phi_i (\geq \phi_i)$. If $x^*(i) < 0$, we are using a variable $x(i)$ with deterministic capacity $\alpha^*(i) (< 0)$. We shall in the following assume that each variable $x(i)$, such that $x^*(i) \neq 0$, has associated with it a random arc capacity ϕ_i^* . If $x^*(i) < 0$, we have $\Pr\{\phi_i^* = \alpha^*(i)\} = 1$, if $x^*(i) > 0$, $\phi^* = \phi + \beta^*(i)$. For each $k = 1, \dots, K$, let $q_k = \sum q(i) x_k^*(i) (\leq 0)$. Sort the primal supports x_k^* such that $q_1 \leq q_2 \leq \dots \leq q_K$. Let $k = 1$, $\rho = 0$ (where ρ will become $EH(\phi)$).

Step 7. Let $P = \{i \mid x_k^*(i) \neq 0\}$. Consider the random variable

$$\beta_k = \max\{0, \min_P \{\phi_i^*/x_k^*(i)\}\}, \quad \beta_k \in [0, \alpha_k].$$

Find $Eq_k \beta_k$. (This work amounts to increasing the capacity of each conformal flow until the first variable capacity is met. This continues on each conformal flow. Details are given for the network case in Wallace [1987b].) Let $\rho = \rho + Eq_k \beta_k$, and $\phi^* = \phi^* - \alpha_k x_k^*$. If $k = K$ or if $q_{k+1} = 0$, stop with $\rho = EH(\phi)$, otherwise let $k = k + 1$ and repeat Step 7.

End.

The value obtained in Algorithm 1 is indeed an upper bound on the expected linear program value.

THEOREM. *The value $SPLU = E_{\xi, \phi}[U(\xi, \phi)]$ obtained in Algorithm 1 is an upper bound on $\mathcal{Q} = E_{\xi, \phi}[Q(\xi, \phi)]$.*

Proof. The proof requires only showing that $x = x^0 + \sum (x^{j+}(\xi_j - \bar{\xi})^+ + x^{j-}(\bar{\xi} - \xi_j)^+) + \sum (\beta_k q_k x_k^*)$ is feasible in $\chi(\xi, \phi^*, 0, c)$. This is obtained by noting that the definitions of $x^{r\pm}$, α^r , and β^r in Steps 0 to 2 maintain feasibility for ϕ^* . \square

The algorithm as described above is our basic version. We prove certain properties of it in the next section. In §5, we present alternative versions of some of the steps in Algorithm 1.

3. Properties of the upper bound. The purpose of this section is to show that the upper bound presented in this paper has some desirable properties and to relate the procedure to other bounding methods.

3.1. Exact bounds for linear problems. All other bounds used in stochastic programming are exact whenever $Q(\xi, \phi)$ is linear in ξ and ϕ over the support of the random variables. This is true of the Madansky upper bound, the piecewise linear upper bound in the pure network case (Wallace [1987b]), the linear upper bound on the expected max flow in a network (Wallace [1987a]), the Jensen lower bound, and the sublinear approximation in Birge and Wets [1986b] (if the random variables have unbounded support). All acceptable bounds should have this property.

PROPERTY 1. *The bound SPLU given by Algorithm 1 is exact if $Q(\xi, \phi)$ is a linear function.*

Proof. Assume $Q(\xi, \phi)$ is linear in ξ and ϕ and that the reduced cost of a non-basic variable is always different from zero (a dual nondegeneracy assumption). Then $EU(\xi, \phi) = EQ(\xi, \phi)$. Of course, if Q is linear, it can be written as

$$Q(\xi, \phi) = Q(\bar{\xi}, 0) + \sum_{k=1}^{m_1} f_k(\xi_k - \bar{\xi}_k) + \sum_j h_j \phi_j.$$

Clearly, Step 0 provides us with $Q(\bar{\xi}, 0)$. Also, if Q is linear, $\alpha^1 + e \leq 0$, $\beta^1 + f \geq 0$ in Step 0, since the basis corresponding to $Q(\bar{\xi}, 0)$ is feasible for all $\xi \in \Xi$. Hence, $f_k = qx^{k+} = -qx^{k-}$. Therefore, if Q is linear, the algorithm will discover the coefficients of ξ in Step 0 and then go to Step 4.

Let us define a variable i to be stochastic if $\phi_i^{\max} > 0$; otherwise, it is deterministic. Consider the conformal realization of $x^* = \sum \alpha_k x_k^*$. First note that x_k^* is an elementary vector (Rockafellar [1984, p.453]). This means that there is no way to split x_k^* into two or more other vectors where at least one has fewer nonzeros than x_k^* .

Assume there exists an elementary vector y such that $y(i) \neq 0$ for more than one stochastic random variable. Then fix the value of ϕ_i at 0 for all variables except for those with $y(i) \neq 0$. Then, Q would not be linear. (Compare with the random variable β in Step 7.) Hence, if Q is linear, there is no elementary vector with more than one stochastic variable.

Now, assume that we have found two elementary vectors y_1 and y_2 , such that they share the stochastic variable i (i.e., $y_1(i) \neq 0, y_2(i) \neq 0$). Also assume that $q_1/y_1(i) \neq q_2/y_2(i)$. (The variable q_i defined as in Step 6.) Let all $\phi_j = 0$ for $i \neq j$. Then Q is not linear in variable i , because the marginal gain of increasing ϕ_i is not the same in both elementary vectors. Hence, two elementary vectors can only share a stochastic variable if $q_1/y_1(i) = q_2/y_2(i)$. (This corresponds to two circuits in a pure network that have the same cost and share an arc with a random capacity.) Of course, $h_i = q_1/y_1(i)$.

Hence, if Q is linear, no elementary vector x_k^* has more than one stochastic variable and two elementary vectors can only share a stochastic variable if they have the same cost (in the sense described above). Since Step 6 only creates elementary vectors x_k^* , the random variable β_k in Step 7 is linear in its single random variable. Hence, our method produces the exact solution. \square

3.2. The bound is polynomial. The Edmundson–Madansky bound requires that $Q(\xi, \phi)$ be solved in all extreme cases of ξ and ϕ . There are $2^{m_1+n_1}$ such points; hence, the method is exponential in the number of stochastic variables. Only for very moderate values of n_1 and m_1 is it possible to apply this bound.

The major goal of this paper is therefore to find a good upper bound that can be computed in a number of operations that is polynomial rather than exponential in the number of random variables.

PROPERTY 2. *Algorithm 1 calculates SPLU in a number of operations that is polynomial in the number of random variables.*

Proof. The amount of work is in the worst case:

Step 0. One linear program (α^1, β^1 can be found by inspection).

Step 1. Two times m_1 linear programs.

Step 4. One linear program to find x^* . The conformal realization is independent of n_1 and m_1 . (The worst case is n linear programs, $n \leq n_1$.)

Step 5. The integration is a constant amount of work for each random variable.

Step 7. Finding $Eq_k \beta_k$ amounts to checking the $m_1 * \max_i \{m_i\}$ (in the worst case) possible values of β_k . The value m_i is the total number of possible values for ϕ_i . This has to be done not more than n times (since the number of zeroes increases by one for each k).

Hence, the algorithm is polynomial in n_1 and m_1 . \square

3.3. Relation to networks. The method presented in this paper is closely related to the network method in Wallace [1987b]. The major difference is in Step 1, where we only solve two networks in the network case and not $2m_1$ as here. Below is a short network interpretation of some of the vectors and scalars used in the algorithm to help in its understanding.

Step 0. The variable x^{i+} shows how the flow changes on the basic arcs as the supply at node i is increased by one unit (or the demand is decreased). Hence,

$$x^{i+}(j) = \begin{cases} +1 & \text{if arc } j \text{ is a forward arc on the path from node } i \text{ to the slack node,} \\ -1 & \text{if arc } j \text{ is a reverse arc,} \\ 0 & \text{if arc } j \text{ is not on the path.} \end{cases}$$

The variable x^{i-} is similarly defined for increased demand (or decreased supply).

The value $\alpha^1(i) > 0$ implies that with the chosen set of paths (x_i^\pm) there are supply/demand combinations that give a negative flow on arc i , even when we disregard node 1.

The value $\beta^1(i) > 0$ implies that with the chosen paths, there are supply/demand combinations that overuse arc i even when we do not consider node 1.

Step 1. x_i^\pm are still paths, but not along a basis. Both basic and nonbasic arcs are used. If Step 1 finishes successfully, we have actually replaced the original network by a star-shaped network (where the slack node is in the center of the star). The arc going from the center node to node i has unit cost qx_i^- , the arc in the other direction has unit cost qx_i^+ . The way we have used α^r and β^r has guaranteed that whatever combination we get of supply and demand, sending that flow along the paths x_i^\pm would be feasible and cost the same as in the star-shaped network.

Hence, we have found an upper bounding simple recourse problem (Wets [1983]). In stochastic programming this approximation depends on the actual value of the first stage decisions (as in the recourse function in the introduction). Hence, in some sense, it is a local approximation.

Step 4. $\alpha^* \leq 0$ shows how much flow can be sent along the original arcs in the negative direction without making that total flow negative (whatever the supply/demand is). Similarly, β^* shows how much is left of the capacity in the arcs in the worst case.

The vector x^* is just a circulation in the network, and x_k^* are circuits of minimal length (in terms of the number of arcs in them). α_k shows how much flow the circuit can take (or, more precisely, how much flow it has been allotted.)

Step 7. β_k is a random variable describing the capacity of circuit k .

3.4. Relation to sublinear approximations. The separable piecewise linear upper bound is also a generalization of the ray function approximation in Birge and Wets [1986a] and its extension in the sublinear approximation in Birge and Wets [1986b]. These procedures find the value of $Q(\xi, \phi)$ in different coordinate directions again to obtain a separable function that can easily be integrated. The approach in Birge and Wets [1986b] uses varying choices of the coordinate system that leads to an extension of the SPLU bound given here. This extension would involve solving for x^{j+} and x^{j-} in different directions so that a variety of bounds could be obtained.

The ray function approximation amounts to solving for

$$q^T x^{j+} = \min\{q^T x \mid Ax = e^j, x \geq 0\}$$

and

$$q^T x^{j-} = \min\{q^T x \mid Ax = -e^j, x \geq 0\}.$$

These values of x^{j+} and x^{j-} are then used in $U(\xi, \phi)$ as in SPLU. The extension is to use the elements of other coordinate systems in place of $\pm e^j$ in the definitions (i.e., use some vectors d^j that form a basis for \mathbb{R}^n). This procedure can be used in Algorithm 1 to obtain an alternative bound.

The sublinear approximation with varying directions has been found to produce accurate approximations in a variety of examples. The advantage of the SPLU bound is that it applies to bounded regions so it may be used on partitions of the support of the random variable in a refinement procedure in solving a stochastic program. Algorithm 1 also incorporates the procedures for handling random bounds that often arise in practical examples.

3.5. Finiteness. There is no guarantee that our upper bound is finite, i.e., that all linear programs that must be solved are feasible. An infinite bound of course results if $EQ(\xi, \phi) = +\infty$, i.e., the problem itself is infeasible, but it can also be that $EQ(\xi, \phi) < +\infty$, whereas $EU(\xi, \phi) = +\infty$. This is not always avoidable. We note that the only other polynomial upper bound, the ray approximation, is never better than our bound (assuming the possible extensions mentioned above), and that exponential bounds may be necessary in some cases.

3.6. Partitioning. When approximations, such as the one in this paper, are used in two-stage stochastic programming, a comparison is made with a lower bound (EL) usually based on Jensen's inequality. Then, if $EU - EL$ is too large (according to some rule), the support rectangle (for independent random variables) is partitioned into smaller rectangles called *cells*, and the bounding procedures are applied to these cells, which in turn are weighted by their probability.

Hence, whenever a partition is called for, one must decide which cell to partition and along which coordinate direction to perform the partition. With an upper bounding method that can take on the value $+\infty$ even for a feasible problem, one should clearly partition the cell where $EU = +\infty$, along the coordinate direction that was

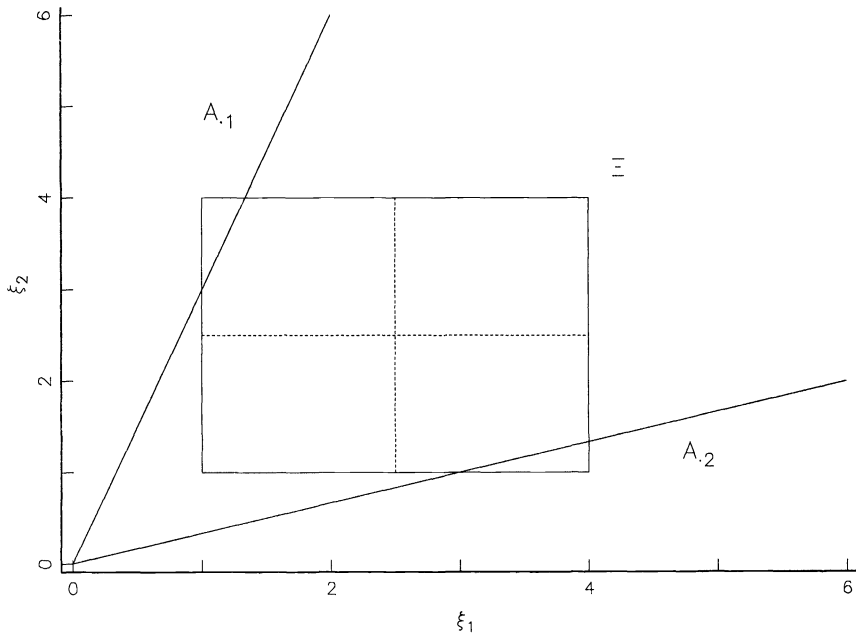


FIG. 1. Example with two random variables.

being treated when the infeasibility was discovered. This provides a dynamic scheme in which the algorithm is applied on each cell until either an infinite value is obtained or the difference between lower and upper bounds is above the acceptable threshold. A partition is made in either instance. Partition strategies are discussed in Birge and Wets [1986a], Birge and Wallace [1986] and Frauendorfer and Kall [1986].

4. Examples. In this section, we first present a small example to illustrate the bound. We then give computational results on a larger problem from energy modeling (Louveaux [1987]).

4.1. A problem with two random variables. The first example is a problem with two random variables and without random capacities. We wish to find bounds on $EQ(\xi)$ where

$$(4.1) \quad Q(\xi) = \min x_1 + x_2 + x_3 + x_4 + 10x_5 + 10x_6$$

subject to

$$(4.2) \quad x_1 + 3x_2 + x_3 - x_5 = \xi_1,$$

$$(4.3) \quad 3x_1 + x_2 + x_4 - x_6 = \xi_2,$$

$$x_1, \dots, x_6 \geq 0$$

where ξ_1 and ξ_2 are uniformly distributed on $[1, 4]$. This problem is illustrated in Fig. 1, where A_i refers to the i th column in the constraint matrix of (4.2-4.3). We follow Algorithm 1 step by step.

Step 0. Find $Q(\bar{\xi}) = 1.25$.

$$x^0 = (0.625, 0.625, 0, 0, 0, 0).$$

$$x^{1+} = (-0.125, 0.375, 0, 0, 0, 0); x^{1-} = (0.125, -0.375, 0, 0, 0, 0).$$

$$x^{2+} = (0.375, -0.125, 0, 0, 0, 0); x^{2-} = (-0.375, 0.125, 0, 0, 0, 0).$$

$$\alpha^1(1) = -0.0625; \alpha^1(2) = -0.4375 \text{ (Note: } \beta^1(i) = +\infty \text{)}.$$

Now $e(1) = 0.1875 > -\alpha^1(1) = 0.0625$, so go to Step 1. Note in Fig. 1 that we have essentially moved along the vertical line through $\bar{\xi}$. The bound α^1 recorded the (negative) minimum multiples of the vectors A_1 and A_2 for points along that line. The value $e(1)$ recorded the greatest change in the multiple of A_1 from the multiple for $\bar{\xi}$ for other points along the horizontal line through $\bar{\xi}$. The function is not linear because this change is greater than the minimal multiple ($\alpha_1(1)$) for movement in the vertical direction.

Step 1. Solve $\min\{q^T x \mid \chi[1.5e_1, 0, (-0.0625, -0.4375, 0, 0, 0, 0), \infty]\} = 1.125 = (0.75) * (1.5) = q^T x^{1+}(\xi_1^{\max} - \bar{\xi}_1)$, where $x^{1+} = (-0.0625, 0.1875, 1.0, 0, 0, 0)$, and $\min\{q^T x \mid \chi[-1.5e_1, 0, (-0.0625, -0.4375, 0, 0, 0, 0), \infty]\} = 1.375 = (0.9167) * (1.5) = q^T x^{1-}(\bar{\xi}_1 - \xi_1^{\min})$, where $x^{1-} = (-0.0625, -0.4375, 0, 0.625, 0.125, 0)$. Next go to Step 4.

Step 4. We can skip this since there are no random bounds. Step 5 then is the terminal step.

Step 5. Here we compute

$$\begin{aligned} EU(\xi) &= Q(\bar{\xi}) + \int_{\xi_1 \geq \bar{\xi}_1} x^{1+}(\xi_1 - \bar{\xi}_1) dF(\xi_1) + \int_{\xi_1 < \bar{\xi}_1} x^{1-}(\bar{\xi}_1 - \xi_1) dF(\xi_1) \\ &\quad + \int_{\xi_2 \geq \bar{\xi}_2} x^{2+}(\xi_2 - \bar{\xi}_2) dF(\xi_2) + \int_{\xi_2 < \bar{\xi}_2} x^{2-}(\bar{\xi}_2 - \xi_2) dF(\xi_2) \\ &= 1.25 + (0.5)(0.75)(0.75) + (0.5)(0.9167)(0.75) \\ &\quad + (0.5)(0.25)(0.75) + (0.5)(-0.25)(0.75) \\ &= 1.875. \end{aligned}$$

End.

So, we have $SPLU = 1.875$. We compare this with the Edmundson–Madansky (EM) bound. In this example, the EM bound assigns equal weights to the values of $Q(\xi)$ at each of the extreme points of Ξ . Hence,

$$EM = 0.25 * (Q(1, 1) + Q(1, 4) + Q(4, 1) + Q(4, 4)) = 1.625.$$

The EM bound is better than the SPLU bound but this difference may be eliminated by refinements of the SPLU bound. We describe possible refinements in §5.

4.2. Computational results for an energy model. The usefulness of the SPLU bound is best demonstrated on a practical example in which the number of random variables varies. We wish specifically to observe the performance of SPLU relative to the EM bound as the number of random variables increases. The performance is measured in the sharpness of the bound and the computational effort. As a practical example, we consider the small energy model in Louveaux [1987]. We do not consider random bounds because that is directly analogous to the network case discussed in Wallace [1987b].

In this example, we have four technologies which can be used to satisfy three demands at varying costs. High cost “backstop” technologies are also available to satisfy demand so the problem is feasible for any demand realization. The randomness occurs in the capacity of the technologies and the demands. This allows from one to seven random variables. The examples were also chosen with varying ranges

TABLE 1
Results for Edmundson-Madansky and SPLU bounds

PROBLEM*	TIME†		VALUE		
	EM	SPLU	EM	SPLU	Jensen
1-NAR	9	10	182.75	182.75	182.75
1-MED	10	18	220.50	220.25	220.00
1-WID	10	21	385.50	341.50	297.50
2-NAR	16	20	183.38	183.06	182.75
2-MED	17	26	220.50	220.50	220.00
2-WID	22	35	389.85	389.10	297.50
3-NAR	22	28	183.38	183.38	182.75
3-MED	25	38	221.38	222.50	220.00
3-WID	33	40	433.60	439.58	297.50
4-NAR	51	41	184.09	185.50	182.75
4-MED	44	41	227.22	255.18	220.00
4-WID	47	45	434.26	469.50	297.50
5-NAR	94	52	184.19	186.44	182.75
5-MED	87	51	227.41	278.38	220.00
5-WID	75	54	434.35	499.18	297.50
6-NAR	163	54	185.58	192.49	182.75
6-MED	193	61	235.91	303.35	220.00
6-WID	149	72	443.52	524.30	297.50
7-MED	366	70	236.27	328.35	220.00
7-WID	304	76	444.12	549.30	297.50

* - Number of random variables - range of random variables

† - CPU milliseconds

(narrow, medium, and wide) on the random variables resulting in twenty-one sets of examples. We assume uniform distributions. This assumption favors bounds (such as the Edmundson-Madansky bound) that place weights at extreme values since other distributions generally have more mass around the center of the support.

The experiments were conducted on the Amdahl 5860 at The University of Michigan Computing Center. The SPLU and EM bounds were both implemented in FORTRAN codes using the same linear programming routine LPM-1 (Pfefferkorn and Tomlin [1976]). Each bound was computed for each of the twenty-one test problems. The Jensen inequality lower bound was also computed to determine the values of the upper bounds relative to the lower bounds. The results are given in Table 1.

The results in Table 1 show that the polynomial bound SPLU does not generally provide as accurate a bound as the EM bound, but that as the number of random

variables increases the computational time in SPLU increases much less rapidly than the time for EM. In these examples, the growth of time for SPLU is indeed approximately linear (gaining ten milliseconds for each random variable), while the time for EM approximately doubles as each new random variable is introduced. This demonstrates that the real advantage of the SPLU bound is in problems with a large number of random variables where the EM bound cannot be computed. These results are comparable with the results in Wallace [1987b] for networks.

Refinements are also possible to reduce the error in SPLU. In §5, a refinement scheme using parametric linear programming is introduced. The use of different coordinate directions is another possibility as mentioned above. For unbounded ranges, the resulting sublinear approximation values were reduced up to thirty percent from the coordinate direction values for a similar set of test problems (Birge and Wets [1986b]).

5. Extensions and conclusions. The SPLU bound can be refined in a variety of ways. The use of other coordinate directions may be possible, but it is best used when linear transformations of the random variables have a known distributional form as is the case for normally distributed random variables. As mentioned above, a common procedure is to partition the support of the random variables and to apply the bound on each of the partitions. Here we give a parametric programming approach that can obtain more accurate results without partitioning the random variables. The following modifications of Algorithm 1 provide this basic bound.

ALGORITHM 2. Substitute the following steps into Algorithm 1 to obtain

$$U'(\xi, \phi) = Q(\bar{\xi}, 0) + H(\phi) + \sum_{\xi_i \geq (<) \bar{\xi}} f_i^{+(-)}((-)\xi_i - (+)\bar{\xi}_i).$$

Step 1'. Solve the parametric linear program

$$\min\{q^T x \mid \chi(\epsilon e_r, 0, \alpha^r, \beta^r)\}$$

for $\epsilon \in [0, \xi_r^{\max} - \bar{\xi}_r]$ (or $\epsilon \in [0, \infty)$ if ξ_r is unbounded). This generates a piecewise linear function $f_r^+(\epsilon e_r)$ with break points $\{0, \epsilon_1, \dots, \epsilon_T\}$, and with slope values, $q^T x_1^{r+}, \dots, q^T x_T^{r+}$. Then, solve the parametric linear program

$$\min\{q^T x \mid \chi(-\epsilon e_r, 0, \alpha^r, \beta^r)\}$$

for $\epsilon \in [0, \bar{\xi}_r - \xi_r^{\min}]$ (or $\epsilon \in [0, \infty)$ if ξ_r unbounded.) We obtain a piecewise linear function $f_r^-(\epsilon e_r)$ with breakpoints $\{0, \epsilon^1, \dots, \epsilon^T\}$, and with slope values, $q^T x_1^{r-}, \dots, q^T x_T^{r-}$.

Step 2' and Step 4'. Substitute $\min_t \{x_t^{j+}(i)\}$ for $x^{j+}(i)$ and $\min_t \{x_t^{j-}(i)\}$ for $x^{j-}(i)$ in the definitions of $\alpha^{r+1}(i)$ and $\alpha^*(i)$ and substitute $\max_t \{x_t^{j+}(i)\}$ for $x^{j+}(i)$ and $\max_t \{x_t^{j-}(i)\}$ for $x^{j-}(i)$ in the definitions of $\beta^{r+1}(i)$ and $\beta^*(i)$.

The changes in Step 1' of Algorithm 2 lead to better bounds if α^r and β^r are the same and $q^T x_1^r < q^T x_T^r$. In Algorithm 2, the approximation obtains as low a value as possible for all ξ_r for changes in the r th direction given the values found for movement in previous directions. In Algorithm 1, the approximation just uses the extreme values of ξ_r .

This difference can be seen in the example from §4.1 which is illustrated in Fig. 2. The dashed line corresponds to the function used in Algorithm 1 by using the extreme values. The solid line corresponds to the functions f_r^+ and f_r^- . The new bound is

$$\text{SPLU}' = E[U'(\xi, \phi)] = 1.449.$$

We note that SPLU' is now below the EM bound value of 1.625.

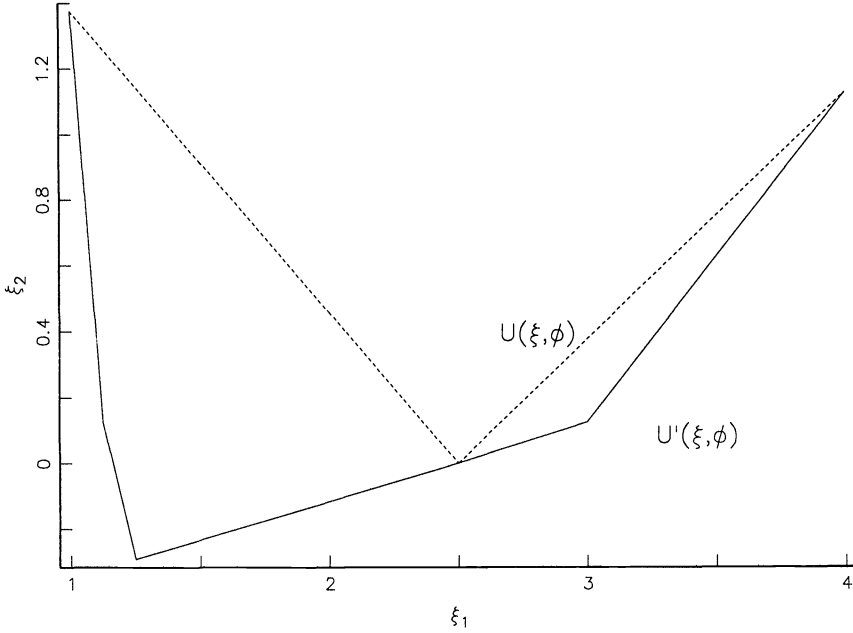


FIG. 2. Parametric linear program bound.

The bound from Algorithm 2 is not always better than SPLU because the bounds may change for different values of r , i.e. α^{r+1} may increase and β^{r+1} may decrease. Although this difference appears to rarely make SPLU' worse than SPLU according to our limited computational experience, it may be advantageous to guarantee that a bound at least as good as SPLU is obtained. This guarantee is accomplished in the following modification of Step 1'.

Step 1''. Solve

$$\min\{q^T x \mid \chi((\xi_r^{\max} - \bar{\xi}_r)e_r, 0, \alpha^r, \beta^r)\} = q^T \bar{x}.$$

Let

$$\alpha_*^r(i) = \begin{cases} \bar{x}(i) & \text{if } \bar{x} < 0, \\ 0 & \text{otherwise} \end{cases}$$

and

$$\beta_*^r(i) = \begin{cases} \bar{x}(i) & \text{if } \bar{x} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then solve the parametric linear program

$$\min\{q^T x \mid \chi((\epsilon e_r)e_r, 0, \alpha_*^r, \beta_*^r)\}$$

to obtain $f_r^+(\epsilon e_r)$ as explained in Step 1'.

This modification of Algorithm 2 results in bounds that are at least as sharp as SPLU and can still benefit from the parametric program as in the example given above. The key benefit of the SPLU bound that the computational effort only grows polynomially with increases in the number of random variables is maintained. We have demonstrated how this improvement results in reduced times on one set of examples and that the greatest value of the SPLU bound may be in cases where the EM and other exponential bounds cannot be reasonably computed. The refinements mentioned above may allow the SPLU bound to be even more useful in the solution of practical stochastic linear programming problems.

REFERENCES

- A. BEN-TAL AND E. HOCHMAN [1972], *More bounds on the expectation of a random variable*, J. Appl. Probab., 9, pp. 803-812.
- J.R. BIRGE [1987], *Exhaustible resource models with uncertain returns from exploration investment*, in Numerical Methods in Stochastic Programming, R. Wets and Y. Ermoliev, eds., Chapter 27, to appear.
- J.R. BIRGE AND S.W. WALLACE [1986], *Refining bounds for stochastic linear programs with linearly transformed independent random variables*, Oper. Research Letters, 5, pp. 73-77.
- J.R. BIRGE AND R. J-B. WETS [1986a], *Designing approximation schemes for stochastic optimization problems, in particular for stochastic programs with recourse*, Math. Programming Stud., 27, pp. 54-102.
- [1986b], *A sublinear approximation method for stochastic programming*, Technical Report 86-24, Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, MI.
- [1987], *Computing bounds for stochastic programming problems by means of a generalized moment problem*, Math. Oper. Res., 12, pp. 149-162.
- T. CIPRA [1985], *Moment problem with given covariance structure in stochastic programming*, Ekonom.-Mat. Obzor, 21, pp. 66-77.
- J. DUPAČOVÁ [1974], *Minimax stochastic programs with nonconvex nonseparable penalty functions*, in Colloq. Math. Soc. János Bolyai, A. Prékopa, ed., North-Holland, Amsterdam, pp. 303-316.
- H.P. EDMUNDSON [1956], *Bounds on the expectation of a convex function of a random variable*, The Rand Corporation, Paper 982, Santa Monica, CA.
- Y. ERMOLIEV, A. GAIVORONSKI AND C. NEDEVA [1985], *Stochastic optimization problems with partially known distribution functions*, this Journal, 23, pp. 696-716.
- A. FERGUSON AND G.B. DANTZIG [1956], *The allocation of aircraft to routes: an example of linear programming under uncertain demands*, Management Sci., 3, pp. 45-73.
- K. FRAUENDORFER [1986], *Solving S.L.P. recourse problems with arbitrary multivariate distributions - the dependent case*, Institute for Operations Research, Universität Zürich, Technical Report.
- K. FRAUENDORFER AND P. KALL [1986], *A Solution Method for SLP recourse problems with arbitrary multivariate distributions - the independent case*, Institute for Operations Research, Universität Zürich, Technical Report.
- H. GASSMANN AND W. T. ZIEMBA [1986], *A tight upper bound for the expectation of a convex function of a multivariate random variable*, Math. Programming Stud., 27, pp. 39-53.
- D. HAUSCH AND W.T. ZIEMBA [1983], *Bounds on the value of information in uncertain decision problems II*, Stochastics, 10, pp. 181-217.
- C. HUANG, W. ZIEMBA AND A. BEN-TAL, *Bounds on the expectation of a convex function of a random variable: with applications to stochastic programming*, Oper. Res., 25, pp. 315-325.
- J.L. JENSEN [1906], *Sur les fonctions convexes et les inégalités entre les valeurs moyennes*, Acta Math., 30, pp. 175-193.
- M. KUSY AND W.T. ZIEMBA [1986], *A bank asset and liability management model*, Oper. Res., 34, pp. 356-376.
- F. LOUVEAUX [1987], *Optimal investment for electricity generation: a stochastic model and a test problem*, in Numerical Methods in Stochastic Programming, R. Wets and Y. Ermoliev, eds., to appear.
- A. MADANSKY [1959], *Bounds on the expectation of a convex function of a multivariate random variable*,

- Ann. Math. Statist., 30, pp. 743–746.
- C.E. PFEFFERKORN AND J.A. TOMLIN [1976], *Design of a linear programming system for ILLIAC IV*, Technical Report SOL 76–8, Systems Optimization Laboratory, Stanford University, Stanford, CA.
- A. PRÉKOPA AND T. SZANTAI [1978], *Flood control reservoir system design using stochastic programming*, Math. Programming Stud., 9, pp. 138–151.
- R.T. ROCKAFELLAR [1984], *Network Flows and Monotropic Programming*, John Wiley, New York.
- S.W. WALLACE [1987a], *Investing in arcs in a network to maximize the expected max flow*, Networks, 17, pp. 87–103.
- [1987b], *A piecewise linear upper bound on the network recourse problem*, Math. Programming, 38, pp. 133–146.
- R. J-B. WETS [1983], *Solving stochastic problems with simple recourse*, Stochastics, 10, pp. 219–242.
- J. ŽÁČKOVÁ [1966], *On minimax solutions of stochastic linear programming problems*, Časopis Pěst. Mat., 91, pp. 423–430.

ERRATUM AND ADDENDUM: THE EXISTENCE OF VALUE AND SADDLE POINT IN GAMES OF FIXED DURATION*

LEONARD D. BERKOVITZ†

Martin Brokate has pointed out an error in the proof of the crucial Lemma 8.3. The point \bar{x}_1 defined on page 186 depends on the strategy Γ_T , and thus the assertion that the right-hand side of (8.7) equals $W^-(t_1, \bar{x}_1)$ is incorrect. Below, we shall present a proof of Lemma 8.3 suggested to us by K. Haji-Ghassemi. Before presenting this proof we shall take this opportunity to improve our definition of strategy.

Let $\mathcal{U}[a, b]$ denote the set of measurable functions u on $[a, b]$ such that $u(t) \in Y$ almost everywhere. Let $\mathcal{X}[a, b]$ denote the set of measurable functions v on $[a, b]$ such that $v(t) \in Z$ almost everywhere.

A strategy Γ for Player I is a choice of a sequence $\Pi = \{\Pi_n\}$ of partitions of $[t_0, T]$ and a choice of a sequence of maps $\Gamma_\Pi = \{\Gamma_{\Pi, n}\}$, where the $\Gamma_{\Pi, n}$ are to be defined below. Thus $\Gamma = (\Gamma_\Pi, \Pi)$. For typographic simplicity, we will suppress the dependence on Π in the notation and write Γ for Γ_Π and $\{\Gamma_n\}$ for $\{\Gamma_{\Pi, n}\}$. We restrict the choice of sequences of partitions to those such that $\|\Pi_n\| \rightarrow 0$, as $n \rightarrow \infty$. Let the partition points of Π_n be $t_0 < t_1 < \dots < t_p = T$. Each map Γ_n is a collection of maps $\Gamma_{n,1}, \dots, \Gamma_{n,p}$ as follows. The map $\Gamma_{n,1}$ selects an element in $\mathcal{U}[t_0, t_1]$. For $2 \leq j \leq p$, the map $\Gamma_{n,j}$ is a map from $\mathcal{U}[t_0, t_{j-1}] \times \mathcal{X}[t_0, t_{j-1}]$ to $\mathcal{U}[t_{j-1}, t_j]$.

A strategy Δ for Player II is a choice of sequence of partitions $\bar{\Pi} = \{\bar{\Pi}_n\}$ of $[t_0, T]$ such that $\|\bar{\Pi}_n\| \rightarrow 0$ as $n \rightarrow \infty$ and a choice of sequence of maps $\{\Delta_n\}$. Each Δ_n is a collection of maps $\Delta_{n,1}, \dots, \Delta_{n,q}$ as follows. If $\bar{\Pi}_n$ has partition points $t_0 = s_0 < s_1 < \dots < s_q = T$, then $\Delta_{n,1}$ selects a function v in $\mathcal{X}[s_0, s_1]$. For $2 \leq j \leq q$, $\Delta_{n,j}$ is a map from $\mathcal{U}[t_0, s_{j-1}] \times \mathcal{X}[t_0, s_{j-1}]$ to $\mathcal{X}[s_{j-1}, s_j]$.

As before, each pair (Γ_n, Δ_n) determines an outcome (u_n, v_n) , and we proceed as before. The change in definition does not involve any major modifications in the arguments of the paper, and in fact simplifies some of the constructions. We note here, however, that in defining the extremal strategies Γ_e and Δ_e the associated sequences of partitions $\{\Pi_{en}\}$ and $\{\bar{\Pi}_{en}\}$ must have the number of partitions at the n th stage equal to n . Thus $p = q = n$.

We now turn to the proof of Lemma 8.3. The first paragraph of the proof stands, and the rest of the proof is replaced by the following.

Let $\Phi_1(u) = \{x_1: x_1 = \psi(t_1, \tau, \xi, u, \zeta): \zeta \text{ an arbitrary relaxed control}\}$. Then $\Phi_1(u) \cap C(v_0) = \phi$. Let $\bar{\Gamma}_n$ denote the constant strategy over $[\tau, t_1]$ with value u . By Lemma 6.1, $\Phi_1(u) = \{x_1: x_1 = \varphi[t_1, \tau, \xi, \bar{\Gamma}_u, \Delta], \Delta \text{ arbitrary over } [\tau, t_1]\}$. By Lemma 6.2, $\Phi_1(u)$ is compact.

Let $v_1 = \inf \{W^-(t_1, x_1): x_1 \in \Phi_1(u)\}$. Since W^- is continuous and $\Phi_1(u)$ is compact, there exists an $\bar{x}_1 \in \Phi_1(u)$ such that $v_1 = W^-(t_1, \bar{x}_1)$. Moreover, since $\Phi_1(u) \cap C(v_0) = \phi$, if $\alpha \equiv v_1 - v_0$, then $\alpha > 0$.

For every $x_1 \in \Phi_1(u)$, there exists a $\Gamma(x_1)$ on $[t_1, T]$ such that for all Δ on $[t_1, T]$ and all motions $\varphi[\cdot, t_1, x_1, \Gamma(x_1), \Delta]$, we have

$$(1) \quad g(\varphi[T, t_1, x_1, \Gamma(x_1), \Delta]) \geq v_1 - \frac{\alpha}{4}.$$

From Lemma 6.5, with $\tau' = \tau$ and θ the identity map, and from the continuity of g we

* SIAM J. Control. Optim., 23 (1985), pp. 172-196.

† Department of Mathematics, Purdue University, West Lafayette, Indiana 47906.

get that for every $x_1 \in \Phi_1(u)$ there exists a $\delta(x_1) > 0$ such that if $|x'_1 - x_1| < \delta(x_1)$, then

$$(2) \quad g(\varphi[T, t_1, x'_1, \Gamma(x_1), \Delta^*]) \geq v_1 - \alpha/2 \geq v_0 + \frac{\alpha}{2}$$

for all Δ^* over $[t_1, T]$.

The open balls $B(x_1, \delta(x_1))$ in R^n with center at x_1 and radius $\delta(x_1) > 0$ cover $\Phi_1(u)$. Since $\Phi_1(u)$ is compact, there exists a finite set of points $\bar{x}_{11}, \dots, \bar{x}_{1,k}$ in $\Phi_1(u)$ such that the balls $B(\bar{x}_{1i}, \delta(\bar{x}_{1i}))$, $i = 1, \dots, k$ cover $\Phi_1(u)$.

The remainder of the proof is devoted to the definition of a strategy $\hat{\Gamma}$ on $[\tau, T]$ such that for all strategies Δ over $[\tau, T]$ and all motions $\varphi[\tau, \xi, \hat{\Gamma}, \Delta]$, we have

$$(3) \quad g(\varphi[T, \tau, \xi, \hat{\Gamma}, \Delta]) \geq v_0 + \frac{\alpha}{2}.$$

This, of course, will lead to a contradiction of the assumption that $(\tau, \xi) \in C(v_0)$, and the lemma will be proved.

Corresponding to each of the points $\bar{x}_{11}, \dots, \bar{x}_{1,k}$ which serve as centers of the balls in the finite open cover of $\Phi_1(u)$ there exists a strategy $\Gamma(\bar{x}_{1i})$ as in (1). Let $\Pi'(i) = \{\Pi'_{i,n}\}$ denote the sequence of partitions of $[t_1, T]$ associated with $\Gamma(\bar{x}_{1i})$, $i = 1, \dots, k$. Let Π'_n be the partition of $[t_1, T]$ that is the common refinement of $\Pi'_{1n}, \dots, \Pi'_{kn}$. Let Π_n be the partition of $[\tau, T]$ such that t_1 is a partition point, the interval $[\tau, t_1]$ is partitioned into n equal subintervals and the interval $[t_1, T]$ is partitioned by Π'_n . We then take $\Pi = \{\Pi_n\}$ to be the sequence of partitions associated with $\hat{\Gamma}$. It follows from the definition of Π that $\|\Pi_n\| \rightarrow 0$ as $n \rightarrow \infty$.

For each $n = 1, 2, \dots$, we now define $\hat{\Gamma}_n = (\hat{\Gamma}_{n,1}, \dots, \hat{\Gamma}_{n,p})$. Let $\tau = \tau_0 < \tau_1 < \dots < \tau_p = T$ be the partition points of Π_n . Let the integer r be such $\tau_r = t_1$. For $i = 1, \dots, r$ we define $\hat{\Gamma}_{n,i}$ to be the mapping that always selects the function u on the interval $[\tau_{i-1}, \tau_i]$. Thus

$$(\Gamma_{n,i}(\mathcal{Y}[\tau_0, \tau_{i-1}] \times \mathcal{Z}[\tau_0, \tau_{i-1}]))(t) = u(t), \quad \tau_{i-1} \leq t < \tau_i$$

for $i = 1, \dots, r$.

For $i = r+1, \dots, p$ we define $\hat{\Gamma}_{n,i}$ as follows. Let $v \in \mathcal{Z}[\tau, t_1]$. The pair of controls (u, v) determine a function φ uniquely as the solution of the differential equation

$$(4) \quad \frac{dx}{dt} = f(t, x, u(t), v(t), \quad x(\tau) = \zeta.$$

If $(t_1, \varphi(t_1))$ does not belong to the cover $\bigcup_{i=1}^k B(\bar{x}_{1i}, \delta(\bar{x}_{1i}))$ of $\Phi_1(u)$, we define $\hat{\Gamma}_{n,i}$, $i = r+1, \dots, p$ to be the map that always selects a fixed element y^* in Y . If $(t_1, \varphi(t_1))$ belongs to the cover there will be a smallest index i_0 , which we take to be 1 for typographic simplicity, such that $(t_1, \varphi(t_1)) \in B(\bar{x}_{11}, \delta(\bar{x}_{11}))$. For $i = r+1, r+2, \dots, p$, we shall take $\hat{\Gamma}_{n,i}$ to be $\Gamma_n(\bar{x}_{11})$ in the following sense.

Let $t_1 = \tau < \tau_{i_1} < \tau_{i_2} < \dots < \tau_p = T$ denote the points in the partition Π_n that are also points in the partition $\Pi'_{1,n}$ for $\Gamma_n(\bar{x}_{11})$. Let u_{n,i_1} be the control that $\Gamma_{n,1}(\bar{x}_{11})$ selects on $[\tau_r, \tau_{i_1}]$. The map $\hat{\Gamma}_{n,r+1}$ then selects the control u_{n,i_1} on $[\tau_{r_1}, \tau_{r+1}]$. For any other integers i such that $r+1 \leq i_1$, the map $\hat{\Gamma}_{n,r+i}$ will select the control $u_{n,i}$, on the interval $[\tau_{r+i-1}, \tau_{r+i}]$. If u_{n,i_2} denotes the control that $\Gamma_{n,2}(\bar{x}_{11})$ selects on $[\tau_{i_1}, \tau_{i_2}]$ then at any partition point of the form τ_{i_1+j} with $i_1+j \leq i_2$, the map $\hat{\Gamma}_{n,i_1+j}$ will select u_{n,i_2} on $[\tau_{i_1+j-1}, \tau_{i_1+j}]$. Proceeding in this fashion we define all of the maps $\hat{\Gamma}_{n,i}$, $i = r+1, \dots, p$. Thus, we have defined $\hat{\Gamma}_n = (\hat{\Gamma}_{n,1}, \dots, \hat{\Gamma}_{n,p})$. We define $\hat{\Gamma} = \{\hat{\Gamma}_n\}$.

Now let Δ be any strategy for Player II over $[\tau, T]$. Then $(\hat{\Gamma}, \Delta)$ will result in a sequence of outcomes (u_n, v_n) , where $u_n(t) = u(t)$ for $\tau \leq t \leq t_1$. Let $\Delta_T = \Delta_T(\Delta)$ be the strategy on $[t_1, T]$ such that the partition of $[t_1, T]$ is that induced by Δ on this interval and such that $\Delta_{T,n}$ is the constant component strategy which always selects v_n at the n th stage.

Let $\varphi[\tau, \tau, \xi, \hat{\Gamma}, \Delta]$ be any motion in the game over $[\tau, T]$ resulting from $(\hat{\Gamma}, \Delta)$. Let $\varphi_n(\tau, \tau, \xi_n, u_n, v_n)$ be the relabeled subsequence of n th stage trajectories converging uniformly to $\varphi[\tau, \tau, \xi, \hat{\Gamma}, \Delta]$. Since $\hat{\Gamma}$ always chooses u over $[\tau, t_1]$ we have that $x_1 \equiv \varphi[t_1, \tau, \xi, \hat{\Gamma}, \Delta]$ belongs to $\Phi_1(u)$. Thus $x_1 \in B(\bar{x}_{1i}, \delta(\bar{x}_{1i}))$ for some $i \in \{1, \dots, k\}$. Let i_0 be the smallest integer for which this is true. To simplify the notation, we again suppose that $i_0 = 1$.

Let

$$x_{1n} = \varphi_n(t_1, \tau, \xi_n, u_n, v_n) = \varphi_n(t_1, \tau, \xi_n, u, v_n),$$

where the equality occurs because $u_n = u$ for $\tau \leq t \leq t_1$. For $t \geq t_1$

$$(5) \quad \varphi_n(t, \tau, \xi_n, u_n, v_n) = \varphi_n(t, t_1, x_{1n}, u_n, v_n),$$

where $\varphi_n(t, t_1, x_{1n}, u_n, v_n)$ is the solution of (4) with $u = u_n$, $v = v_n$ and initial conditions $x(t_1) = x_{1n}$. Since $x_1 \in B(\bar{x}_{11}, \delta(\bar{x}_{11}))$ and $x_{1n} \rightarrow x_1$, there exists an integer N such that for $n > N$, $x_{1n} \in B(\bar{x}_{11}, \delta(\bar{x}_{11}))$. Thus, for $n > N$, the pair (u_n, v_n) , which is the outcome of $(\hat{\Gamma}_n, \Delta_n)$ on $[\tau, T]$, when restricted to the interval $[t_1, T]$ is the outcome of $\Gamma_n(\bar{x}_{11})$ and $\Delta_{T,n}$. If we let $n \rightarrow \infty$ in (5) we can conclude that there exists a motion $\varphi[t_1, t_1, x_1, \Gamma(\bar{x}_{11}), \Delta_T(\Delta)]$ such that for $t_1 \leq t \leq T$

$$\varphi[t, \tau, \xi, \hat{\Gamma}, \Delta] = \varphi[t, t_1, x_1, \Gamma(\bar{x}_{11}), \Delta_T(\Delta)].$$

Since $x_1 \in B(\bar{x}_{11}, \delta(\bar{x}_{11}))$ it follows from (2) that (3) holds for the given Δ and the given motion. However, Δ and the motion were arbitrary, so (3) holds for all Δ and all motions $\varphi[\tau, \tau, \xi, \hat{\Gamma}, \Delta]$. The lemma is proved.

THE IDENTIFICATION OF A DISTRIBUTED PARAMETER MODEL FOR A FLEXIBLE STRUCTURE*

H. T. BANKS[†], S. S. GATES[‡], I. G. ROSEN[§], AND Y. WANG[¶]

Abstract. We develop a computational method for the estimation of parameters in a distributed model for a flexible structure. The structure we consider (part of the "RPL experiment") consists of a cantilevered beam with a thruster and linear accelerometer at the free end. The thruster is fed by a pressurized hose whose horizontal motion effects the transverse vibration of the beam. We use the Euler-Bernoulli theory to model the vibration of the beam and treat the hose-thruster assembly as a lumped or point-mass-dashpot-spring system at the tip. Using measurements of linear acceleration at the tip, we estimate the hose parameters (mass, stiffness, damping) and a Voigt-Kelvin viscoelastic structural damping parameter for the beam using a least-squares fit to the data.

We consider spline based approximations to the hybrid (coupled ordinary and partial differential equations) system; theoretical convergence results and numerical studies with both simulation and actual experimental data obtained from the structure are presented and discussed.

Key words. identification, distributed parameter systems, approximation, flexible structures

AMS(MOS) subject classifications. 35, 65, 73D50, 93B30

1. Introduction. The difficulties involved in the design of practical and efficient control laws for large flexible spacecraft (e.g., the inherent infinite dimensionality of the system, a large number of closely spaced modal frequencies, high flexibility, light damping, a fuel-limited, hostile, highly variable environment, etc.) have stimulated research into the development of system identification and parameter estimation procedures which will yield high fidelity models. A particular area of interest involves schemes for the estimation of material parameters describing, for example, mass, inertia, and stiffness or damping properties in distributed models for the vibration of viscoelastic systems—specifically, mechanical beams, plates and the like. In addition, since the resulting inverse problems are often infinite-dimensional, substantial attention has been focused on approximation; see, for example, [1]–[4], [8], and [12]. In these treatments, the parameter estimation problem is formulated as a least-squares fit to measurements of either displacement or velocity which in many cases involves a

* Received by the editors September 2, 1986; accepted for publication May 8, 1987. The authors' research scheme was tested with data obtained from an experimental structure designed and built at the Charles Stark Draper Laboratory, Cambridge, Massachusetts with funding provided by the U.S. Air Force Rocket Propulsion Laboratory, Edwards Air Force Base, California 93523.

[†] Center for Control Sciences, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The work of this author was supported in part by the National Science Foundation under grant MCS-8504316, the Air Force Office of Scientific Research under contract AFOSR-84-0398, and the National Aeronautics and Space Administration under grant NAG-1-517. Part of this research was carried out while this author was a visiting scientist at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, Virginia 23368, which is operated under NASA contracts NAS1-17070 and NAS1-18107.

[‡] The Charles Stark Draper Laboratory, Inc., Cambridge, Massachusetts 02139.

[§] Department of Mathematics, University of Southern California, Los Angeles, California 90089. The work of this author was supported in part by the Air Force Office of Scientific Research under contract AFOSR-84-0393. Part of this research was carried out while this author was a visiting scientist at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, Virginia 23368, which is operated under NASA contracts NAS1-17070 and NAS1-18107.

[¶] Department of Mathematics, University of Southern California, Los Angeles, California 90089. The work of this author was supported in part by the Air Force Office of Scientific Research under contract AFOSR-84-0393.

problem with *bounded* observations of the state variables. Although significant gains have been made in the development of instrumentation to measure displacement and velocity (e.g., laser technology, etc.), one of the least expensive, most reliable, and most commonly used sensors is the linear accelerometer. While in principle it is possible to integrate acceleration measurements once or twice to obtain respectively velocity or displacement data, in practice this task can pose significant challenges. For example, integration of the signal could result in the amplification of low frequency measurement noise or dynamic effects which have not been included in the underlying model. In light of this, we have undertaken to show here, both theoretically and computationally, that a scheme in the spirit of those developed in the previously cited references can also be effectively used with acceleration measurements. Since such schemes entail inverse procedures for problems with *unbounded* state observations, their theoretical foundation involves a *nontrivial* extension of the familiar variational arguments which have been used to demonstrate the convergence of the finite element state approximations upon which the identification schemes are based.

The other primary motivation for the present effort is that while these methods have been extensively tested and evaluated with simulation data, they have never been tried with actual experimental data. We have tested our scheme with data obtained from an experimental structure which was designed and constructed at the Charles Stark Draper Laboratory in Cambridge, Massachusetts, with funding provided by the United States Air Force Rocket Propulsion Laboratory (RPL). The RPL structure (as we will henceforth refer to it) was designed to serve as a test bed for the implementation and evaluation of control algorithms for large angle slewing of spacecraft with flexible appendages. The structure was specifically designed to exhibit structural modes and damping characteristics representative of realistic large flexible space structures.

In § 2 we describe the RPL structure (its geometry, instrumentation, etc.) and formulate an inverse problem involving a distributed system. In § 3, we use the resulting infinite-dimensional estimation problem to motivate the development of a finite-dimensional, finite-element-based approximation scheme. We also discuss our theoretical convergence results. In § 4 we present numerical findings.

We use standard notation throughout. For X a normed linear space, $L(X)$ denotes the space of bounded linear operators from X into X . For Ω an interval and $k = 0, 1, 2, \dots$, $C^k(\Omega; X)$ denotes the space of functions from Ω into X which are k times continuously strongly differentiable on Ω . When $k = 0$ we shall simply write $C(\Omega; X)$. A function f from Ω into X will be said to belong to $L_2(\Omega; X)$ if $\int_{\Omega} |f(t)|_X^2 dt < \infty$. For $k = 0, 1, 2, \dots$, $H^k(\Omega; X)$ denotes the completion of $C^k(\Omega; X)$ with respect to the norm

$$|f|_k = \left(\sum_{j=0}^k \int_{\Omega} |f^{(j)}(t)|_X^2 dt \right)^{1/2}.$$

If, in addition, X is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_X$, then $H^k(\Omega; X)$ is a Hilbert space with inner product

$$\langle f, g \rangle_k = \sum_{j=0}^k \int_{\Omega} \langle f^{(j)}(t), g^{(j)}(t) \rangle_X dt.$$

When $X = \mathbb{R}$, we use the abbreviated notation $C^k(\Omega)$, $L_2(\Omega)$ and $H^k(\Omega)$. Note that $H^0(\Omega) = L_2(\Omega)$ and $\langle \cdot, \cdot \rangle_0$ is the standard inner product on $L_2(\Omega)$.

2. The identification problem. The RPL structure (see Fig. 2.1 below) consists of four flexible appendages which are cantilevered at right angles to one another from a

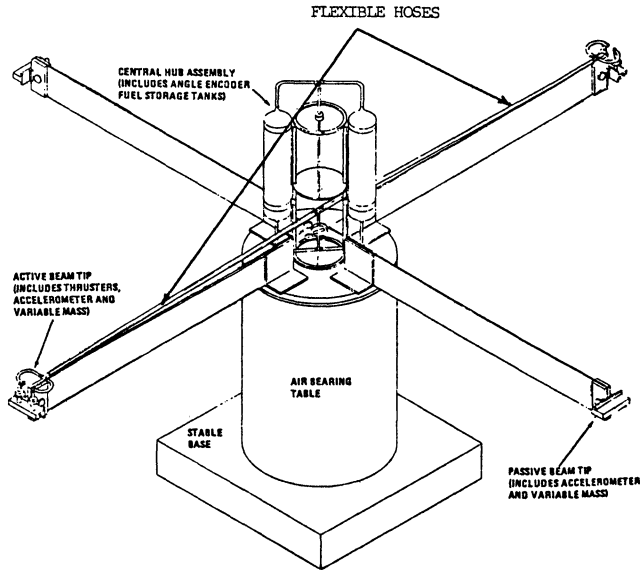


FIG. 2.1

rigid central hub. The hub is mounted on an air bearing table thus permitting the near frictionless rotation of the structure about the vertical axis.

Two of the appendages (which are mounted to the hub 180° apart) are "active"; each has two nitrogen cold gas thrusters mounted in opposing directions at its tip. The remaining two appendages are "passive" with only counterbalancing masses affixed to their free ends. The presence of the tip masses on the passive arms serves to preserve the overall symmetry of the structure. Nitrogen gas from tanks mounted on the central hub is supplied to the thrusters via two stainless steel mesh-wrapped high pressure hoses. The expulsion of propellant from the thruster nozzles is controlled by electro-mechanical or solenoidal valves. Each of the four appendages is equipped with a sensor in the form of a linear accelerometer attached at its tip. Data from the accelerometers is processed and recorded and control input signals to the thrusters are generated by a MINC 11/23 microcomputer. A detailed description of the structure's design specifications can be found in [6] and [15].

The problem which is of primary concern to us here involves the modeling of the effects of the nitrogen supply hoses on the transverse vibration of the active members. We consider therefore, the structure with the central hub immobilized and look only at the vibration of one of the active appendages and view it as a simple cantilevered beam (see Fig. 2.2).

We treat the thruster assembly as a point mass that is rigidly attached to the beam at the tip and propose a model for the hose effects in the form of a proof mass which reacts against the tip mass. In effect, we consider the idealized, simplified structure depicted in Fig. 2.3 below involving a single, cantilevered, flexible, uniform beam with a two-mass-dashpot-spring system affixed to its free end.

In formulating a mathematical model for the structure shown in Fig. 2.3, we assume that the beam is of length l with uniform rectangle cross section of height h and width b . We let $u(t, x)$ and $y(t)$ denote, respectively, the transverse displacement of the beam at position x along its span and the displacement of the proof or hose mass, each at time t . Both are measured relative to the x -axis in the coordinate frame

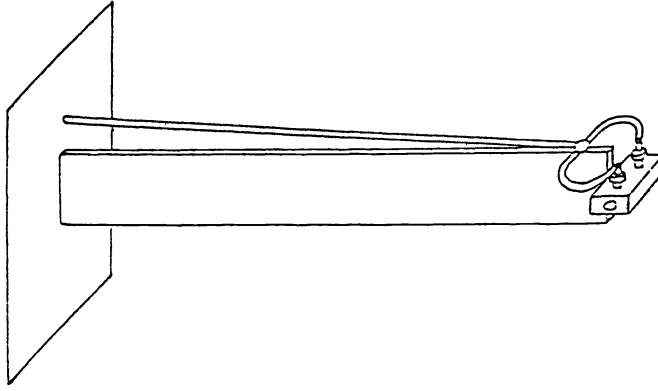


FIG. 2.2

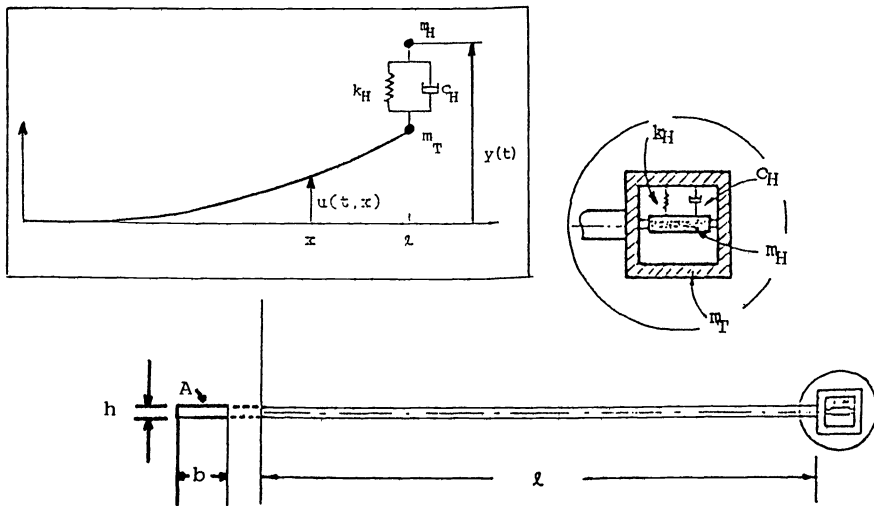


FIG. 2.3

determined by the longitudinal axis of the beam in its undeformed state with origin located at the beam's root or fixed end. Assuming the beam undergoes only small deformations (i.e., $|u(t, x)| \ll l$ and $|(\partial u / \partial x)(t, x)| \ll 1$) and has a small height to span length ratio, the Euler-Bernoulli theory (see [5]) including Voigt-Kelvin viscoelastic structural damping (see [10]) yields the partial differential equation

$$(2.1) \quad \rho \frac{\partial^2 u}{\partial t^2}(t, x) + c_D I \frac{\partial^4 u}{\partial x^4} \frac{\partial u}{\partial t}(t, x) + EI \frac{\partial^4 u}{\partial x^4}(t, x) = 0, \quad 0 < x < l, \quad t > 0$$

where ρ is the linear mass density of the beam, E is the modulus of elasticity, c_D is the coefficient of viscosity, and I is the second moment or moment of inertia of the cross-sectional area A about the neutral axis. For the beam we consider here with constant rectangular cross-section, $I = bh^3/12$. Since the beam is assumed to be uniform, the parameters ρ , E , and c_D are taken to be constant in time and space.

Balancing forces at the free end, elementary Newtonian mechanics yields the equations of motion

$$(2.2) \quad \begin{aligned} m_T \frac{\partial^2 u}{\partial t^2}(t, l) - c_D I \frac{\partial^3}{\partial x^3} \frac{\partial u}{\partial t}(t, l) - EI \frac{\partial^3 u}{\partial x^3}(t, l) \\ = c_H \left(\frac{dy}{dt}(t) - \frac{\partial u}{\partial t}(t, l) \right) + k_H (y(t) - u(t, l)) + f(t), \quad t > 0, \end{aligned}$$

$$(2.3) \quad m_H \frac{d^2 y}{dt^2}(t) + c_H \left(\frac{dy}{dt}(t) - \frac{\partial u}{\partial t}(t, l) \right) + k_H (y(t) - u(t, l)) = 0, \quad t > 0$$

for the tip and hose masses m_T and m_H , respectively. Here k_H is the hose stiffness, c_H is the hose damping coefficient, and $f(t)$ is the externally applied force at time t due to the firing of the thrusters mounted at the tip.

Making the assumption that the rotatory inertia of the proof mass system is negligible, we can express rotational equilibrium at the tip as

$$(2.4) \quad c_D I \frac{\partial^2}{\partial x^2} \frac{\partial u}{\partial t}(t, l) + EI \frac{\partial^2 u}{\partial x^2}(t, l) = 0, \quad t > 0.$$

The zero displacement and zero slope constraints at the fixed end are given by

$$(2.5) \quad u(t, 0) = 0 \quad \text{and} \quad \frac{\partial u}{\partial x}(t, 0) = 0, \quad t > 0,$$

respectively. Taking the structure to be initially at rest, we have the initial conditions

$$(2.6) \quad u(0, x) = 0 \quad \text{and} \quad \frac{\partial u}{\partial t}(0, x) = 0, \quad 0 \leq x \leq l,$$

$$(2.7) \quad y(0) = 0 \quad \text{and} \quad \frac{dy}{dt}(0) = 0.$$

In the mathematical model given by (2.1)–(2.7) above the parameters ρ , m_T and I can be measured or computed directly. The modulus of elasticity E is typically determined in the laboratory. For the most commonly used materials (including aluminium, which is the material from which the structure of interest to us here is made) its value can be readily looked up in standard engineering tables. The parameters c_D , m_H , c_H , and k_H , on the other hand, must be determined experimentally; that is, they will have to be identified according to the observed response of the structure to a given input disturbance. This is one class of inverse problems which we formulate and consider below. In the system of equations (2.1)–(2.7) we explicitly modeled (albeit, in a rather simple fashion) the dynamical effects of the hose. The unknown hose parameters are then determined as the solution to an inverse problem.

An alternative approach to obtaining a model which exhibits a reasonable degree of fidelity involves a technique sometimes referred to as model adjustment. We start with a simple model and the parameters are then “adjusted” so as to compensate for unmodeled dynamics. The choice of parameters to be adjusted and the resulting variations may or may not be motivated by physical considerations.

In our problem, for example, we might consider a simple cantilevered beam with tip mass (i.e., $m_H = c_H = k_H = 0$) and then adjust the theoretical or measured values of E and m_T to compensate for the dynamical effects which result from the hose mass and motion. A value for the parameter c_D could also be identified if damping effects

are considered significant. Model adjustment was used in [6] to obtain a model for the RPL structure upon which control design could be based.

We define an inverse problem which encompasses both of the general approaches which have been outlined above. We assume that an input disturbance described by the function $f(t)$, $t \in [0, T]$ is applied to the structure via the tip thrusters, and that the linear acceleration at the free end of the beam, $z(t)$, is measured and recorded for each $t \in [t_0, t_1]$, where $0 \leq t_0 \leq t_1 \leq T$. (Of course, in actual practice, z could in fact only be sampled discretely.) Let R_+ denote the positive real numbers and let Q be a closed and bounded subset of R_+^6 . We seek a $\bar{q} \in Q$ which minimizes

$$J(q) = \int_{t_0}^{t_1} \left| \frac{\partial^2 u}{\partial t^2}(t, l; q) - z(t) \right|^2 dt$$

where $u(\cdot, \cdot; q)$ denotes the solution to the initial-boundary value problem (2.1)–(2.7) corresponding to $q = (m_T, E, c_D, m_H, c_H, k_H) \in Q$.

Our primary concerns in the next section will include well-posedness of the system (2.1)–(2.7), existence of a minimizer for J , and development of approximation techniques to find this minimizer.

3. Approximation theory. A computational method for the solution of the estimation problem posed above will invariably involve finite-dimensional approximation of the initial-boundary value problem (2.1)–(2.7). We have been successful in solving inverse problems for distributed parameter models for flexible structures (see, for example, [1]–[4], [12]) using spline-based Ritz–Galerkin techniques. These previous efforts, however, involve problems with displacement and/or velocity observations which can be essentially treated as bounded in the state variables in many cases. For the problems under consideration here, the state variables are $\hat{u}(t) = (y(t), u(t, l), u(t, \cdot))$ and thus even when the state equations are rewritten as a first-order system, tip acceleration observations (i.e., observations of $u_{tt}(t, l)$) involve *unbounded* observations for which theoretical approximation results are not readily obtained. Nonetheless, we show below (Lemma 3.1 is the significant new theoretical result) that a theoretical framework can be developed for finite element approximations based upon an abstract Hilbert space formulation of the hybrid system of ordinary and partial differential equations and boundary conditions given in (2.1)–(2.7). This abstract formulation is also useful in establishing existence, uniqueness, and necessary regularity results for solutions. We briefly outline the essential features of our general approach (including theoretical convergence results) in the context of the particular problem of interest to us here.

Let $H = R^2 \times L_2(0, l)$ be endowed with the usual product spaces inner product

$$\langle (\zeta, \eta, \phi), (\lambda, \mu, \psi) \rangle_H = \zeta\lambda + \eta\mu + \langle \phi, \psi \rangle_0$$

and let

$$V = \{(\zeta, \eta, \phi) \in H : \phi \in H^2(0, l), \phi(0) = D\phi(0) = 0, \eta = \phi(l)\}$$

be endowed with the inner product

$$\langle (\zeta, \phi(l), \phi), (\lambda, \psi(l), \psi) \rangle_V = (\zeta - \phi(l))(\lambda - \psi(l)) + \langle D^2\phi, D^2\psi \rangle_0$$

where the symbol D is used here and below to denote the spatial differentiation operator d/dx . The space V together with the inner product $\langle \cdot, \cdot \rangle_V$ form a Hilbert space which is densely and compactly embedded in H .

We rewrite the system (2.1)–(2.7) as the abstract second-order initial value problem in H

$$(3.1) \quad M\hat{u}_t(t) + C\hat{u}_t(t) + K\hat{u}(t) = F(t), \quad t > 0,$$

$$(3.2) \quad \delta(\hat{u}(t) + \varepsilon\hat{u}_t(t)) = 0, \quad t > 0,$$

$$(3.3) \quad \hat{u}(0) = 0, \quad \hat{u}_t(0) = 0$$

in the states $\hat{u}(t) = (y(t), u(t, l), u(t, \cdot))$. The operators $M \in L(H)$, $C : \mathcal{D} \subset H \rightarrow H$ and $K : \mathcal{D} \subset H \rightarrow H$ are given by

$$(3.4) \quad \begin{aligned} M(\zeta, \eta, \phi) &= (m_H \zeta, m_T \eta, \rho \phi), \\ C(\zeta, \eta, \phi) &= (c_H(\zeta - \eta), c_H(\eta - \zeta) - c_D ID^3 \phi(l), c_D ID^4 \phi), \\ K(\zeta, \eta, \phi) &= (k_H(\zeta - \eta), k_H(\eta - \zeta) - EID^3 \phi(l), EID^4 \phi) \end{aligned}$$

where $\mathcal{D} = \{(\zeta, \eta, \phi) \in V : \phi \in H^4(0, l)\}$. For each $t > 0$, $F(t) = (0, f(t), 0) \in H$, $\delta : \mathcal{D} \subset H \rightarrow R$ is given by $\delta((\zeta, \eta, \phi)) = D^2 \phi(l)$ and $\varepsilon = c_D/E$.

The restrictions \tilde{C} and \tilde{K} of the operators C and K that appear in (3.1) above to $\mathcal{N}(\delta)$, the null space of the operator δ , have natural extensions to bounded operators from V (which is the V -closure of $\mathcal{N}(\delta)$) into V' , the dual of V . The extensions are defined in terms of the bilinear forms $c(\cdot, \cdot) : V \times V \rightarrow R$ and $k(\cdot, \cdot) : V \times V \rightarrow R$ given by

$$(3.5) \quad (\tilde{C}\hat{\phi})(\hat{\psi}) = c(\hat{\phi}, \hat{\psi}) = c_H(\zeta - \phi(l))(\lambda - \psi(l)) + c_D I \langle D^2 \phi, D^2 \psi \rangle_0,$$

$$(3.6) \quad (\tilde{K}\hat{\phi})(\hat{\psi}) = k(\hat{\phi}, \hat{\psi}) = k_H(\zeta - \phi(l))(\lambda - \psi(l)) + EI \langle D^2 \phi, D^2 \psi \rangle_0$$

for $\hat{\phi} = (\zeta, \phi(l), \phi) \in V$ and $\hat{\psi} = (\lambda, \psi(l), \psi) \in V$.

The finite element method we develop below could be derived from standard energy considerations. While this is not the approach we take, it is worth noting that the usual energy expressions can be given in terms of the forms, operators, and inner products defined above. The kinetic energy is given by

$$\mathcal{T}_0 = \frac{1}{2} \langle M\hat{u}_t(t), \hat{u}_t(t) \rangle_H,$$

the potential or strain energy by

$$\mathcal{U}_0 = \frac{1}{2} k(\hat{u}(t), \hat{u}(t)),$$

and the Rayleigh dissipation function by

$$\mathcal{F}_0 = \frac{1}{2} c(\hat{u}_t(t), \hat{u}_t(t)).$$

Written in its weak, variational, or distributional form

$$(3.7) \quad \langle M\hat{u}_t(t), \hat{\phi} \rangle_H + c(\hat{u}_t(t), \hat{\phi}) + k(\hat{u}(t), \hat{\phi}) = \langle F(t), \hat{\phi} \rangle_H, \quad t > 0, \quad \hat{\phi} \in V,$$

$$(3.8) \quad \hat{u}(0) = 0, \quad \hat{u}_t(0) = 0$$

the initial value problem (3.1)–(3.2) in H becomes an initial value problem in V' . If we assume that $f \in L_2(0, T)$ and rewrite (3.7), (3.8) as an equivalent first order vector system, the theory of abstract parabolic systems (see [9], [14]) yields the existence of a unique mapping

$$\hat{u} \in C([0, T]; V) \cap H^1((0, T); V) \cap C^1([0, T]; H) \cap H^2((0, T); V')$$

which satisfies (3.7), (3.8). If we are willing to assume further that f is Hölder continuous then there exists a

$$(3.9) \quad \hat{u} \in C([0, T]; V) \cap C^1((0, T]; V) \cap C^1([0, T]; H) \cap C^2((0, T]; H)$$

with $\hat{u}(t) + \varepsilon \hat{u}_t(t) \in \mathcal{D}$, $t > 0$ which uniquely satisfies the initial value problem (3.1)–(3.3).

In order to demonstrate the convergence of the approximation schemes we develop below, we shall require a somewhat more regular solution to the initial value problem (3.7), (3.8) than either of the conditions of f stated above can guarantee. In addition to (3.9), we shall require that $\hat{u} \in H^2((0, T); V)$. This can be guaranteed (see [7]) if we assume that $f \in H^1(-\tau, T)$ for some $\tau > 0$ with $f(t) = 0$, $t < 0$ and we modify our original mathematical model so that

$$(3.10) \quad F(t) = f(t)\hat{\theta}, \quad t \in [-\tau, T]$$

for some $\hat{\theta} = (0, \theta(l), \theta)$, a fixed element in V . We note that with $\hat{\theta}$ chosen appropriately in V , F given by (3.10) may in fact represent an improved model of reality when compared with our present choice of F , where $\hat{\theta} = (0, 1, 0) \in H$.

Central to our approach is a cubic spline, based Galerkin approximation to the initial value problem (3.7), (3.8). For each $N = 1, 2, \dots$, let Δ^N denote the uniform mesh $\{0, l/N, 2l/N, \dots, l\}$ on $[0, l]$ and let $\{B_j^N\}_{j=-1}^{N+1}$ denote the usual cubic B-splines defined with respect to the nodal set Δ^N (see [11], [13]). Briefly, each B_j^N is a C^2 function on $[0, l]$ which is a cubic polynomial on each subinterval $[(k-1) \times (l/N), k(l/N)]$, $k = 1, 2, \dots, N$. The support of B_j^N is $[(j-2)(l/N), (j+2)(l/N)] \cap [0, l]$ with $B_j^N(j(l/N)) = 4$, $DB_j^N(j(l/N)) = 0$, $B_j^N((j \pm 1)(l/N)) = 1$, and $DB_j^N((j \pm 1)(l/N)) = \mp N/l$. Defining $\{\beta_j^N\}_{j=1}^{N+1}$ by $\beta_1^N = B_0^N - 2B_1^N - 2B_{-1}^N$ and $\beta_j^N = B_j^N$, $j = 2, 3, \dots, N+1$, we have $\beta_j^N(0) = DB_j^N(0) = 0$, $j = 1, 2, \dots, N+1$. With $\hat{\beta}_0 = (1, 0, 0)$ and $\hat{\beta}_j^N = (0, \beta_j^N(l), \beta_j^N)$, $j = 1, 2, \dots, N+1$, $V^N = \text{span}\{\beta_j^N\}_{j=0}^{N+1}$ is an $(N+2)$ -dimensional subspace of V .

The Galerkin equations in V^N corresponding to (3.7), (3.8) for $\hat{u}^N(t) \in V^N$ are given by

$$(3.11) \quad \begin{aligned} & \langle M\hat{u}_t^N(t), \hat{\beta}_j^N \rangle_H + c(\hat{u}_t^N(t), \hat{\beta}_j^N) + k(\hat{u}^N(t), \hat{\beta}_j^N) \\ & = \langle F(t), \hat{\beta}_j^N \rangle_H, \quad t > 0, \quad j = 0, 1, 2, \dots, N+1, \end{aligned}$$

$$(3.12) \quad \hat{u}^N(0) = 0, \quad \hat{u}_t^N(0) = 0.$$

We set

$$\hat{u}^N(t) = \sum_{j=0}^{N+1} w_j^N(t) \hat{\beta}_j^N, \quad t \geq 0,$$

and the initial value problem (3.11), (3.12) in V^N is equivalent to the linear, non-homogeneous, second-order $(N+2)$ -vector system

$$(3.13) \quad M^N \frac{d^2 w^N}{dt^2}(t) + C^N \frac{dw^N}{dt}(t) + K^N w^N(t) = F^N(t), \quad t > 0,$$

$$(3.14) \quad w^N(0) = 0, \quad \frac{dw^N}{dt}(0) = 0$$

where $w^N(t) = (w_0^N(t), w_1^N(t), \dots, w_{N+1}^N(t))^T$. The entries in the $(N+2) \times (N+2)$ matrices M^N , C^N , and K^N are given by

$$M_{i,j}^N = \langle M\hat{\beta}_i^N, \hat{\beta}_j^N \rangle_H, \quad C_{i,j}^N = c(\hat{\beta}_i^N, \hat{\beta}_j^N), \quad K_{i,j}^N = k(\hat{\beta}_i^N, \hat{\beta}_j^N),$$

$i, j = 0, 1, 2, \dots, N+1$, respectively. For each $t > 0$ the components in the $N+2$ -vector $F^N(t)$ are given by $F_i^N(t) = \langle F(t), \hat{\beta}_i^N \rangle_H = f(t)\beta_i^N(l)$ or, if we recall (3.10), by

$$F_i^N(t) = f(t)\langle \hat{\theta}, \hat{\beta}_i^N \rangle_H = f(t)(\theta(l)\beta_i^N(l) + \langle \theta, \beta_i^N \rangle_0), \quad i = 0, 1, 2, \dots, N+1.$$

We consider the sequence of approximation finite-dimensional identification problems which consist of finding $\bar{q}^N \in Q$ which minimizes

$$(3.15) \quad J^N(q) = \int_{t_0}^{t_1} \left| \frac{\partial^2 u^N}{\partial t^2}(t, l; q) - z(t) \right|^2 dt$$

where for each $q \in Q$, $\hat{u}^N(t; q) = (y^N(t; q), u^N(t, l; q), u^N(t, \cdot; q))$ is the unique solution to the initial value problem (3.11), (3.12) in V^N corresponding to $q = (m_T, E, c_D, m_H, c_H, k_H) \in Q$. In actual practice, for a given $q \in Q$, $J^N(q)$ is computed as

$$J^N(q) = \int_{t_0}^{t_1} |w_{N-1}^N(t; q) + 4w_N^N(t; q) + w_{N+1}^N(t; q) - z(t)|^2 dt$$

where $w^N(\cdot; q) = (w_0^N(\cdot; q), \dots, w_{N+1}^N(\cdot; q))^T$ is the unique solution to the $(N+2)$ -vector system (3.13), (3.14) corresponding to $q \in Q$.

With finite-dimensional state constraints, the solution of the N th estimation problem above is, *at least in principle*, routine. For inverse problems which are closely related to the one we treat here, our earlier numerical studies have shown that satisfactory results can be obtained using any one of a number of standard computational techniques for least squares minimization (for example, Newton's method, conjugate gradient, steepest descent, Levenberg-Marquardt, etc.; see [2]).

Our fundamental theoretical result is that each of the approximating identification problems and the original problem have solutions. Moreover, we show that the solutions to the approximating problems, in some sense, approximate solutions to the original problem. We require the following lemma.

LEMMA 3.1. Suppose $\{q^N\} \subset Q$ with $q^N \rightarrow q^0$ as $N \rightarrow \infty$. Let $\hat{u}^N(\cdot; q^N)$ denote the unique solution to the initial value problem (3.11), (3.12) corresponding to q^N and let $\hat{u}(\cdot; q^0)$ denote the unique solution to the initial value problem (3.7), (3.8) corresponding to q^0 . If $u(\cdot; q^N) \in H^2((0, T); V)$ then

$$(3.16) \quad \int_0^T \left| \hat{u}_u^N(t; q^N) - \hat{u}_u(t; q^0) \right|_H^2 dt \rightarrow 0$$

as $N \rightarrow \infty$.

Proof. For each $N = 1, 2, \dots$, let P^N denote the orthogonal projection of H onto V^N defined with respect to the standard inner product on H , $\langle \cdot, \cdot \rangle_H$. Using the approximation theoretic properties of interpolatory splines, we can argue with little difficulty that (see [3])

$$(3.17) \quad \lim_{N \rightarrow \infty} |(P^N - I)(\zeta, \eta, \phi)|_H = 0$$

for each $(\zeta, \eta, \phi) \in H$ and that

$$(3.18) \quad \lim_{N \rightarrow \infty} |(P^N - I)\hat{\phi}|_V = 0$$

for each $\hat{\phi} \in V$.

For $q = (m_T, E, c_D, m_H, c_H, k_H) \in Q$ it is immediately clear that $M, c(\cdot, \cdot)$, and $k(\cdot, \cdot)$, the operator and forms defined in (3.4)–(3.6), respectively, depend upon q . For $q^0 = (m_T^0, E^0, c_D^0, m_H^0, c_H^0, k_H^0) \in Q$ and $q^N = (m_T^N, E^N, c_D^N, m_H^N, c_H^N, k_H^N) \in Q$ we

adopt the shorthand notation $M^0 = M(q^0)$, $c^0(\cdot, \cdot) = c(q^0)(\cdot, \cdot)$, $k^0(\cdot, \cdot) = k(q^0)(\cdot, \cdot)$, $M^N = M(q^N)$, $c^N(\cdot, \cdot) = c(q^N)(\cdot, \cdot)$, and $k^N(\cdot, \cdot) = k(q^N)(\cdot, \cdot)$. Similarly, we denote $\hat{u}(\cdot; q^0)$ and $\hat{u}^N(\cdot; q^N)$ by \hat{u}^0 and \hat{u}^N , respectively.

From (3.17), the assumption that $\hat{u}^0 \in H^2((0, T); V)$ and the inequality

$$\int_0^T |\hat{u}_t^N(t) - \hat{u}_t^0(t)|_H^2 dt \leq 2 \int_0^T |\hat{u}_t^N(t) - P^N \hat{u}_t^0(t)|_H^2 dt + 2 \int_0^T |(I - P^N) \hat{u}_t^0(t)|_H^2 dt$$

it is clear we need only to consider the first term on the right-hand side of the above estimate.

Letting $\hat{v}^N(t) = \hat{u}^N(t) - P^N \hat{u}^0(t)$ for $t \geq 0$, and using (3.7), (3.8), (3.11), (3.12), and $V^N \subset V$, we find that

$$\begin{aligned} & \langle M^N \hat{v}_t^N, \hat{\phi}^N \rangle_H + c^N(\hat{v}_t^N, \hat{\phi}^N) + k^N(\hat{v}_t^N, \hat{\phi}^N) \\ (3.19) \quad &= \langle M^N (I - P^N) \hat{u}_t^0, \hat{\phi}^N \rangle_H + \langle (M^0 - M^N) \hat{u}_t^0, \hat{\phi}^N \rangle_H + c^N((I - P^N) \hat{u}_t^0, \hat{\phi}^N) \\ &+ c^0(\hat{u}_t^0, \hat{\phi}^N) - c^N(\hat{u}_t^0, \hat{\phi}^N) + k^N((I - P^N) \hat{u}_t^0, \hat{\phi}^N) \\ &+ k^0(\hat{u}_t^0, \hat{\phi}^N) - k^N(\hat{u}_t^0, \hat{\phi}^N), \quad t > 0, \quad \hat{\phi}^N \in V^N, \end{aligned}$$

$$(3.20) \quad \hat{v}^N(0) = 0, \quad \hat{v}_t^N(0) = 0.$$

Choosing $\hat{\phi}^N = \hat{v}_t^N(t) \in V^N$, from (3.19) we obtain

$$\begin{aligned} & \langle M^N \hat{v}_t^N, \hat{v}_t^N \rangle_H + c^N(\hat{v}_t^N, \hat{v}_t^N) = \langle M^N (I - P^N) \hat{u}_t^0, \hat{v}_t^N \rangle_H + \langle (M^0 - M^N) \hat{u}_t^0, \hat{v}_t^N \rangle_H \\ &+ \frac{d}{dt} c^N((I - P^N) \hat{u}_t^0, \hat{v}_t^N) - c^N((I - P^N) \hat{u}_t^0, \hat{v}_t^N) \\ &+ \frac{d}{dt} \{c^0(\hat{u}_t^0, \hat{v}_t^N) - c^N(\hat{u}_t^0, \hat{v}_t^N)\} \\ &- \{c^0(\hat{u}_t^0, \hat{v}_t^N) - c^N(\hat{u}_t^0, \hat{v}_t^N)\} \\ &+ \frac{d}{dt} k^N((I - P^N) \hat{u}_t^0, \hat{v}_t^N) - k^N((I - P^N) \hat{u}_t^0, \hat{v}_t^N) \\ &+ \frac{d}{dt} \{k^0(\hat{u}_t^0, \hat{v}_t^N) - k^N(\hat{u}_t^0, \hat{v}_t^N)\} \\ &- \{k^0(\hat{u}_t^0, \hat{v}_t^N) - k^N(\hat{u}_t^0, \hat{v}_t^N)\} \\ &- \frac{d}{dt} k^N(\hat{v}_t^N, \hat{v}_t^N) + k^N(\hat{v}_t^N, \hat{v}_t^N), \quad t > 0. \end{aligned}$$

Integrating the above expression from 0 to t and recalling (3.20), we find

$$\begin{aligned} & \int_0^t \langle M^N \hat{v}_{ss}^N, \hat{v}_{ss}^N \rangle_H ds + \frac{1}{2} c^N(\hat{v}_t^N, \hat{v}_t^N) \\ (3.21) \quad &= \int_0^t \{ \langle M^N (I - P^N) \hat{u}_{ss}^0, \hat{v}_{ss}^N \rangle_H - \langle (M^0 - M^N) \hat{u}_{ss}^0, \hat{v}_{ss}^N \rangle_H \\ &- c^N((I - P^N) \hat{u}_{ss}^0, \hat{v}_s^N) - (c^0(\hat{u}_{ss}^0, \hat{v}_s^N) - c^N(\hat{u}_{ss}^0, \hat{v}_s^N)) \\ &- k^N((I - P^N) \hat{u}_{ss}^0, \hat{v}_s^N) - (k^0(\hat{u}_{ss}^0, \hat{v}_s^N) - k^N(\hat{u}_{ss}^0, \hat{v}_s^N)) + k^N(\hat{v}_s^N, \hat{v}_s^N) \} ds \\ &+ c^N((I - P^N) \hat{u}_t^0, \hat{v}_t^N) + (c^0(\hat{u}_t^0, \hat{v}_t^N) - c^N(\hat{u}_t^0, \hat{v}_t^N)) \\ &+ k^N((I - P^N) \hat{u}_t^0, \hat{v}_t^N) + (k^0(\hat{u}_t^0, \hat{v}_t^N) - k^N(\hat{u}_t^0, \hat{v}_t^N)) - k^N(\hat{v}_t^N, \hat{v}_t^N). \end{aligned}$$

We recall that Q has been assumed to be a closed and bounded subset of R_+^6 and observe therefore that the forms $c^0(\cdot, \cdot)$, $c^N(\cdot, \cdot)$, $k^0(\cdot, \cdot)$ and $k^N(\cdot, \cdot)$ are uniformly bounded. These two facts, together with the repeated application of the inequality

$$\langle a, b \rangle \leq |a| |b| \leq \alpha |a|^2 + \frac{1}{4\alpha} |b|^2, \quad \alpha > 0$$

in (3.21), yield the estimate

$$\begin{aligned} & \int_0^t |\hat{v}_{ss}^N|_H^2 ds + |\hat{v}_t^N|_V^2 \\ & \leq \gamma_0 \left\{ \int_0^t \left(\frac{1}{4\alpha} |(I - P^N) \hat{u}_{ss}^0|_H^2 + \alpha |\hat{v}_{ss}^N|_H^2 \right. \right. \\ & \quad + \frac{1}{4\alpha} (|m_H^N - m_H^0|^2 + |m_T^N - m_T^0|^2) |\hat{u}_{ss}^0|_H^2 + \alpha |\hat{v}_{ss}^N|_H^2 + |(I - P^N) \hat{u}_{ss}^0|_V^2 \\ & \quad + |\hat{v}_s^N|_V^2 + (|c_H^N - c_H^0|^2 + |c_D^N - c_D^0|^2) |\hat{u}_{ss}^0|_V^2 + |\hat{v}_s^N|_V^2 + |(I - P^N) \hat{u}_s^0|_V^2 \\ & \quad + |\hat{v}_s^N|_V^2 + (|k_H^N - k_H^0|^2 + |E^N - E^0|^2) |\hat{u}_s^0|_V^2 + |\hat{v}_s^N|_V^2 + |\hat{v}_s^N|_V^2 \Big) ds \\ & \quad + \frac{1}{4\alpha} |(I - P^N) \hat{u}_t^0|_V^2 + \alpha |\hat{v}_t^N|_V^2 \\ & \quad + \frac{1}{4\alpha} (|c_H^N - c_H^0|^2 + |c_D^N - c_D^0|^2) |\hat{u}_t^0|_V^2 + \alpha |\hat{v}_t^N|_V^2 + \frac{1}{4\alpha} |(I - P^N) \hat{u}_t^0|_V^2 \\ & \quad + \alpha |\hat{v}_t^N|_V^2 + \frac{1}{4\alpha} (|k_H^N - k_H^0|^2 + |E^N - E^0|^2) |\hat{u}_t^0|_V^2 + \alpha |\hat{v}_t^N|_V^2 \\ & \quad \left. \left. + \frac{1}{4\alpha} |\hat{v}^N|_V^2 + \alpha |\hat{v}_t^N|_V^2 \right\} \end{aligned}$$

where γ_0 is a positive constant. Choosing $\alpha > 0$ sufficiently small, we find

$$(3.22) \quad \int_0^t |\hat{v}_{ss}^N(s)|_H^2 ds + |\hat{v}_t^N(t)|_V^2 \leq \sigma_0(t) + \int_0^t \sigma_1(s) ds + \gamma_1 \int_0^t |\hat{v}_s^N(s)|_V^2 ds$$

where

$$\begin{aligned} \sigma_0(t) &= \gamma_2 \{ |(I - P^N) \hat{u}^0(t)|_V^2 + |(I - P^N) \hat{u}_t^0(t)|_V^2 \\ & \quad + |q^N - q^0|^2 (|\hat{u}^0(t)|_V^2 + |\hat{u}_t^0(t)|_V^2) + |\hat{v}^N(t)|_V^2 \}, \\ \sigma_1(t) &= \gamma_3 \{ |(I - P^N) \hat{u}_t^0(t)|_V^2 + |(I - P^N) \hat{u}_{tt}^0(t)|_V^2 + |q^N - q^0|^2 (|\hat{u}_t^0(t)|_V^2 + |\hat{u}_{tt}^0(t)|_V^2) \} \end{aligned}$$

and γ_i , $i = 1, 2, 3$, are positive constants which do not depend on N .

If we choose $\hat{\phi}^N = \hat{v}_t^N(t) \in V^N$ in (3.19), arguments similar to those used above (see [2], [3]) yield

$$(3.23) \quad \lim_{N \rightarrow \infty} |\hat{v}^N(t)|_V^2 = 0$$

for each $t \in [0, T]$. Using $\hat{u}^0 \in H^2((0, T); V)$, (3.18), and an application of the Gronwall inequality to (3.22) we obtain the desired result.

We note that we also obtain

$$(3.24) \quad \lim_{N \rightarrow \infty} |\hat{v}_t^N(t)|_V^2 = 0$$

for each $t \in [0, T]$. From (3.23) and (3.24) we find $|\hat{u}^N(t; q^N) - \hat{u}(t; q^0)|_V \rightarrow 0$ and $|\hat{u}_t^N(t; q^N) - \hat{u}_t(t; q^0)| \rightarrow 0$ as $N \rightarrow \infty$ for each $t \in [0, T]$.

We remark that it is the L_2 convergence (more precisely, H convergence) in (3.16) which necessitates, at least in theory, that we be provided with distributed time observations (i.e., observations which are continuous in time). It is clear from (3.23) and (3.24) that for fits based upon displacement, velocity, or slope, time-sampled measurements are sufficient. Of course when the approximating optimization problems are solved, the integral least squares performance indices (3.15) are discretized. Consequently, in practice, only discrete measurements of linear acceleration at the tip are required.

THEOREM 3.1. *Each of the approximating identification problems has a solution \bar{q}^N . The sequence $\{\bar{q}^N\} \subset Q$ admits a convergent subsequence $\{\bar{q}^{N_j}\}$ with $\bar{q}^{N_j} \rightarrow \bar{q} \in Q$ as $j \rightarrow \infty$. If for each $q \in Q$, $\hat{u}(\cdot; q)$, the unique solution to the initial value problem (3.7), (3.8) corresponding to q is an element in $H^2((0, T); V)$, then \bar{q} is a solution to the original identification problem. In addition, the limit point of any convergent subsequence of $\{\bar{q}^N\}$ is a solution to the original identification problem as well.*

Proof. Standard continuous dependence results for linear ordinary differential equations, the fact that Q has been assumed to be a closed and bounded subset of R^6 , and the form of J^N are sufficient to conclude that a solution $\bar{q}^N \in Q$ to the N th approximating identification problem exists. Once again since Q is a closed and bounded (and therefore compact) subset of R^6 , the sequence $\{\bar{q}^N\} \subset Q$ admits a convergent subsequence. If $\{\bar{q}^{N_j}\} \subset \{\bar{q}^N\}$ with $\bar{q}^{N_j} \rightarrow \bar{q} \in Q$ as $j \rightarrow \infty$ and q is any point in Q , then two applications of Lemma 3.1 (the second one with the constant sequence $\{q\}$) yield

$$J(\bar{q}) = \lim_{j \rightarrow \infty} J^{N_j}(\bar{q}^{N_j}) \leq \lim_{j \rightarrow \infty} J^{N_j}(q) = J(q)$$

and the theorem is proved.

Although Theorem 3.1 guarantees only subsequential convergence, in all test and simulation examples we have considered, we in fact observe the convergence of the sequence $\{\bar{q}^N\}$ itself to the optimal parameters \bar{q} . Also, it is not difficult to verify that with only minor modification (see [2]) the approximation scheme reported on here (together with the convergence theory outlined in the lemma and theorem above) could be applied to inverse problems involving the estimation of spatially varying parameters (such as linear mass density ρ , flexural stiffness EI , or damping coefficient $c_D I$) which appear in (2.1)–(2.4). We note, of course, that when either EI or $c_D I$ are spatially varying, the Euler-Bernoulli equation and corresponding boundary conditions are of a slightly different form than those given in (2.1)–(2.4) (see [3]).

4. Numerical results. We used our scheme to attempt to solve the inverse problem which was posed above with data obtained from an experiment on the RPL structure. We report on our findings and observations here.

All computer codes were written in Fortran and run on the IBM 3081 at the University of Southern California. The approximating finite-dimensional least-squares minimization problems were solved using the IMSL implementation of the Levenberg-Marquardt algorithm (routine ZXSSQ), an iterative Newton-method steepest-descent hybrid (see [2]). The second-order $(N+2)$ -vector systems (3.13), (3.14) were solved (integrated) in each iteration (for the evaluation of J^N and its gradient) using Gear's method for stiff systems (IMSL routine DGEAR). The integral least squares performance index was approximated by a discrete sum over a uniform mesh on $[t_0, t_1]$. The integral inner products in the definitions of the matrices M^N , C^N , and K^N were computed using a composite two-point Gauss-Legendre quadrature rule.

The second-time derivative of w^N , or generalized acceleration, (d^2w^N/dt^2) , was computed using a second-order centered difference on the generalized displacement,

$$(4.1) \quad \frac{d^2w^N}{dt^2}(t) \approx \frac{w^N(t+\Delta) - 2w^N(t) + w^N(t-\Delta)}{\Delta^2}.$$

We found this to be a somewhat more stable method for computing acceleration (an unbounded measurement) than was a first-order centered difference on the generalized velocity,

$$(4.2) \quad \frac{dw^N}{dt}(t) \approx \frac{(dw^N/dt)(t+(\Delta/2)) - (dw^N/dt)(t-(\Delta/2))}{\Delta}.$$

Using either of the time-differencing formulas (4.1) or (4.2) proved to be significantly more stable than solving the differential equation (3.13) directly to compute $(d^2w^N/dt^2)(t)$ via an inversion of M^N . As to why this was so, we can only offer the conjecture that the time differencing provided, at least to a certain extent, some filtration of the signal.

Before turning our attention to the experimental data, we tested our scheme with simulated data. "True" values for the unknown parameters c_D (actually $c_D I$), m_H , c_H , and k_H were chosen and a quintic spline-based semidiscrete Galerkin scheme applied to the initial value problem (3.7), (3.8) was used to generate data.

We set $\rho = .03$, $m_T = .15$, $EI = 80.0$, $l = 4.0$, and

$$f(t) = \begin{cases} 1.0, & 0 \leq t \leq 0.05, \\ 0.0, & 0.05 < t \leq 5.0; \end{cases}$$

the fit was carried out based upon observations of linear acceleration at the tip at times $t_i = 0.1i$, $i = 2, 3, \dots, 50$. We note that this is equivalent to taking $t_0 = 0.1$, $t_1 = 5.0$ and using a standard rectangle rule with uniform mesh spacing 0.1 to discretize the integral appearing in the definition of the least-squares performance index J^N . The initial estimates $c_D I = 0.0035$, $m_H = 0.035$, and $k_H = 0.4$ were used to start the interactive optimization procedure. In (4.1), Δ was taken to be 0.1. Our results are summarized in Table 4.1.

TABLE 4.1

N	$\bar{c}_D^N I$	\bar{m}_H^N	\bar{c}_H^N	\bar{k}_H^N	$J^N(\bar{q}^N)$
2	0.037537	0.039471	0.003428	0.298626	2.57×10^{-1}
3	0.066997	0.039485	0.003907	0.298875	4.37×10^{-2}
4	0.005063	0.39777	0.003997	0.299455	5.06×10^{-3}
5	0.005667	0.039899	0.003971	0.299787	7.66×10^{-4}
6	0.005049	0.040035	0.004006	0.300087	4.63×10^{-5}
True value	0.005000	0.040000	0.004000	0.300000	
Initial estimate	0.003500	0.035000	0.003500	0.400000	

The experiment which we describe below was carried out for us on the RPL structure by Dr. Michel A. Floyd, formerly of the Control and Flight Dynamics Division of the Charles Stark Draper Laboratory and the Department of Aeronautics and Astronautics, MIT.

The air bearing table was clamped so that the central hub could not rotate. The thruster lines for one of the active appendages was set to 300 psi and the thruster was fired for 0.05 seconds (50 milliseconds). With the appendage initially at rest, the firing

of the thruster was assumed to have begun at time $t = 0$. Linear acceleration at the tip was observed over the time interval 0 to 5 seconds. With a sampling period of 0.005 seconds (5 milliseconds) a total of 1000 measurements were recorded. The data is plotted in Fig. 4.1. The scale factor for the accelerometer is 5 volts/g ($g = 32 \text{ ft/sec}^2$).

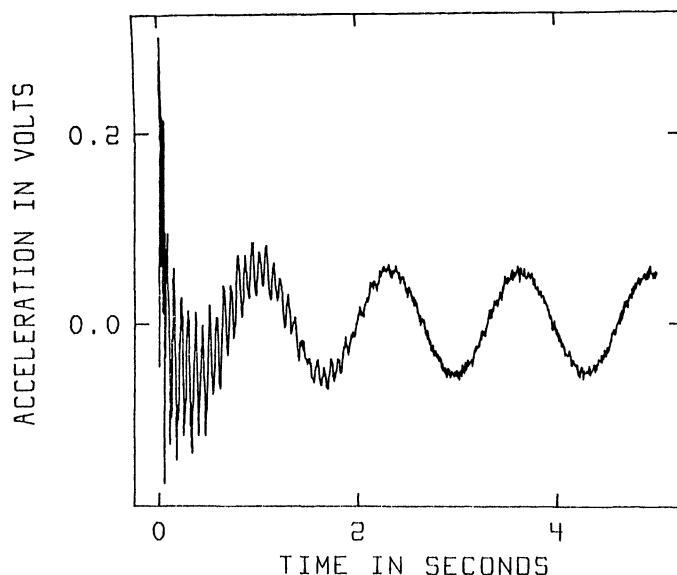


FIG. 4.1

The noticeably higher frequency ($\approx 14 \text{ Hz}$) component of the data is a torsional mode of the arm excited by the motion of the thruster valve mechanisms and inertial and elastic forces applied to the tip of the arm by the nitrogen supply hose. The opening or closing of the solenoidal valve in the thruster generates an inertial force which acts as a torque on the tip of the arm. Consequently, torsional modes are excited. Also, in addition to modifying transverse bending characteristics, since the hose is attached to the top of the arm, its horizontal motion will tend to generate torques which have a "twisting" effect. Although the accelerometer is mounted at the center of the arm (and therefore on a nodal line of the longitudinal torsional modes, if we assume vertical symmetry), as the arm twists, the accelerometer picks up a component of the earth's gravitational force. Since the first torsional mode has a much higher frequency than either of the first two flexible modes (0.75 Hz and 7.5 Hz, as identified from an FFT of the data) and since it is rapidly damped, we neglected its contribution to the accelerometer signal, treating it as white noise, and left it unmodeled. A detailed discussion of the causes of the excitation of the torsional modes and its effect on the transverse bending characteristics of the active appendages can be found in [6].

The physical characteristics of the structure are as follows. The arm is made of aluminum and is 4 feet in length, 6 inches in width and 0.125 inches in height. From this we obtain $l = 4.0 \text{ ft}$, $\rho = 0.027 \text{ slug/ft}$ and $I = 4.71 \times 10^{-8} (\text{ft})^4$. The theoretically predicted value for E is $15.84 \times 10^8 \text{ lb/(ft)}^2$. The mass of the thruster assembly was determined to be $m_T = 0.149 \text{ slug}$. From the calibration table in [6], we find that a hose pressure of 300 psi is equivalent to a force of 0.297 lb. Therefore we set

$$f(t) = \begin{cases} 0.297 \text{ lb}, & 0 \leq t \leq 0.05, \\ 0.0, & 0.05 < t \leq 5.0. \end{cases}$$

To obtain a basis for comparison, we neglected the hose effects and structural damping (i.e., we chose $c_D = m_H = c_H = k_H = 0$) and used the standard Euler-Bernoulli model with the parameters ρ , E , I , and m_T and input f as specified above to generate the plot of linear acceleration at the tip given, Fig. 4.2.

The plot was obtained by integrating the initial value problem (3.13), (3.14) with $N=4$ and then using (4.1) to compute the acceleration at the free end. The residuals $((\partial^2 u / \partial t^2)(t, l) - (\partial^2 u^N / \partial t^2)(t, l))$ over the time interval $[0, 5]$ are plotted in Fig. 4.3. The sum of the squares of the residuals (at intervals of 0.1 seconds) was found to be 3.03.

Using the data on the interval 3.0 to 5.0 (where the contribution from the torsional modes has been significantly damped) with a sampling period of 0.1 seconds, we used our scheme with $N=4$ to obtain optimal estimates for the coefficient of viscosity c_D

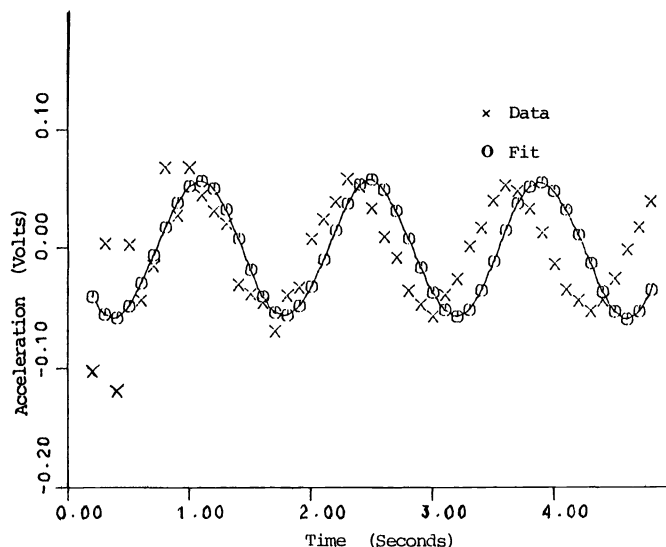


FIG. 4.2

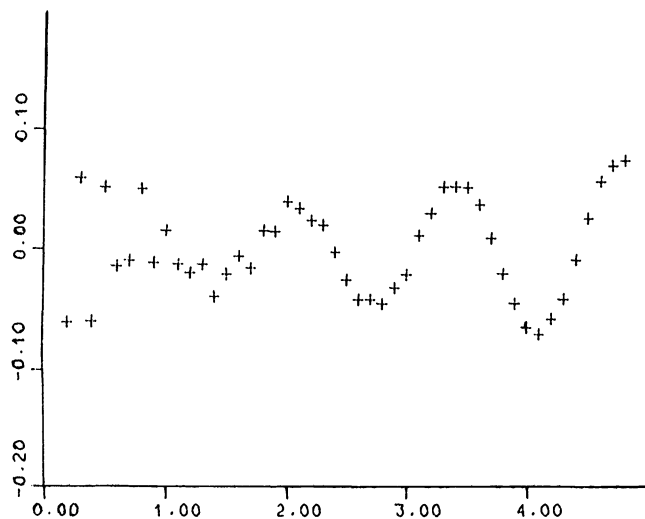


FIG. 4.3

and the hose parameters m_H , c_H , and k_H . In the set of runs we are about to describe the values of E and m_T were held fixed at their theoretically predicted values. A rough calculation, based upon "matching" the first two observed natural frequencies of the data with the first two modal frequencies of the model, was used to obtain a crude initial estimate for the ratio k_H/m_H . Then, using our scheme to minimize over the parameters m_H and k_H only, we obtained the optimal values shown in Table 4.2. Integrating the system (3.13), (3.14) over the time interval $[0, 5]$ with m_H and k_H set to the values in the table and $c_D = c_H = 0$, we found the sum of the squares of the residuals (at intervals of 0.1 seconds) to be 0.73.

Next, holding m_H and k_H fixed at the values shown in Table 4.2, we carried out a search on c_H (the initial estimate for c_H was taken to be zero and c_D was held fixed at zero). Then we used the resulting values of m_H , c_H , and k_H as initial estimates and performed a fit over all three parameters. The result is shown in Table 4.3. The sum of the squares of the residuals was found to be 0.728.

Continuing to use the same procedure to generate "start-up" values, we eventually used our scheme to search over all four parameters c_D , m_H , c_H , and k_H simultaneously obtaining the values given in Table 4.4 and the fit plotted in Fig. 4.4. The residuals are plotted in Fig. 4.5. The sum of their squares was computed to be 0.70.

In designing a controller for the RPL experiment, Floyd [6] used model adjustment to tune a simple, undamped, cantilevered beam with tip mass model for the active arms (i.e., the arms with the hoses) of the structure. He used the following procedure: The air-bearing table was locked in a stationary position. With the hose depressurized, an impulsive force was applied to the beam and linear acceleration at the tip was measured and recorded. Based upon the physical assumption that with the hose depressurized, the presence of the hose serves only to add mass to the tip of the arm, the parameter m_T was adjusted so that the first mode or frequency of the model agreed with the first observed cantilever mode (obtained via an FFT) of the data. Then, with the hose pressurized, the same experimental procedure was carried out. This time, however, the modulus of elasticity E of the beam was adjusted to compensate for the variation in stiffness which results from the presence of the hose. The adjusted values of the tip mass \bar{m}_T , and the modulus of elasticity \bar{E} , obtained by Floyd are given in

TABLE 4.2

m_H (slug)	k_H (lb/ft)
0.039269	0.339935

TABLE 4.3

m_H (slug)	c_H (lb · sec/ft)	k_H (lb/ft)
0.043431	0.004056	0.351385

TABLE 4.4

C_D (lb · sec/(ft) ²)	m_H (slug)	c_H (lb · sec/ft)	k_H (lb/ft)
127.40	0.0801	0.007804	0.412977

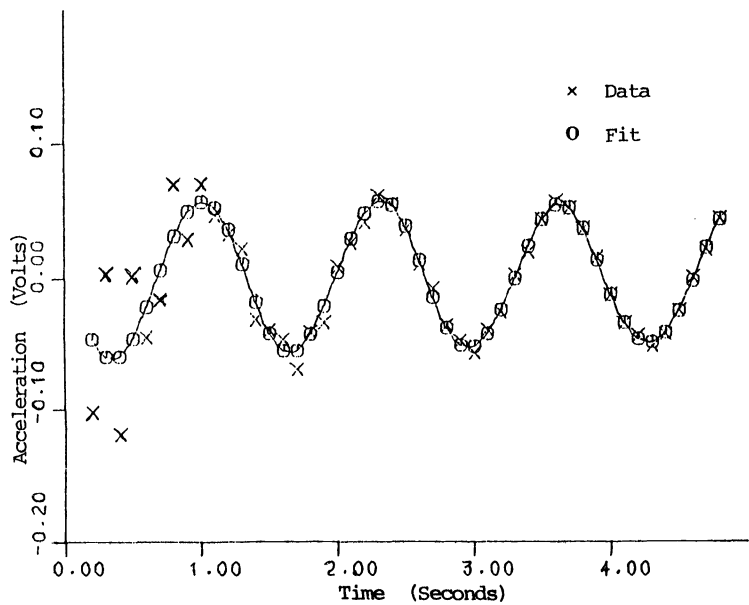


FIG. 4.4

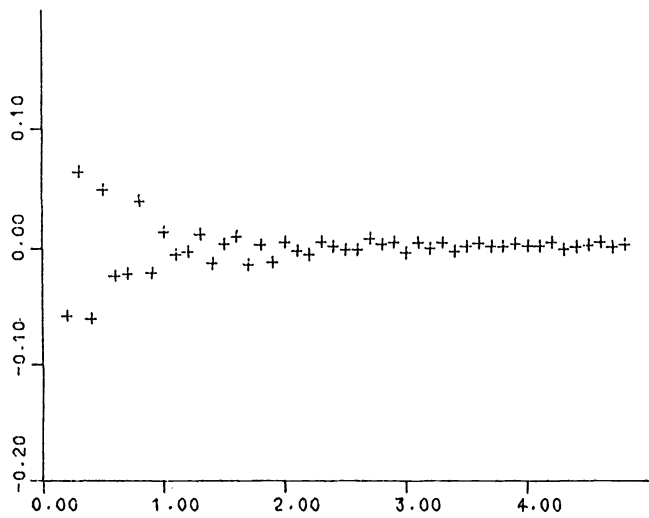


FIG. 4.5

TABLE 4.5

\bar{m}_T (slug)	\bar{E} (lb/(ft) ²)
0.254	17.31×10^8

Table 4.5. We integrated the system (3.13), (3.14) using the adjusted values of m_T and E given in the table (and $c_D = m_H = c_H = k_H = 0$) and obtained the plot shown in Fig. 4.6. The corresponding residuals are plotted in Fig. 4.7. The sum of the squares of the residuals was computed to be 5.1.

Starting with the same basic model, we used our scheme to determine the values of m_T and E which minimize the sum of the squares of the residuals over the time interval $[3.0, 5.0]$ with a sampling period of 0.1 seconds. Taking the theroretically predicted values of m_T and E ($m_T = 0.149$ slug, $E = 15.84 \times 10^8$ lb/(ft)²) as start-up values for the optimization routine yielded the results given in Table 4.6. The corresponding fit and residuals are plotted in Figs. 4.8 and 4.9, respectively. The sum of the squares of the residuals (over the interval $[0, 5]$) was computed to be 0.73.

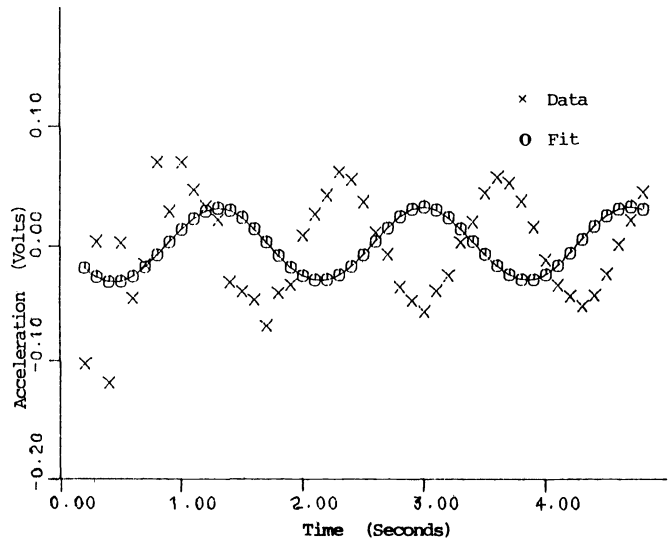


FIG. 4.6

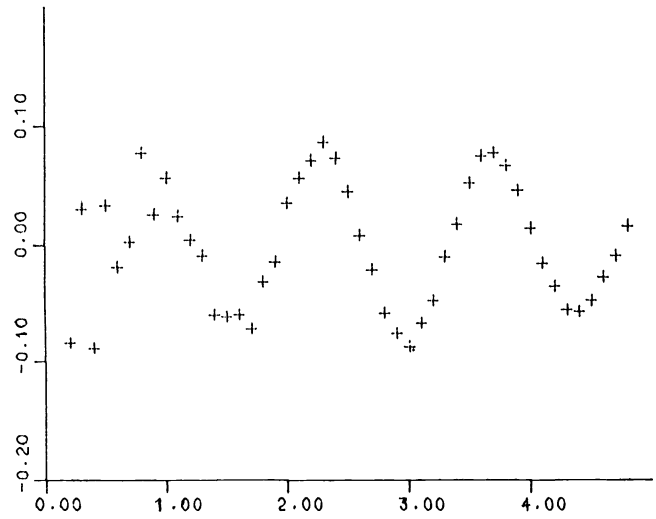


FIG. 4.7

TABLE 4.6

m_T (slug)	E (lb/(ft) ²)
0.185	21.95×10^8

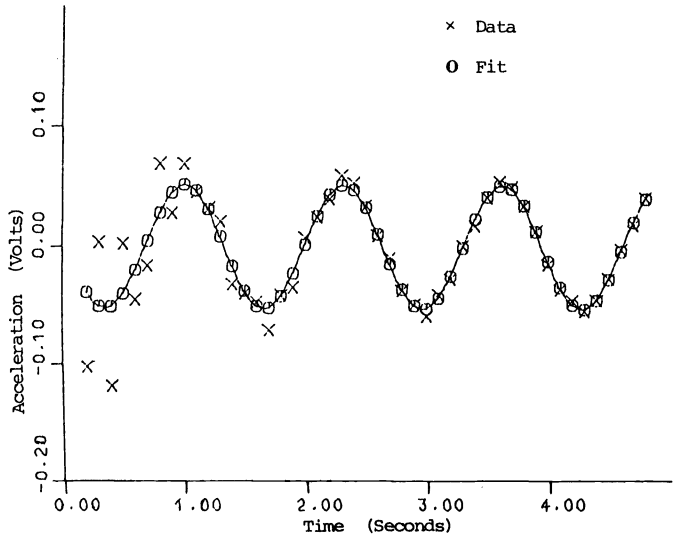


FIG. 4.8

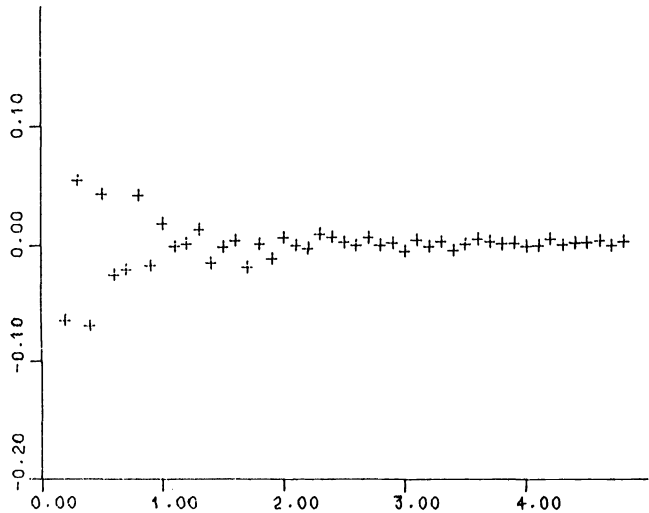


FIG. 4.9

In summary, we have seen that analysis of the RPL experimental data can be carried out in several ways with a number of different models. Our techniques can be used to provide reasonable fits of the data to models with or without hose and/or beam damping. Even if one attempts to leave the physics of the hose-beam dynamic interaction unmodeled and perform “model adjustment” (by adjusting the values of the tip mass m_T and beam modulus of elasticity E), our estimation techniques provide

a much better fit than that obtained using "modal matching" methods common in engineering practice.

One of the primary objectives of our effort here was to demonstrate the efficacy of our scheme and in particular, to assess its effectiveness when provided with actual experimental data. While we are pleased with the results obtained for the RPL data, we are careful to point out that to provide a fair and complete evaluation of the usefulness of our models for the RPL experimental structure, a more complete and in-depth study involving extensive experimental work and statistical analysis would necessarily be required.

Acknowledgment. The authors thank Dr. Michel A. Floyd of Integrated Systems, Inc. in Palo Alto, California for his willingness to discuss the technical details of the RPL structure and for providing us with the experimental data upon which this research was based.

REFERENCES

- [1] H. T. BANKS AND J. M. CROWLEY, *Parameter identification in continuum models*, J. Astronaut. Sci., 33 (1985), pp. 85-94.
- [2] H. T. BANKS, J. M. CROWLEY, AND I. G. ROSEN, *Methods for the identification of material parameters in distributed models for flexible structures*, ICASE Report 84-66, Institute for Computer Applications in Science and Engineering, NASA-Langley Research Center, Hampton, VA, 1984; Mat. Apl. Comput., 5 (1986), pp. 139-186.
- [3] H. T. BANKS AND I. G. ROSEN, *A Galerkin method for the estimation of parameters in hybrid systems governing the vibration of flexible beams with tip bodies*, ICASE Report No. 85-7, Institute for Computer Applications in Science and Engineering, NASA-Langley Research Center, Hampton, VA, 1985.
- [4] ———, *Computational methods for the identification of spatially varying stiffness and damping in beams*, ICASE Report No. 86-70, Inst. for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA, 1986; Control-Theory and Adv. Tech. 3 (1987), pp. 1-32.
- [5] R. W. CLOUGH AND J. PENZIEN, *Dynamics of Structures*, McGraw-Hill, New York, 1975.
- [6] M. A. FLOYD, *Single-step optimal control of large space structures*, Ph.D. thesis, Dept. of Aeronautics and Astronautics, Massachusetts Inst. of Technology, Cambridge, MA, 1984; Report CSDL-T-840, The Charles Stark Draper Laboratory, Massachusetts Inst. of Technology, Cambridge, MA, 1984.
- [7] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
- [8] K. KUNISCH AND E. GRAIF, *Parameter estimation for the Euler-Bernoulli beam*, Mat. Apl. Comput., 4 (1985), pp. 95-124.
- [9] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, Berlin, 1971.
- [10] P. P. POPOV, *Introduction to Mechanics of Solids*, Prentice-Hall, Englewood Cliffs, NJ, 1968.
- [11] P. M. PRENTER, *Splines and Variational Methods*, Wiley-Interscience, New York, 1975.
- [12] I. G. ROSEN, *A numerical scheme for the identification of hybrid systems describing the vibration of flexible beams with tip bodies*, J. Math. Anal. Appl., 116 (1986), pp. 262-288.
- [13] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [14] R. E. SHOWALTER, *Hilbert Space Methods for Partial Differential Equations*, Pitman, London, 1977.
- [15] R. STRUNCE, *Verification of RCS Controller Methods for Flexible Spacecraft (RPL-EXP)*, Report CSDL-P-1653, The Charles Stark Draper Laboratory, Massachusetts Inst. of Technology, Cambridge, MA, 1982.

SHAPE SENSITIVITY ANALYSIS OF BOUNDARY OPTIMAL CONTROL PROBLEMS FOR PARABOLIC SYSTEMS*

JAN SOKOŁOWSKI†

Abstract. This paper is concerned with the shape sensitivity analysis of a class of boundary control, constrained, optimal control problems for parabolic systems. The notion of Euler and Lagrange derivatives of a boundary optimal control in the direction of a vector field is introduced. The derivatives are obtained in the form of optimal solutions of auxiliary optimal control problems. The method of sensitivity analysis used in this paper is based on related results on the differential stability of metric projections in Hilbert space onto a convex, closed subset, combined with the material derivative method of shape sensitivity analysis.

Parabolic initial-boundary value problems with Dirichlet and Neumann boundary conditions are considered in this paper.

Key words. shape sensitivity analysis, boundary control, Euler derivative, Lagrange derivative, optimality system, metric projection

AMS(MOS) subject classifications. 49A22, 49A52, 49B50

1. Introduction. This paper deals with differential sensitivity analysis of control constrained boundary optimal control problems with respect to the deformations of the domain of integration. We use the method of sensitivity analysis proposed by the author [31]–[35] combined with the material derivative method [42] throughout this paper.

This method of sensitivity analysis is based on the concept of the conical differentiability of metric projection in Hilbert space onto a convex and closed set [6], [8], [21], [31]. In addition, the material derivative method is used to handle the sensitivity analysis with respect to the perturbation of the domain of integration of the state equation.

New results obtained for convex, boundary optimal control problems with constraints on control are presented in this paper. It is shown that the right-derivative of an optimal control in the direction of a vector field is given by a unique optimal solution of an auxiliary convex, control constrained optimal control problem, or, equivalently, by a unique solution of an optimality system.

In [35] we presented the method of sensitivity analysis for control constrained optimal control problems for distributed parameter systems. An example of the shape sensitivity analysis for distributed control problems for the Laplace equation is provided in [35]. However, the method proposed in [35] cannot be directly applied to the shape sensitivity analysis of boundary optimal control problems. Therefore, new results on shape sensitivity analysis of optimal controls are derived in this paper for systems described by parabolic initial-boundary value problems with Neumann and Dirichlet boundary conditions. The notion of the so-called Euler and Lagrange derivatives of an optimal boundary control in the direction of a vector field is introduced in this paper. We prove the existence and we derive the form of the Euler and Lagrange derivatives. We refer the reader to [1], [5], [11], [13], [28]–[30], [36] for the related results on the sensitivity analysis of optimization problems. The differential stability

* Received by the editors May 5, 1986; accepted for publication (in revised form) June 5, 1987. This research was sponsored by Research Project R.P. 1/1.02/2.5.

† Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Poland. The paper was completed while the author was visiting the Department of Mathematics, University of Florida, Gainesville, Florida 32611.

of optimal solutions of optimal control problems for distributed parameter systems is considered in [12], [20], [31]–[35]. The related results on shape sensitivity analysis of unilateral problems are presented in [37]–[40]. We refer the reader to [14], [15], [17]–[19] for the results on optimal control of distributed parameter systems and, in particular, on boundary control problems for systems described by partial differential equations of parabolic type.

The shape sensitivity of solutions of partial differential equations has been studied extensively in the literature, e.g. [2]–[4], [7], [9], [10], [22]–[27], [37]–[43]. Several applications of shape sensitivity analysis in the domain of structural optimization are presented in, e.g., [3], [9], [10], [23], [25].

We refer the reader to, e.g., [42] for the results on the material derivative method which we use throughout the paper.

The outline of the paper is as follows. Section 2 is devoted to the shape sensitivity analysis of a boundary optimal control problem for a parabolic equation with Neumann boundary conditions. The form of the Euler and Lagrange derivatives of an optimal control in the direction of a vector field is derived. Section 3 describes the results obtained for a boundary optimal control problem for a parabolic equation with Dirichlet boundary conditions.

Notation. Standard notation is used throughout this paper [16]. Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with smooth boundary $\Gamma = \partial\Omega$.

The Sobolev space $H^1(\Omega)$ is a Hilbert space [16] with scalar product

$$(1.1) \quad (u, v)_{H^1(\Omega)} = \int_{\Omega} \{u(x)v(x) + \nabla u(x) \cdot \nabla v(x)\} dx \quad \forall u, v \in H^1(\Omega).$$

$H_0^1(\Omega)$ is a subspace of space $H^1(\Omega)$ of the form

$$(1.2) \quad H_0^1(\Omega) = \{v \in H^1(\Omega) \mid v|_{\Gamma} = 0\};$$

here $v|_{\Gamma} \in H^{1/2}(\Gamma) \subset L^2(\Gamma)$ denotes the trace [16] of an element $v \in H^1(\Omega)$ on the boundary Γ of domain Ω . The space $H^{1/2}(\Gamma)$ is defined in [16].

$W(0, T)$ is a space of elements

$$\phi(x, t), \quad x \in \Omega, \quad t \in (0, T)$$

defined as follows:

$$(1.3) \quad W(0, T) = \left\{ \phi \in L^2(0, T; H^1(\Omega)) \mid \frac{\partial \phi}{\partial t} \in L^2(0, T; (H^1(\Omega))') \right\}.$$

We denote by $W_{\varepsilon}(0, T)$, $\varepsilon \in [0, \delta)$ a space of the form (1.3) with the domain Ω replaced in (1.3) by a domain Ω_{ε} , $\varepsilon \in [0, \delta)$.

We denote

$$(1.4) \quad H^{2,1}(Q) = \left\{ \phi \in L^2(Q) \mid \frac{\partial \phi}{\partial t}, \frac{\partial \phi}{\partial x_i}, \frac{\partial^2 \phi}{\partial x_i \partial x_j} \in L^2(Q), i, j = 1, \dots, n \right\}$$

where $Q = \Omega \times (0, T)$.

The Sobolev space $H^{1,1}(\Sigma) = H^1(0, T; L^2(\partial\Omega)) \cap L^2(0, T; H^1(\partial\Omega))$ is a Hilbert space with the scalar product

$$(1.5) \quad (u, v)_{H^{1,1}(\Sigma)} = \int_{\Sigma} \left\{ uv + \frac{\partial u}{\partial t} \frac{\partial v}{\partial t} + \nabla_{\Gamma} u \cdot \nabla_{\Gamma} v \right\} d\Sigma \quad \forall u, v \in H^{1,1}(\Sigma).$$

$C_0^1(\partial\Omega)$ is the vector space of all functions ϕ which have compact support in $\partial\Omega$ and together with their gradients $\nabla_{\Gamma} \phi$ are continuous on $\Gamma = \partial\Omega$.

$\mathcal{D}_1(\partial\Omega)$ denotes the space $C_0^1(\partial\Omega)$ with the standard definition of the convergence, i.e., $\phi_n \rightarrow \phi$ in the sense of $\mathcal{D}_1(\partial\Omega)$ provided the following conditions are satisfied: there exists a compact set $A \subset \partial\Omega$ such that $\text{supp}(\phi_n - \phi) \subset A$ for every n ; furthermore

$$\lim_{n \rightarrow \infty} \phi_n(x) = \phi(x), \quad \lim_{n \rightarrow \infty} \nabla_\Gamma \phi_n = \nabla_\Gamma \phi$$

uniformly on A .

$\mathcal{D}'_1(\partial\Omega)$ is the dual of the space $\mathcal{D}_1(\partial\Omega)$.

2. Neumann boundary conditions. This section is concerned with an optimal control problem for a system described by a parabolic equation with Neumann boundary conditions. The form of the so-called Euler and Lagrange derivatives of an optimal control in the direction of a vector field $V(\cdot, \cdot) \in C(0, \delta; C^1(R^n, R^n))$ is derived.

Let $\Omega \subset R^n$ be a given domain with smooth boundary $\Gamma = \partial\Omega$; we denote $Q = \Omega \times (0, T)$, $\Sigma = \Gamma \times (0, T)$, where $T > 0$ is a given constant. For a given control $u \in L^2(\Sigma)$, the state $y = y(u; x, t)$, $x \in \Omega$, $t \in (0, T)$ is given by a unique solution of the following parabolic equation:

$$(2.1) \quad \frac{\partial y}{\partial t} - \Delta y = 0 \quad \text{in } Q,$$

$$(2.2) \quad \frac{\partial y}{\partial n} = u \quad \text{on } \Sigma,$$

$$(2.3) \quad y(u; x, 0) = 0 \quad \text{in } \Omega;$$

here $\partial y / \partial n = \langle \nabla y, \mathbf{n} \rangle_{R^n}$, \mathbf{n} is a unit outward normal vector on Γ .

It is well known that there exists a unique weak solution to the system (2.1)–(2.3) such that

$$(2.4) \quad y \in L^2(0, T; H^1(\Omega)),$$

$$(2.5) \quad \frac{\partial y}{\partial t} \in L^2(0, T; (H^1(\Omega))'),$$

$$(2.6) \quad y|_\Sigma \in H^{1,1/2}(\Sigma)$$

for any control $u \in L^2(\Sigma)$. Furthermore, the element $y = y(u)$ satisfies the following integral identity:

$$(2.7) \quad \int_\Omega \left\{ \frac{\partial y}{\partial t} \phi + \nabla y \cdot \nabla \phi \right\} dx = \int_{\partial\Omega} u \phi d\Gamma \quad \forall \phi \in H^1(\Omega) \quad \text{for a.e. } t \in (0, T).$$

From (2.4) it follows [16] that

$$(2.8) \quad y \in C(0, T; L^2(\Omega)).$$

Therefore, the following cost functional:

$$(2.9) \quad J(u) = \frac{1}{2} \int_\Omega [y(u; x, T) - z_d(x)]^2 dx + \frac{\alpha}{2} \int_\Sigma (u)^2 d\Sigma, \quad \alpha > 0$$

is well defined for any control $u \in L^2(\Sigma)$. We denote by $K = K(\Sigma)$ the set of admissible controls defined by

$$(2.10) \quad K = \{u \in L^2(\Sigma) \mid 0 \leq u(x, t) \leq M \text{ for a.e. } (x, t) \in \Sigma\}$$

where $M > 0$ is a given constant.

Let us consider the following optimal control problem.

Problem (P). Find an element $u \in K(\Sigma)$ such that

$$(2.11) \quad J(u) \leq J(v) \quad \forall v \in K(\Sigma).$$

We will consider the differential stability of the solution of Problem (P) with respect to the perturbations of the domain $\Omega \subset R^n$. To this end we introduce control problems (P_ε) defined in the cylinder $Q_\varepsilon = \Omega_\varepsilon \times (0, T)$, $\varepsilon \in [0, \delta]$; ε is the parameter.

First we define a family of domains $\{\Omega_\varepsilon\} \subset R^n$, $\varepsilon \in [0, \delta]$.

2.1. Family of domains $\{\Omega_\varepsilon\}$. In order to derive the form of the so-called Euler and Lagrange derivatives of an optimal control u in the direction of a vector field

$$(2.12) \quad V(\cdot, \cdot) \in C(0, \delta; C^1(R^n; R^n))$$

we introduce a family of domains $\{\Omega_\varepsilon\} \subset R^n$, $\varepsilon \in [0, \delta]$ which is defined in the following way.

Let

$$(2.13) \quad T_\varepsilon: R^n \rightarrow R^n, \quad \varepsilon \in [0, \delta]$$

be a mapping of the form

$$(2.14) \quad T_\varepsilon(X) = x(\varepsilon), \quad X \in R^n, \quad \varepsilon \in [0, \delta]$$

where

$$(2.15) \quad \begin{aligned} \frac{dx}{ds}(s) &= V(s, x(s)), \quad s \in (0, \delta), \\ x(0) &= X. \end{aligned}$$

We denote

$$(2.16) \quad \Omega_\varepsilon = T_\varepsilon(\Omega) = \{x \in R^n \mid \exists X \in \Omega \text{ such that } x(0) = X, x(\varepsilon) = x\}.$$

Let us note that

$$\Omega_0 = T_0(\Omega) = \Omega$$

for any vector field $V(\cdot, \cdot)$. We will use the following notation: $DT_\varepsilon(x)$ denotes the Jacobian matrix of the mapping (2.13) evaluated at a point $x \in R^n$, $DT_\varepsilon^{-1}(x)$ is the inverse of matrix $DT_\varepsilon(x)$ and ${}^*DT_\varepsilon^{-1}(x)$ is the transpose of matrix $DT_\varepsilon^{-1}(x)$. Furthermore, we denote

$$(2.17) \quad \gamma_\varepsilon(x) = \det(DT_\varepsilon(x)), \quad x \in \bar{\Omega},$$

$$(2.18) \quad A_\varepsilon(x) = \gamma_\varepsilon(x) DT_\varepsilon^{-1}(x) \cdot {}^*DT_\varepsilon^{-1}(x), \quad x \in \bar{\Omega},$$

$$(2.19) \quad \sigma_\varepsilon(x) = \|\gamma_\varepsilon(x) {}^*DT_\varepsilon^{-1}(x) \cdot \mathbf{n}(x)\|_{R^n}, \quad x \in \partial\Omega.$$

It can be verified [41] that for ε small enough

$$(2.20) \quad \gamma_\varepsilon = 1 + \varepsilon \dot{\gamma} + o(\varepsilon) \quad \text{in } C(\bar{\Omega}),$$

$$(2.21) \quad A_\varepsilon = I + \varepsilon \dot{A} + o(\varepsilon) \quad \text{in } C(\bar{\Omega}; R^{n^2}),$$

$$(2.22) \quad \sigma_\varepsilon = 1 + \varepsilon \dot{\sigma} + o(\varepsilon) \quad \text{in } C(\partial\Omega),$$

where in (2.20)–(2.22), $\|o(\varepsilon)\|/\varepsilon \rightarrow 0$ with $\varepsilon \rightarrow 0$, in an appropriate norm, and elements $\dot{\gamma}$, \dot{A} , $\dot{\sigma}$ are given by

$$(2.23) \quad \dot{\gamma}(x) = \operatorname{div} V(0, x), \quad x \in \bar{\Omega},$$

$$(2.24) \quad \dot{A}(x) = \operatorname{div} V(0, x) I - {}^*DV(0, x) - DV(0, x), \quad x \in \bar{\Omega},$$

$$(2.25) \quad \dot{\sigma}(x) = \operatorname{div} V(0, x) - \langle DV(0, x) \cdot \mathbf{n}(x), \mathbf{n}(x) \rangle_{R^n}, \quad x \in \partial\Omega.$$

2.2. Optimal control problem (P_ε). Let a family of domains $\{\Omega_\varepsilon\} \in R^n$ defined by (2.16) be given. We define an optimal control problem (P_ε) in the cylinder $Q_\varepsilon = \Omega_\varepsilon \times (0, T)$, $\varepsilon \in [0, \delta)$. To this end, we first define the state equation, the cost functional and the set of admissible controls of the following form.

State equation. Find an element $y = y(u; x, t)$, $u \in L^2(\Sigma_\varepsilon)$, $(x, t) \in Q_\varepsilon$, such that

$$(2.26) \quad \frac{\partial y}{\partial t} - \Delta y = 0 \quad \text{in } Q_\varepsilon,$$

$$(2.27) \quad \frac{\partial y}{\partial n_\varepsilon} = u \quad \text{on } \Sigma_\varepsilon = \partial\Omega_\varepsilon \times (0, T),$$

$$(2.28) \quad y(u; x, 0) = 0 \quad \text{in } \Omega_\varepsilon.$$

Here n_ε , $\varepsilon \in [0, \delta)$, is a unit outward normal vector on $\partial\Omega_\varepsilon$.

Cost functional.

$$(2.29) \quad J_\varepsilon(u) = \frac{1}{2} \int_{\Omega_\varepsilon} [y(u; x, t) - z_d(x)]^2 dx + \frac{\alpha}{2} \int_{\Sigma_\varepsilon} (u(x, t))^2 d\Sigma, \quad \alpha > 0, \quad u \in L^2(\Sigma_\varepsilon).$$

Here $z_d(\cdot) \in H^1(R^n)$ is a given element.

Set of admissible controls.

$$(2.30) \quad K(\Sigma_\varepsilon) = \{u \in L^2(\Sigma_\varepsilon) \mid 0 \leq u(x, t) \leq M \text{ for a.e. } (x, t) \in \Sigma_\varepsilon\}.$$

Let us consider an optimal control (P_ε) defined in Q_ε , $\varepsilon \in [0, \delta)$.

Problem (P_ε). Find an element $u_\varepsilon \in K(\Sigma_\varepsilon)$ such that

$$(2.31) \quad J_\varepsilon(u_\varepsilon) \leq J_\varepsilon(u) \quad \forall u \in K(\Sigma_\varepsilon).$$

There exists a unique optimal solution $u_\varepsilon \in L^2(\Sigma_\varepsilon)$ of Problem (P_ε) which is given by a unique solution of the following optimality system.

Optimality system for Problem (P_ε). Find $(u_\varepsilon, y_\varepsilon, p_\varepsilon) \in K(\Sigma_\varepsilon) \times W_\varepsilon(0, T) \times W_\varepsilon(0, T)$ which satisfy the following system.

State equation.

$$(2.32) \quad \frac{\partial y_\varepsilon}{\partial t} - \Delta y_\varepsilon = 0 \quad \text{in } Q_\varepsilon,$$

$$(2.33) \quad \frac{\partial y_\varepsilon}{\partial n_\varepsilon} = u_\varepsilon \quad \text{on } \Sigma_\varepsilon,$$

$$(2.34) \quad y_\varepsilon(x, 0) = 0 \quad \text{in } \Omega_\varepsilon.$$

Adjoint state equation.

$$(2.35) \quad -\frac{\partial p_\varepsilon}{\partial t} - \Delta p_\varepsilon = 0 \quad \text{in } Q_\varepsilon,$$

$$(2.36) \quad \frac{\partial p_\varepsilon}{\partial n_\varepsilon} = 0 \quad \text{on } \Sigma_\varepsilon,$$

$$(2.37) \quad p_\varepsilon(x, T) = y_\varepsilon(x, T) - z_d(x) \quad \text{in } \Omega_\varepsilon.$$

Optimality conditions.

$$(2.38) \quad \begin{aligned} u_\varepsilon &\in K(\Sigma_\varepsilon), \\ \int_{\Sigma_\varepsilon} (\alpha u_\varepsilon - p_\varepsilon)(u - u_\varepsilon) d\Sigma &\geq 0 \quad \forall u \in K(\Sigma_\varepsilon). \end{aligned}$$

In order to differentiate an optimal solution u_ε with respect to the parameter ε at $\varepsilon = 0^+$, we transport the optimal solution u_ε to the fixed domain using mapping (2.13), i.e., we denote

$$(2.39) \quad u^\varepsilon = u_\varepsilon \circ T_\varepsilon \in L^2(\Sigma) \quad \forall \varepsilon \in [0, \delta).$$

By the change of variables in optimality system (2.32)–(2.38) we obtain Lemma 2.1.

LEMMA 2.1. *The element $u^\varepsilon \in L^2(\Sigma)$ is given by a unique solution of the following optimal control problem.*

Problem (P $^\varepsilon$). Find an element $u^\varepsilon \in K(\Sigma)$ which minimizes the following cost functional:

$$(2.40) \quad \begin{aligned} J^\varepsilon(u) = & \frac{1}{2} \int_{\Omega} [\eta^\varepsilon(u; x, T) - z_d^\varepsilon(x)]^2 \gamma_\varepsilon(x) dx \\ & + \frac{\alpha}{2} \int_{\Sigma} (u(x, t))^2 \sigma_\varepsilon(x) d\Sigma \quad \text{over the set } K(\Sigma). \end{aligned}$$

Here $\eta^\varepsilon(u)$ is given by a unique solution of the state equation:

$$(2.41) \quad \gamma_\varepsilon(x) \frac{\partial \eta^\varepsilon}{\partial t}(u; x, t) - \operatorname{div}(A_\varepsilon(x) \cdot \nabla \eta^\varepsilon(u; x, t)) = 0 \quad \text{in } Q,$$

$$(2.42) \quad \langle A^\varepsilon(x) \cdot \nabla \eta^\varepsilon(u; x, t), \mathbf{n}(x) \rangle_{R^n} = \sigma_\varepsilon(x) u(x, t) \quad \text{on } \Sigma,$$

$$(2.43) \quad \eta^\varepsilon(u; x, 0) = 0 \quad \text{in } \Omega.$$

We show that the optimal solution of Problem (P $^\varepsilon$) is Lipschitz continuous with respect to the parameter.

THEOREM 2.1. *For $\varepsilon > 0$, ε small enough*

$$(2.44) \quad \|u^\varepsilon - u^0\|_{L^2(\Sigma)} \leq C\varepsilon.$$

Proof. The unique optimal solution $u^\varepsilon \in K(\Sigma)$ of Problem (P $^\varepsilon$), $\varepsilon \in [0, \delta)$, is given by the unique solution of the following variational inequality:

Find $u^\varepsilon \in K(\Sigma)$ such that

$$(2.45) \quad a^\varepsilon(u^\varepsilon, u - u^\varepsilon) \geq \langle f^\varepsilon, u - u^\varepsilon \rangle \quad \forall u \in K(\Sigma)$$

where

$$(2.46) \quad \begin{aligned} a^\varepsilon(u, v) \stackrel{\text{def}}{=} & \int_{\Omega} \gamma_\varepsilon(x) \eta^\varepsilon(u; x, T) \eta^\varepsilon(v; x, T) dx \\ & + \alpha \int_{\Sigma} \sigma_\varepsilon(x) u(x, t) v(x, t) d\Sigma \quad \forall u, v \in L^2(\Sigma), \end{aligned}$$

$$(2.47) \quad \langle f^\varepsilon, v \rangle \stackrel{\text{def}}{=} \int_{\Omega} \gamma_\varepsilon(x) \eta^\varepsilon(v; x, T) z_d^\varepsilon(x) dx \quad \forall v \in L^2(\Sigma).$$

From (2.62)–(2.64), in view of (2.41)–(2.43), it follows that

$$(2.48) \quad \forall u \in L^2(\Sigma): \quad \eta^\varepsilon(u) = \eta^0(u) + \varepsilon \dot{\eta}(u) + o(\varepsilon) \quad \text{in } W(0, T)$$

where $\|o(\varepsilon)\|_{W(0,T)}/\varepsilon \rightarrow 0$ with $\varepsilon \rightarrow 0$. An element $\dot{\eta}(u) \in W(0, T)$ is given by a unique solution of the following parabolic equation:

$$(2.49) \quad \frac{\partial \dot{\eta}(u)}{\partial t} - \Delta \dot{\eta}(u) + \dot{\gamma}(x) \frac{\partial \eta^0(u)}{\partial t} - \operatorname{div}(\dot{A} \cdot \nabla \eta^0(u)) = 0 \quad \text{in } 0,$$

$$(2.50) \quad \frac{\partial \dot{\eta}}{\partial n}(u) + \langle \dot{A} \cdot \nabla \eta^0(u), \mathbf{n} \rangle_{R^n} = \dot{\sigma} u \quad \text{on } \Sigma,$$

$$(2.51) \quad \dot{\eta}(u; x, 0) = 0 \quad \text{in } \Omega.$$

In view of (2.48), it follows that the bilinear form $a^\varepsilon(\cdot, \cdot)$ is differentiable with respect to the parameter, i.e., there exists a continuous bilinear form $\dot{a}(\cdot, \cdot)$ such that

$$(2.52) \quad \lim_{\varepsilon \downarrow 0} \sup_{\substack{\|u\|_{L^2(\Sigma)} \leq 1 \\ \|v\|_{L^2(\Sigma)} \leq 1}} |(a^\varepsilon(u, v) - a^0(u, v))/\varepsilon - \dot{a}(u, v)| = 0.$$

The bilinear form $\dot{a}(\cdot, \cdot): L^2(\Sigma) \times L^2(\Sigma) \rightarrow R$ is defined by (2.56). On the other hand, since

$$(2.53) \quad z_d^\varepsilon = z_d + \varepsilon \nabla z_d \cdot V(0) + o(\varepsilon) \quad \text{in } L^2(\Omega)$$

in view of (2.20), (2.48) it follows that the element f^ε in (2.47) is differentiable with respect to the parameter ε ; i.e., there exists the element \dot{f} such that

$$(2.54) \quad \lim_{\varepsilon \downarrow 0} \sup_{\|v\|_{L^2(\Sigma)} \leq 1} |\langle f^\varepsilon - f^0, v \rangle / \varepsilon - \langle \dot{f}, v \rangle| = 0.$$

Since for $\varepsilon > 0$, ε small enough the bilinear form (2.46) is coercive:

$$(2.55) \quad a^\varepsilon(u, u) \geq \frac{\alpha}{2} \|u\|_{L^2(\Sigma)}^2 \quad \forall u \in L^2(\Sigma);$$

therefore by a standard argument [35], from (2.45), in view of (2.52) and (2.54), (2.44) follows. The bilinear form $\dot{a}(\cdot, \cdot)$ and the element \dot{f} are given by

$$(2.56) \quad \begin{aligned} \dot{a}(u, v) \stackrel{\text{def}}{=} & \int_{\Omega} \{ \dot{\gamma}(x) \eta^0(u; x, T) \eta^0(v; x, T) + \dot{\eta}(u; x, T) \eta^0(v; x, T) \\ & + \eta^0(u; x, T) \dot{\eta}(v; x, T) \} dx \end{aligned}$$

$$+ \alpha \int_{\Sigma} \dot{\sigma}(x) u(x, t) v(x, t) d\Sigma \quad \forall u, v \in L^2(\Sigma),$$

$$(2.57) \quad \begin{aligned} \langle \dot{f}, v \rangle \stackrel{\text{def}}{=} & \int_{\Omega} \{ \dot{\eta}(v; x, T) z_d(x) + \dot{\gamma}(x) \eta^0(v; x, T) z_d(x) \\ & + \eta^0(v; x, T) (\nabla z_d(x) \cdot V(0, x)) \} dx \quad \forall v \in L^2(\Sigma), \end{aligned}$$

respectively.

Using (2.44) we show that an optimal control u^ε is actually right-differentiable with respect to parameter ε , at $\varepsilon = 0$.

THEOREM 2.2. *For $\varepsilon > 0$, ε small enough*

$$(2.58) \quad u^\varepsilon = u^0 + \varepsilon \dot{u}(\Sigma) + o(\varepsilon) \quad \text{in } L^2(\Sigma)$$

where $\|o(\varepsilon)\|_{L^2(\Sigma)}/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$.

The Euler derivative $\dot{u} = \dot{u}(\Sigma)$ of an optimal control u^0 in the direction of a vector field $V(\cdot, \cdot)$ is given by a unique solution of the following optimality system.

Find $(\dot{u}, z, w) \in S(\Sigma) \times W(0, T) \times W(0, T)$ which satisfy the following system.

State equation. Find $z = z(x, t), (x, t) \in Q$.

$$\begin{aligned} & \int_{\Omega} \left\{ \frac{\partial z}{\partial t} \phi + \nabla z \cdot \nabla \phi \right\} dx \\ (2.59) \quad &= \int_{\partial\Omega} \dot{u} \phi \, d\Gamma - \int_{\partial\Omega} v \cdot \nabla_{\Gamma} u_0 \phi \, d\Gamma + \int_{\partial\Omega} v_n \left\{ \left(H u_0 - \frac{\partial y^0}{\partial t} \right) \phi - \nabla_{\Gamma} y^0 \cdot \nabla_{\Gamma} \phi \right\} d\Gamma \\ & \text{for a.e. } t \in (0, T) \quad \forall \phi \in H^2(\Omega), \quad \frac{\partial \phi}{\partial n} = 0, \\ (2.60) \quad & z(x, 0) = 0 \quad \text{in } \Omega. \end{aligned}$$

Adjoint state equation. Find $w = w(x, t), (x, t) \in Q$.

$$\begin{aligned} (2.61) \quad & \int_{\Omega} \left\{ \frac{\partial w}{\partial t} \phi + \nabla w \cdot \nabla \phi \right\} dx = \int_{\partial\Omega} v_n \left\{ \frac{\partial p^0}{\partial t} \phi - \nabla_{\Gamma} p^0 \cdot \nabla_{\Gamma} \phi \right\} d\Gamma \\ & \text{for a.e. } t \in (0, T) \quad \forall \phi \in H^2(\Omega), \quad \frac{\partial \phi}{\partial n} = 0, \\ (2.62) \quad & w(x, T) = z(x, T) \quad \text{in } \Omega. \end{aligned}$$

Optimality conditions.

$$\begin{aligned} (2.63) \quad & \dot{u} \in S(\Sigma), \\ & \int_{\Sigma} (\alpha \dot{u} - w)(u - \dot{u}) \, d\Sigma \geq 0 \quad \forall u \in S(\Sigma) \end{aligned}$$

where

$$\begin{aligned} (2.64) \quad & S(\Sigma) = \{u \in L^2(\Sigma) \mid u(x, t) \geq 0 \text{ for a.e. } (x, t) \in \Xi_1^0, \\ & u(x, t) = 0 \text{ for a.e. } (x, t) \in \Xi_1^+ \cup \Xi_2^+, \\ & u(x, t) \leq 0 \text{ for a.e. } (x, t) \in \Xi_2^0\}. \end{aligned}$$

Here we denote

$$(2.65) \quad \Xi_1 = \{(x, t) \in \Sigma \mid u^0(x, t) = 0\},$$

$$(2.66) \quad \Xi_2 = \{(x, t) \in \Sigma \mid u^0(x, t) = M\},$$

$$(2.67) \quad \Xi_i^0 = \{(x, t) \in \Xi_i \mid \alpha u^0(x, t) = p_0(x, t)\}, \quad i = 1, 2,$$

$$(2.68) \quad \Xi_i^+ = \Xi_i \setminus \Xi_i^0, \quad i = 1, 2.$$

Remark 2.1. Let us consider the following optimal problem.

Problem (P̄). Find an element $\dot{u} \in S(\Sigma)$ which minimizes the cost functional

$$\begin{aligned} I(u) = & \frac{1}{2} \int_{\Omega} [y(u; x, T) - \xi(x, T)]^2 dx \\ & + \frac{1}{2} \int_{\Sigma} v_n \left\{ \frac{\partial p^0}{\partial t}(x, t) y(u; x, t) - \nabla_{\Gamma} p^0(x, t) \cdot \nabla_{\Gamma} y(u; x, t) \right\} d\Sigma \\ & + \frac{\alpha}{2} \int_{\Sigma} [u(x, t)]^2 d\Sigma \end{aligned}$$

over the set $S(\Sigma) \subset L^2(\Sigma)$.

Here $y(u; \cdot, \cdot)$, $u \in L^2(\Sigma)$ is the solution of state equation (2.1)–(2.3), $\xi(x, t)$, $(x, t) \in Q$ denotes the solution of problem (2.59), (2.60) for $\dot{u} = 0$. Let us observe that the optimality system for Problem (P) takes the form (2.59)–(2.63).

Proof of Theorem 2.2. By Theorem 2.1 it follows that there exists an element $q \in L^2(\Sigma)$ such that for $\varepsilon > 0$, ε small enough

$$(2.69) \quad u^\varepsilon = u^0 + \varepsilon q + r(\varepsilon) \quad \text{in } L^2(\Sigma)$$

where $r(\varepsilon)/\varepsilon \rightarrow 0$ weakly in $L^2(\Sigma)$, with $\varepsilon \downarrow 0$.

We denote $y^\varepsilon = y_\varepsilon \circ T_\varepsilon$; after the change of variables in (2.32)–(2.34) it follows that the element y^ε satisfies the following state equation in the form of integral identity:

$$(2.70) \quad \int_{\Omega} \left\{ \gamma_\varepsilon(x) \frac{\partial y^\varepsilon}{\partial t}(x, t) \psi(x) + \langle A_\varepsilon(x) \cdot \nabla y^\varepsilon(x, t), \nabla \phi(x) \rangle_{R^n} \right\} dx \\ = \int_{\partial\Omega} \sigma_\varepsilon(x) u^\varepsilon(x, t) \psi(x) d\Gamma \quad \forall \psi \in H^1(\Omega) \quad \text{for a.e. } t \in (0, T),$$

$$(2.71) \quad y^\varepsilon(x, 0) = 0 \quad \text{in } \Omega.$$

Furthermore, we denote

$$(2.72) \quad p^\varepsilon = p_\varepsilon \circ T_\varepsilon \in W(0, T) \quad \forall \varepsilon \in [0, \delta),$$

and from (2.35)–(2.37), after the change of variables, we obtain

$$(2.73) \quad \int_{\Omega} \left\{ -\gamma_\varepsilon(x) \frac{\partial p^\varepsilon}{\partial t}(x, t) \psi(x) + \langle A_\varepsilon(x) \cdot \nabla p^\varepsilon(x, t), \nabla \psi(x) \rangle_{R^n} \right\} dx = 0 \\ \forall \psi \in H^1(\Omega) \quad \text{for a.e. } t \in (0, T),$$

$$(2.74) \quad p^\varepsilon(x, T) = y^\varepsilon(x, T) - z_d^\varepsilon(x) \quad \text{in } \Omega.$$

Here $z_d^\varepsilon(x) = (z_d \circ T_\varepsilon)(x)$, $x \in \Omega$.

Finally we change the variables in the optimality conditions (2.38). This leads to the following variational inequality:

$$(2.75) \quad u^\varepsilon \in K(\Sigma), \\ \int_{\Sigma} (\alpha u^\varepsilon - p^\varepsilon)(u - u^\varepsilon) \sigma^\varepsilon d\Sigma \geq 0 \quad \forall u \in K(\Sigma).$$

From (2.69), in view of (2.70), (2.71) and (2.73), (2.74), it follows that

$$(2.76) \quad y^\varepsilon = y^0 + \varepsilon z + r(\varepsilon) \quad \text{in } W(0, T),$$

$$(2.77) \quad p^\varepsilon = p^0 + \varepsilon w + r(\varepsilon) \quad \text{in } W(0, T)$$

and

$$(2.78) \quad p^\varepsilon|_{\Sigma} = p^0|_{\Sigma} + \varepsilon w|_{\Sigma} + o(\varepsilon) \quad \text{in } L^2(\Sigma)$$

where $\|o(\varepsilon)\|_{L^2(\Sigma)}/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$.

The elements z, w are given by unique solutions of the following parabolic problems:

$$(2.79) \quad \frac{\partial z}{\partial t} - \Delta z + \dot{\gamma} \frac{\partial y^0}{\partial t} - \operatorname{div}(\dot{A} \cdot \nabla y^0) = 0 \quad \text{in } Q,$$

$$(2.80) \quad \frac{\partial z}{\partial n} + \langle \dot{A} \cdot \nabla y^0, n \rangle_{R^n} = q + \sigma u^0 \quad \text{on } \Sigma,$$

$$(2.81) \quad z(x, 0) = 0 \quad \text{in } \Omega,$$

$$(2.82) \quad -\frac{\partial w}{\partial t} - \Delta w - \dot{\gamma} \frac{\partial p^0}{\partial t} - \operatorname{div} (\dot{A} \cdot \nabla p^0) = 0 \quad \text{in } Q,$$

$$(2.83) \quad \frac{\partial w}{\partial n} + \langle \dot{A} \cdot \nabla p^0, \mathbf{n} \rangle_{R^n} = 0 \quad \text{on } \Sigma,$$

$$(2.84) \quad w(x, T) = z(x, T) - \nabla z_d \cdot V(0) \quad \text{in } \Omega,$$

respectively.

It follows by Lemma 4.2 in the Appendix that, in view of (2.78), the solution u^ε of variational inequality (2.75) can be represented for $\varepsilon > 0$, ε small enough, in the form

$$(2.85) \quad u^\varepsilon = u^0 + \varepsilon \dot{u} + o(\varepsilon) \quad \text{in } L^2(\Sigma)$$

where $\|o(\varepsilon)\|_{L^2(\Sigma)}/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$. Furthermore by Lemma 4.2 an element \dot{u} in (2.85) is given by a unique solution of the following variational inequality:

$$(2.86) \quad \begin{aligned} \dot{u} &\in S(\Sigma), \\ \int_{\Sigma} (\alpha \dot{u} + \alpha \dot{\sigma} u^0 - w - p^0 \dot{\sigma})(u - \dot{u}) \, d\Sigma &\geq 0 \quad \forall u \in S(\Sigma) \end{aligned}$$

where the cone $S(\Sigma)$ is defined by (2.64). We can simplify (2.86), since by definition of the set $S(\Sigma)$ it follows that

$$(2.87) \quad \int_{\Sigma} (\alpha u^0 - p^0) \phi \, d\Sigma = 0 \quad \forall \phi \in S(\Sigma);$$

therefore

$$\int_{\Sigma} (\alpha \dot{\sigma} u^0 - \dot{\sigma} p^0)(u - \dot{u}) \, d\Sigma = 0 \quad \forall u \in S(\Sigma).$$

Hence (2.86) takes the form

$$(2.88) \quad \begin{aligned} \dot{u} &\in S(\Sigma), \\ \int_{\Sigma} (\alpha \dot{u} - w)(u - \dot{u}) \, d\Sigma &\geq 0 \quad \forall u \in S(\Sigma). \end{aligned}$$

From (2.85), in view of (2.69), it follows that we actually have the equality

$$(2.89) \quad \dot{u} = q.$$

Therefore, an element \dot{u} is uniquely determined by the following optimality system.

Find $(\dot{u}, \dot{y}, \dot{p}) \in S(\Sigma) \times W(0, T) \times W(0, T)$ such that the following system is satisfied.

State equation.

$$(2.90) \quad \frac{\partial \dot{y}}{\partial t}(x, t) - \Delta \dot{y}(x, t) + \dot{\gamma}(x) \frac{\partial y^0}{\partial t}(x, t) - \operatorname{div} (\dot{A}(x) \cdot \nabla y^0(x, t)) = 0 \quad \text{in } Q,$$

$$(2.91) \quad \frac{\partial \dot{y}}{\partial n}(x, t) + \langle \dot{A}(x) \cdot \nabla y^0(x, t), \mathbf{n}(x) \rangle_{R^n} = \dot{u}(x, t) + \dot{\sigma}(x) u^0(x, t) \quad \text{on } \Sigma,$$

$$(2.92) \quad \dot{y}(x, 0) = 0 \quad \text{in } \Omega.$$

Adjoint state equation.

$$(2.93) \quad -\frac{\partial \dot{p}}{\partial t}(x, t) - \Delta \dot{p}(x, t) - \dot{\gamma}(x) \frac{\partial p^0}{\partial t}(x, t) - \operatorname{div}(\dot{A}(x) \cdot \nabla p^0(x, t)) = 0 \quad \text{in } Q,$$

$$(2.94) \quad \frac{\partial \dot{p}}{\partial n}(x, t) + \langle \dot{A}(x) \cdot \nabla p^0(x, t), \mathbf{n}(x) \rangle_{R^n} = 0 \quad \text{on } \Sigma,$$

$$(2.95) \quad \dot{p}(x, T) = \dot{\gamma}(x, T) - \nabla z_d(x) \cdot V(0, x) \quad \text{in } \Omega.$$

Optimality conditions.

$$(2.96) \quad \dot{u} \in S(\Sigma), \quad \int_{\Sigma} (\alpha \dot{u}(x, t) - \dot{p}(x, t))(u(x, t) - \dot{u}(x, t)) \, d\Sigma \geq 0 \quad \forall u \in S(\Sigma).$$

In order to obtain (2.59)–(2.63) from (2.90)–(2.96) let us observe that if a vector field $V(\cdot, \cdot)$ is selected in such a way that the normal component satisfies

$$(2.97) \quad v_n(x) = \langle V(0, x), \mathbf{n}(x) \rangle_{R^n} = 0, \quad x \in \partial\Omega$$

Then for $\varepsilon > 0$, ε small enough $\Omega_\varepsilon = \Omega$; here we use the assumption that $\partial\Omega$ is smooth. Therefore $u_\varepsilon = u_0$ for such a vector field $V(\cdot, \cdot)$ and it follows that

$$(2.98) \quad \dot{u}(x) = \langle \nabla_\Gamma u_0(x), \mathbf{v}_\tau(x) \rangle_{R^n}, \quad x \in \partial\Omega$$

where \mathbf{v}_τ is a tangent vector on $\partial\Omega$,

$$(2.99) \quad \mathbf{v}_\tau(x) = V(0, x) - \mathbf{n} \langle V(0, x), \mathbf{n}(x) \rangle_{R^n}$$

and $\nabla_\Gamma u_0$ is a tangential component of the gradient on $\partial\Omega$. Condition (2.98) provides an additional regularity of an optimal control u_0 , since by (2.88) it follows that

$$\dot{u} \in L^2(\Sigma) \quad \text{for any vector field } V(\cdot, \cdot) \in C(0, \delta; C^1(R^n; R^n));$$

hence (2.98) implies that

$$(2.100) \quad \nabla_\Gamma u_0 \in L^2(\Sigma; R^n),$$

so we obtain the following regularity of an optimal control:

$$(2.101) \quad u_0 \in L^2(0, T; H^1(\partial\Omega)).$$

Let us assume now that

$$(2.102) \quad v_n(x) = 0, \quad \mathbf{v}_\tau(x) = 0, \quad x \in \partial\Omega;$$

then by (2.98) it follows that

$$(2.103) \quad \dot{u}(x) = 0 \quad \text{for a.e. } x \in \partial\Omega.$$

On the other hand, the element $\dot{u} \in S(\Sigma)$ is given by a unique solution of the following variational inequality:

$$(2.104) \quad \dot{u} \in S(\Sigma): \int_{\Sigma} (\dot{u} - G(V))(u - \dot{u}) \, d\Sigma \geq 0 \quad \forall u \in S(\Sigma).$$

Here we denote $G(V) = (1/\alpha)\bar{p}|_{\Sigma}$, $V = V(0) = V(0, \cdot)$, and $V = V(0)$ satisfies the conditions (2.102). By (2.103), (2.104) it follows that

$$(2.105) \quad - \int_{\Sigma} G(V)u \, d\Sigma \geq 0 \quad \forall u \in S(\Sigma).$$

The mapping $V \rightarrow G(V)$ is linear; so taking $\pm V$ in (2.105), we obtain

$$(2.106) \quad \int_{\Sigma} G(V)u \, d\Sigma = 0 \quad \forall u \in S(\Sigma)$$

for any vector field $V = V(0)$ such that conditions (2.102) are satisfied.

Equation (2.106) provides Green's formula for optimality system (2.90)–(2.96). It allows us to simplify the state equation and the adjoint state equation in the optimality system. In particular, by (2.106) it follows that for any $\phi \in H^1(\Omega)$ there exist distributions

$$g_n(\phi), h_n(\phi) \in \mathcal{D}'_1(\partial\Omega), g_\tau(\phi), h_\tau(\phi) \in \mathcal{D}'_1(\partial\Omega; R^n) = \mathcal{D}'_1(\partial\Omega)$$

such that the state equation (2.90)–(2.92) in the optimality system (2.90)–(2.96) can be replaced by the state equation of the following form:

Find $z = z(x, t)$, $(x, t) \in Q$, such that

$$(2.107) \quad \int_{\Omega} \left\{ \frac{\partial z}{\partial t} \phi + \nabla z \cdot \nabla \phi \right\} dx = \int_{\partial\Omega} \dot{u} \phi \, d\Gamma + \langle g_n(\phi), v_n \rangle_{\mathcal{D}'_1(\partial\Omega) \times \mathcal{D}_1(\partial\Omega)} + \langle g_\tau(\phi), v_\tau \rangle_{\mathcal{D}'_1(\partial\Omega) \times \mathcal{D}_1(\partial\Omega)}$$

for a.e. $t \in (0, T) \quad \forall \phi \in H^1(\Omega)$,

$$(2.108) \quad z(x, 0) = 0 \quad \text{in } \Omega.$$

Similarly the adjoint state equation (2.93)–(2.95) in the optimality system (2.90)–(2.96) can be replaced by the following adjoint state equation:

Find $w = w(x, t)$, $(x, t) \in Q$, such that

$$(2.109) \quad \int_{\Omega} \left\{ -\frac{\partial w}{\partial t} \phi + \nabla w \cdot \nabla \phi \right\} dx = \langle h_n(\phi), v_n \rangle_{\mathcal{D}'_1(\partial\Omega) \times \mathcal{D}_1(\partial\Omega)} + \langle h_\tau(\phi), v_\tau \rangle_{\mathcal{D}'_1(\partial\Omega) \times \mathcal{D}_1(\partial\Omega)}$$

for a.e. $t \in (0, T) \quad \forall \phi \in H^1(\Omega)$.

The representation of distributions $g_n(\phi)$, $g_\tau(\phi)$, $h_n(\phi)$, $h_\tau(\phi)$ is obtained by integration by parts in the weak formulations of problems (2.90), (2.91) and (2.93), (2.94), respectively, with a test function $\phi \in H^2(\Omega)$ such that $\partial\phi/\partial n = 0$ on $\partial\Omega$.

Finally, we obtain the following representation of distributions: $g_n(\phi)$, $h_n(\phi)$, $g_\tau(\phi)$, $h_\tau(\phi)$:

$$(2.110) \quad g_n(\phi) = \left(Hu_0 - \frac{\partial y^0}{\partial t} \right) \phi - \nabla_\Gamma y^0 \cdot \nabla_\Gamma \phi,$$

$$(2.111) \quad g_\tau(\phi) = -\phi \nabla_\Gamma u_0,$$

$$(2.112) \quad h_n(\phi) = \phi \frac{\partial p^0}{\partial t} - \nabla_\Gamma p^0 \cdot \nabla_\Gamma \phi,$$

$$(2.113) \quad h_\tau(\phi) = 0$$

for $\phi \in H^1(\partial\Omega)$. Here H is the mean curvature of the boundary $\partial\Omega$ of domain Ω and $\nabla_\Gamma \phi$ is the tangential gradient on $\partial\Omega$ of an element $\phi \in H^1(\partial\Omega)$.

This concludes the proof of Theorem 2.2. \square

Now we are in a position to present the main result of this section. Let \tilde{u}_ε denote an extension of an optimal control $u_\varepsilon \in L^2(0, T; H^1(\partial\Omega_\varepsilon))$ to an open neighborhood of $\partial\Omega_\varepsilon$ such that $\tilde{u}_\varepsilon|_\Sigma$ is a well-defined element of space $L^2(\Sigma)$.

We denote by $u' \in L^2(\Sigma)$ the Lagrange derivative of an optimal control $u_0 = u^0$ in the direction of a vector field $V(\cdot, \cdot)$:

$$(2.114) \quad u'(x, t) = \dot{u}(x, t) - v_\tau(x) \cdot \nabla_\Gamma u_0, \quad (x, t) \in \Sigma.$$

THEOREM 2.3. For $\varepsilon > 0$, ε small enough

$$(2.115) \quad \tilde{u}_\varepsilon|_\Sigma = u_0 + \varepsilon u' + o(\varepsilon) \quad \text{in } L^2(\Sigma)$$

where $\|o(\varepsilon)\|_{L^2(\Sigma)}/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$.

The Lagrange derivative $u' \in L^2(\Sigma)$ of an optimal control $u_0 \in L^2(\Sigma)$ in the direction of a vector field $V(\cdot, \cdot)$ is given by a unique solution of the following optimality system.

Find $(u', z, w) \in S(\Sigma) \times W(0, T) \times W(0, T)$ such that the following system is satisfied.

State equation. Find $z = z(x, t)$, $(x, t) \in Q$.

$$(2.116)$$

$$\int_\Omega \left\{ \frac{\partial z}{\partial t} \phi + \nabla z \cdot \nabla \phi \right\} dx = \int_{\partial\Omega} u' \phi \, d\Gamma + \int_{\partial\Omega} v_n \left\{ \left(H u_0 - \frac{\partial y^0}{\partial t} \right) \phi - \nabla_\Gamma y^0 \cdot \nabla_\Gamma \phi \right\} d\Gamma$$

for a.e. $t \in (0, T) \quad \forall \phi \in H^2(\Omega), \quad \frac{\partial \phi}{\partial n} = 0,$

$$(2.117) \quad z(x, 0) = 0 \quad \text{in } \Omega.$$

Adjoint state equation. Find $w = w(x, t)$, $(x, t) \in Q$.

$$(2.118) \quad \int_\Omega \left\{ -\frac{\partial w}{\partial t} \phi + \nabla w \cdot \nabla \phi \right\} dx = \int_{\partial\Omega} v_n \left\{ \frac{\partial p^0}{\partial t} \phi - \nabla_\Gamma p^0 \cdot \nabla_\Gamma \phi \right\} d\Gamma$$

for a.e. $t \in (0, T) \quad \forall \phi \in H^2(\Omega), \quad \frac{\partial \phi}{\partial n} = 0,$

$$(2.119) \quad w(x, T) = z(x, T) \quad \text{in } \Omega.$$

Optimality conditions. $u' \in S(\Sigma)$:

$$(2.120) \quad \int_\Sigma (\alpha u' - w)(u - u') \, d\Sigma \geq 0 \quad \forall u \in S(\Sigma).$$

Proof. The Lagrange derivative of an optimal control $u_0 = u^0$ is defined by (2.114), where $\dot{u} \in L^2(\Sigma)$ is the Euler derivative of an optimal control u_0 . The element $\dot{u} \in L^2(\Sigma)$ is given by a unique solution of (2.59)–(2.63). Let us note that from (2.59), (2.60), in view of (2.114), we obtain (2.116), (2.117); therefore (2.118), (2.119) follow from (2.61) and (2.62). We shall show that (2.120) follows from (2.63). Since $u_0 \in L^2(0, T; H^1(\partial\Omega))$ and

$$(2.121) \quad \nabla_\Gamma u_0(x, t) = 0 \quad \text{for a.e. } (x, t) \in \Xi_i, \quad i = 1, 2$$

where sets Ξ_i , $i = 1, 2$ are defined by (2.65), (2.66), respectively, then in view of (2.64), it follows that

$$\mathbf{v}_\tau \cdot \nabla_\Gamma u_0 \in S(\Sigma) \subset L^2(\Sigma) \quad \text{for any vector field } V(\cdot, \cdot) \in C(0, \delta; C^1(\mathbb{R}^n; \mathbb{R}^n))$$

furthermore we have

$$(2.122) \quad u' = \dot{u} - \mathbf{v}_\tau \cdot \nabla_\Gamma u_0 \in S(\Sigma),$$

$$(2.123) \quad \eta = u - \mathbf{v}_\tau \cdot \nabla_\Gamma u_0 \in S(\Sigma) \quad \forall u \in S(\Sigma).$$

Hence by (2.63), (2.120) follows, which completes the proof. \square

Let us observe that, in view of (2.116)–(2.120), it follows that an element $u' \in L^2(\Sigma)$ can be defined in the form of an optimal control for an auxiliary optimal control problem.

COROLLARY 2.1. *The Lagrange derivative $u' \in L^2(\Sigma)$ of an optimal control u_0 in the direction of a vector field $V(\cdot, \cdot)$ is given by a unique solution of the following optimal control problem.*

Problem (P'). Find an element $u' \in L^2(\Sigma)$ which minimizes the cost functional

$$(2.124) \quad J(u) = \frac{1}{2} \int_{\Omega} [z(u; x, T)]^2 dx + \frac{\alpha}{2} \int_{\Sigma} [u(x, t)]^2 d\Sigma \\ + \int_{\Sigma} v_n(x) \left\{ \frac{\partial p^0}{\partial t}(x, t) z(u; x, t) - \nabla_\Gamma p^0(x, t) \cdot \nabla_\Gamma z(u; x, t) \right\} d\Sigma$$

over the set $S(\Sigma)$ of admissible controls.

Here we denote by $z = z(u; x, t)$, $u \in L^2(\Sigma)$, $(x, t) \in Q$, a unique solution of the following state equation:

$$(2.125) \quad \int_{\Omega} \left\{ \frac{\partial z}{\partial t} \phi + \nabla z \cdot \nabla \phi \right\} dx = \int_{\partial\Omega} u \phi d\Gamma + \int_{\partial\Omega} v_n \left\{ \left(H u_0 - \frac{\partial y^0}{\partial t} \right) \phi \nabla_\Gamma y^0 \cdot \nabla_\Gamma \phi \right\} d\Gamma \\ \text{for a.e. } t \in (0, T) \quad \forall \phi \in H^2(\Omega), \quad \frac{\partial \phi}{\partial n} = 0,$$

$$(2.126) \quad z(x, 0) = 0 \quad \text{in } \Omega.$$

3. Dirichlet boundary conditions. This section is concerned with the shape sensitivity analysis of an optimal control problem for a parabolic equation with the Dirichlet boundary conditions.

Let us consider the following parabolic equation defined in a cylinder $Q = \Omega \times (0, T)$.

For $u \in L^2(\Sigma)$, find an element $y = y(u; x, t)$, $(x, t) \in Q$ which satisfies the following parabolic initial-boundary value problem:

$$(3.1) \quad \frac{\partial y}{\partial t} - \Delta y = 0 \quad \text{in } Q,$$

$$(3.2) \quad y = u \quad \text{on } \Sigma,$$

$$(3.3) \quad y(u; x, 0) = 0 \quad \text{in } \Omega.$$

There exists a unique weak solution $y \in L^2(Q)$ of the problem (3.1)–(3.3) which verifies the following integral identity [19]:

$$(3.4) \quad \int_Q y \left(\frac{\partial \phi}{\partial t} + \Delta \phi \right) dQ = \int_{\Sigma} u \frac{\partial \phi}{\partial n} d\Sigma$$

$$\forall \phi \in H^{2,1}(Q) \cap L^2(0, T; H_0^1(\Omega)), \quad \phi(x, T) = 0 \quad \text{in } \Omega,$$

and it is known [19] that for any $u \in L^2(\Sigma)$

$$(3.5) \quad y \in L^2(0, T; H^{1/2}(\Omega)),$$

$$(3.6) \quad \frac{\partial y}{\partial t} \in L^2(0, T; H^{-3/2}(\Omega)).$$

We refer the reader to [16] for the definition of Sobolev spaces $H^s(\Omega)$, $s = \frac{1}{2}, -\frac{3}{2}$. We assume that the set of admissible controls $K(\Sigma)$ is given by (2.10) and we define a cost functional of the form

$$(3.7) \quad I(v) = \frac{1}{2} \int_{\Omega} (y(v; x, T) - z_d(x))^2 dx + \frac{\alpha}{2} \int_{\Sigma} (v(x, t))^2 d\Sigma, \quad z_d(\cdot) \in L^2(\Omega)$$

which is well defined for any control $v \in K(\Sigma) \subset L^\infty(\Sigma)$; hence there exists a unique optimal control $u \in K(\Sigma)$ such that

$$(3.8) \quad I(u) \leq I(v) \quad \forall v \in K(\Sigma).$$

It seems, however, that the optimal control u fails to be differentiable in the direction of a vector field $V(\cdot, \cdot) \in C^1(0, \delta; C^2(R^n; R^n))$.

Therefore we will consider the following optimal control problem.

Problem (P). Find an element $u \in K(\Sigma) \cap H^{1,1}(\Sigma)$ which minimizes the cost functional

$$(3.9) \quad J(u) = \frac{1}{2} \int_{\Omega} (y(u; x, T) - z_d(x))^2 dx + \frac{\alpha}{2} \|u\|_{H^{1,1}(\Sigma)}^2, \quad \alpha > 0,$$

over the set $K_1(\Sigma) = K(\Sigma) \cap H^{1,1}(\Sigma)$.

Here $z_d \in H^1(R^n)$ is a given element; an element $y = y(u; x, t)$ is given by a unique solution of parabolic initial-boundary value problem (3.1)–(3.3), $u \in H^{1,1}(\Sigma)$ is the control.

Remark 3.1. The Sobolev space $H^{1,1}(\Sigma)$ can be replaced in (3.9) by a Sobolev space $H^{r,s}(\Sigma)$, with some s, r , $0 < s < r < 1$. For the sake of simplicity, we do not consider here the optimal choice of space $H^{r,s}(\Sigma)$ in (3.9).

In order to derive the form of the Euler and Lagrange derivatives of the solution of Problem (P) let us consider an optimal control problem defined in a cylinder $Q_\varepsilon = \Omega_\varepsilon \times (0, T)$, where the domain Ω_ε , $\varepsilon \in [0, \delta)$ is defined by (2.16).

Problem (P_ε). Find an element $u_\varepsilon \in K_1(\Sigma_\varepsilon)$ which minimizes the cost functional

$$(3.10) \quad J_\varepsilon(u) = \frac{1}{2} \int_{\Omega_\varepsilon} [\eta_\varepsilon(u; x, t) - z_d(x)]^2 dx + \frac{\alpha}{2} \|u\|_{H^{1,1}(\Sigma_\varepsilon)}^2, \quad \alpha > 0$$

over the set $K_1(\Sigma_\varepsilon)$.

Here $\eta_\varepsilon = \eta_\varepsilon(u; x, t)$, $u \in H^{1,1}(\Sigma_\varepsilon)$, $(x, t) \in Q_\varepsilon$, denotes a unique solution of the following parabolic equation:

$$(3.11) \quad \frac{\partial \eta_\varepsilon}{\partial t} - \Delta \eta_\varepsilon = 0 \quad \text{in } Q_\varepsilon,$$

$$(3.12) \quad \eta_\varepsilon = u \quad \text{on } \Sigma_\varepsilon,$$

$$(3.13) \quad \eta_\varepsilon(u; x, 0) = 0 \quad \text{in } \Omega_\varepsilon.$$

There exists an optimal solution $u_\varepsilon \in K_1(\Sigma_\varepsilon)$ or Problem (P_ε) which is given by a unique solution of the following optimality system.

Optimality system for Problem (P_ε) . Find $(u_\varepsilon, y_\varepsilon, p_\varepsilon) \in K_1(\Sigma_\varepsilon) \times W_\varepsilon(0, T) \times W_\varepsilon(0, T)$ which satisfy the following system.

State equation.

$$(3.14) \quad \frac{\partial y_\varepsilon}{\partial t} - \Delta y_\varepsilon = 0 \quad \text{in } Q_\varepsilon,$$

$$(3.15) \quad y_\varepsilon = u_\varepsilon \quad \text{in } \Sigma_\varepsilon,$$

$$(3.16) \quad y_\varepsilon(x, 0) = 0 \quad \text{in } \Omega_\varepsilon.$$

Adjoint state equation.

$$(3.17) \quad -\frac{\partial p_\varepsilon}{\partial t} - \Delta p_\varepsilon = 0 \quad \text{in } Q_\varepsilon,$$

$$(3.18) \quad p_\varepsilon = 0 \quad \text{on } \Sigma_\varepsilon,$$

$$(3.19) \quad p_\varepsilon(x, T) = y_\varepsilon(x, T) - z_d(x) \quad \text{in } \Omega_\varepsilon.$$

Optimality conditions.

$$(3.20) \quad u_\varepsilon \in K_1(\Sigma_\varepsilon): \alpha(u_\varepsilon, u - u_\varepsilon)_{H^{1,1}(\Sigma_\varepsilon)} \geq \int_{\Sigma_\varepsilon} \frac{\partial p_\varepsilon}{\partial n_\varepsilon} (u - u_\varepsilon) d\Sigma \quad \forall u \in K_1(\Sigma_\varepsilon).$$

In order to derive the form of Euler derivative \dot{u} of an optimal control $u_\varepsilon \in K_1(\Sigma_\varepsilon)$ we transport the optimality system (3.14)–(3.20) to the fixed cylinder $Q = \Omega \times (0, T)$.

Let us denote

$$(3.21) \quad u^\varepsilon = u_\varepsilon \circ T_\varepsilon \in H^{1,1}(\Sigma),$$

$$(3.22) \quad y^\varepsilon = y_\varepsilon \circ T_\varepsilon \in W(0, T),$$

$$(3.23) \quad p^\varepsilon = p_\varepsilon \circ T_\varepsilon \in W(0, T)$$

for $\varepsilon \in [0, \delta]$; then the change of variables in (3.14)–(3.20) leads to the following optimality system.

Find $(u^\varepsilon, y^\varepsilon, p^\varepsilon) \in K_1(\Sigma) \times W(0, T) \times W(0, T)$ which satisfy the following system.

State equation.

$$(3.24) \quad \gamma_\varepsilon \frac{\partial y^\varepsilon}{\partial t} - \operatorname{div}(A_\varepsilon \cdot \nabla y^\varepsilon) = 0 \quad \text{in } Q,$$

$$(3.25) \quad y^\varepsilon = u^\varepsilon \quad \text{on } \Sigma,$$

$$(3.26) \quad y^\varepsilon(x, 0) = 0 \quad \text{in } \Omega.$$

Adjoint state equation.

$$(3.27) \quad -\gamma_\varepsilon \frac{\partial p^\varepsilon}{\partial t} - \operatorname{div}(A_\varepsilon \cdot \nabla p^\varepsilon) = 0 \quad \text{in } Q,$$

$$(3.28) \quad p^\varepsilon = 0 \quad \text{on } \Sigma,$$

$$(3.29) \quad p^\varepsilon(x, T) = y^\varepsilon(x, T) - z_d^\varepsilon(x) \quad \text{in } \Omega.$$

Optimality conditions. $u^\varepsilon \in K_1(\Sigma)$:

$$(3.30) \quad a^\varepsilon(u^\varepsilon, \phi - u^\varepsilon) \geq \langle f^\varepsilon, \phi - u^\varepsilon \rangle \quad \forall \phi \in K_1(\Sigma).$$

Here $a^\varepsilon(\cdot, \cdot)$, $\langle f^\varepsilon, \cdot \rangle$ are defined by (4.28), (4.37), respectively.

THEOREM 3.1. *For $\varepsilon > 0$, ε small enough*

$$(3.31) \quad \|u^\varepsilon - u^0\|_{H^{1,1}(\Sigma)} \leq C\varepsilon.$$

Proof. It can be verified that an element $u^\varepsilon \in H^{1,1}(\Sigma)$, $\varepsilon \in [0, \delta)$ is an optimal solution of the following optimal control problem.

Problem (P^ε). Find an element $u^\varepsilon \in K_1(\Sigma)$ which minimizes the following cost functional:

$$(3.32) \quad J^\varepsilon(u) = \frac{1}{2} \int_{\Omega} \gamma_\varepsilon [\eta^\varepsilon(u; x, T) - z_d^\varepsilon(x)]^2 dx + \frac{\alpha}{2} a^\varepsilon(u, u),$$

over the set $K_1(\Sigma) \subset H^{1,1}(\Sigma)$.

Here the element $\eta^\varepsilon = \eta^\varepsilon(u)$ is given by unique solution of the following state equation:

$$(3.33) \quad \gamma_\varepsilon \frac{\partial \eta^\varepsilon}{\partial t} - \operatorname{div}(A_\varepsilon \nabla \eta_\varepsilon) = 0 \quad \text{in } Q,$$

$$(3.34) \quad \eta^\varepsilon = u \quad \text{on } \Sigma,$$

$$(3.35) \quad \eta^\varepsilon(u; x, 0) = 0 \quad \text{in } \Omega.$$

Taking into account Lemmas 4.3–4.5 in the Appendix, the estimation (3.31) follows by the same argument as in the proof of Theorem 2.1.

From (3.31) it follows that there exists an element

$$(3.36) \quad q \in H^{1,1}(\Sigma)$$

such that

$$(3.37) \quad u^\varepsilon = u^0 + \varepsilon q + r(\varepsilon) \quad \text{in } H^{1,1}(\Sigma)$$

where $r(\varepsilon)/\varepsilon \rightarrow 0$ weakly in $H^{1,1}(\Sigma)$ with $\varepsilon \downarrow 0$. Hence, in view of (3.24)–(3.29) it follows that

$$(3.38) \quad y^\varepsilon = y^0 + \varepsilon \dot{y} + o(\varepsilon) \quad \text{in } W(0, T),$$

$$(3.39) \quad p^\varepsilon = p^0 + \varepsilon \dot{p} + o(\varepsilon) \quad \text{in } W(0, T)$$

where elements \dot{y}, \dot{p} are given by unique solutions of auxiliary parabolic initial-boundary value problems. It is important to observe here that by (3.37)–(3.39) it follows that assumption (4.38) of Lemma 4.5 is verified. On the other hand, it can be shown, using the same argument as in the proof of Theorem 2.2, that the element q in (3.37) is given by a unique solution of an optimality system. This leads to the following result.

THEOREM 3.2. *For $\varepsilon > 0$, ε small enough*

$$(3.40) \quad u^\varepsilon = u^0 + \varepsilon \dot{u} + o(\varepsilon) \quad \text{in } H^{1,1}(\Sigma)$$

where $\|o(\varepsilon)\|_{H^{1,1}(\Sigma)}/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$.

The Euler derivative \dot{u} of an optimal control $u^0 = u_0$ in the direction of a vector field $V(\cdot, \cdot)$ is given by a unique solution of the following optimality system.

Find $(\dot{u}, \dot{y}, \dot{p}) \in S_1(\Sigma) \times W(0, T) \times W(0, T)$ such that the following system is satisfied.

State equation.

$$(3.41) \quad \frac{\partial \dot{y}}{\partial t} - \Delta \dot{y} = -\dot{\gamma} \frac{\partial y^0}{\partial t} + \operatorname{div}(\dot{A} \cdot \nabla y^0) \quad \text{in } Q,$$

$$(3.42) \quad \dot{y} = \dot{u} \quad \text{on } \Sigma,$$

$$(3.43) \quad \dot{y}(x, 0) = 0 \quad \text{in } \Omega.$$

Adjoint state equation.

$$(3.44) \quad -\frac{\partial \dot{p}}{\partial t} - \Delta \dot{p} = \dot{\gamma} \frac{\partial p^0}{\partial t} + \operatorname{div}(\dot{A} \cdot \nabla p^0) \quad \text{in } Q,$$

$$(3.45) \quad \dot{p} = 0 \quad \text{on } \Sigma,$$

$$(3.46) \quad \dot{p}(x, t) = \dot{y}(x, T) - \nabla z_d \cdot V(0) \quad \text{in } \Omega.$$

Optimality conditions.

$$(3.47) \quad \begin{aligned} \dot{u} &\in S_1(\Sigma), \\ \alpha(\dot{u}, \phi - \dot{u})_{H^{1,1}(\Sigma)} &\cong \langle f', \phi - \dot{u} \rangle - \alpha a'(u^0, \phi - \dot{u}) \quad \forall \phi \in S_1(\Sigma) \end{aligned}$$

where the cone $S_1(\Sigma) \subset H^{1,1}(\Sigma)$ is defined by (4.11).

Remark 3.2. The system (3.41)–(3.47) coincides with the optimality system for the following optimal control problem.

Problem (P). Find an element $\dot{u} \in S_1(\Sigma)$ which minimizes the following cost functional:

$$\begin{aligned} I(u) = & \frac{1}{2} \int_{\Omega} [\xi(u; x, T) - \nabla z_d(x) \cdot V(0, x)]^2 dx \\ & + \int_Q \left\{ \dot{\gamma}(x) \frac{\partial p^0}{\partial t}(x, t) + \operatorname{div}(\dot{A}(x) \cdot \nabla p^0(x, t)) \right\} \xi(u; x, t) dQ \\ & + \frac{\alpha}{2} \|\dot{u}\|_{H^{1,1}(\Sigma)}^2 \end{aligned}$$

over the set $S_1(\Sigma) \subset H^{1,1}(\Sigma)$.

Here we denote by $\xi(u; x, t)$, $u \in H^{1,1}(\Sigma)$, $(x, t) \in Q$ the solution of (3.41)–(3.43) for $\dot{u} = u$.

The proof of Theorem 3.2 is similar to the proof of Theorem 2.2 and is omitted here. The bilinear form $a'(\cdot, \cdot)$ and the linear form $\langle f', \cdot \rangle$ in (3.47) are defined by (4.30), (4.41), respectively.

In order to derive the explicit form of the right-hand side term in inequality (3.47) we use (4.30), (4.41) and take into account the following condition:

$$(3.48) \quad (u^0, \phi)_{H^{1,1}(\Sigma)} = \langle f^0, \phi \rangle \quad \forall \phi \in S_1(\Sigma)$$

which is satisfied by definition (4.11) of the cone $S_1(\Sigma) \subset H^{1,1}(\Sigma)$. In view of (3.48) we have

$$\begin{aligned} \alpha(u^0, \dot{\sigma}\phi)_{H^{1,1}(\Sigma)} &= \alpha \int_{\Sigma} \left\{ u^0(\phi \dot{\sigma}) + \frac{\partial u^0}{\partial t} \frac{\partial}{\partial t}(\phi \dot{\sigma}) + \nabla_{\Gamma} u^0 \cdot \nabla_{\Gamma}(\dot{\sigma}\phi) \right\} d\Sigma \\ (3.49) \quad &= \int_{\Sigma} \frac{\partial p^0}{\partial n}(\phi \dot{\sigma}) d\Sigma \quad \forall \phi \in S_1(\Sigma); \end{aligned}$$

here $\dot{\sigma}(x) = \operatorname{div} V(0, x) - \langle DV(0, x) \cdot \mathbf{n}(x), \mathbf{n}(x) \rangle_{R^n}$, $x \in \partial\Omega$. Therefore

$$\begin{aligned}
 \langle f', \phi \rangle - \alpha a'(u^0, \phi) &= \int_{\Sigma} \frac{\partial \dot{p}}{\partial n} \phi \, d\Sigma \\
 &+ \alpha \int_{\Sigma} \{ \phi \nabla_{\Gamma} u^0 \cdot \nabla_{\Gamma} \sigma' + \langle (DV + {}^*DV) \cdot \nabla_{\Gamma} u^0, \nabla_{\Gamma} \phi \rangle_{R^n} \} \, d\Sigma \\
 &- \int_{\Sigma} \phi \left\{ \langle DV \cdot \mathbf{n}, \mathbf{n} \rangle_{R^n} \frac{\partial p^0}{\partial n} + \langle (DV + {}^*DV) \cdot \nabla_{\Gamma} p^0, \mathbf{n} \rangle_{R^n} \right\} \, d\Sigma \\
 &\quad \forall \phi \in S_1(\Sigma).
 \end{aligned}
 \tag{3.50}$$

Now, if for a given vector field $V(\cdot, \cdot)$ the condition $v_n = \langle V(0), \mathbf{n} \rangle_{R^n} = 0$ is satisfied on $\partial\Omega$ then the optimality system (3.41)–(3.47) is identically verified by $\dot{u} = \langle \nabla_{\Gamma} u^0, \mathbf{v}_{\tau} \rangle_{R^n}$.

We recall that the Lagrange derivative of an optimal control u^0 in the direction of a vector field $V(\cdot, \cdot)$ is defined as follows:

$$u' = \dot{u} - \langle \nabla_{\Gamma} u^0, \mathbf{v}_{\tau} \rangle_{R^n}.$$

THEOREM 3.3. *The Lagrange derivative $u' \in H^{1,1}(\Sigma)$ of an optimal control $u^0 = u_0$ in the direction of a vector field $V(\cdot, \cdot) \in C^1(0, \delta; C^2(R^n; R^n))$ is given by a unique solution of the following optimality system.*

Find $(u', z, w) \in S_1(\Sigma) \times W(0, T) \times W(0, T)$ which satisfy the following system.

State equation.

$$\frac{\partial z}{\partial t} - \Delta z = 0 \quad \text{in } Q, \tag{3.51}$$

$$z = u' - v_n \frac{\partial y^0}{\partial n} \quad \text{on } \Sigma, \tag{3.52}$$

$$z(x, 0) = 0 \quad \text{in } \Omega. \tag{3.53}$$

Adjoint state equation.

$$-\frac{\partial w}{\partial t} - \Delta w = 0 \quad \text{in } Q, \tag{3.54}$$

$$w = -v_n \frac{\partial p^0}{\partial n} \quad \text{on } \Sigma, \tag{3.55}$$

$$w(x, T) = z(x, T) \quad \text{in } \Omega. \tag{3.56}$$

Optimality conditions.

$$u' \in S_1(\Sigma):$$

$$\begin{aligned}
 \alpha(u', \phi - u')_{H^{1,1}(\Sigma)} &\cong \int_{\Sigma} \frac{\partial w}{\partial n} (\phi - u') \, d\Sigma + \alpha \int_{\Sigma} \nabla_{\Gamma} u^0 \cdot \nabla_{\Gamma} (v_n H) (\phi - u') \, d\Sigma \\
 &- \int_{\Sigma} \left[\frac{\partial p^0}{\partial n} \frac{\partial v_n}{\partial n} + \nabla_{\Gamma} v_n \cdot \nabla_{\Gamma} p^0 \right] (\phi - u') \, d\Sigma \quad \forall \phi \in S_1(\Sigma).
 \end{aligned}
 \tag{3.57}$$

The proof of Theorem 3.3 is similar to the proof of Theorem 2.3 and is omitted here.

Let us observe that from (3.51)–(3.57) it follows that the Lagrange derivative $u' \in H^{1,1}(\Sigma)$ can be equivalently characterized as the unique solution of the following optimal control problem.

Problem (P'). Find an element $u' \in S_1(\Sigma)$ which minimizes the following cost functional:

$$(3.58) \quad \begin{aligned} I(u) = & \frac{1}{2} \int_{\Omega} [z(u; x, T)]^2 dx - \int_{\Sigma} v_n(x) \frac{\partial p^0}{\partial n}(x, t) z(u; x, t) d\Sigma + \frac{\alpha}{2} \|u\|_{H^{1,1}(\Sigma)}^2 \\ & + \int_{\Sigma} \left[\frac{\partial p^0}{\partial n}(x, t) \frac{\partial v_n}{\partial n}(x) + \nabla_{\Gamma} v_n(x) \cdot \nabla_{\Gamma} p^0(x, t) \right] u(x, t) d\Sigma \\ & - \alpha \int_{\Sigma} \{ \nabla_{\Gamma} u^0(x, t) \cdot \nabla_{\Gamma} (v_n(x) H(x)) \} u(x, t) d\Sigma \end{aligned}$$

over the set $S_1(\Sigma) \subset H^{1,1}(\Sigma)$.

Here $z(u; x, t)$, $u \in H^{1,1}(\Sigma)$, $(x, t) \in Q$ denotes the solution of Problem (3.51)–(3.53) for $u' = u$.

4. Appendix.

4.1. Metric projection $P_K(\cdot)$ in $L^2(\Sigma)$ onto $K(\Sigma)$. We recall briefly the properties of the metric projection $P_K(\cdot)$ in the space $L^2(\Sigma)$ onto the set $K(\Sigma)$ of the form (2.10) which are used in this paper. Suppose there are given elements $f, h \in L^2(\Sigma)$; we denote

$$f_{\varepsilon} = f + \varepsilon h + o(\varepsilon) \quad \text{in } L^2(\Sigma), \quad \varepsilon \in [0, \delta),$$

and let $u_{\varepsilon} = P_K(f_{\varepsilon})$, i.e.,

$$(4.1) \quad \begin{aligned} u_{\varepsilon} & \in K(\Sigma), \\ \int_{\Sigma} (u_{\varepsilon} - f_{\varepsilon})(u - u_{\varepsilon}) d\Sigma & \geq 0 \quad | u \in K(\Sigma). \end{aligned}$$

LEMMA 4.1. For $\varepsilon > 0$, ε small enough

$$(4.2) \quad u_{\varepsilon} = u_0 - \varepsilon P_S(h) + o(\varepsilon) \quad \text{in } L^2(\Sigma)$$

where $\|o(\varepsilon)\|_{L^2(\Sigma)}/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$.

The element $q = P_S(h)$ is given by a unique solution of the following variational inequality:

$$(4.3) \quad \begin{aligned} q & \in S(\Sigma), \\ \int_{\Sigma} (q - h)(\phi - q) d\Sigma & \geq 0 \quad \forall \phi \in S(\Sigma), \end{aligned}$$

i.e., q is the metric projection of h in space $L^2(\Sigma)$ onto set $S(\Sigma)$. For an elementary proof of Lemma 4.1 we refer the reader to [32].

Now, let an element $u_{\varepsilon} \in L^2(\Sigma)$, $\lambda \in [0, \delta)$ be given a unique solution of the following variational inequality:

$$(4.4) \quad \begin{aligned} u_{\varepsilon} & \in K(\Sigma), \\ \int_{\Sigma} (a_{\varepsilon} u_{\varepsilon} - f_{\varepsilon})(\phi - u_{\varepsilon}) d\Sigma & \geq 0 \quad \forall \phi \in K(\Sigma); \end{aligned}$$

here $a_{\varepsilon} \in L^{\infty}(\Sigma)$, $\varepsilon \in [0, \delta)$ is a given element such that

$$(4.5) \quad \begin{aligned} a_{\varepsilon} = a_0 + \varepsilon a_1 + o(\varepsilon) \quad \text{in } L^{\infty}(\Sigma) \quad \text{where } \|o(\varepsilon)\|_{L^{\infty}(\Sigma)}/\varepsilon & \rightarrow 0 \\ \text{with } \varepsilon \downarrow 0, a_0, a_1 & \in L^{\infty}(\Sigma) \end{aligned}$$

are given elements such that $a_0(x, t) \geq c > 0$ for almost everywhere $(x, t) \in \Sigma$.

LEMMA 4.2. For $\varepsilon > 0$, ε small enough

$$(4.6) \quad u_\varepsilon = u_0 + \varepsilon q + o(\varepsilon) \quad \text{in } L^2(\Sigma)$$

where $\|o(\varepsilon)\|_{L^2(\Sigma)}/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$. The element $q \in L^2(\Sigma)$ is given by a unique solution of the following variational inequality:

$$(4.7) \quad \begin{aligned} q &\in S(\Sigma), \\ \int_{\Sigma} (a_0 q + a_1 u_0 - h)(\phi - q) d\Sigma &\geq 0 \quad \forall \phi \in S(\Sigma). \end{aligned}$$

The proof of Lemma 4.2 follows by Theorem 1 in [35] and is therefore omitted here.

4.2. Metric projection in $H^{1,1}(\Sigma)$. We present some results on the differential stability of the metric projection in space $H^{1,1}(\Sigma)$ onto the set $K_1(\Sigma) = K(\Sigma) \cap H^{1,1}(\Sigma)$.

For a given element $f \in H^{1,1}(\Sigma)$, we denote by Pf its metric projection in space $H^{1,1}(\Sigma)$ onto set $K_1(\Sigma)$, i.e., the element Pf is given by a unique solution of the following variational inequality:

$$(4.8) \quad \begin{aligned} Pf &\in K_1(\Sigma), \\ (Pf - f, \phi - Pf)_{H^{1,1}(\Sigma)} &\geq 0 \quad \forall \phi \in K_1(\Sigma). \end{aligned}$$

It can be shown, using the results of Mignot [21], that for $\varepsilon > 0$, ε small enough

$$(4.9) \quad \forall h \in H^{1,1}(\Sigma): \quad P(f + \varepsilon h) = Pf + \varepsilon P'h + o(\varepsilon) \quad \text{in } H^{1,1}(\Sigma)$$

where $\|o(\varepsilon)\|_{H^{1,1}(\Sigma)}/\varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$.

The mapping $P': H^{1,1}(\Sigma) \rightarrow H^{1,1}(\Sigma)$ denotes the metric projection in space $H^{1,1}(\Sigma)$ onto a cone $S_1(\Sigma) \subset H^{1,1}(\Sigma)$, i.e.,

$$(4.10) \quad \begin{aligned} P'h &\in S_1(\Sigma), \\ (P'h - h, \phi - P'h)_{H^{1,1}(\Sigma)} &\geq 0 \quad \forall \phi \in S_1(\Sigma) \end{aligned}$$

where

$$(4.11) \quad \begin{aligned} S_1(\Sigma) &= \{\phi \in H^{1,1}(\Sigma) \mid \phi(x, t) \geq 0, \text{ q.e. on } \Xi_1, \phi(x, t) \leq 0, \text{ q.e. on } \Xi_2, (Pf - f, \phi)_{H^{1,1}(\Sigma)} = 0\} \\ &= \{\phi \in H^{1,1}(\Sigma) \mid \phi(x, t) \geq 0, \text{ q.e. on } \Xi_1^0, \phi(x, t) \leq 0, \text{ q.e. on } \Xi_2^0, \phi(x, t) = 0, \\ &\quad \text{q.e. on } \Xi_1^+ \cup \Xi_2^+\}. \end{aligned}$$

Here we denote

$$(4.12) \quad \Xi_1 = \{(x, t) \in \Sigma \mid u_0(x, t) = 0\}, \quad u_0 = Pf \in H^{1,1}(\Sigma),$$

$$(4.13) \quad \Xi_2 = \{(x, t) \in \Sigma \mid u_0(x, t) = M\},$$

$$(4.14) \quad \Xi_i^0 = \{(x, t) \in \Xi_i \mid u_0(x, t) = f(x, t)\},$$

$$(4.15) \quad \Xi_i^+ = \Xi_i \setminus \Xi_i^0, \quad i = 1, 2.$$

Remark 4.1. Let us recall [21], that in (4.11) “q.e.” means everywhere with possible exception of a set of capacity zero.

Let us consider the differential stability with respect to the parameter of solutions of the following variational inequality:

$$(4.16) \quad \begin{aligned} u^\varepsilon &\in K_1(\Sigma), \\ a^\varepsilon(u^\varepsilon, \phi - u^\varepsilon) &\geq \langle f^\varepsilon, \phi - u^\varepsilon \rangle \quad \forall \phi \in K_1(\Sigma) \end{aligned}$$

where $a^\varepsilon(\cdot, \cdot): H^{1,1}(\Sigma) \times H^{1,1}(\Sigma) \rightarrow R$ is a bilinear form and $\langle f^\varepsilon, \cdot \rangle: H^{1,1}(\Sigma) \rightarrow R$ is a linear form for $\varepsilon \in [0, \delta]$. We assume that the following conditions are fulfilled:

$$(4.17) \quad a^0(u, v) = (u, v)_{H^{1,1}(\Sigma)} \quad \forall u, v \in H^{1,1}(\Sigma)$$

and there exists a continuous bilinear form

$$(4.18) \quad a'(\cdot, \cdot): H^{1,1}(\Sigma) \times H^{1,1}(\Sigma) \rightarrow R$$

such that

$$(4.19) \quad \lim_{\varepsilon \rightarrow 0} \sup_{\|u\|, \|v\| \leq 1} |(a^\varepsilon(u, v) - a^0(u, v))/\varepsilon - a'(u, v)| = 0;$$

here we denote $\|\cdot\| = \|\cdot\|_{H^{1,1}(\Sigma)}$. Furthermore, we assume that there exists an element

$$(4.20) \quad f' \in H' = (H^{1,1}(\Sigma))'$$

such that

$$(4.21) \quad f^\varepsilon = f^0 + \varepsilon f' + o(\varepsilon) \quad \text{in } H'$$

where $\|o(\varepsilon)\|_{H'} / \varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$.

LEMMA 4.3. For $\varepsilon > 0$, ε small enough

$$(4.22) \quad u^\varepsilon = u^0 + \varepsilon q + o(\varepsilon) \quad \text{in } H^{1,1}(\Sigma)$$

where $\|o(\varepsilon)\|_{H^{1,1}(\Sigma)} / \varepsilon \rightarrow 0$ with $\varepsilon \downarrow 0$. The element $q \in H^{1,1}(\Sigma)$ is given by a unique solution of the following variational inequality:

$$(4.23) \quad \begin{aligned} q &\in S_1(\Sigma), \\ (q, \phi - q)_{H^{1,1}(\Sigma)} &\geq \langle f, \phi - q \rangle - a'(u^0, \phi - q) \quad \forall \phi \in S_1(\Sigma). \end{aligned}$$

The proof of Lemma 4.3 follows, in view of (4.9), by Theorem 1 in [35] and is therefore omitted here.

4.3. Transport of surface integrals. We recall the results concerning the transport of surface integrals defined on Σ_ε to the fixed surface Σ using the mapping T_ε .

Let us denote

$$(4.24) \quad a_\varepsilon(u, v) \stackrel{\text{def}}{=} (u, v)_{H^{1,1}(\Sigma_\varepsilon)} \quad \forall u, v \in H^{1,1}(\Sigma_\varepsilon).$$

We define

$$(4.25) \quad a^\varepsilon(u, v) \stackrel{\text{def}}{=} a_\varepsilon(u \circ T_\varepsilon^{-1}, v \circ T_\varepsilon^{-1}) \quad \forall u, v \in H^{1,1}(\Sigma).$$

For any $\phi \in H^1(\partial\Omega_\varepsilon)$ we will write $\nabla_\Gamma \phi$ instead of $\nabla_{\Gamma_\varepsilon} \phi$.

LEMMA 4.4. The bilinear form (4.25) verifies condition (4.19).

Proof. We first find the explicit form of $a^\varepsilon(\cdot, \cdot)$. We have for all $u, v \in H^{1,1}(\Sigma_\varepsilon)$:

$$(4.26) \quad \begin{aligned} a_\varepsilon(u, v) &= \int_{\Sigma_\varepsilon} \left\{ uv + \frac{\partial u}{\partial t} \frac{\partial v}{\partial t} + \langle \nabla_\Gamma u, \nabla_\Gamma v \rangle_{R^n} \right\} d\Sigma \\ &= a^\varepsilon(u \circ T_\varepsilon, v \circ T_\varepsilon) \\ &= \int_\Sigma \sigma_\varepsilon \left\{ (u \circ T_\varepsilon)(v \circ T_\varepsilon) + \frac{\partial}{\partial t} (u \circ T_\varepsilon) \frac{\partial}{\partial t} (v \circ T_\varepsilon) \right. \\ &\quad \left. + \langle (\nabla_\Gamma u) \circ T_\varepsilon, (\nabla_\Gamma v) \circ T_\varepsilon \rangle_{R^n} \right\} d\Sigma. \end{aligned}$$

Here we change the variables, using the mapping T_ε .

It can be shown that for all $\phi \in H^1(\partial\Omega_\varepsilon)$

$$(4.27) \quad (\nabla_\Gamma \phi) \circ T_\varepsilon = {}^*DT_\varepsilon^{-1} \cdot \nabla_\Gamma(\phi \circ T_\varepsilon) \\ - \|{}^*DT_\varepsilon^{-1} \cdot \mathbf{n}\|_{R^n}^{-2} \langle {}^*DT_\varepsilon^{-1} \cdot \nabla_\Gamma(\phi \circ T_\varepsilon), {}^*DT_\varepsilon^{-1} \cdot \mathbf{n} \rangle_{R^n} {}^*DT_\varepsilon^{-1} \cdot \mathbf{n};$$

therefore

$$(4.28) \quad a^\varepsilon(u, v) = \int_\Sigma \sigma_\varepsilon \\ \cdot \left\{ uv + \frac{\partial u}{\partial t} \frac{\partial v}{\partial t} + ({}^*DT_\varepsilon^{-1} \cdot \nabla_\Gamma u - \|{}^*DT_\varepsilon^{-1} \cdot \mathbf{n}\|_{R^n}^{-2} \langle {}^*DT_\varepsilon^{-1} \cdot \nabla_\Gamma u, {}^*DT_\varepsilon^{-1} \cdot \mathbf{n} \rangle_{R^n} {}^*DT_\varepsilon^{-1} \cdot \mathbf{n}) \right. \\ \left. \cdot ({}^*DT_\varepsilon^{-1} \cdot \nabla_\Gamma v - \|{}^*DT_\varepsilon^{-1} \cdot \mathbf{n}\|_{R^n}^{-2} \langle {}^*DT_\varepsilon^{-1} \cdot \nabla_\Gamma v, {}^*DT_\varepsilon^{-1} \cdot \mathbf{n} \rangle_{R^n} {}^*DT_\varepsilon^{-1} \cdot \mathbf{n}) \right\} d\Sigma \\ \forall u, v \in H^{1,1}(\Sigma).$$

Since [42]

$$(4.29) \quad {}^*DT_\varepsilon^{-1} = I - \varepsilon {}^*DV(0) + o(\varepsilon) \quad \text{in } C(\bar{\Omega}; R^n),$$

then from (4.28), in view of (2.22), it follows that the condition (4.19) is satisfied where the bilinear form $a'(\cdot, \cdot)$ is given by

$$(4.30) \quad a'(u, v) = \int_\Sigma \left\{ \dot{\sigma} \left[uv + \frac{\partial u}{\partial t} \frac{\partial v}{\partial t} + \langle \nabla_\Gamma u, \nabla_\Gamma v \rangle_{R^n} \right] \right. \\ \left. - \langle (DV(0) + {}^*DV(0)) \cdot \nabla_\Gamma u, \nabla_\Gamma v \rangle_{R^n} \right\} d\Sigma \quad \forall u, v \in H^{1,1}(\Sigma).$$

Remark 4.2. If the tangential component $\mathbf{v}_\tau(x)$, $x \in \partial\Omega$, of a given vector field $V(0)$ satisfies

$$(4.31) \quad \mathbf{v}_\tau(x) = 0, \quad x \in \partial\Omega,$$

then it can be shown that

$$(4.32) \quad \dot{\sigma} = v_n H;$$

furthermore for $\Omega \subset R^2$, $\nabla_\Gamma = \boldsymbol{\tau} \partial / \partial \tau$,

$$(4.33) \quad \langle (DV(0) + {}^*DV(0)) \boldsymbol{\tau}, \boldsymbol{\tau} \rangle_{R^2} = 2v_n \langle \boldsymbol{\tau}, \partial \mathbf{n} / \partial \tau \rangle_{R^2}$$

on $\partial\Omega$; here H is the mean curvature of $\partial\Omega$. Finally, let us consider the linear form

$$(4.34) \quad \langle f_\varepsilon, \cdot \rangle: H^{1,1}(\Sigma_\varepsilon) \rightarrow R \quad \text{given by}$$

$$\langle f_\varepsilon, \phi \rangle \stackrel{\text{def}}{=} \int_{\Sigma_\varepsilon} \frac{\partial p_\varepsilon}{\partial n_\varepsilon} \phi \, d\Sigma, \quad \phi \in H^{1,1}(\Sigma_\varepsilon)$$

where $\partial p_\varepsilon / \partial n_\varepsilon = \langle \nabla p_\varepsilon, \mathbf{n}_\varepsilon \rangle_{R^n} \in (H^{1,1}(\Sigma_\varepsilon))'$ is a given element.

We define an element $f^\varepsilon \in (H^{1,1}(\Sigma))'$

$$(4.35) \quad \langle f^\varepsilon, v \rangle \stackrel{\text{def}}{=} \langle f_\varepsilon, v \circ T_\varepsilon^{-1} \rangle \quad \forall v \in H^{1,1}(\Sigma)$$

and after the change of variables in (4.34), we obtain

$$(4.36) \quad \langle f^\varepsilon, \phi \circ T_\varepsilon \rangle = \int_\Sigma \sigma_\varepsilon \left(\frac{\partial p_\varepsilon}{\partial n_\varepsilon} \circ T_\varepsilon \right) \phi \circ T_\varepsilon \, d\Sigma \quad \forall \phi \in H^{1,1}(\Sigma_\varepsilon).$$

Thus simple calculations show that

$$(4.37) \quad \langle f^\varepsilon, v \rangle = \int_\Sigma v \langle A_\varepsilon \nabla p^\varepsilon, \mathbf{n} \rangle_{R^n} \, d\Sigma \quad \forall v \in H^{1,1}(\Sigma)$$

where we denote $p^\varepsilon = p_\varepsilon \circ T_\varepsilon$.

LEMMA 4.5. *Let us assume that the element p^ε is differentiable with respect to ε at $\varepsilon = 0^+$, i.e., there exists the strong limit*

$$(4.38) \quad \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \langle A_\varepsilon (\nabla p^\varepsilon - \nabla p^0), \mathbf{n} \rangle_{R^n} = \frac{\partial \dot{p}}{\partial n} \quad \text{in } (H^{1,1}(\Sigma))'.$$

Then the element $f^\varepsilon \in (H^{1,1}(\Sigma))'$ defined by (4.37) satisfies the condition (4.21).

Proof. We have

$$(4.39) \quad A_0 = I,$$

and hence

$$(4.40) \quad \langle f^0, v \rangle = \int_\Sigma v \frac{\partial p^0}{\partial n} d\Sigma.$$

In view of (4.38), (2.21), (2.24), it follows that

$$(4.41) \quad \begin{aligned} \frac{1}{\varepsilon} \langle f^\varepsilon - f^0, v \rangle &= \frac{1}{\varepsilon} \int_\Sigma v \langle A_\varepsilon (\nabla p^\varepsilon - \nabla p^0), \mathbf{n} \rangle_{R^n} d\Sigma \\ &\quad + \frac{1}{\varepsilon} \int_\Sigma v \langle (A_\varepsilon - I) \cdot \nabla p^0, \mathbf{n} \rangle_{R^n} d\Sigma \rightarrow \int_\Sigma v \frac{\partial \dot{p}}{\partial n} d\Sigma \\ &\quad + \int_\Sigma v \left\{ \operatorname{div} V(0) \frac{\partial p^0}{\partial n} - \langle (DV(0) + *DV(0)) \cdot \nabla p^0, \mathbf{n} \rangle_{R^n} \right\} d\Sigma \\ &= \langle f', v \rangle \quad \forall v \in H^{1,1}(\Sigma) \quad \text{for } \varepsilon \downarrow 0. \end{aligned}$$

REFERENCES

- [1] M. P. BENDSØE, N. OLHOFF, AND J. SOKOŁOWSKI, *Sensitivity analysis of problems of elasticity with unilateral constraints*, J. Structural Mech., 13 (1985), pp. 201–222.
- [2] D. CHENAIS, *Sur une famille de variétés à bord lipschitziennes: application a un problème d'identification de domaines*, Ann. Inst. Fourier, 27 (1977), pp. 201–231.
- [3] K. DEMS AND Z. MROZ, *Variational approach by means of adjoint systems to structural optimization and sensitivity analysis, Part 2. Structure shape variations*, Internat. J. Solids and Structures, 20 (1984), pp. 527–552.
- [4] A. DERVIEUX, *Resolution de problèmes à frontiere libre*, Thèse d'état, l'Université Paris, 1981.
- [5] A. V. FIANCO, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, New York, 1983.
- [6] S. FITZPATRICK AND R. R. PHELPS, *Differentiability of the metric projection in Hilbert space*, Trans. Amer. Math. Soc., 270 (1982), pp. 483–501.
- [7] J. HADAMARD, *Mémoire un: Le problème de l'analyse relatif à l'équilibre des plaques élastiques encastrées*, in *Mémoire des savants étrangers*, 1908.
- [8] A. HARAUX, *How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities*, J. Math. Soc. Japan, 29 (1977), pp. 615–631.
- [9] E. J. HAUG, K. K. CHOI, AND V. KOMKOV, *Design Sensitivity Analysis of Structural Systems*, Academic Press, New York 1986.
- [10] E. J. HAUG AND J. CÉA, *Optimization of Distributed Parameter Structures*, Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981.
- [11] R. B. HOLMES, *Smoothness of certain metric projections on Hilbert space*, Trans. Amer. Math. Soc., 184 (1973), pp. 87–100.
- [12] P. HOLNICKI, J. SOKOŁOWSKI, AND A. ZOCHOWSKI, *Sensitivity analysis of an optimal control problem arising from air quality control in urban areas*, in *System Modelling and Optimization*, Proc. 12th IFIP Conference, Budapest, Hungary, Springer-Verlag, Berlin, 1986. As [34] pp. 854–865.
- [13] K. JITTORNTRUM, *Solution point differentiability without strict complementarity in nonlinear programming*, in *Mathematical Programming Studies* 21, A. V. Fianco, ed., North-Holland, Amsterdam, 1984, pp. 127–138.
- [14] I. LASIECKA AND R. TRIGGIANI, *Dirichlet boundary control problems for parabolic equations with quadratic cost: analyticity and Riccati's feedback synthesis*, SIAM J. Control Optim., 21 (1984), pp. 41–67.

- [15] J. L. LIONS, *Contrôle optimal de systems gouvernes par des equations aux derivees partielles*, Dunod, Paris, 1968.
- [16] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, Vol. 1, Dunod, Paris, 1968.
- [17] J. L. LIONS, *Perturbations singulières dans les problèmes aux limites et en contrôle optimal*, Lecture Notes in Mathematics 323, Springer-Verlag, Berlin, 1973.
- [18] ———, *Some Methods in the Mathematical Analysis of Systems and Their Control*, Gordon and Breach, New York, 1981.
- [19] ———, *Contrôle des systems distribues singuliers*, Dunod, Paris, 1983.
- [20] K. MALANOWSKI AND J. SOKOŁOWSKI, *Sensitivity of solutions to convex, control constrained optimal control problems for distributed parameter systems*, J. Math. Anal. Appl., 120 (1986), pp. 240–263.
- [21] F. MIGNOT, *Contrôle dans les inéquations variationnelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130–185.
- [22] F. MURAT AND J. SIMON, *Sur le contrôle par un domaine géométrique*, Université de Paris 6, Publication du Laboratoire d'Analyse Numérique 189, 1976.
- [23] O. PIRONNEAU, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, New York, 1984.
- [24] B. ROUSSELET, *Etude de la régularité des valeurs propres par rapport à des déformations lipschitziennes de domaines*, C.R. Acad. Sci. Paris Sér. I, 283 (1976), p. 507.
- [25] B. ROUSSELET, *Quelques résultats en optimisation de domaines*, Thèse, Université de Nice, 1982.
- [26] ———, *Shape sensitivity of a membrane*, J. Optim. Theory Appl., 40 (1983), pp. 595–623.
- [27] J. SIMON, *Variation par rapport au domaine dans des problèmes aux limites*, publication du Laboratoire d'Analyse Numérique 189, Université de Paris 6, 1980.
- [28] J. SOKOŁOWSKI, *Sensitivity analysis for a class of variational inequalities*, in Optimization of Distributed Parameter Structures, Vol. 2, E. J. Haug and J. Céa, eds., Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981, pp. 1600–1609.
- [29] ———, *Optimal control in coefficients of boundary value problems with unilateral constraints*, Bull. Polish Acad. Sci. Tech. Sci., 31 (1983), pp. 71–81.
- [30] ———, *Sensitivity analysis of Signorini variational inequality*, in Banach Center Publications, B. Bojarski, ed., Polish Scientific Publishers, Warsaw, 1987.
- [31] ———, *Differential stability of solutions to constrained optimization problems*, Appl. Math. Optim., 13 (1985), pp. 97–115.
- [32] ———, *Differential stability of control constrained optimal control problems for distributed parameter systems*, in Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Lecture Notes in Control and Information Sciences 75, Springer-Verlag, Berlin, 1985, pp. 382–399.
- [33] ———, *Sensitivity analysis and parametric optimization of optimal control problems for distributed parameter systems*, Zeszyty Nauk. Politech. Warszawskiej, seria Elektronika 73, 1985, 152 pages. (In Polish.)
- [34] ———, *Differential stability of solutions to boundary optimal control problems for parabolic systems*, in System Modelling and Optimization, Proc. 12th IFIP Conference, Budapest, Hungary, Lecture Notes in Control and Information Sciences 84, Springer-Verlag, Berlin, 1986, pp. 854–865.
- [35] ———, *Sensitivity analysis of control constrained optimal control problems for distributed parameter systems*, SIAM J. Control Optim., 25 (1987), pp. 1542–1556.
- [36] ———, *Sensitivity analysis of contact problems with prescribed friction*, Appl. Math. Optim., to appear.
- [37] J. SOKOŁOWSKI AND J. P. ZOLÉSIO, *Dérivée par rapport au domaine de la solution d'un problème unilateral*, C.R. Acad. Sci. Paris Sér. I, 301 (1985), pp. 103–106.
- [38] ———, *Shape sensitivity analysis of unilateral problems*, Publication Mathématiques 67, Université de Nice, 1985; SIAM J. Math. Anal., 18 (1987), pp. 1416–1437.
- [39] ———, *Shape sensitivity analysis of an elastic-plastic torsion problems*, Bull. Polish Acad. Sci. Tech. Sci., 33 (1985), pp. 579–586.
- [40] J. P. ZOLÉSIO, *Shape controllability for free boundaries*, in System Modelling and Optimization, P. Thoft-Christensen, ed., Lecture Notes in Control and Information Sciences 59, Springer-Verlag, Berlin 1984, pp. 354–361.
- [41] ———, *Identification de domaines par déformations*, Thèse d'Etat, Université de Nice, 1979.
- [42] ———, *The material derivative (or speed) method for shape optimization*, in Optimization of Distributed Parameter Structures, Vol. 2, E. J. Haug and J. Céa, eds., Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981, pp. 1089–1151.
- [43] J. SIMON, *Differentiability with respect to the domain in boundary value problems*, Numer. Funct. Anal. Optim., 2 (1980), pp. 649–687.

A QP-FREE, GLOBALLY CONVERGENT, LOCALLY SUPERLINEARLY CONVERGENT ALGORITHM FOR INEQUALITY CONSTRAINED OPTIMIZATION*

ELIANE R. PANIER†, ANDRÉ L. TITS†, AND JOSÉ N. HERSKOVITS‡

Abstract. An algorithm is proposed for the minimization of a smooth function subject to smooth inequality constraints. Unlike sequential quadratic programming type methods, this algorithm does not involve the solution of quadratic programs, but merely that of linear systems of equations. Locally the iteration can be viewed as a perturbation of a quasi-Newton iteration on both the primal and dual variables for the solution of the equalities in the Kuhn-Tucker first order conditions of optimality. It is observed that, provided the current iterate is feasible and the current multiplier estimates are strictly positive, the primal component of the quasi-Newton direction is a direction of descent for the objective function. This fact is used to induce global convergence, without the need of a surrogate merit function. A careful "bending" of the search direction prevents any Maratos-like effect, and local superlinear convergence is proven. While the algorithm requires that an initial feasible point be available, the successive iterates it constructs are feasible as well, a valuable property in the context of engineering system design.

Key words. constrained optimization, quasi-Newton iteration, global convergence, superlinear convergence, engineering design

AMS(MOS) subject classifications. 90C30, 65K10

1. Introduction. By analogy with a technique widely used in equality constrained optimization, it is proposed to solve the inequality constrained problem

$$(1.1) \quad \min f(x) \quad \text{s.t. } g(x) \leq 0,$$

where $f: \mathbb{R}^n \rightarrow \mathbb{R}$ and $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$ are smooth, using a quasi-Newton iteration applied to the solution of the equalities in the Kuhn-Tucker first order necessary conditions of optimality

$$\begin{aligned} \nabla_x L(x, \lambda) &= 0, \\ \lambda_i g_i(x) &= 0, \quad i = 1, \dots, m \end{aligned}$$

where $L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i g_i(x)$ is the Lagrangian associated with (1.1). Specifically, consider the linear system in (d^0, λ^0)

$$(1.2) \quad H d^0 + \nabla_x L(x, \lambda^0) = 0,$$

$$(1.3) \quad \mu_i \langle \nabla g_i(x), d^0 \rangle + \lambda_i^0 g_i(x) = 0, \quad i = 1, \dots, m$$

where H is an estimate of the Hessian of L , x the current estimate of a solution x^* , $x + d^0$ the next estimate, μ the current estimate of the Kuhn-Tucker multiplier vector associated with x^* , and λ^0 the next estimate of this vector. Locally, this iteration is a higher order perturbation of the sequential quadratic programming (SQP) iteration and a stabilization scheme to force global convergence has been proposed in the form

* Received by the editors November 17, 1986; accepted for publication (in revised form) July 3, 1987. This research was supported by the National Science Foundation under grants DMC-84-20740 and CDR-85-00108, and by the FINEP (Brazilian Research and Development Foundation). It was initiated when the second author visited Coordenação dos Programas de Pós-Graduação do Engenharia on a travel grant from the Partners of the Americas.

† Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, Maryland 20742.

‡ Programa de Engenharia Mecânica, Coordenação dos Programas de Pós-Graduação do Engenharia/Universidade Federal do Rio de Janeiro, 21945 Rio de Janeiro, Brazil.

of a compound algorithm involving the solution of a quadratic (or a linear) program at each one of the early iterations [14]. In this scheme, the local iteration is used if the step it yields is "small enough" according to a prespecified linear convergence rate. Although this algorithm is proven to converge, it is subject to "false starts" and repeated switching that are bound to reduce its efficiency.

In [5], [6], an iteration similar to (1.2)–(1.3) is considered, although μ is not interpreted as the current multiplier estimate. There it is observed that, if H is positive definite, if μ has strictly positive components, and if x satisfies the constraints as *strict* inequalities, d^0 is a descent direction for both f and $L(\cdot, \lambda^0)$. To allow a reasonably long step to be taken without leaving the feasible set, a modified direction d is computed by replacing the right-hand side of (1.3) by a suitably small negative quantity. With such a search direction and with a stepsize rule based on the decrease of f and of the g_i 's for which the estimated multiplier λ_i is negative, global convergence to Kuhn–Tucker points is proven [6]. However, local superlinear convergence of the quasi-Newton iteration is lost due to truncation of the step. Thus it is proposed in [5] to use instead a line search based on a decrease of $L(\cdot, \lambda^0)$. Under the assumption that the sequence of iterates converges, it is proven that the limit point x^* is still a Kuhn–Tucker point. However, since λ^0 (and thus the merit function $L(\cdot, \lambda^0)$) changes at each iteration, oscillations between several accumulation points are clearly possible. It is then proven in [5] that, under the previous assumption and provided the components of μ corresponding to constraints that are active at the solution x^* are forced to converge to the Kuhn–Tucker multipliers at x^* , local superlinear convergence occurs. However, the question of devising a suitable updating rule for μ is not addressed and, as global convergence requires that the components of μ corresponding to constraints that are active at the solution be uniformly bounded from below, knowledge of a lower bound for the corresponding multipliers is required.

The algorithm proposed in this paper overcomes the shortcomings just pointed out, namely, (i) by using a stepsize rule based on a decrease of f , it avoids oscillations; sufficient descent is achieved by carefully limiting the size of the perturbation defining d , (ii) by using a further correction \tilde{d} and a search along an arc [9], [13], it prevents the step from being truncated close to the solution, and (iii) it includes an updating rule for the vector μ which, while preserving global convergence, ensures convergence of the components of that vector to the true multipliers at the solution, thus ensuring superlinear convergence. While the need to provide an initial feasible point may be felt as a liability, the fact that all iterates are feasible, coupled with the decrease of the objective function at each iteration, can be of great value, for instance in the context of engineering design [12]. To our knowledge, the only existing superlinearly convergent algorithm enjoying these properties is the one in [13], and it requires the solution of two quadratic programs at each iteration. It is clear that the proposed algorithm can be extended to handle equality constraints as well, since for equality constrained problems the SQP iteration itself amounts to solving a system of linear equations. Obviously, however, feasibility of the successive iterates with respect to equality constraints cannot be preserved, and a merit function taking these into account must be used. A scheme such as the one used in [10] and in [6] may be appropriate. Finally, Coleman and Conn [1], [2] and Spellucci [17] have also proposed superlinearly convergent algorithms in which computation of a search direction requires only the solution of linear systems of equations. Unlike ours, these schemes are based on active set strategies.

Thus, let d^0 be as defined in (1.2)–(1.3). Suppose that $\mu > 0$ and $g(x) < 0$, and assume that H is positive definite. Clearly, d^0 is a descent direction for f at x since

(1.2) and (1.3) yield

$$\begin{aligned}
 \langle \nabla f(x), d^0 \rangle &= -\langle d^0, Hd^0 \rangle - \sum_{i=1}^m \lambda_i^0 \langle \nabla g_i(x), d^0 \rangle \\
 (1.4) \qquad &= -\langle d^0, Hd^0 \rangle + \sum_{i=1}^m \frac{(\lambda_i^0)^2}{\mu_i} g_i(x) \\
 &\leq -\langle d^0, Hd^0 \rangle \\
 (1.5) \qquad &\leq -\sigma \|d^0\|^2
 \end{aligned}$$

for some positive σ . However, as pointed out above, d^0 is not entirely suitable as a search direction. Indeed, if any $g_i(x)$ becomes very close to 0, (1.3) forces d^0 to tend to a direction tangent to the feasible set. Since feasibility of all iterates is required, this may result in a collapse of the step length, and convergence to a nonstationary point may occur. This effect can be avoided if one substitutes in the right-hand side of (1.3) a negative number $-\varepsilon$. In order for the convergence properties of the quasi-Newton iteration to be preserved, ε would have to tend to zero faster than d^0 , say, $\varepsilon = \|d^0\|^\nu$ for some $\nu > 1$. Thus, after d^0 is obtained, a new direction d^1 would be computed by solving the linear system (see also [5], [6], [13])

$$\begin{aligned}
 Hd^1 + \nabla_x L(x, \lambda^1) &= 0, \\
 (1.6) \qquad \mu_i \langle \nabla g_i(x), d^1 \rangle + \lambda_i^1 g_i(x) &= -\mu_i \|d^0\|^\nu, \quad i = 1, \dots, m.
 \end{aligned}$$

The introduction of μ_i in the right-hand side of (1.6) reflects the fact that the correction is necessary only close to the constraint boundaries. Unfortunately, the new direction d^1 may no longer be a descent direction for f . Indeed, (1.5) now becomes

$$\langle \nabla f(x), d^1 \rangle \leq -\sigma \|d^1\|^2 + \|d^0\|^\nu \sum_i \lambda_i.$$

We thus propose to use a search direction of the form

$$d = (1 - \rho)d^0 + \rho d^1$$

and the corresponding approximate multiplier vector

$$\lambda = (1 - \rho)\lambda^0 + \rho\lambda^1$$

with $\rho \in [0, 1]$ as large as possible subject to the constraint that

$$\langle \nabla f(x), d \rangle \leq \theta \langle \nabla f(x), d^0 \rangle$$

for a prespecified $\theta \in (0, 1)$. It turns out that, with such a search direction, the basic convergence properties of the quasi-Newton iteration are preserved.

Equation (1.4) suggests that linear system (1.2)–(1.3) may be singular if any of the component of μ vanishes (since d^0 may then tend to infinity), and that chronic ill-conditioning is likely to occur if some of these components are small. While, as shown below, bounding the components of μ away from zero ensures convergence to Kuhn-Tucker points, such strategy may prevent superlinear convergence as μ may not converge to the true multiplier vector at the solution. It thus seems reasonable to bound μ away from 0 unless convergence to a Kuhn-Tucker point is detected, i.e., to use an update rule of the type (superscript “+” indicates the next iteration)

$$\mu_i^+ = \max \{\lambda_i, \|d\|\}, \quad i = 1, \dots, m$$

if all multipliers λ_i seem to converge to nonnegative numbers, and

$$\mu_i^+ = \mu_{0,i}, \quad i = 1, \dots, m$$

otherwise, where $\mu_{0,i}$ are given positive numbers. The multipliers λ_i can be considered not to converge to nonnegative numbers if, e.g., they belong to the set

$$J = \{i \text{ s.t. } \lambda_i \leq g_i(x)\}.$$

Indeed, assuming that strict complementarity holds, J will be empty in the vicinity of a Kuhn–Tucker point. With such an updating scheme, components of μ may become very small only in the vicinity of a Kuhn–Tucker pair (x^*, μ^*) , where (1.2)–(1.3) is well behaved whenever second order sufficiency conditions of optimality hold.

Maratos [8] pointed out that, in the case of the SQP iteration (direction referred to below as d_{SQP}) with line search based on the decrease of an exact penalty function, the unit step may not be accepted close to a solution. Mayne and Polak [9] later showed that, although $x + d_{\text{SQP}}$ may not be “lower” than x , $x + d_{\text{SQP}} + \tilde{d}$ will, if \tilde{d} solves

$$(1.7) \quad \min \frac{1}{2} \|\tilde{d}\|^2 \quad \text{s.t. } g_i(x + d_{\text{SQP}}) + \langle \nabla g_i(x), \tilde{d} \rangle = 0 \quad \forall i \in I(x^*)$$

where $I(x^*)$ is the index set of active constraints at x^* . To combine the global convergence properties associated with d_{SQP} and the local convergence properties associated with $d_{\text{SQP}} + \tilde{d}$, Mayne and Polak then suggest that the search be performed along the arc $x + td_{\text{SQP}} + t^2\tilde{d}$.¹ Correction (1.7) can be viewed as a second quasi-Newton iteration (with $\nabla g_i(x + d_{\text{SQP}})$ estimated by $\nabla g_i(x)$), analogous to an equality constraint “restoration” phase, aiming at better approximating the solution x^* by more closely approaching the boundaries of the constraints thought to be active at the solution. With a strict complementarity assumption in mind, we estimate the set $I(x^*)$ by

$$I = \{i \text{ s.t. } g_i(x) \geq -\lambda_i\}.$$

It turns out that, if d is close enough to d_{SQP} , since the penalty function used in [9] is exactly $f(x)$ when x is in the feasible set, a correction \tilde{d} obtained with d substituted in (1.7) does result in $f(x + d + \tilde{d})$ being sufficiently lower than $f(x)$, close to x^* . As \tilde{d} is a second order correction on d , the condition is that $d - d_{\text{SQP}}$ be of order higher than 2, which can be ensured by selecting $\nu > 2$ in (1.6). A remaining difficulty is that $x + d + \tilde{d}$ may fail to be in the feasible set, resulting in the unit step being rejected due to infeasibility. To prevent such an occurrence, a further “bending” of the search direction will be used here (see [13]), i.e., the correction \tilde{d} will be obtained by solving

$$\min \frac{1}{2} \|\tilde{d}\|_H^2 \quad \text{s.t. } g_i(x + d) + \langle \nabla g_i(x), \tilde{d} \rangle = -\psi \quad \forall i \in I$$

where the norm $\|\tilde{d}\|_H = \langle \tilde{d}, H\tilde{d} \rangle^{1/2}$ is now used for the sake of uniformity. The positive quantity ψ has to be carefully chosen, as excessive bending would jeopardize the descent property on f we just obtained. Also, if \tilde{d} is too large, even the unit stepsize may not yield superlinear convergence.

A careful implementation of the ideas just put forth does indeed allow achievement of the properties claimed. The proposed algorithm is precisely stated in § 2. In § 3, it is shown that, under mild assumptions, this algorithm is convergent irrespective of the initial guess. Rate of convergence analysis is the object of § 4, where conditions for superlinear convergence are given. In § 5, implementation issues are discussed. In particular, it is shown that computational requirements can be reduced, in the worst case, to little more than two $m \times m$ matrix decompositions per iteration (in addition to function evaluations).

¹ A similar idea had previously been used by Maratos [8], McCormick [11], and Gabay and Luenberger [3], among others.

Throughout the paper, we will denote the feasible set by

$$X = \{x \in \mathbb{R}^n \text{ s.t. } g_i(x) \leq 0, \ i = 1, \dots, m\}$$

and the strictly feasible set by

$$X_0 = \{x \in \mathbb{R}^n \text{ s.t. } g_i(x) < 0, \ i = 1, \dots, m\}.$$

The set of active indices at a point x is defined as

$$I(x) = \{i \in \{1, \dots, m\} \text{ s.t. } g_i(x) = 0\}.$$

The Euclidean norm of any vector v will be denoted by $\|v\|$, the corresponding induced norm of a matrix M by $\|M\|$ and the cardinality of any finite set I by $|I|$, and the empty set will be written as \emptyset . A point x is said to be *stationary* for (1.1) if it is feasible and satisfies the equalities

$$\nabla f(x) + \sum_{i=1}^m \lambda_i \nabla g_i(x) = 0,$$

$$\lambda_i g_i(x) = 0, \quad i = 1, \dots, m$$

for some multipliers $\lambda_i, i = 1, \dots, m$. If, moreover, the multipliers are all nonnegative, the point x is said to be a *Kuhn-Tucker point* associated with problem (1.1).

2. The algorithm. The following algorithm is proposed for solving problem (1.1).

ALGORITHM A.

Parameters. $\alpha \in (0, \frac{1}{2})$, $\beta \in (0, 1)$, $\theta \in (0, 1)$, $\nu > 2$, $\tau \in (2, 3)$, $\kappa \in (0, 1)$, $\bar{\mu} > 0$.

Data. $x_0 \in X_0$; $H_0 \in \mathbb{R}^{n \times n}$, symmetric positive definite matrix; $\mu_{0,i}$ scalars satisfying $0 < \mu_{0,i} \leq \bar{\mu}, i = 1, \dots, m$.

Step 0. Initialization. $k = 0$.

Step 1. Computation of a search direction.

(i) Compute d_k^0 and λ_k^0 solving the linear system in (d, λ)

$$(2.1) \quad (\text{LS1}) \quad \begin{cases} H_k d + \nabla f(x_k) + \sum_{i=1}^m \lambda_i \nabla g_i(x_k) = 0, \\ \mu_{k,i} \langle \nabla g_i(x_k), d \rangle + \lambda_i g_i(x_k) = 0, \quad i = 1, \dots, m. \end{cases}$$

$$(2.2) \quad \begin{cases} \mu_{k,i} \langle \nabla g_i(x_k), d \rangle + \lambda_i g_i(x_k) = 0, \quad i = 1, \dots, m. \end{cases}$$

If $d_k^0 = 0$, stop.

(ii) Compute d_k^1 and λ_k^1 solving the linear system in (d, λ)

$$(2.3) \quad (\text{LS2}) \quad \begin{cases} H_k d + \nabla f(x_k) + \sum_{i=1}^m \lambda_i \nabla g_i(x_k) = 0, \\ \mu_{k,i} \langle \nabla g_i(x_k), d \rangle + \lambda_i g_i(x_k) = -\mu_{k,i} \|d_k^0\|^\nu, \quad i = 1, \dots, m. \end{cases}$$

$$(2.4) \quad \begin{cases} \mu_{k,i} \langle \nabla g_i(x_k), d \rangle + \lambda_i g_i(x_k) = -\mu_{k,i} \|d_k^0\|^\nu, \quad i = 1, \dots, m. \end{cases}$$

(iii) Compute the search direction d_k and the approximate multiplier vector λ_k according to

$$(2.5) \quad d_k = (1 - \rho_k) d_k^0 + \rho_k d_k^1, \quad \lambda_k = (1 - \rho_k) \lambda_k^0 + \rho_k \lambda_k^1$$

where

$$(2.6) \quad \rho_k = \begin{cases} 1 & \text{if } \langle \nabla f(x_k), d_k^1 \rangle \leq \theta \langle \nabla f(x_k), d_k^0 \rangle, \\ (1 - \theta) \frac{\langle \nabla f(x_k), d_k^0 \rangle}{\langle \nabla f(x_k), d_k^0 - d_k^1 \rangle} & \text{otherwise.} \end{cases}$$

(iv) Compute the set of indices of “active” constraints

$$(2.7) \quad I_k = \{i \text{ s.t. } g_i(x_k) \geq -\lambda_{k,i}\}$$

and the set of indices of constraints with multipliers of “wrong sign”

$$(2.8) \quad J_k = \{i \text{ s.t. } \lambda_{k,i} \leq g_i(x_k)\}.$$

(v) If $J_k = \emptyset$, compute a correction \tilde{d}_k , solution of the linear least squares problem in d

$$(LS3) \quad \min_{\frac{1}{2} \|d\|_{H_k}^2} \quad \text{s.t. } g_i(x_k + d_k) + \langle \nabla g_i(x_k), d \rangle = -\psi_k \quad \forall i \in I_k$$

where

$$(2.9) \quad \psi_k = \max \left\{ \|d_k\|^\tau, \max_{i \in I_k} \left| \frac{\mu_{k,i}}{\lambda_{k,i}} - 1 \right|^\kappa \|d_k\|^2 \right\}.$$

If $J_k \neq \emptyset$ or if (LS3) has no solution or if $\|\tilde{d}_k\| > \|d_k\|$, set $\tilde{d}_k = 0$.

Step 2. Line search. Compute t_k , the first number t of the sequence $\{1, \beta, \beta^2, \dots\}$ satisfying

$$(2.10) \quad f(x_k + td_k + t^2 \tilde{d}_k) \leq f(x_k) + \alpha t \langle \nabla f(x_k), d_k \rangle,$$

$$(2.11) \quad g_i(x_k + td_k + t^2 \tilde{d}_k) \leq g_i(x_k), \quad i \in J_k,$$

$$(2.12) \quad g_i(x_k + td_k + t^2 \tilde{d}_k) < 0, \quad i \notin J_k.$$

Step 3. Updates. Compute a new symmetric definite positive approximation H_{k+1} to the Hessian matrix. If $J_k \neq \emptyset$, set

$$(2.13a) \quad \mu_{k+1,i} = \mu_{0,i}, \quad i = 1, \dots, m;$$

otherwise, set

$$(2.13b) \quad \mu_{k+1,i} = \min \{ \max \{ \lambda_{k,i}, \|d_k\| \}, \bar{\mu} \}, \quad i = 1, \dots, m.$$

Set $x_{k+1} = x_k + t_k d_k + t_k^2 \tilde{d}_k$. Set $k = k + 1$. Go back to Step 1. \square

Condition (2.11) in the line search has not been discussed so far. By forcing a decrease of the constraints with multiplier of “wrong” sign, it ensures that stationary points that are not Kuhn–Tucker points will be avoided. This will always be possible since, as is easily checked, d_k is a descent direction for these constraints.

3. Global convergence. In this section, it is first shown that Algorithm A is well defined, i.e., that (LS1) and (LS2) always have a unique solution and that the line search (Step 2) completes successfully. This is the object of Lemmas 3.1 and 3.2 and of Proposition 3.3. The case when the algorithm stops after finitely many iterations is then considered in Proposition 3.4. Finally, Lemmas 3.5–3.9, Proposition 3.10 and Theorem 3.11 constitute the actual convergence analysis.

The following is assumed throughout the paper.

Assumptions.

A1. The strictly feasible set X_0 is nonempty.

A2. The objective f is continuously differentiable.

A3. The constraints g_i , $i = 1, \dots, m$ are continuously differentiable.

A4. The set $X \cap \{x \text{ such that } f(x) \leq f(x_0)\}$ is compact.

A5. For all $x \in X$, the vectors $\nabla g_i(x)$, $i \in I(x)$ are linearly independent.

A6. There exist $\sigma_1, \sigma_2 > 0$ such that $\sigma_1 \|d\|^2 \leq \langle d, H_k d \rangle \leq \sigma_2 \|d\|^2$, for all k , for all $d \in \mathbb{R}^n$.

Our first task is to show that the algorithm is well defined. That (LS1) and (LS2) have a unique solution follows from Lemma 3.1. This lemma will be of further help later on, in the convergence analysis.

LEMMA 3.1. *Given any vector $x \in X$, any positive definite matrix $H \in \mathbb{R}^{n \times n}$ and any nonnegative vector $\mu \in \mathbb{R}^m$ such that $\mu_i > 0$ if $g_i(x) = 0$, the matrix $F(x, H, \mu)$ defined by*

$$F(x, H, \mu) = \begin{pmatrix} H & \nabla g_1(x) & \cdots & \nabla g_m(x) \\ \mu_1 \nabla g_1(x)^T & g_1(x) & & \\ \vdots & & \ddots & \\ \mu_m \nabla g_m(x)^T & 0 & & g_m(x) \end{pmatrix}$$

is nonsingular.

Proof. It is enough to show that the only solution (d, λ) of the homogeneous system

$$(3.1) \quad Hd + \sum_{i=1}^m \lambda_i \nabla g_i(x) = 0,$$

$$(3.2) \quad \mu_i \langle \nabla g_i(x), d \rangle + \lambda_i g_i(x) = 0, \quad i = 1, \dots, m$$

is $(0, 0)$. Scalar multiplication of both sides of (3.1) by d yields

$$(3.3) \quad \langle d, Hd \rangle + \sum_{i=1}^m \lambda_i \langle \nabla g_i(x), d \rangle = 0.$$

On the other hand, it follows from (3.2) and the assumption on μ that

$$\langle \nabla g_i(x), d \rangle = 0 \quad \forall i \in I(x).$$

Similarly,

$$(3.4) \quad \lambda_i = 0 \quad \forall i \text{ s.t. } \mu_i = 0.$$

Therefore,

$$\sum_{i=1}^m \lambda_i \langle \nabla g_i(x), d \rangle = \sum_{\{i \notin I(x) \text{ s.t. } \mu_i > 0\}} \lambda_i \langle \nabla g_i(x), d \rangle.$$

Again using (3.2), we get

$$(3.5) \quad \sum_{i=1}^m \lambda_i \langle \nabla g_i(x), d \rangle = - \sum_{\{i \notin I(x) \text{ s.t. } \mu_i > 0\}} \frac{\lambda_i^2}{\mu_i} g_i(x).$$

Relationship (3.3) thus yields

$$\langle d, Hd \rangle - \sum_{\{i \notin I(x) \text{ s.t. } \mu_i > 0\}} \frac{\lambda_i^2}{\mu_i} g_i(x) = 0.$$

Since H is positive definite and $g_i(x) \leq 0$, $i = 1, \dots, m$, the last inequality implies $d = 0$ and $\lambda_i = 0$ for all $i \notin I(x)$ such that $\mu_i > 0$. In view of (3.4), Assumption A5 together with (3.1) then implies that $(d, \lambda) = (0, 0)$. \square

Thus, (d_k^0, λ_k^0) and (d_k^1, λ_k^1) are well defined. The descent properties of d_k follow from the next lemma.

LEMMA 3.2. *The search direction d_k satisfies*

$$(3.6) \quad \langle \nabla f(x_k), d_k \rangle \leq \theta \langle \nabla f(x_k), d_k^0 \rangle \leq -\theta \langle d_k^0, H_k d_k^0 \rangle.$$

Moreover,

$$(3.7) \quad \langle \nabla g_i(x_k), d_k \rangle < 0 \quad \forall i \in J_k.$$

Proof. Scalar multiplication of both sides of (2.1) by d_k^0 yields, using (2.2),

$$\langle \nabla f(x_k), d_k^0 \rangle = -\langle d_k^0, H_k d_k^0 \rangle + \sum_{i=1}^m \left(\frac{\lambda_{k,i}^0}{\mu_{k,i}} \right)^2 g_i(x_k).$$

In view of the feasibility of iterate x_k , we have

$$\langle \nabla f(x_k), d_k^0 \rangle \leq -\langle d_k^0, H_k d_k^0 \rangle$$

and (3.6) follows from (2.5) and (2.6). Finally, (2.2), (2.4), and the definition of d_k in (2.5) yield

$$\langle \nabla g_i(x_k), d_k \rangle = -\frac{\lambda_{k,i}}{\mu_{k,i}} g_i(x_k) - \rho_k \|d_k^0\|^\nu, \quad i = 1, \dots, m$$

so that, in view of (2.8),

$$\langle \nabla g_i(x_k), d_k \rangle < -\frac{1}{\mu_{k,i}} (g_i(x_k))^2 - \rho_k \|d_k^0\|^\nu < 0 \quad \forall i \in J_k. \quad \square$$

These descent properties are the key to proving that the line search is always completed, and thus that the algorithm is well defined.

PROPOSITION 3.3. *The algorithm is well defined.*

Proof. We only need to show that the line search yields a step $t_k = \beta^j$ for some finite $j = j(k)$. Since $g(x_k) < 0$, in view of regularity Assumption A3, it is enough to show that (2.10) and (2.11) hold for t small enough. The Mean Value Theorem yields

$$f(x_k + td_k + t^2 \tilde{d}_k) = f(x_k) + t \langle \nabla f(x_k + \xi d_k + \xi^2 \tilde{d}_k), d_k + 2\xi \tilde{d}_k \rangle$$

for some $\xi \in [0, t]$. Since f is continuously differentiable, $\langle \nabla f(x_k), d_k \rangle < 0$ (see (3.6) and Assumption A6) and $\alpha < 1$, there exists $t_f > 0$ such that

$$f(x_k + td_k + t^2 \tilde{d}_k) \leq f(x_k) + \alpha t \langle \nabla f(x_k), d_k \rangle \quad \forall t \in [0, t_f].$$

For $i \in \{1, \dots, m\}$ it also holds that

$$g_i(x_k + td_k + t^2 \tilde{d}_k) = g_i(x_k) + t \langle \nabla g_i(x_k + \xi_i d_k + \xi_i^2 \tilde{d}_k), d_k + 2\xi_i \tilde{d}_k \rangle$$

for some $\xi_i \in [0, t]$. If i belongs to J_k , in view of (3.7) and of the fact that g_i is continuously differentiable, there exists $t_i > 0$ such that

$$g_i(x_k + td_k + t^2 \tilde{d}_k) \leq g_i(x_k) \quad \forall t \in [0, t_i].$$

Define $t = \min \{t_f, t_i, i \in J_k\}$. The line search conditions (2.10) and (2.11) are satisfied for all t in $[0, t]$. \square

The particular case when the algorithm stops at an iterate x_k at Step 1(i) is considered next.

PROPOSITION 3.4. *If the algorithm stops at a point x_k with $d_k^0 = 0$, then $\nabla f(x_k) = 0$.*

Proof. If the solution d_k^0 of (LS1) is zero, from (2.2) and the strict feasibility of the iterate x_k , it follows that $\lambda_{k,i}^0 = 0, i = 1, \dots, m$, so that, in view of (2.1), $\nabla f(x_k) = 0$. \square

Thus, a point at which the algorithm stops is a particular Kuhn–Tucker point. From now on, it is assumed that the algorithm never stops.

The next two lemmas, direct consequences of Lemma 3.1, give asymptotic properties of the solutions of (LS1) and (LS2) as $\{x_k\}$ approaches a limit x^* , not necessarily a stationary point. Lemma 3.5 asserts that d_k is a good approximation to d_k^0 whenever d_k^0 is small, provided a certain condition, related to strict complementarity, holds at x^* . Lemma 3.6 establishes that, under the same assumptions, the solutions of both problems are well behaved as x_k approaches x^* . These lemmas will be of use in proving global convergence of the algorithm as well as in proving superlinear convergence in § 4.

LEMMA 3.5. *Let $K \subset \mathbb{N}$ be a subset of indices such that, for some x^* and μ^* ,*

$$\{x_k\} \xrightarrow[k \rightarrow \infty]{k \in K} x^*, \quad \{\mu_k\} \xrightarrow[k \rightarrow \infty]{k \in K} \mu^*.$$

Suppose moreover that $\mu_i^ > 0$ if $g_i(x^*) = 0$. Then there exists C such that for all $k \in K$,*

$$\|d_k - d_k^0\| \leq C \|d_k^0\|^\nu.$$

Proof. Let $C = \bar{\mu} \sup \{\|F(x_k, H_k, \mu_k)^{-1}\| \text{ such that } k \in K\}$, where $F(x_k, H_k, \mu_k)$ is as in Lemma 3.1. Under the present assumptions and in view of Assumption A6 and Lemma 3.1, C is finite. We now define $\Delta\lambda_k = \lambda_k - \lambda_k^0$ and $\Delta d_k = d_k - d_k^0$. The vector $(\Delta d_k, \Delta\lambda_k)$ is the solution of

$$F(x_k, H_k, \mu_k) \begin{pmatrix} \Delta d_k \\ \Delta\lambda_k \end{pmatrix} = \begin{pmatrix} 0 \\ -\rho_k \mu_k \|d_k^0\|^\nu \end{pmatrix}.$$

Since the coefficients ρ_k are bounded from above by 1 it follows that

$$(\|\Delta d_k\|^2 + \|\Delta\lambda_k\|^2)^{1/2} \leq C \|d_k^0\|^\nu$$

so that

$$\|\Delta d_k\| \leq C \|d_k^0\|^\nu. \quad \square$$

LEMMA 3.6. *Let $K \subset \mathbb{N}$ be a subset of indices such that, for some x^* , H^* , ρ^* and μ^* ,*

$$(3.8) \quad \{x_k\} \xrightarrow[k \rightarrow \infty]{k \in K} x^*,$$

$$(3.9) \quad \{H_k\} \xrightarrow[k \rightarrow \infty]{k \in K} H^*,$$

$$(3.10) \quad \{\rho_k\} \xrightarrow[k \rightarrow \infty]{k \in K} \rho^*,$$

$$(3.11) \quad \{\mu_k\} \xrightarrow[k \rightarrow \infty]{k \in K} \mu^*.$$

Suppose moreover that $\mu_i^ > 0$ for all i such that $g_i(x^*) = 0$. Then*

(i) $\{d_k^0\} \rightarrow_{k \in K, k \rightarrow \infty} d^{0*}$ and $\{\lambda_k^0\} \rightarrow_{k \in K, k \rightarrow \infty} \lambda^{0*}$ where (d^{0*}, λ^{0*}) is the solution of the nonsingular linear system in (d, λ)

$$H^* d + \nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla g_i(x^*) = 0,$$

$$\mu_i^* \langle \nabla g_i(x^*), d \rangle + \lambda_i g_i(x^*) = 0, \quad i = 1, \dots, m.$$

(ii) $\{d_k^1\} \rightarrow_{k \in K, k \rightarrow \infty} d^{1*}, \{d_k\} \rightarrow_{k \in K, k \rightarrow \infty} d^*, \{\lambda_k^1\} \rightarrow_{k \in K, k \rightarrow \infty} \lambda^{1*}$,
and $\{\lambda_k\} \rightarrow_{k \in K, k \rightarrow \infty} \lambda^*$, where (d^{1*}, λ^{1*}) solves the linear system in (d, λ) ,

$$H^*d + \nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla g_i(x^*) = 0,$$

$$\mu_i^* \langle \nabla g_i(x^*), d \rangle + \lambda_i g_i(x^*) = -\mu_i^* \|d^{0*}\|^v, \quad i = 1, \dots, m,$$

and $d^* = (1 - \rho^*)d^{0*} + \rho^*d^{1*}$, $\lambda^* = (1 - \rho^*)\lambda^{0*} + \rho^*\lambda^{1*}$.

(iii) $d^{0*} = 0$ if and only if $d^* = 0$.

Proof. The first assertion follows from Lemma 3.1 and the Implicit Function Theorem. The second assertion follows from Lemma 3.1, the Implicit Function Theorem, and the definition of d_k in (2.5) and (2.6). The “only if” part of the third assertion follows from Lemma 3.5. Finally, by continuity, (3.6) implies that

$$\langle \nabla f(x^*), d^* \rangle \leq -\theta \langle d^{0*}, H^*d^{0*} \rangle$$

so that, in view of Assumption A6, the “if” part of the third assertion also holds. \square

In the balance of this section, it is shown that any accumulation point of the sequence $\{x_k\}$ constructed by Algorithm A is a Kuhn–Tucker point. This result follows easily under the assumption that the directions d_k tend to zero and that the sets J_k of indices of constraints with multiplier of wrong sign are empty, even when the assumptions of Lemma 3.6 are not satisfied. This is the object of the next lemma.

LEMMA 3.7. *Let x^* be an accumulation point of the sequence generated by Algorithm A and suppose that $\{x_k\} \rightarrow_{k \in K, k \rightarrow \infty} x^*$. If $\{d_k\} \rightarrow_{k \in K, k \rightarrow \infty} 0$, then x^* is a stationary point and $\{\lambda_k\} \rightarrow_{k \in K, k \rightarrow \infty} \lambda^*$ where λ^* is the multiplier vector associated with x^* . If moreover $J_k = \emptyset$ for all $k \in K$ then x^* is a Kuhn–Tucker point. If in addition $\lambda_i^* \leq \bar{\mu}$ for all i , then $\{\mu_{k+1}\} \rightarrow_{k \in K, k \rightarrow \infty} \lambda^*$.*

Proof. In view of (3.6) and Assumption A6, if a subsequence of directions $\{d_k\}_{k \in K}$ converges to zero, the corresponding subsequence $\{d_k^0\}_{k \in K}$ must also converge to zero. Now, the quantities λ_k and d_k satisfy

$$(3.12) \quad H_k d_k + \nabla f(x_k) + \sum_{i=1}^m \lambda_{k,i} \nabla g_i(x_k) = 0,$$

$$(3.13) \quad \mu_{k,i} \langle \nabla g_i(x_k), d_k \rangle + \lambda_{k,i} g_i(x_k) = -\rho_k \mu_{k,i} \|d_k^0\|^v, \quad i = 1, \dots, m.$$

Let us show that the coefficients $\lambda_{k,i}$ must be bounded on K for all i . By contradiction, suppose that there exists $K' \subset K$ and i_0 such that

$$|\lambda_{k,i_0}| = \max_i |\lambda_{k,i}| \quad \forall k \in K' \quad \text{and} \quad |\lambda_{k,i_0}| \xrightarrow[k \rightarrow \infty]{k \in K'} \infty.$$

Dividing (3.12) by $|\lambda_{k,i_0}|$ gives

$$(3.14) \quad \frac{1}{|\lambda_{k,i_0}|} H_k d_k + \frac{1}{|\lambda_{k,i_0}|} \nabla f(x_k) + \sum_{i=1}^m \frac{\lambda_{k,i}}{|\lambda_{k,i_0}|} \nabla g_i(x_k) = 0.$$

The scalars $\hat{\lambda}_{k,i} = \lambda_{k,i}/|\lambda_{k,i_0}|$, $i = 1, \dots, m$, satisfy $-1 \leq \hat{\lambda}_{k,i} \leq 1$, $i = 1, \dots, m$, so that there exists $K'' \subset K'$ such that $\{\hat{\lambda}_{k,i}\} \rightarrow_{k \in K'', k \rightarrow \infty} \hat{\lambda}_i$ for some $\hat{\lambda}_i$, $i = 1, \dots, m$. Taking the limit on K'' in (3.14), we thus get, in view of Assumptions A2, A3, A6, and the fact that $\{d_k\}$ is bounded on K ,

$$(3.15) \quad \sum_{i=1}^m \hat{\lambda}_i \nabla g_i(x^*) = 0.$$

Now, from (3.13), we have

$$\frac{\mu_{k,i}}{|\lambda_{k,i_0}|} \langle \nabla g_i(x_k), d_k \rangle + \frac{\lambda_{k,i}}{|\lambda_{k,i_0}|} g_i(x_k) = \frac{-\rho_k \mu_{k,i} \|d_k^0\|^\nu}{|\lambda_{k,i_0}|}, \quad i = 1, \dots, m.$$

Taking the limit on K'' in the last equality, we obtain

$$\hat{\lambda}_i g_i(x^*) = 0, \quad i = 1, \dots, m,$$

i.e., $\hat{\lambda}_i = 0$ for all i such that $g_i(x^*) < 0$. Therefore, (3.15) becomes

$$\sum_{i \in I(x^*)} \hat{\lambda}_i \nabla g_i(x^*) = 0.$$

In view of Assumption A5, this implies that $\hat{\lambda}_i = 0$, $i = 1, \dots, m$, which is a contradiction since $\hat{\lambda}_{i_0} = \lim_{k \in K'', k \rightarrow \infty} \lambda_{k,i_0} / |\lambda_{k,i_0}| \in \{-1, 1\}$. Thus, the coefficients $\lambda_{k,i}$ are bounded on K . Let us now consider an accumulation point $\hat{\lambda}$ of $\{\lambda_k\}_{k \in K}$ and $K' \subset K$ such that $\{\lambda_k\} \rightarrow_{k \in K', k \rightarrow \infty} \hat{\lambda}$. Taking the limit on K' in (3.12) and (3.13) yields, in view of Assumptions A2, A3 and A6, since $\{d_k^0\} \rightarrow_{k \in K, k \rightarrow \infty} 0$ and $\{d_k\} \rightarrow_{k \in K, k \rightarrow \infty} 0$,

$$\nabla f(x^*) + \sum_{i=1}^m \hat{\lambda}_i \nabla g_i(x^*) = 0,$$

$$\hat{\lambda}_i g_i(x^*) = 0, \quad i = 1, \dots, m.$$

Since x^* is feasible, as the limit of a sequence of feasible points, it follows that it is a stationary point. Also, in view of Assumption A5, the accumulation point $\hat{\lambda}$ is equal to the unique multiplier vector λ^* at x^* . If the sets J_k are empty on K , i.e., if

$$\lambda_{k,i} > g_i(x_k), \quad i = 1, \dots, m \quad \forall k \in K,$$

it follows by continuity that

$$\lambda_i^* \geq g_i(x^*), \quad i = 1, \dots, m$$

so that the multipliers associated with active constraints are nonnegative, and x^* is a Kuhn-Tucker point. Finally, the last result comes directly from the definition of μ_{k+1} in (2.13). \square

For the overall proof, assuming that the sequence $\{x_k\}$ has an accumulation point x^* corresponding to some $K \subset \mathbb{N}$, two cases will be considered, according to whether or not the assumptions in Lemma 3.7 are satisfied at the *previous* iteration, i.e., whether or not $\{d_{k-1}\} \rightarrow_{k \in K, k \rightarrow \infty} 0$ and $J_{k-1} = \emptyset$ for all $k \in K$. If this condition is satisfied, it is a simple consequence of Lemma 3.7 that x^* is still a Kuhn-Tucker point (Lemma 3.7 asserts that limit points of $\{x_{k-1}\}$ on K are such). This is established in Lemma 3.8. If the condition is not satisfied, in view of the update formula (2.13) it can be assumed, without loss of generality, that the components of μ_k are bounded away from zero on K , so Lemma 3.6 can be used. It will be shown that, in such case, $\{d_k\}$ must tend to zero on K (Lemma 3.9) so that, in view of Lemma 3.7, x^* must be stationary (Proposition 3.10). Under an additional mild assumption, Theorem 3.11 will establish convergence to Kuhn-Tucker points.

LEMMA 3.8. *Let x^* be an accumulation point of the sequence $\{x_k\}$ generated by Algorithm A and let $K \subset \mathbb{N}$ be such that $\{x_k\} \rightarrow_{k \in K, k \rightarrow \infty} x^*$. If*

$$(3.16) \quad \{d_{k-1}\} \xrightarrow[k \in K, k \rightarrow \infty]{} 0$$

and

$$(3.17) \quad J_{k-1} = \emptyset \quad \forall k \in K,$$

then x^* is a Kuhn-Tucker point.

Proof. Reducing K , if necessary, to a subset of indices we may assume, in view of Assumption A4, that the sequence $\{x_{k-1}\}$ converges on K to some vector x^{**} . In view of Lemma 3.7 and relationships (3.16) and (3.17), x^{**} is a Kuhn–Tucker point. We show that $x^* = x^{**}$. For $k \in K$, we have

$$x_k = x_{k-1} + t_{k-1}d_{k-1} + t_{k-1}^2\tilde{d}_{k-1}.$$

Therefore,

$$\|x_k - x_{k-1}\| \leq t_{k-1}\|d_{k-1}\| + t_{k-1}^2\|\tilde{d}_{k-1}\| \leq 2\|d_{k-1}\|.$$

Since $\{d_{k-1}\}$ tends to zero on K , this implies $\{\|x_k - x_{k-1}\|\} \rightarrow_{k \in K, k \rightarrow \infty} 0$, thus $x^* = x^{**}$. \square

LEMMA 3.9. *Let $K \subset \mathbb{N}$ be a subset of indices such that either $\inf_{k \in K} \|d_{k-1}\| > 0$ or $J_{k-1} \neq \emptyset$ for all $k \in K$. Then $\{d_k\} \rightarrow_{k \in K, k \rightarrow \infty} 0$.*

Proof. Suppose by contradiction that there exists a subset K' of indices such that $\inf_{k \in K'} \|d_k\| > 0$. In view of Assumptions A4 and A6 and the definitions of the quantities ρ_k and μ_k in (2.6) and (2.13) respectively, it can be assumed, without loss of generality, that (3.8)–(3.11) hold on K' for some x^* , H^* , ρ^* and μ^* . Also, from the definition of the multiplier vector estimate (2.13), it follows, under the present assumptions, that all the components of μ^* are strictly positive, so that, from Lemma 3.6(iii), the sequence $\{\|d_k^0\|\}$, $k \in K'$, is bounded away from zero. Let $\underline{d} > 0$ be a lower bound on $\{\|d_k^0\|\}$, $k \in K'$.

We first show that there exists $\underline{t} > 0$ such that, for all $t \in [0, \underline{t}]$ and $k \in K'$ large enough, conditions (2.10), (2.11), and (2.12) of the line search are satisfied. It follows from (3.6) and Assumption A6 that, for $k \in K'$,

$$(3.18) \quad \langle \nabla f(x_k), d_k \rangle \leq -\theta\sigma_1 \underline{d}^2.$$

Since, using the Mean Value Theorem, we can write

$$f(x_k + td_k + t^2\tilde{d}_k) = f(x_k) + \int_0^1 \langle \nabla f(x_k + t\xi d_k + t^2\xi^2\tilde{d}_k), td_k + 2t^2\xi\tilde{d}_k \rangle d\xi,$$

it follows that, for $k \in K'$ large enough,

$$\begin{aligned} & f(x_k + td_k + t^2\tilde{d}_k) - f(x_k) - \alpha t \langle \nabla f(x_k), d_k \rangle \\ & \leq t \left\{ \int_0^1 [\langle \nabla f(x_k + t\xi d_k + t^2\xi^2\tilde{d}_k), d_k + 2t\xi\tilde{d}_k \rangle - \langle \nabla f(x_k), d_k \rangle] d\xi - (1 - \alpha)\theta\sigma_1 \underline{d}^2 \right\} \\ & \leq t \left\{ \sup_{\xi \in [0,1]} \|\nabla f(x_k + t\xi d_k + t^2\xi^2\tilde{d}_k) - \nabla f(x_k)\| \|d_k\| \right. \\ & \quad \left. + 2t \sup_{\xi \in [0,1]} \|\nabla f(x_k + t\xi d_k + t^2\xi^2\tilde{d}_k)\| \|\tilde{d}_k\| - (1 - \alpha)\theta\sigma_1 \underline{d}^2 \right\}. \end{aligned}$$

Since, from Lemma 3.6, d_k and \tilde{d}_k are bounded on K' (by definition $\|\tilde{d}_k\| \leq \|d_k\|$) and since f is continuously differentiable, this ensures that there exists $\underline{t}_f > 0$ independent of k , such that, for $k \in K'$ large enough,

$$(3.19) \quad f(x_k + td_k + t^2\tilde{d}_k) - f(x_k) - \alpha t \langle \nabla f(x_k), d_k \rangle \leq 0 \quad \forall t \in [0, \underline{t}_f].$$

Similarly, for $i \in \{1, \dots, m\}$, $k \in K'$ large enough, we have

$$g_i(x_k + td_k + t^2 \tilde{d}_k) \leq g_i(x_k) + t \left\{ \sup_{\xi \in [0,1]} \|\nabla g_i(x_k + t\xi d_k + t^2 \xi^2 \tilde{d}_k) - \nabla g_i(x_k)\| \|d_k\| \right. \\ \left. + 2t \sup_{\xi \in [0,1]} \|\nabla g_i(x_k + t\xi d_k + t^2 \xi^2 \tilde{d}_k)\| \|\tilde{d}_k\| + \langle \nabla g_i(x_k), d_k \rangle \right\}.$$

Let us define

$$u_{k,i}(t) = \sup_{\xi \in [0,1]} \|\nabla g_i(x_k + t\xi d_k + t^2 \xi^2 \tilde{d}_k) - \nabla g_i(x_k)\| \|d_k\| \\ + 2t \sup_{\xi \in [0,1]} \|\nabla g_i(x_k + t\xi d_k + t^2 \xi^2 \tilde{d}_k)\| \|\tilde{d}_k\|, \quad i = 1, \dots, m.$$

In view of (3.19), our first claim will hold provided we can find a uniform range of t 's for which we have

$$(3.20) \quad \gamma_{k,i} g_i(x_k) + t \{u_{k,i}(t) + \langle \nabla g_i(x_k), d_k \rangle\} < 0, \quad i = 1, \dots, m$$

with

$$\gamma_{k,i} = \begin{cases} 0 & \text{if } i \in J_k, \\ 1 & \text{if } i \notin J_k. \end{cases}$$

Using (2.2), (2.4), and (2.5), we can write

$$(3.21) \quad \langle \nabla g_i(x_k), d_k \rangle = -\frac{\lambda_{k,i}}{\mu_{k,i}} g_i(x_k) - \rho_k \|d_k^0\|^\nu, \quad i = 1, \dots, m.$$

Since, in view of (3.6) and Assumption A6, d_k^0 satisfies $\langle \nabla f(x_k), d_k^0 \rangle \leq -\sigma_1 d^2$, it follows from the boundedness of the sequences $\{x_k\}$, $\{d_k^0\}$ and $\{d_k^1\}$ on K' , from Assumption A2 and from definition (2.6) of ρ_k that there exists $\underline{\rho} > 0$ such that $\rho_k \geq \underline{\rho}$ for all $k \in K'$. Relationship (3.21) thus yields, for $k \in K'$,

$$\langle \nabla g_i(x_k), d_k \rangle \leq -\frac{\lambda_{k,i}}{\mu_{k,i}} g_i(x_k) - \underline{\rho} d^\nu, \quad i = 1, \dots, m.$$

In order for (3.20) to be satisfied, it is thus sufficient to have, for $k \in K'$ large enough,

$$(3.22) \quad \gamma_{k,i} g_i(x_k) + t \left\{ u_{k,i}(t) - \frac{\lambda_{k,i}}{\mu_{k,i}} g_i(x_k) - \underline{\rho} d^\nu \right\} < 0, \quad i = 1, \dots, m.$$

Let λ^* be as defined in Lemma 3.6. For i such that $\lambda_i^* = 0$, in view of the definition of $u_{k,i}(t)$, the boundedness of $\{d_k\}$ and $\{\tilde{d}_k\}$ on K' , the convergence of $\{\mu_k\}$ to a vector with positive components on K' , and the fact that g_i is continuously differentiable, there exists $\underline{t}_i > 0$, independent of k , such that, for $t \in [0, \underline{t}_i]$, $k \in K'$ large enough,

$$u_{k,i}(t) - \frac{\lambda_{k,i}}{\mu_{k,i}} g_i(x_k) - \underline{\rho} d^\nu < 0,$$

thus (3.22) holds. Now, let i be such that $\lambda_i^* \neq 0$. In order for (3.22) to be satisfied, it is sufficient to have,

$$(3.23) \quad u_{k,i}(t) - \underline{\rho} d^\nu < 0$$

and

$$(3.24) \quad \gamma_{k,i} - t \frac{\lambda_{k,i}}{\mu_{k,i}} \geq 0.$$

Inequality (3.23) will obviously hold for all $t \geq 0$ small enough and $k \in K'$ large enough, and similarly for (3.24) if $\lambda_i^* < 0$. Suppose $\lambda_i^* > 0$. For $k \in K'$ large enough, $\lambda_{k,i} > 0$

and, in view of the strict feasibility of the iterates, $i \notin J_k$. Therefore, $\gamma_{k,i} = 1$ and (3.24) reduces to $t \leq \mu_{k,i}/\lambda_{k,i}$. The latter ratio is bounded away from 0 on K' . Therefore, there exists again a number $\underline{t}_i > 0$, independent of k , such that, for $t \in [0, \underline{t}_i]$, $k \in K'$ large enough, inequality (3.22) is satisfied. Thus the first claim holds.

From the line search rule (2.10) and relationship (3.18), it follows that, for $k \in K'$ large enough,

$$f(x_k + t_k d_k + t_k^2 \tilde{d}_k) \leq f(x_k) - \alpha \underline{t} \theta \sigma_1 \underline{d}^2,$$

which, in view of the monotonic decrease of f on the entire sequence $\{x_k\}$, contradicts the convergence of $\{f(x_k)\}$ on K' . \square

We have thus established the following proposition.

PROPOSITION 3.10. *Let x^* be an accumulation point of the sequence generated by Algorithm A. Then, x^* is a stationary point.* \square

To conclude this section we show that, under an additional mild assumption, accumulation points of the sequence $\{x_k\}$ are Kuhn–Tucker points. The proof relies on condition (2.11) of the line search, which forces decrease of the constraints with multiplier of “wrong sign.”

A7. The number of stationary points is finite.

THEOREM 3.11. *All the accumulation points of the sequence $\{x_k\}$ generated by Algorithm A are Kuhn–Tucker points.*

Proof. From Proposition 3.10, every accumulation point of the sequence generated by the algorithm is a stationary point. Thus, in view of Assumption A7, there exists $\varepsilon > 0$ such that every accumulation point is unique in a ball of radius ε about it. Consider an accumulation point x^* and, proceeding by contradiction, suppose that x^* is not a Kuhn–Tucker point.

We first show that the entire sequence converges to x^* . Let us suppose, by contradiction, that, infinitely many times, the sequence leaves the ε -ball about x^* at some points $\{x_k\} \rightarrow_{k \in K, k \rightarrow \infty} x^*$, jumping into the neighborhood of another stationary point. A contradiction argument based on Lemmas 3.8 and 3.9 shows that $\{d_k\} \rightarrow_{k \in K, k \rightarrow \infty} 0$. Then, the subsequence cannot possibly jump outside of the ε -ball about x^* at $k \in K$ since $\|\tilde{d}_k\| \leq \|d_k\|$ and $t_k \leq 1$ and the new iterate is defined by $x_{k+1} = x_k + t_k d_k + t_k^2 \tilde{d}_k$. This contradiction establishes that $\{x_k\} \rightarrow_{k \rightarrow \infty} x^*$. We then suppose now that x^* is a stationary point, but not a Kuhn–Tucker point, and we show that this leads to a contradiction. Let us denote by λ^* the multiplier vector associated with x^* (which is unique in view of Assumption A5). From the stationary conditions, it holds, $\lambda_i^* g_i(x^*) = 0$, $i = 1, \dots, m$. Therefore $\lambda_i^* = 0$ if $g_i(x^*) < 0$, $i = 1, \dots, m$. The fact that x^* is not a Kuhn–Tucker point thus means that there exists an index $i_0 \in \{1, \dots, m\}$ such that $g_{i_0}(x^*) = 0$ and $\lambda_{i_0}^* < 0$. Again, a contradiction argument based on Lemmas 3.8 and 3.9 shows that the sequence $\{d_k\}$ converges to zero so that, from Lemma 3.7, $\{\lambda_k\}$ converges to λ^* . It follows that, for k large enough, say $k \geq \underline{k}$, $\lambda_{k,i_0} \leq g_{i_0}(x_k)$, i.e., $i_0 \in J_k$. Thus, in view of (2.11),

$$g_{i_0}(x_k) \leq g_{i_0}(x_{k-1}) \leq \dots \leq g_{i_0}(x_{\underline{k}+1}) \leq g_{i_0}(x_{\underline{k}}) < 0$$

for k large enough. This contradicts the fact that $g_{i_0}(x^*) = 0$. \square

4. Rate of convergence. We now strengthen the regularity assumptions on the functions involved. Assumptions A2 and A3 are replaced by

A2'. The objective function f is three times continuously differentiable.

A3'. The constraints g_i , $i = 1, \dots, m$, are three times continuously differentiable. We also make the following additional assumption, which supersedes Assumption A7.

A8. The sequence generated by the algorithm possesses an accumulation point x^* (in view of Theorem 3.11, a Kuhn–Tucker point) at which (i) second-order sufficiency condition and strict complementary slackness hold, i.e., the Kuhn–Tucker multiplier vector $\lambda^* \in \mathbb{R}^m$ (unique in view of Assumption A5) is such that

$$\lambda_i^* > 0 \quad \forall i \in I(x^*)$$

and the Hessian of the Lagrangian function $\nabla_{xx}L(x^*, \lambda^*)$ is positive definite on the subspace $\{h \text{ such that } \langle \nabla g_i(x^*), h \rangle = 0 \text{ for all } i \in I(x^*)\}$, and (ii) the multiplier vector λ^* satisfies $\lambda_i^* \leq \bar{\mu}$, $i = 1, \dots, m$.

The first task is to show that, under these additional assumptions, the entire sequence $\{x_k\}$ converges to x^* , the sequences $\{\lambda_k\}$ and $\{\mu_k\}$ converge to λ^* and the sets I_k and J_k eventually become identical to $I(x^*)$ and \emptyset respectively. This is the object of Proposition 4.2 below. It makes use of the following result.

LEMMA 4.1. *If there exists a subset of indices $K \subset \mathbb{N}$ such that $\{x_{k-1}\} \rightarrow_{k \in K, k \rightarrow \infty} x^*$ and $\{x_k\} \rightarrow_{k \in K, k \rightarrow \infty} x^*$, where x^* is as in Assumption A8, then $\{d_k\} \rightarrow_{k \in K, k \rightarrow \infty} 0$.*

Proof. Let $K \subset \mathbb{N}$ be as stated and, proceeding by contradiction, suppose that there exists a subset of indices $K' \subset K$ such that $\inf_{k \in K'} \|d_k\| > 0$. In that case, in view of Lemma 3.9, we may assume, without loss of generality, that $\{d_{k-1}\} \rightarrow_{k \in K', k \rightarrow \infty} 0$ and $J_{k-1} = \emptyset$ for all $k \in K'$. From Lemma 3.7 and Assumption A8(ii), it follows that $\{\mu_k\} \rightarrow_{k \in K', k \rightarrow \infty} \lambda^*$. By possibly reducing K' to a subset of indices, it can also be assumed, in view of Assumption A6 and the definition of ρ_k in (2.6), that the subsequences $\{H_k\}_{k \in K'}$ and $\{\rho_k\}_{k \in K'}$ converge respectively to some matrix H^* and some value ρ^* . Therefore, due to Assumption A8(i), the assumptions of Lemma 3.6 are satisfied, so that the matrix of the limit system in (d, λ)

$$(4.1) \quad H^*d + \nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla g_i(x^*) = 0,$$

$$(4.2) \quad \lambda_i^* \langle \nabla g_i(x^*), d \rangle + \lambda_i g_i(x^*) = 0, \quad i = 1, \dots, m$$

is invertible, and the sequences $\{d_k^0\}$ and $\{d_k\}$ converge on K' to the solution 0 of (4.1) and (4.2), a contradiction. \square

PROPOSITION 4.2. *The entire sequence $\{x_k\}$ converges to the point x^* given by Lemma 4.1. Moreover, it holds, for $k \rightarrow \infty$,*

- (i) $\{d_k\} \rightarrow 0$ and $\{\lambda_k\} \rightarrow \lambda^*$,
- (ii) $J_k = \emptyset$ and $I_k = I(x^*)$,
- (iii) $\{\mu_k\} \rightarrow \lambda^*$.

Proof. Consider a ball of radius $\varepsilon > 0$ about x^* where there is no Kuhn–Tucker point other than x^* (in view of Assumption A8, such an ε exists). Let us suppose by contradiction that the sequence $\{x_k\}$ leaves that ball infinitely often. Consider the subsequence $\{x_k\}_{k \in K}$ of points at which the ball is entered and define the related subsequence $\{x_{l(k)}\}_{k \in K}$ of points at which the sequence is about to exit the ball. We first show that $l(k) = k$ for all $k \in K$. Indeed, if this were not the case, there would exist a subset K' of indices such that $\{x_{l(k)-1}\} \rightarrow_{k \in K', k \rightarrow \infty} x^*$ and $\{x_{l(k)}\} \rightarrow_{k \in K', k \rightarrow \infty} x^*$. In order for the sequence to leave the ball at iteration $l(k)$, since $0 \leq t_{l(k)} \leq 1$ and $\|\tilde{d}_{l(k)}\| \leq \|d_{l(k)}\|$, it is necessary that $\|d_{l(k)}\| \geq \varepsilon/4$ for all $k \in K'$ large enough. But this contradicts Lemma 4.1. Therefore, $l(k) = k$. Let us now show that this leads to a contradiction. In order to enter the ε -ball without creating another accumulation point in that ball we must have $\|d_{k-1}\| > \varepsilon/4$ for all $k \in K$ large enough. But in that case, in view of Lemma 3.9, it follows $\{d_k\} \rightarrow_{k \in K, k \rightarrow \infty} 0$. This contradicts the fact that, for $k \in K$, x_{k+1} is outside the ball. Convergence of the sequence $\{x_k\}$ to x^* is thus established.

Claim (i) then follows by successive applications of Lemmas 4.1 and 3.7. Claim (ii) is a direct consequence of definitions (2.7) and (2.8), and strict complementarity Assumption A8(i). The last result follows from Lemma 3.7, making use of Assumption A8(ii). \square

Since Algorithm A is based on a quasi-Newton iteration with higher-order corrections, we can expect that it will exhibit superlinear convergence if $\{H_k\}$ converges to the Hessian of the Lagrangian at the solution. In the context of SQP type algorithms, it has been shown [16] that two-step superlinear convergence occurs under the weaker condition that the Hessian of the Lagrangian be well approximated along the projection of the search direction in the tangent subspace to the active constraints. This is fortunate since, under Assumption A8, it allows the possibility of keeping H_k positive definite. Since the iteration in Algorithm A is a higher-order modification of the SQP iteration, one can hope that a similar property will hold. We thus make the following assumption.

A9. The sequence of matrices $\{H_k\}$ satisfies

$$\frac{\|P_k(H_k - \nabla_{xx}^2 L(x^*, \lambda^*))P_k d_k\|}{\|d_k\|} \rightarrow 0$$

where

$$P_k = [I - R_k(R_k^T R_k)^{-1} R_k^T]$$

with

$$R_k = [\nabla g_i(x_k) \text{ s.t. } i \in I(x^*)] \in \mathbb{R}^{n \times |I(x^*)|}.$$

It will be shown below (Theorem 4.6) that under this assumption, two-step superlinear convergence is ensured, provided the full quasi-Newton step is taken close to the solution. Proving that the latter will indeed occur is the object of Lemmas 4.3 and 4.4 and of Proposition 4.5. The proof will rely crucially on the properties of the estimate H_k . It turns out that Assumption A9 is exactly what will be needed there.

Under Assumption A9, H_k is a good approximation to the Hessian as far as its action on the component of d_k in the tangent subspace to the active constraints is concerned. The following lemma gives bounds on the component of d_k off this subspace. These bounds will be of use in Proposition 4.5 below.

LEMMA 4.3. *For k large enough, the direction d_k can be decomposed into*

$$d_k = P_k d_k + \hat{d}_k$$

with

$$\|\hat{d}_k\| = O\left(\sum_{i \in I(x^*)} \left(\frac{\lambda_{k,i}}{\mu_{k,i}} g_i(x_k)\right)^2\right)^{1/2} + o(\|d_k^0\|^2)$$

where P_k is as defined in Assumption A9.

Proof. By definition, the direction d_k satisfies, for $i = 1, \dots, m$,

$$\mu_{k,i} \langle \nabla g_i(x_k), d_k \rangle + \lambda_{k,i} g_i(x_k) = -\mu_{k,i} \rho_k \|d_k^0\|^\nu.$$

In particular,

$$R_k^T d_k = h_k$$

where h_k is a $|I(x^*)|$ -vector whose components are the numbers

$$-\frac{\lambda_{k,i}}{\mu_{k,i}} g_i(x_k) - \rho_k \|d_k^0\|^\nu, \quad i \in I(x^*).$$

Therefore, $d_k = P_k d_k + \hat{d}_k$ with

$$\hat{d}_k = R_k (R_k^T R_k)^{-1} h_k$$

so that

$$\|\hat{d}_k\| = O(\|h_k\|). \quad \square$$

The introduction of \tilde{d}_k is essential in order to obtain a unit step close to the solution. However its size, directly related to the size of ψ_k (see (2.9)), must be closely monitored to avoid undesirable effects. Suitable bounds, to be used in Proposition 4.5, are established in the following lemma.

LEMMA 4.4. *For k large enough, the correction \tilde{d}_k is obtained as the solution of (LS3) (this solution exists) and it satisfies*

$$\|\tilde{d}_k\| = O\left(\max\left\{\|d_k\|^2, \max_{i \in I(x^*)} \left|\frac{\mu_{k,i}}{\lambda_{k,i}} - 1\right| \|d_k\|\right\}\right) = o(\|d_k\|).$$

Proof. For k large enough, in view of Proposition 4.2(ii), the correction \tilde{d}_k is first computed by attempting to solve the linear least squares problem in d

$$\min \frac{1}{2} \|d\|_{H_k}^2 \quad \text{s.t. } \langle \nabla g_i(x_k), d \rangle = -\psi_k - g_i(x_k + d_k) \quad \forall i \in I(x^*)$$

or equivalently the linear system in (d, λ)

$$H_k d + \sum_{i \in I(x^*)} \lambda_i \nabla g_i(x_k) = 0,$$

$$\langle \nabla g_i(x_k), d \rangle = -\psi_k - g_i(x_k + d_k) \quad \forall i \in I(x^*).$$

First-order expansion of $g_i(x_k + d_k)$ yields

$$-\psi_k - g_i(x_k + d_k) = -\psi_k - g_i(x_k) - \langle \nabla g_i(x_k), d_k \rangle + O(\|d_k\|^2) \quad \forall i \in I(x^*)$$

and, from (2.2), Proposition 4.2, Assumption A8, and Lemma 3.5, and since $\nu > 2$, we have

$$-\psi_k - g_i(x_k + d_k) = -\psi_k + \left(\frac{\mu_{k,i}}{\lambda_{k,i}} - 1\right) \langle \nabla g_i(x_k), d_k \rangle + O(\|d_k\|^2) \quad \forall i \in I(x^*).$$

Since, from (2.9), the definition of τ ($\tau > 2$) and of κ ($\kappa > 0$), Proposition 4.2, and strict complementarity Assumption A8(i), it follows that $\psi_k = o(\|d_k\|^2)$, the result is a consequence of Lemma 3.1. \square

We are now ready to show that the use of correction \tilde{d}_k on the search direction causes the line search to accept a unit step close to the solution, thus avoiding any Maratos-like effect.

PROPOSITION 4.5. *For k large enough, the step $t_k = 1$ is accepted by the line search.*

Proof. In this proof, we will constantly make use of Lemma 3.5 without explicit mention. The assumptions of Lemma 3.5 are satisfied on the entire set \mathbb{N} of indices (see Proposition 4.2 and strict complementarity Assumption A8(i)). Throughout the proof, the phrase “for k large enough” will be implicit. According to Step 2 of Algorithm A, and in view of Proposition 4.2(ii), two conditions are needed for the line search to yield a unit stepwise, namely strict feasibility of the resulting point (2.12) and sufficient decrease (2.10). Thus, this proof will be in two parts.

(i) We first show that a step of 1 brings a sufficient decrease on f . In view of Lemma 4.4 and Assumption A2', we can write

$$(4.3) \quad f(x_k + d_k + \tilde{d}_k) = f(x_k) + \langle \nabla f(x_k), d_k + \tilde{d}_k \rangle + \frac{1}{2} \langle d_k, \nabla_{xx}^2 f(x_k) d_k \rangle + o(\|d_k\|^2).$$

From the definitions of d_k and \tilde{d}_k and Proposition 4.2(ii) and Lemma 4.4, it follows that

$$(4.4) \quad H_k d_k + \nabla f(x_k) + \sum_{i=1}^m \lambda_{k,i} \nabla g_i(x_k) = 0,$$

$$(4.5) \quad \mu_{k,i} \langle \nabla g_i(x_k), d_k \rangle + \lambda_{k,i} g_i(x_k) = -\mu_{k,i} \rho_k \|d_k^0\|^\nu, \quad i = 1, \dots, m,$$

$$(4.6) \quad g_i(x_k + d_k) + \langle \nabla g_i(x_k), \tilde{d}_k \rangle = -\psi_k, \quad i \in I(x^*).$$

From (4.4), we get

$$(4.7) \quad \langle \nabla f(x_k), d_k \rangle = -\langle d_k, H_k d_k \rangle - \sum_{i=1}^m \lambda_{k,i} \langle \nabla g_i(x_k), d_k \rangle$$

and, again using (4.4), we obtain, in view of Lemma 4.4 and Assumption A6,

$$(4.8) \quad \langle \nabla f(x_k), \tilde{d}_k \rangle = - \sum_{i=1}^m \lambda_{k,i} \langle \nabla g_i(x_k), \tilde{d}_k \rangle + o(\|d_k\|^2).$$

In view of (4.7) and (4.8), (4.3) yields

$$(4.9) \quad \begin{aligned} f(x_k + d_k + \tilde{d}_k) &= f(x_k) + \frac{1}{2} \langle \nabla f(x_k), d_k \rangle + \frac{1}{2} \langle d_k, (\nabla_{xx}^2 f(x_k) - H_k) d_k \rangle \\ &\quad - \frac{1}{2} \sum_{i=1}^m \lambda_{k,i} \langle \nabla g_i(x_k), d_k \rangle - \sum_{i=1}^m \lambda_{k,i} \langle \nabla g_i(x_k), \tilde{d}_k \rangle + o(\|d_k\|^2). \end{aligned}$$

For $i \in I(x^*)$, expanding (4.6), we obtain, using Assumption A3' and the identity $\psi_k = o(\|d_k\|^2)$,

$$g_i(x_k) + \langle \nabla g_i(x_k), d_k \rangle + \frac{1}{2} \langle d_k, \nabla_{xx}^2 g_i(x_k) d_k \rangle + \langle \nabla g_i(x_k), \tilde{d}_k \rangle = o(\|d_k\|^2).$$

Multiplying by $\lambda_{k,i}$ and summing for $i \in I(x^*)$ we get, using (4.5),

$$(4.10) \quad \begin{aligned} & -\frac{1}{2} \sum_{i \in I(x^*)} \lambda_{k,i} \langle \nabla g_i(x_k), d_k \rangle - \sum_{i \in I(x^*)} \lambda_{k,i} \langle \nabla g_i(x_k), \tilde{d}_k \rangle \\ &= \sum_{i \in I(x^*)} \left(\lambda_{k,i} - \frac{1}{2} \frac{\lambda_{k,i}^2}{\mu_{k,i}} \right) g_i(x_k) \\ &\quad + \frac{1}{2} \sum_{i \in I(x^*)} \lambda_{k,i} \langle d_k, \nabla_{xx}^2 g_i(x_k) d_k \rangle + o(\|d_k\|^2). \end{aligned}$$

We also have, for $i \notin I(x^*)$, using (4.5) and the convergence to zero of the multipliers associated with the nonactive constraints,

$$(4.11) \quad \lambda_{k,i} \langle \nabla g_i(x_k), d_k \rangle = -\frac{\lambda_{k,i}^2}{\mu_{k,i}} g_i(x_k) - \lambda_{k,i} \langle d_k, \nabla_{xx}^2 g_i(x_k) d_k \rangle + o(\|d_k\|^2)$$

where we have added a term bounded by $o(\|d_k\|^2)$. Also, for $i \notin I(x^*)$,

$$(4.12) \quad \lambda_{k,i} \langle \nabla g_i(x_k), \tilde{d}_k \rangle = o(\|d_k\|^2)$$

since, from Lemma 4.4, $\tilde{d}_k = o(\|d_k\|)$ and since, from (4.5), the convergence of the numbers $\mu_{k,i}$ to zero (Proposition 4.2(iii)), and the fact that $-g_i(x_k)$ is bounded away

from zero, $\lambda_{k,i} = o(\|d_k\|)$. From (4.11) and (4.12), we obtain, summing for all $i \notin I(x^*)$,

$$\begin{aligned}
 & -\frac{1}{2} \sum_{i \notin I(x^*)} \lambda_{k,i} \langle \nabla g_i(x_k), d_k \rangle - \sum_{i \notin I(x^*)} \lambda_{k,i} \langle \nabla g_i(x_k), \tilde{d}_k \rangle \\
 (4.13) \quad & = \frac{1}{2} \sum_{i \notin I(x^*)} \frac{\lambda_{k,i}^2}{\mu_{k,i}} g_i(x_k) \\
 & + \frac{1}{2} \sum_{i \notin I(x^*)} \lambda_{k,i} \langle d_k, \nabla_{xx}^2 g_i(x_k) d_k \rangle + o(\|d_k\|^2).
 \end{aligned}$$

In view of (4.10) and (4.13), (4.9) can be rewritten as

$$\begin{aligned}
 f(x_k + d_k + \tilde{d}_k) &= f(x_k) + \frac{1}{2} \langle \nabla f(x_k), d_k \rangle \\
 (4.14) \quad & + \frac{1}{2} \langle d_k, (\nabla_{xx}^2 f(x_k) + \sum_{i=1}^m \lambda_{k,i} \nabla_{xx}^2 g_i(x_k) - H_k) d_k \rangle \\
 & + \sum_{i \in I(x^*)} \left(\lambda_{k,i} - \frac{1}{2} \frac{\lambda_{k,i}^2}{\mu_{k,i}} \right) g_i(x_k) + \frac{1}{2} \sum_{i \notin I(x^*)} \frac{\lambda_{k,i}^2}{\mu_{k,i}} g_i(x_k) + o(\|d_k\|^2).
 \end{aligned}$$

If in (4.14) we substitute for d_k its decomposition obtained in Lemma 4.3, we obtain,

$$\begin{aligned}
 f(x_k + d_k + \tilde{d}_k) &= f(x_k) + \frac{1}{2} \langle \nabla f(x_k), d_k \rangle \\
 & + \frac{1}{2} \langle d_k, P_k (\nabla_{xx}^2 f(x_k) + \sum_{i=1}^m \lambda_{k,i} \nabla_{xx}^2 g_i(x_k) - H_k) P_k d_k \rangle \\
 (4.15) \quad & + o \left(\sum_{i \in I(x^*)} \left(\frac{\lambda_{k,i}}{\mu_{k,i}} g_i(x_k) \right)^2 \right)^{1/2} \\
 & + \sum_{i \in I(x^*)} \left(\lambda_{k,i} - \frac{1}{2} \frac{\lambda_{k,i}^2}{\mu_{k,i}} \right) g_i(x_k) + \frac{1}{2} \sum_{i \notin I(x^*)} \frac{\lambda_{k,i}^2}{\mu_{k,i}} g_i(x_k) + o(\|d_k\|^2).
 \end{aligned}$$

Now, since by construction $\mu_{k,i} > 0$, we have

$$\frac{1}{2} \sum_{i \notin I(x^*)} \frac{\lambda_{k,i}^2}{\mu_{k,i}} g_i(x_k) \leq 0.$$

Also, since for $i \in I(x^*)$, $\lambda_i^* > 0$ and $g_i(x^*) = 0$, and since, for all k , $g(x_k) < 0$, Proposition 4.2(i), (iii) and the continuity of g imply that

$$o \left(\sum_{i \in I(x^*)} \left(\frac{\lambda_{k,i}}{\mu_{k,i}} g_i(x_k) \right)^2 \right)^{1/2} + \sum_{i \in I(x^*)} \left(\lambda_{k,i} - \frac{1}{2} \frac{\lambda_{k,i}^2}{\mu_{k,i}} \right) g_i(x_k) \leq 0.$$

Thus (4.15) yields

$$\begin{aligned}
 f(x_k + d_k + \tilde{d}_k) &\leq f(x_k) + \frac{1}{2} \langle \nabla f(x_k), d_k \rangle \\
 & + \frac{1}{2} \langle d_k, P_k \left(\nabla_{xx}^2 f(x_k) + \sum_{i=1}^m \lambda_{k,i} \nabla_{xx}^2 g_i(x_k) - H_k \right) P_k d_k \rangle + o(\|d_k\|^2)
 \end{aligned}$$

or

$$\begin{aligned} f(x_k + d_k + \tilde{d}_k) &\leq f(x_k) + \frac{1}{2} \langle \nabla f(x_k), d_k \rangle \\ &\quad + \frac{1}{2} \|d_k\|^2 \frac{\|P_k(\nabla_{xx}^2 f(x_k) + \sum_{i=1}^m \lambda_{k,i} \nabla_{xx}^2 g_i(x_k) - H_k)P_k d_k\|}{\|d_k\|} + o(\|d_k\|^2) \end{aligned}$$

and, in view of Assumption A9,

$$f(x_k + d_k + \tilde{d}_k) \leq f(x_k) + \frac{1}{2} \langle \nabla f(x_k), d_k \rangle + o(\|d_k\|^2).$$

Now, from (3.5) and Assumption A6, $\langle \nabla f(x_k), d_k \rangle \leq -\sigma_1 \|d_k\|^2 + o(\|d_k\|^2)$ and condition (2.10) of the line search obviously holds for $t = 1$, since $\alpha < \frac{1}{2}$.

(ii) We now show that a step of 1 preserves feasibility. Since $J_k = \emptyset$ (Proposition 4.2(ii)), the line search only requires $g_i(x_k + d_k + \tilde{d}_k) < 0$, $i = 1, \dots, m$. For $i \in I(x^*)$, this is always satisfied since the sequences $\{d_k\}$ and $\{\tilde{d}_k\}$ both converge to zero. Consider $i \in I(x^*)$. We have

$$\begin{aligned} g_i(x_k + d_k + \tilde{d}_k) &= g_i(x_k + d_k) + \langle \nabla g_i(x_k + d_k), \tilde{d}_k \rangle + O(\|\tilde{d}_k\|^2) \\ &= g_i(x_k + d_k) + \langle \nabla g_i(x_k), \tilde{d}_k \rangle + O(\|d_k\| \|\tilde{d}_k\|). \end{aligned}$$

From the definition of the solution \tilde{d}_k of (LS3), we get

$$g_i(x_k + d_k + \tilde{d}_k) = -\psi_k + O(\|d_k\| \|\tilde{d}_k\|)$$

and from Lemma 4.4,

$$(4.16) \quad g_i(x_k + d_k + \tilde{d}_k) = -\psi_k + O\left(\max\left\{\|d_k\|^3, \max_{i \in I(x^*)} \left|\frac{\mu_{k,i}}{\lambda_{k,i}} - 1\right| \|d_k\|^2\right\}\right).$$

Thus, since $\tau < 3$ and $\kappa < 1$, (4.16) and (2.9) yield $g_i(x_k + d_k + \tilde{d}_k) < 0$. \square

THEOREM 4.6. *Under the stated assumptions, the convergence is two-step superlinear, i.e.,*

$$\lim_{k \rightarrow \infty} \frac{\|x_{k+2} - x^*\|}{\|x_k - x^*\|} = 0.$$

Proof. In view of Lemma 4.5, the line search gives a step $t_k = 1$, for k large enough. Two successive iterates are thus related by

$$x_{k+1} - x_k = d_k + \tilde{d}_k$$

so that, in view of Proposition 4.2, Assumption A8(ii) and Lemmas 3.5 and 4.4,

$$(4.17) \quad x_{k+1} - x_k = d_k^0 + o(\|d_k^0\|).$$

Let us now consider the quadratic program in d ,

$$(4.18) \quad \begin{aligned} &\min f(x_k) + \langle \nabla f(x_k), d \rangle + \frac{1}{2} d^T H_k d \\ &\text{s.t. } \langle \nabla g_i(x_k), d \rangle + g_i(x_k) \leq 0, \quad i = 1, \dots, m. \end{aligned}$$

In [16], the two-step superlinear convergence of the sequence defined by

$$x_{k+1} = x_k + d_k^e$$

where d_k^e solves (4.18), is proven under assumptions similar to those of this theorem. A straightforward but lengthy analysis shows that, if $x_{k+1} - x_k = d_k^e + o(\|d_k^e\|)$, this result will still hold. Thus, in view of (4.17), it is enough to show that

$$(4.19) \quad d_k^0 = d_k^e + o(\|d_k^0\|).$$

Let us then show, to conclude, that (4.19) is satisfied. From the optimality conditions associated with the solution d_k^e of (4.18) and in view of Lemma 1 of [16], the direction d_k^e satisfies

$$(4.20) \quad H_k d_k^e + \nabla f(x_k) + \sum_{i \in I(x^*)} \lambda_{k,i}^e \nabla g_i(x_k) = 0,$$

$$(4.21) \quad \langle \nabla g_i(x_k), d_k^e \rangle + g_i(x_k) = 0, \quad i \in I(x^*)$$

for some coefficients $\lambda_{k,i}^e, i \in I(x^*)$, for k large enough. Now, by definition, the direction d_k^0 satisfies

$$(4.22) \quad H_k d_k^0 + \nabla f(x_k) + \sum_{i=1}^m \lambda_{k,i}^0 \nabla g_i(x_k) = 0,$$

$$(4.23) \quad \mu_{k,i} \langle \nabla g_i(x_k), d_k^0 \rangle + \lambda_{k,i}^0 g_i(x_k) = 0, \quad i = 1, \dots, m.$$

For $i \notin I(x^*)$, we have, from (4.23) and Proposition 4.2(iii), $\lambda_{k,i}^0 = o(\|d_k^0\|)$. Thus, (4.22) yields

$$(4.24) \quad H_k d_k^0 + \nabla f(x_k) + \sum_{i \in I(x^*)} \lambda_{k,i}^0 \nabla g_i(x_k) = o(\|d_k^0\|).$$

In view of Proposition 4.2(i), (iii), (4.23) yields

$$(4.25) \quad \langle \nabla g_i(x_k), d_k^0 \rangle + g_i(x_k) = o(\|d_k^0\|), \quad i \in I(x^*).$$

Now, the values $d_k^e, \lambda_{k,i}^e, i \in I(x^*)$ are solutions of a linear system with invertible matrix (4.20), (4.21) and d_k^0, λ_k^0 are solutions of the same system (4.24), (4.25) with right-hand side perturbed by $o(\|d_k^0\|)$. We then must have $d_k^0 = d_k^e + o(\|d_k^0\|)$. \square

5. Implementation issues. The first issue to be addressed is that of selecting an updating rule for H_k (in Step 3 of Algorithm A). The natural choice is to use the BFGS updating formula with Powell's modification [15] to ensure positive definiteness, although it is not clear whether Assumption A9 will then always be satisfied. Computational requirements are minimized if, instead of H_k itself, the Cholesky factor is updated (see, e.g., [4]).

The next question is that of efficiently solving (LS1)–(LS3). A key is given by Propositions 5.1 and 5.2 below, where the following notation is used:

$$A_k = [\nabla g_1(x_k), \nabla g_2(x_k), \dots, \nabla g_m(x_k)] \in \mathbb{R}^{n \times m},$$

$$\tilde{A}_k = [\nabla g_i(x_k) \text{ s.t. } i \in I_k] \in \mathbb{R}^{n \times |I_k|},$$

$$G_k = \text{diag}(g_i(x_k), \quad i = 1, \dots, m) \in \mathbb{R}^{m \times m},$$

$$M_k = \text{diag}(\mu_{k,i}, \quad i = 1, \dots, m) \in \mathbb{R}^{m \times m},$$

$$\tilde{g}_k = (g_i(x_k + d_k) + \psi_k \text{ s.t. } i \in I_k) \in \mathbb{R}^{|I_k|},$$

$$l = (1, 1, \dots, 1)^T \in \mathbb{R}^m.$$

PROPOSITION 5.1 (see [18]). *The matrix $B_k = M_k A_k^T H_k^{-1} A_k - G_k$ is invertible and the solutions of (LS1) and (LS2) can be expressed as*

$$(5.1) \quad \lambda_k^0 = -B_k^{-1} M_k A_k^T H_k^{-1} \nabla f(x_k),$$

$$(5.2) \quad d_k^0 = -H_k^{-1} (\nabla f(x_k) + A_k \lambda_k^0),$$

$$(5.3) \quad \lambda_k^1 = \lambda_k^0 + \|d_k^0\|^\nu M_k B_k^{-1} l,$$

$$(5.4) \quad d_k^1 = -H_k^{-1} (\nabla f(x_k) + A_k \lambda_k^1).$$

Proof. In order to show that the matrix B_k is invertible, it is sufficient to prove that the matrix $A_k^T H_k^{-1} A_k - M_k^{-1} G_k$ is positive definite. The latter holds due to positive definiteness of H_k , strict positiveness of the coefficients $\mu_{k,i}$ and strict feasibility of the iterate x_k . Now, with the notation just introduced, (LS1) can be written as

$$(5.5) \quad H_k d_k^0 + \nabla f(x_k) + A_k \lambda_k^0 = 0,$$

$$(5.6) \quad M_k A_k^T d_k^0 + G_k \lambda_k^0 = 0.$$

Also, (d_k^1, λ_k^1) satisfies

$$(5.7) \quad H_k d_k^1 + \nabla f(x_k) + A_k \lambda_k^1 = 0,$$

$$(5.8) \quad M_k A_k^T d_k^1 + G_k \lambda_k^1 = -\|d_k^0\|^\nu \mu_k.$$

From (5.5), we have

$$A_k^T d_k^0 = -A_k^T H_k^{-1} [\nabla f(x_k) + A_k \lambda_k^0]$$

and, from (5.6),

$$M_k^{-1} G_k \lambda_k^0 = A_k^T H_k^{-1} [\nabla f(x_k) + A_k \lambda_k^0],$$

i.e.,

$$(A_k^T H_k^{-1} A_k - M_k^{-1} G_k) \lambda_k^0 = -A_k^T H_k^{-1} \nabla f(x_k).$$

Thus,

$$\lambda_k^0 = -(A_k^T H_k^{-1} A_k - M_k^{-1} G_k)^{-1} A_k^T H_k^{-1} \nabla f(x_k),$$

i.e.,

$$\lambda_k^0 = -M_k B_k^{-1} A_k^T H_k^{-1} \nabla f(x_k).$$

Expression (5.2) is obtained directly from (5.5). Expressions (5.3) and (5.4) can be obtained similarly by (5.7) and (5.8). \square

For the computation of the search direction, which is a convex combination of d_k^0 and d_k^1 (see (2.5), (2.6)), we essentially need to compute the matrix B_k defined in Proposition 5.1, and its inverse. Since, from the BFGS update, the Cholesky factor associated with H_k is known, matrix B_k can be easily evaluated. Its inverse is not computed explicitly. Rather, an LU decomposition is performed. The computation of the search direction d_k thus essentially requires the decomposition of an $m \times m$ matrix.

The next proposition addresses the computation of the correction \tilde{d}_k .

PROPOSITION 5.2. *If the matrix $\tilde{B}_k = \tilde{A}_k^T H_k^{-1} \tilde{A}_k$ is invertible, the solution \tilde{d}_k of (LS3) can be expressed as*

$$(5.9) \quad \tilde{d}_k = -H_k^{-1} \tilde{A}_k \tilde{B}_k^{-1} \tilde{g}_k.$$

Proof. The solution \tilde{d}_k of (LS3) satisfies the optimality conditions

$$(5.10) \quad H_k \tilde{d}_k + \tilde{A}_k \tilde{\lambda}_k = 0,$$

$$(5.11) \quad \tilde{A}_k^T \tilde{d}_k + \tilde{g}_k = 0$$

for some multiplier vector $\tilde{\lambda}_k \in \mathbb{R}^{|I_k|}$. Relationship (5.10) yields

$$(5.12) \quad \tilde{d}_k = -H_k^{-1} \tilde{A}_k \tilde{\lambda}_k.$$

Multiplying by \tilde{A}_k^T and using (5.11) gives

$$\tilde{g}_k = \tilde{A}_k^T H_k^{-1} \tilde{A}_k \tilde{\lambda}_k.$$

Therefore, the vector $\tilde{\lambda}_k$ is given by

$$\tilde{\lambda}_k = (\tilde{A}_k^T H_k^{-1} \tilde{A}_k)^{-1} \tilde{g}_k.$$

Substituting this expression into (5.12) yields (5.9). \square

In order to save computation, evaluation of the correction \tilde{d}_k should be attempted only when the iterate is close to a solution of problem (1.1) (d_k small); otherwise, \tilde{d}_k should be set to 0. Since, close to a solution, the matrix \tilde{B}_k is always invertible due to linear independence of the gradients of the active constraints (Assumption A5), the computation of the correction \tilde{d}_k essentially requires the Cholesky factorization of the matrix \tilde{B}_k of dimension at most equal to $m \times m$.

Thus the total work per iteration (in addition to function evaluations) is essentially that associated with two Cholesky factorizations of size at most m , the total number of constraints. Note however that Algorithm A can be modified in a straightforward manner so as to ignore the "obviously inactive" constraints in the computation of the search direction.

The strict feasibility of the successive iterates is essential in order to identify the sign of the multipliers associated with the active constraints. A value $g_i(x_k)$ close to zero could possibly prevent the sign identification from occurring. Thus, rule (2.12) in Step 2 of Algorithm A may lead to numerical difficulties. An alternate rule could be

$$g_i(x_k + td_k + t^2 \tilde{d}_k) \leq \max \left\{ \frac{g_i(x_k)}{2}, -\frac{\psi_k}{2} \right\}, \quad i = 1, \dots, m.$$

In view of (2.9) and (4.16), such a modification would not affect the convergence properties of Algorithm A.

Algorithm A, in its present form, may present some instability problems. First of all, independence of the active constraints is essential for the computation of a search direction. Near dependence may lead to numerical problems. Linear system (2.1), (2.2) may also become very ill-conditioned if some multiplier μ_i corresponding to a nearly active constraint g_i becomes very small. This may occur close to a solution of problem (1.1) at which the strict complementarity conditions are not satisfied. In order to avoid that kind of problem, we could bound from below the approximate multipliers associated with nearly active constraints by a suitably small positive number.

Finally, it is clear that scaling can be (and *should* be) introduced at various places in Algorithm A, in particular in the definitions of I_k and J_k , in expression (2.9) for ψ_k and in the update formula (2.13b) for μ_k . We could consider, for example, redefining I_k as

$$I_k = \{i \text{ s.t. } g_i(x_k) \geq -\lambda_{k,i} \|\nabla g_i(x_k)\|^2 \|x_k - x_0\|^2\},$$

and similarly for J_k , so that asymptotically, it would be insensitive to the scaling of g .

Preliminary numerical experiments have been performed using test problems from [7]. The algorithm parameters were given values $\alpha = 0.3$, $\beta = 0.8$, $\theta = 0.8$, $\nu = 3$, $\tau = 2.5$, $\kappa = 0.5$, $\bar{\mu} = \infty$. To improve the behavior of the algorithm in the early iterations, $\|d_k^0\|^\nu$ in the right-hand side of (2.4) was replaced by $\min(10^{-2} \|d_k^0\|, \|d_k^0\|^\nu)$. On most problems, the performance of Algorithm A was roughly comparable to that of VF02AD as reported in [7] and to that of the algorithm in [13]. The behavior of the algorithm appeared to be relatively insensitive to changes in the values of the algorithm parameters, as could be expected from the asymptotic convergence properties.

REFERENCES

- [1] T. F. COLEMAN AND A. R. CONN, *Nonlinear programming via an exact penalty function: asymptotic analysis*, Math. Programming, 24 (1982), pp. 123–136.
- [2] ———, *Nonlinear programming via an exact penalty function: global analysis*, Math. Programming, 24 (1982), pp. 137–161.
- [3] D. GABAY AND D. G. LUENBERGER, *Efficiently converging minimization methods based on the reduced gradient*, SIAM J. Control Optim., 14 (1976), pp. 42–61.
- [4] P. E. GILL, W. MURRAY, AND M. A. SAUNDERS, *Methods for computing and modifying the LDV factors of a matrix*, Math. Computation, 29 (1975), pp. 1051–1077.
- [5] J. HERSKOVITS, *A superlinearly convergent two-stage feasible direction algorithm for nonlinear constrained optimization*, Research Report, Coordenação dos Programas de Pós-Graduação de Engenharia/Universidade Federal do Rio de Janeiro, INRIA, Le Chesnay, France, 1982.
- [6] ———, *A two-stage feasible direction algorithm for nonlinear constrained optimization*, Math. Programming, 36 (1986), pp. 19–38.
- [7] W. HOCK AND K. SCHITTKOWSKI, *Test Examples for Nonlinear Programming Codes*, Lecture Notes in Econom. and Math. Systems 187, Springer-Verlag, Berlin, New York, 1981.
- [8] N. MARATOS, *Exact penalty function algorithms for finite dimensional and optimization problems*, Ph.D. thesis, Imperial College of Science and Technology, London, 1978.
- [9] D. Q. MAYNE AND E. POLAK, *A superlinearly convergent algorithm for constrained optimization problems*, Math. Programming Stud., 16 (1982), pp. 45–61.
- [10] ———, *Feasible Directions Algorithms for Optimization Problems with Equality and Inequality Constraints*, Math. Programming, 11 (1976), pp. 67–80.
- [11] G. P. MCCORMICK, *An arc method for nonlinear programming*, SIAM J. Control Optim., 13 (1975), pp. 1194–1216.
- [12] W. T. NYE AND A. L. TITS, *An application-oriented, optimization-based methodology for interactive design of engineering systems*, Internat. J. Control, 43 (1986), pp. 1693–1721.
- [13] E. R. PANIER AND A. L. TITS, *A superlinearly convergent feasible method for the solution of inequality constrained optimization problems*, SIAM J. Control Optim., 25 (1987), pp. 934–950.
- [14] E. POLAK AND A. L. TITS, *On globally stabilized quasi-Newton methods for inequality constrained optimization problems*, in Proc. 10th Internat. Federation of Information Processing Societies Conference on System Modeling and Optimization, New York, August–September 1981, Lecture Notes in Control and Information Sciences 38, R. F. Drenick and F. Kozin, eds., Springer-Verlag, Berlin, New York, 1982, pp. 539–547.
- [15] M. J. D. POWELL, *A fast algorithm for nonlinearly constrained optimization calculations*, in Numerical Analysis, Proceedings, Biennial Conference, Dundee 1977, Lecture Notes in Math. 630, G. A. Watson, ed., Springer-Verlag, Berlin, New York, 1978, pp. 144–157.
- [16] ———, *The convergence of variable metric methods for nonlinearly constrained optimization calculations*, in Nonlinear Programming 3, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1978, pp. 27–63.
- [17] P. SPELLUCCI, *Han's method without solving QP*, in Optimization and Control, A. Auslender, W. Oettli, and J. Stoer, eds., Lecture Notes in Control and Information Sciences 30, Springer-Verlag, Berlin, New York, Tokyo, 1981, pp. 123–141.
- [18] R. A. TAPIA, *Diagonalized multiplier methods and quasi-Newton methods for constrained optimization*, J. Optim. Theory Appl., 22 (1977), pp. 135–194.

FEEDBACK STABILIZATION OF LINEAR DYNAMIC SYSTEMS WITH MULTIRATE SAMPLED OUTPUT*

TZYH-JONG TARN[†], XIAOMING ZENG[‡], AND JOHN R. ZAVGREN, JR.[§]

Abstract. In this paper a new feedback control technique called multirate sampled output feedback is introduced and investigated in the context of finite-dimensional and infinite-dimensional linear control systems. This control technique has three outstanding properties: the feedback gain function which transforms sampled output information into a control function is independent of the choice of the sample sequence, the sample sequence may be chosen randomly, and the gain function values are only needed on a finite interval (they can be reused cyclicly as in the case of periodic output feedback). It is shown that linear control systems satisfying certain rather weak assumptions can be stabilized by multirate sampled output feedback. An algorithm is presented for designing multirate sampled output feedback controls.

Key words. multirate sampled output feedback, feedback stabilization, infinite dimensional systems

AMS(MOS) subject classifications. 93D15, 93C20, 93C57

1. Introduction. Feedback stabilization of infinite-dimensional systems is a relatively new field of research which has seen major advances in the last few years. The results obtained to date are concerned with one of three traditional feedback schemes: state feedback, dynamic feedback, or direct output feedback. The work of Chammas and Leondes concerning periodic output feedback control of finite-dimensional systems [1]–[4] was extended to a large class of infinite-dimensional systems by Tarn, Zavgren, and Zeng [5]. This novel form of control, a departure from the three traditional feedback schemes, holds promise as a useful yet general stabilization technique. In this paper we introduce a generalization of periodic output feedback control—first presented in Zeng [6]—which we will call multirate sampled output feedback control. The main results of this paper prove that multirate sampled output feedback is a control scheme of considerable versatility because (1) it can stabilize systems under rather weak assumptions, (2) it uses sampled output information only, and (3) the sample times may be chosen randomly. This control scheme should prove to have many practical advantages and it will be quite easy to implement.

Before we develop the details of multirate sampled output feedback control, a little background information on the general subject of feedback control of infinite-dimensional systems is provided to help put this work in perspective. As mentioned before, feedback control schemes normally fit into one of three categories: state feedback, dynamic output feedback, or direct output feedback. State feedback control is an interesting area of research which generates useful results from a theoretical viewpoint. From a practical viewpoint these results are of limited value because we cannot measure the state of an infinite-dimensional system. The quantities which have physical meaning are those that can be observed, namely, the output of the system. Direct output feedback may be considered as a step toward utilizing output information in stabilizing a control system, but this method of stabilization is fraught with difficulty, even in the finite-dimensional realm, due to the fact that only a relatively small subset

* Received by the editors May 5, 1986; accepted for publication (in revised form) September 8, 1987. This research was partially supported by National Science Foundation grants ECS-8515899 and INT-8519654 at Washington University, St. Louis, Missouri.

[†] Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130.

[‡] Department of Mathematics, Glassboro State College, Glassboro, New Jersey 08028.

[§] Bolt, Barenek, and Newman Laboratory, 10 Moulton St., Cambridge, Massachusetts 02138.

of stabilizable systems can be stabilized by this technique. We do not consider that this stabilization scheme holds much promise for stabilization of infinite-dimensional systems in general. Dynamic feedback is an especially useful stabilization method because it utilizes information solely obtainable from the output of the system. Even though this stabilization scheme is consistent with the philosophy of utilizing only available information, it has another shortcoming. The compensator we develop for an infinite-dimensional system is itself usually an infinite-dimensional system: this can present difficulties in implementation.

To date, two significant results have been obtained which guarantee the existence of a finite-dimensional compensator. Schumacher [7] and Kobayashi [8] obtained sufficient conditions for special classes of infinite-dimensional systems which guarantee the existence of an observer of finite order such that the closed-loop system is exponentially stable. Kobayashi's work is especially interesting in that the author uses only sampled output values. In practice, output information is not always available as a continuous stream. In fact, output information is often only available at discrete time instants. These results are important steps toward the development of practical feedback stabilization methods for infinite-dimensional systems. A feedback stabilization scheme which only uses sampled data, is easy to implement, and is capable of stabilizing a wide range of infinite-dimensional systems is an attractive choice from a practical viewpoint.

In 1978 Chammas and Leondes introduced a new feedback control scheme, for finite-dimensional time invariant linear systems which they called periodic output feedback control [1]–[4]. To define periodic output feedback control consider the linear time invariant system

$$(1.1) \quad \begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) \end{aligned}$$

where $x \in R^n$, $u \in R^p$, $y \in R^m$, and A , B , C are constant matrices of the appropriate size. A periodic output feedback control scheme for this system is defined by

$$(1.2) \quad \begin{aligned} u(t) &= K(t)y(n\tau), \quad n\tau \leq t < (n+1)\tau, \quad n = 0, 1, 2, \dots, \\ K(t) &= K(t + \tau) \end{aligned}$$

where the time-varying periodic gain matrix $K(\cdot)$ takes values in $R^{p \times m}$. The primary result obtained in their four papers is that if system (1.1) is stabilizable and detectable, then there exists¹ $\tau > 0$ and a periodic function $K(\cdot)$ with period τ such that the use of the feedback control scheme (1.2), in conjunction with the system (1.1), achieves the result that the state of the closed-loop system approaches zero on the sampling points, i.e.,

$$\lim_{n \rightarrow \infty} x(n\tau) = 0.$$

These papers contain the basic ideas which we extended to infinite-dimensional systems [5].

Periodic output feedback is a new departure from the control schemes just mentioned and should, at least in principle, be an attractive alternative in light of the following observations.

¹ The choice of τ is limited only by a technical consideration which must be made only if two or more eigenvalues of the matrix A have the same real parts.

(1) This control scheme uses sampled output data only, i.e., it does not require continuous system monitoring which is needed in the case of state feedback or output feedback.

(2) The feedback gain function, $K(\cdot)$, is periodic; therefore it is enough to specify its values on a finite interval of width τ and then reuse these values cyclicly as we progress down the real line.

This control scheme can be thought of as a hybrid of open-loop and closed-loop control in the following sense: the control is a closed-loop control at the sampling times and it is an open-loop control between the sampling times. The control function $u(\cdot)$ has no information concerning the state of the system on the open intervals $(n\tau, (n+1)\tau)$ for $n = 0, 1, 2, \dots$.

After some technical preliminaries which serve to define the class of infinite-dimensional systems we are to investigate, we will generalize the concept of periodic output feedback to a new type of control scheme which we call multirate sampled output feedback control. This new control scheme differs from periodic output feedback control in that the sample times are not equally spaced; they can even be randomly spaced.

2. Technical preliminaries. The infinite-dimensional systems we considered in [5] and in this paper are modeled by evolution equations of the form

$$(2.1) \quad \begin{aligned} z_t &= T(t)z_0 + \int_0^t T(t-s)Bu(s) ds, \\ y(t) &= Cz_t \end{aligned}$$

where z_t is an element of the Banach space Z , $\{T(t); t \geq 0\}$ is a strongly continuous semigroup of bounded linear operators on Z , and B and C are bounded linear operators from R^p to Z and from Z to R^m , respectively.

DEFINITION 2.1. A closed linear operator A mapping a Banach space Z into Z is said to satisfy the spectrum decomposition assumption if the spectrum, $\sigma(A)$, contains a bounded part, σ_1 , separated from the rest, σ_2 , in such a way that a rectifiable simple closed curve, Γ (or, more generally, a finite number of such curves), can be drawn so as to enclose an open set containing σ_1 in its interior and σ_2 in its exterior.

THEOREM 2.1 [9]. *Let A be a closed linear operator mapping the Banach space Z into Z which satisfies the spectrum decomposition assumption; then*

(1) *The operator A can be decomposed in accordance with the space decomposition $Z = Z_1 \oplus Z_2$ in the following sense:*

$$P_1 D(A) \subset D(A), \quad AZ_1 \subset Z_1, \quad AZ_2 \subset Z_2$$

where P_1 is the projection on Z_1 along Z_2 , and is given by the expression

$$P_1 = \frac{1}{2\pi i} \int_{\Gamma} (\lambda I - A)^{-1} d\lambda$$

and the A invariant subspaces Z_1 and Z_2 are defined by $P_1 Z \triangleq Z_1$ and $(I - P_1)Z \triangleq Z_2$.

(2) $\sigma_1 = \sigma(A_1)$, $\sigma_2 = \sigma(A_2)$, where A_1 and A_2 are the restrictions of A to Z_1 and Z_2 , respectively, i.e., $A_1 \triangleq A|_{Z_1} = AP_1$, $A_2 \triangleq A|_{Z_2} = A(I - P_1)$.

(3) A_1 is bounded.

(4) P_1 and $(I - P_1)$ commute with A , i.e., $P_1 A \subset AP_1$, $(I - P_1)A \subset A(I - P_1)$.

COROLLARY 2.1. *The following results are direct consequences of Theorem 2.1:*

(1) Because $T(t)$ commutes with P_1 and $(I - P_1)$ we have

$$T(t)Z_1 \subset Z_1, \quad T(t)Z_2 \subset Z_2,$$

i.e., Z_i is $T(t)$ invariant for any $t \geq 0$ for $i = 1$ and $i = 2$.

(2) The restriction of $\{T(t); t \geq 0\}$ to Z_i is a strongly continuous semigroup which is given by $\{T_i(t) \triangleq T(t)P_i; t \geq 0\}$ with infinitesimal generator A_i , respectively, for $i = 1$ and $i = 2$, where $P_2 \triangleq I - P_1$.

(3) $D(A_1) = Z_1$ and $\{T_1(t); t \geq 0\}$ is a uniformly continuous group given by

$$T_1(t) = \exp(A_1 t) = \sum_{n=0}^{\infty} A_1^n \frac{t^n}{n!}.$$

When the spectrum of A has an isolated point, say λ , $\sigma(A)$ can be divided into two separated parts σ_1 and σ_2 such that σ_1 consists of the single point λ . In this special case any closed Jordan curve enclosing λ but no other point of $\sigma(A)$ may be chosen as Γ . If the subspace Z_1 is finite-dimensional then λ is an eigenvalue of A , and the dimension of Z_1 is called the algebraic multiplicity of λ . If λ is a pole of the resolvent of A of order m , then λ is in the point spectrum of A . In this case we have

$$Z_1 = \{z \in Z \mid (\lambda I - A)^m z = 0\} = N((\lambda I - A)^m),$$

$$Z = R((\lambda I - A)^m) \oplus N((\lambda I - A)^m).$$

Z_1 is often called the generalized eigenspace of A associated with λ and any $z \in Z_1$ of nonzero norm is called a generalized eigenvector associated with the eigenvalue λ .

Fundamental assumptions and definitions of terms. In the remainder of this paper the following conditions will be imposed on system (2.1).

(A1) The infinitesimal generator A of the strongly continuous semigroup $\{T(t); t \geq 0\}$ satisfies the spectrum decomposition assumption in the sense that $\sigma_1 = \{\lambda \in \sigma(A) \mid \operatorname{Re} \lambda \geq -\delta\}$ has the property that for some $\delta > 0$ the dimension of the union of the generalized eigenspaces generated from points in σ_1 , Z_1 , is finite. For any such δ we define the following subspaces, semigroups, infinitesimal generators, and operators:

$$Z_u \triangleq Z_1, \quad Z_s \triangleq Z_2,$$

$$T_u(t) \triangleq T_1(t), \quad T_s(t) \triangleq T_2(t),$$

$$A_u \triangleq A_1, \quad A_s \triangleq A_2,$$

$$P_u \triangleq P_1, \quad P_s \triangleq P_2.$$

(A2) $\{T_s(t); t \geq 0\}$ satisfies the spectrum-determined growth condition, i.e.,

$$\sup \operatorname{Re} \sigma(A_s) = \inf_{t>0} \frac{\ln \|T_s(t)\|}{t} = \lim_{t \rightarrow \infty} \frac{\ln \|T_s(t)\|}{t}.$$

(A3') The pair (A_u, B_u) is controllable on Z_u , where B_u is the restriction of B to Z_u , i.e., $B_u \triangleq P_u B$.

(A3) $\overline{R(T_0)} \supset Z_u$, where $R(T_0)$ is the set of all states reachable from the initial condition $z(0) = 0$ on the interval $[0, T_0]$ for some $T_0 > 0$ (the overbar indicates closure).

(A4) (A_u, C_u) is an observable pair on Z_u , where the operator $C_u \triangleq C|_{Z_u}$.

Let $W = L^2([0, T_0]; R^p)$. If we define the mapping $G: W \rightarrow Z$ by

$$Gu = \int_0^{T_0} T(T_0 - s)Bu(s) ds,$$

then $R(T_0) = R(G)$. The following theorem is a useful tool for verifying that a particular system satisfies Assumption (A3).

THEOREM 2.2. *Assumption (A3) holds if and only if $\text{Ker}(G^*) \subset \bigcap_{z \in Z_u} \text{Ker}(z^{**})$ where $*$ and $**$ denote the conjugate and the second conjugate space or operator respectively. (z^{**} is the canonical image of z in the second conjugate space Z^{**} .)*

Proof. If $\overline{R(G)} \supset Z_u$ and $G^*z^* = 0$, then $\langle Gu, z^* \rangle = \langle u, G^*z^* \rangle = 0$ for all $u \in W$. Therefore $\langle v, z^* \rangle = 0$ for all $v \in Z_u$ and $z^* \in \bigcap_{z \in Z_u} \text{Ker}(z^{**})$. Conversely, if there is a point $v \in Z_u \setminus \overline{R(G)}$ then by the Hahn-Banach Theorem, there exists a nonzero element $z^* \in Z^*$ such that $\overline{R(G)} \subset \text{Ker}(z^*)$, i.e., $\langle u, G^*z^* \rangle = \langle Gu, z^* \rangle = 0$ for all $u \in W$, but $v \notin \text{Ker}(z^*)$, i.e., $\langle v, z^* \rangle \neq 0$. This implies that $z^* \in \text{Ker}(G^*) \setminus \bigcap_{z \in Z_u} \text{Ker}(z^{**})$.

Remark. In 1975, Triggiani [10] investigated infinite-dimensional linear systems of the form (2.1) under the Assumptions (A1), (A2), and (A3') and proved that such systems are stabilizable by state feedback control of the form $u(t) = Dz(t)$, $t \geq 0$, where D is a bounded linear operator from Z to R^p . In [5] the authors used assumptions (A1), (A2), (A4) and the slightly different assumption (A3') to prove the following theorem which is a direct extension of the original result of Chammas and Leondes on stabilization of finite-dimensional linear systems.

THEOREM 2.3. *Suppose the linear system (2.1) satisfies Assumptions (A1)–(A4); then there exists $\tau > 0$ and a periodic $p \times m$ matrix-valued function $K(t)$ with period τ such that the closed-loop system with the periodic sampled output feedback control*

$$u(t) = K(t)y(n\tau), \quad n\tau \leq t < (n+1)\tau$$

is stable in the sense that

$$\lim_{t \rightarrow \infty} \|z(t, z_0, u(t))\| = 0$$

for all initial conditions $z_0 \in Z$.

3. Multirate sampled output feedback control. In the previous theorem, we assumed that the output information of the system which was used in the feedback control is sampled at equally spaced times $\{n\tau\}_{n=0,1,2,\dots}$, i.e., if the sampling time sequence is denoted by $\{t_n\}_{n=0,1,\dots}$, then

$$(3.1) \quad t_{n+1} - t_n = \tau$$

for all $n = 0, 1, 2, \dots$.

In this section we will extend the work of Chammas and Leondes in another direction. We will relax condition (3.1) and consider the more general sampling sequence $\{t_n\}_{n=0,1,\dots}$, i.e., we will consider the stabilization of linear dynamic systems using multirate sampled output feedback control.

First of all, the “natural” condition

$$(3.2) \quad m \leq \tau_n \triangleq t_{n+1} - t_n \leq M$$

should be satisfied, i.e., the length of the interval between two sampling times must be uniformly bounded from below and above.

We will now prove by example that this relatively weak condition is too general even for finite-dimensional systems.

More specifically we shall consider finite-dimensional linear systems of the form

$$(3.3) \quad \begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) \end{aligned}$$

in conjunction with a feedback control of the form

$$u(t) = K(t)y(t_n), \quad t_n \leq t < t_{n+1}$$

where

$$K(t) \in L^2_{\text{loc}}(0, \infty; R^{p \times m}).$$

For such systems the state of the closed-loop system at the sampling points can be expressed by the following formula:

$$(3.4) \quad x(t_n) = \prod_{i=1}^{n-1} \left(\exp(A\tau_i) + \int_0^{\tau_i} T(\tau_i - s)BK(s + t_{i-1}) ds \right) x_0.$$

If (A, C) is an observable pair and we choose the sequence $\{t_n\}_{n=0,1,\dots}$ correctly [11], then $(\exp(A\tau_i), C)$ will be an observable pair for each i . The only action we can take to ensure the sequence (3.4) converges is to control the eigenvalues of the matrices

$$(3.5) \quad \exp(A\tau_i) + G_i C$$

by the choice of the sequence of matrices $\{G_i\}$, where

$$(3.6) \quad G_i \triangleq \int_0^{\tau_i} T(\tau_i - s)BK(s + t_{i-1}) ds.$$

Suppose all the eigenvalues of $\exp(A\tau_i) + G_i C$ are located inside the open disk centered about the origin in the complex plane with radius, $r < 1$ for every i . Then for fixed i ,

$$\|(\exp(A\tau_i) + G_i C)^k\| \leq \mu^k$$

for sufficiently large k , where $\mu \in (0, 1)$. The smaller r is, the more quickly $\|(\exp(A\tau_i) + G_i C)^k\|$ converges to zero. In the extreme case where $r = 0$, every matrix is nilpotent and its norm will be zero after a finite number of steps. As the next example will illustrate, the nilpotency of every element of the sequence $\{\exp(A\tau_i) + G_i C\}$ is not a sufficient condition for the convergence of the sequence (3.4).

Example 3.1. For system (1.1) let $x \in R^2$, $y \in R^1$, $u \in R^1$, $C \triangleq (1, 0)$, $x_0 = (0, 1)^T$,

$$A \triangleq \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}, \quad B \triangleq \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

It is easy to verify that (A, B) is a controllable pair and (A, C) is an observable pair. Using the fact that

$$\exp(A\tau) \triangleq \begin{bmatrix} 1 & \tau \\ 0 & 1 \end{bmatrix} \exp(\lambda\tau)$$

we can easily verify that $(\exp(A\tau), C)$ is an observable pair if $\tau \neq 0$. Let $G \triangleq (g_1 g_2)^T$, then

$$\exp(A\tau) + GC = \begin{bmatrix} \exp(\lambda\tau) + g_1 & \tau \exp(\lambda\tau) \\ g_2 & \exp(\lambda\tau) \end{bmatrix}.$$

For the particular choices $g_1 = -2 \exp(\lambda\tau)$, and $g_2 = -(1/\tau) \exp(\lambda\tau)$, this matrix becomes

$$\exp(\lambda\tau) \begin{bmatrix} -1 & \tau \\ -1/\tau & 1 \end{bmatrix},$$

which can be easily verified to be a nilpotent matrix. Now, for any two choices of τ , say τ_1 and τ_2 , define G_1 and G_2 the way G was defined above for τ . In this case

$$\begin{aligned} E &\triangleq (\exp(A\tau_1) + G_1 C)(\exp(A\tau_2) + G_2 C) \\ &= \exp(\lambda(\tau_1 + \tau_2)) \begin{bmatrix} 1 - \tau_1/\tau_2 & \tau_1 - \tau_2 \\ 1/\tau_1 - 1/\tau_2 & 1 - \tau_2/\tau_1 \end{bmatrix}. \end{aligned}$$

For the particular choices $\tau_1 = 1$, $\tau_2 = 3$ and $\lambda = \frac{1}{4} \ln \frac{3}{2}$ we have

$$E = \begin{bmatrix} 1 & -3 \\ 1 & -3 \end{bmatrix}$$

and E has a spectrum consisting of the points $\{0, -2\}$. It can be shown that

$$E^n x_0 = \begin{bmatrix} \frac{3}{2}(-2)^n \\ \frac{3}{2}(-2)^n \end{bmatrix}$$

which has no limit as n goes to infinity. This result has the implication that if we choose the sampling sequence $\{t_n\}$ for the system (3.3), such that $\tau_n \triangleq t_{n+1} - t_n = 1$ for all even n , and $\tau_n = 3$ for all odd n and $t_1 = 3$, and if we use a feedback gain function $K(\cdot)$ such that

$$\sigma \left(\exp(A\tau_i) + \int_0^{\tau_i} T(\tau_i - s) B K(s + t_{i-1}) ds C \right) = \{0\},$$

for all i , the state of the closed-loop system with initial condition x_0 may be divergent. This means that constraints must be put on the sampling time sequence to be used in multirate output feedback if stability is to be achieved. For infinite-dimensional systems of the form (2.1) we have the following result.

THEOREM 3.1. *Assume that an infinite-dimensional system of the form (2.1) satisfies Assumptions (A1)–(A4) and; that the sequence of reals $\{r_i\}_{i=0}^q$, $0 = r_0 < r_1 < r_2 < \dots < r_q < \infty$ satisfies the following two conditions:*

(i) $r_{i+1} - r_i \geq T_0$ (T_0 as in Assumption (A3)), $i = 1, 2, \dots, q$, and $r_1 \geq \tau^*$ for some $\tau^* > 0$;

(ii) $(\exp(A_u r_i), C_u)$ is an observable pair for each i .

Then there exist integers m_1, m_2, \dots, m_q and a function $\hat{K} \in L^2(0, r_q; R^{p \times m})$ such that for any sampling time sequence $\{t_n\}_{n=0,1,\dots}$ which satisfies

(i) $\tau_n \triangleq t_{n+1} - t_n \in \{r_1, \dots, r_q\}$ for all integers $n = 0, 1, \dots$;

(ii) $\tau_n = r_i \neq \tau_{n-1}$ implies $\tau_{n+k} = r_i$ for all $k < m_i$,

the closed-loop system with the feedback control

$$(3.7) \quad \begin{aligned} u(t) &= K(t)y(t_n), & t_n \leq t < t_{n+1}, \\ K(t) &= \hat{K}(t - t_n) \end{aligned}$$

is stable in the sense that $\lim_{t \rightarrow \infty} \|z(t, z_0, u(t))\| = 0$ for all initial conditions $z_0 \in Z$.

We shall break up the proof of this theorem into several lemmas.

LEMMA 3.1.1. *Suppose that Assumptions (A1) and (A2) are satisfied and $\mu' \in (0, 1)$. Then there exists $\tau^* > 0$, such that for any $\tau > \tau^*$ which gives rise to an observable pair*

$(\exp(A_u\tau), C_u)$ on Z_u , there exists $n_0 > 0$, and a linear operator G mapping R^m to Z_u such that

$$\|(T(\tau) + GC)^n\| \leq (\mu')^n$$

for all integers $n \geq n_0$.

Proof. Choose $\mu'' \in (0, \mu')$. From Assumption (A1), Theorem 2.1, and Corollary 2.1 we know that $\sup \{\operatorname{Re}(\lambda) \mid \lambda \in \sigma(A_s)\} \leq -\delta$. Because $\{T_s(t); t \geq 0\}$ satisfies the spectrum-determined growth condition, there exist $M > 0$ and $\delta' > 0$, $\delta' < \delta$, such that

$$\|T_s(t)\| \leq M \exp(-\delta't)$$

for all $t \geq 0$. This implies the existence of a lower bound $\tau^* > 0$ such that

$$\|T_s(t)\| \leq \mu''$$

for all $t \geq \tau^*$. Since $(\exp(A_u\tau), C_u)$ is an observable pair and $\tau \geq \tau^*$, we can find a linear operator G from R^p to Z_u such that

$$\sigma(\exp(A_u\tau) + GC_u) \subset \left\{ \lambda \in \mathbb{C} \mid |\lambda| < \frac{\mu''}{2} \right\}.$$

Therefore, there exists an integer $n' > 0$, such that

$$\|(\exp(A_u\tau) + GC_u)^n\| \leq (\mu'')^n$$

for all $n \geq n'$. Now let us define the following operators:

$$H \triangleq \exp(A_u\tau) + GC_u,$$

$$F \triangleq GC_s \quad (C_s \triangleq C|_{Z_s}),$$

$$R \triangleq T_s(\tau).$$

Then the mapping $(T(\tau) + GC)^n$ can be expressed in an alternate form as

$$(T(\tau) + GC)^n = \begin{bmatrix} H & F \\ 0 & R \end{bmatrix}^n = \begin{bmatrix} H^n & \sum_{k=1}^n H^{k-1} F R^{n-k} \\ 0 & R^n \end{bmatrix}.$$

Now we can derive the following inequalities:

$$\begin{aligned} \|(T(\tau) + GC)^n\| &\leq \|H^n\| + \|R^n\| + \left\| \sum_{k=1}^n H^{k-1} F R^{n-k} \right\| \\ &\leq \|H^n\| + \|R\|^n + \sum_{k=1}^n \|H^{k-1} F R^{n-k}\|. \end{aligned}$$

If $n > n'$ then

$$\begin{aligned} \sum_{k=1}^n \|H^{k-1} F R^{n-k}\| &= \sum_{k=1}^{n'} \|H^{k-1} F R^{n-k}\| + \sum_{k=n'+1}^n \|H^{k-1} F R^{n-k}\| \\ &\leq \sum_{k=1}^{n'} \|H^{k-1} F\| (\mu'')^{n-k} + \sum_{k=n'+1}^n \|F\| (\mu'')^{n-1} \\ &= \left(\sum_{k=1}^{n'} \|H^{k-1} F\| (\mu'')^{-k} + (n - n') \|F\| (\mu'')^{-1} \right) (\mu'')^n. \end{aligned}$$

Because $\mu'' < \mu'$,

$$\lim_{n \rightarrow \infty} (n(\mu''/\mu')^n) = \lim_{n \rightarrow \infty} (M(\mu''/\mu')^n) = 0$$

for any constant $M > 0$. From the inequality

$$\|(T(\tau) + GC)^n\| \leq \left(2 + \sum_{k=1}^{n'} \|H^{k-1}F\|(\mu'')^{-k} + (n - n')\|F\|(\mu'')^{-1} \right) (\mu'')^n,$$

we get the result

$$\lim_{n \rightarrow \infty} \|(T(\tau) + GC)^n\|/(\mu')^n = 0.$$

This implies that there exists an integer n_0 such that

$$\|(T(\tau) + GC)^n\|/(\mu')^n \leq 1$$

for all $n \geq n_0$. If we multiply both sides of this inequality by $(\mu')^n$ we have the desired result, i.e.,

$$\|(T(\tau) + GC)^n\| \leq (\mu')^n$$

for all $n \geq n_0$. \square

LEMMA 3.1.2. *Let A be any linear bounded operator from Z to Z satisfying $\|A^{n_0}\| \leq \mu' < 1$ for some integer $n_0 > 0$; then for any $\mu'' < 1$, $\mu'' > \mu'$, there exists $\varepsilon_0 > 0$, such that for any linear bounded operator from Z to Z , say B , with $\|B\| \leq \varepsilon_0$, we have*

$$\|(A + B)^{n_0 k}\| \leq (\mu'')^k$$

for all integers $k > 0$.

Proof. Define the term $M(\varepsilon)$ by

$$M(\varepsilon) \triangleq \sum_{j=1}^{n_0} \binom{n_0}{j} \|A\|^{n_0-j} \varepsilon^j.$$

Then $M(\varepsilon)$ is a polynomial in ε without a constant term and all its coefficients are positive. This implies that $M(\varepsilon)$ is monotonic, nondecreasing, and $M(\varepsilon) \rightarrow 0$ as $\varepsilon \rightarrow 0^+$. This means there exists $\varepsilon_0 > 0$, such that

$$0 < \mu' + M(\varepsilon) \leq \mu''$$

for all $\varepsilon \in [0, \varepsilon_0]$. Therefore,

$$\begin{aligned} \|(A + B)^{n_0 k}\| &\leq \|(A + B)^{n_0}\|^k \\ &\leq \left(\|A^{n_0}\| + \sum_{j=1}^{n_0} \binom{n_0}{j} \|A\|^{n_0-j} \|B\|^j \right)^k \\ &\leq \left(\mu' + \sum_{j=1}^{n_0} \binom{n_0}{j} \|A\|^{n_0-j} \varepsilon_0^j \right)^k \\ &\leq (\mu'')^k \end{aligned}$$

for all integers k . \square

Proof of Theorem 3.1. Let τ^* be as in Lemma 3.1.1 and $\mu \in (0, 1)$; then, by hypothesis, $(\exp(A_u r_i), C_u)$ is an observable pair for each $i = 1, 2, \dots, q$. From the proof of Lemma 3.1.1 we know there exists for each i an integer n'_i and a linear operator G_i mapping Y to Z_u such that

$$\|(T(r_i) + G_i C)^n\| \leq (\mu')^n$$

for all integers $n \geq n'_i$, for any given μ' satisfying $0 < \mu' < \mu$.

Let the linear operator A from Z to Z be defined by $A \triangleq T(r_i) + G_i C$; then A satisfies the property

$$\|A^{n'_i}\| \leq \mu'.$$

As in the proof of Lemma 3.1.2, given $\mu'' \in (\mu', \mu)$ there exists ε_i such that for any linear operator B from Z to Z with $\|B\| \leq \varepsilon_i$, we have

$$\|(A+B)^{kn'_i}\| \leq (\mu'')^k$$

for all integers $k > 0$. Consequently there is a positive constant, M , such that

$$\|(A+B)^n\| \leq M(\mu'')^k$$

if $n = kn'_i + e$, for an integer, e , $0 \leq e < n'_i$. The inequality

$$\frac{\|(A+B)^n\|}{\mu^k} \leq M(\mu''/\mu)^k$$

implies that there exists an integer m_i for each integer i such that $\|(A+B)^n\| \leq \mu^k$ for all $n \geq m_i$. Let

$$\varepsilon \triangleq \frac{\min(\varepsilon_1, \dots, \varepsilon_q)}{\|C\|},$$

$G_0 = 0$ and $t_0 = 0$. Because of the assumption, $r_{i+1} - r_i \geq T_0$, and Assumption (A3), there exists a function

$$K^i(t) \in L^2(r_{i-1}, r_i; R^{p \times m})$$

such that

$$\left\| \int_{r_{i-1}}^{r_i} T(r_i - s)BK^i(s) ds - G_i + T(r_i - r_{i-1})G_{i-1} \right\| < \frac{\varepsilon}{2(2M)^{q-i}}$$

where $M > 1$ is a constant satisfying $\|T(t)\| \leq M$, $t \in [0, r_q]$. We will define $\hat{K}(t)$ on $[0, r_q]$ by $\hat{K}(t) = K^i(t)$ for $t \in [r_{i-1}, r_i]$. Now we have the following inequality:

$$\left\| \int_0^{r_1} T(r_1 - s)B\hat{K}(s) ds - G_1 \right\| < \frac{\varepsilon}{2(2M)^{q-1}} < \frac{\varepsilon}{(2M)^{q-1}}.$$

Assume that for i arbitrary but fixed

$$\left\| \int_0^{r_{i-1}} T(r_{i-1} - s)B\hat{K}(s) ds - G_{i-1} \right\| < \frac{\varepsilon}{(2M)^{q-(i-1)}};$$

then

$$\begin{aligned} & \left\| \int_0^{r_i} T(r_i - s)B\hat{K}(s) ds - G_i \right\| \\ &= \left\| T(r_i - r_{i-1}) \int_0^{r_{i-1}} T(r_{i-1} - s)B\hat{K}(s) ds - T(r_i - r_{i-1})G_{i-1} \right. \\ & \quad \left. + \int_{r_{i-1}}^{r_i} T(r_i - s)B\hat{K}(s) ds - G_i + T(r_i - r_{i-1})G_{i-1} \right\| \\ &< \|T(r_i - r_{i-1})\| \left\| \int_0^{r_{i-1}} T(r_{i-1} - s)B\hat{K}(s) ds - G_{i-1} \right\| + \frac{\varepsilon}{2(2M)^{q-i}} \\ &< \frac{M\varepsilon}{(2M)^{q-(i-1)}} + \frac{\varepsilon}{2(2M)^{q-i}} \\ &< \frac{\varepsilon}{2(2M)^{q-i}} + \frac{\varepsilon}{2(2M)^{q-i}} \\ &= \frac{\varepsilon}{(2M)^{q-i}}. \end{aligned}$$

By use of the principle of induction, we know that the inequality

$$\left\| \int_0^{r_i} T(r_i - s) B \hat{K}(s) ds - G_i \right\| < \frac{\varepsilon}{(2M)^{q-i}} \leq \varepsilon$$

holds for all integers $i = 1, 2, \dots, q$.

Because of the way we have chosen the sampling time sequence, $\{t_n\}_{n=0,1,\dots}$ the state of the closed-loop system with feedback control (3.7) at the sampling time t_{n+1} can be expressed by

$$\begin{aligned} z(t_{n+1}) &= \prod_{j=1}^n \left(T(\tau_j) + \int_0^{\tau_j} T(\tau_j - s) B \hat{K}(s) ds C \right) z_0 \\ &= \prod_{j=1}^l \left(T(a_j) + \int_0^{a_j} T(a_j - s) B \hat{K}(s) ds \right)^{n_j} z_0 \end{aligned}$$

where $n = n_1 + n_2 + \dots + n_l$, and for each integer j there exists an integer i such that $a_j = r_i$ and $n_j \geq m_i$ (with the possible exception of $j = l$). Hence for each integer $j < l$

$$\begin{aligned} &\left(T(a_j) + \int_0^{a_j} T(a_j - s) B \hat{K}(s) ds C \right)^{n_j} \\ &= \left(T(r_i) + \int_0^{r_i} T(r_i - s) B \hat{K}(s) ds C \right)^{n_j} \\ &= \left[\left(T(r_i) + G_i C \right) + \left(\int_0^{r_i} T(r_i - s) B \hat{K}(s) ds - G_i \right) C \right]^{n_j}. \end{aligned}$$

Because

$$\begin{aligned} \left\| \left(\int_0^{r_i} T(r_i - s) B \hat{K}(s) ds - G_i \right) C \right\| &\leq \left\| \int_0^{r_i} T(r_i - s) B \hat{K}(s) ds - G_i \right\| \|C\| \\ &\leq \varepsilon \|C\| \\ &\leq \varepsilon_i, \end{aligned}$$

we have

$$\left\| \left(T(a_j) + \int_0^{a_j} T(a_j - s) B \hat{K}(s) ds C \right)^{n_j} \right\| \leq \mu^{k_j}$$

where $n_j = k_j n'_i + e_j$ and $0 \leq e_j < n'_i$. This implies that

$$\|z(t_{n+1})\| \leq M \mu^{k_1 + k_2 + \dots + k_{l-1}},$$

for some constant $M > 0$. Because $k_1 + k_2 + \dots + k_{l-1} \rightarrow \infty$ as $n \rightarrow \infty$ we have the result

$$\|z(t_{n+1})\| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

For any $t \in [0, \infty)$, we can find an n such that $t_n \leq t < t_{n+1}$ which implies, as $t \rightarrow \infty$, $n \rightarrow \infty$. Given any t , $z(t)$ can be expressed as

$$\begin{aligned} z(t) &= T(t - t_n) z(t_n) + \int_{t_n}^t T(t - s) B u(s) ds \\ &= \left(T(t - t_n) + \int_{t_n}^t T(t - s) B K(s) ds C \right) z(t_n) \\ &= \left(T(t - t_n) + \int_0^{t-t_n} T(t - t_n - s) B \hat{K}(s) ds C \right) z(t_n). \end{aligned}$$

Because $0 \leq t - t_n < r_q$ there exists a constant $M > 0$ such that

$$\left\| T(t - t_n) + \int_0^{t - t_n} T(t - t_n - s) B \hat{K}(s) ds C \right\| \leq M$$

for all $t \geq 0$. This implies that $\|z(t)\|$ is bounded above by a monotonically decreasing function with limit zero as t goes to infinity, and thus by the “sandwich theorem”

$$\lim_{t \rightarrow \infty} \|z(t)\| = 0. \quad \square$$

Example 3.2. Let $Z \triangleq L^2[0, 1]$ and $W \triangleq L^2_{\text{loc}}[0, \infty); \mathbb{R}$. Consider the controlled diffusion equation

$$(3.8) \quad \begin{aligned} \frac{\partial z}{\partial t} &= \frac{\partial^2 z}{\partial \theta^2} + z + u(t), \quad t \geq 0, \quad 0 \leq \theta \leq 1, \\ \frac{\partial z}{\partial \theta} \Big|_{(0,t)} &= \frac{\partial z}{\partial \theta} \Big|_{(1,t)} = 0, \quad z(\cdot, 0) = z_0 \in Z, \\ y(t) &= \int_0^1 z(\theta, t) d\theta, \quad z(\cdot, t): [0, \infty) \rightarrow Z, \quad u(t) \in W. \end{aligned}$$

When we define the mapping $A: Z \rightarrow Z$ by

$$\begin{aligned} Az &= z'' + z, \\ D(A) &= \{z \mid z \in Z, z'' \in Z, z'(0) = z'(1) = 0\}, \end{aligned}$$

the mapping $B: R \rightarrow Z$ by

$$B\gamma = b(x)\gamma \quad \text{where } b(x) \equiv 1 \in Z,$$

and the mapping $C: Z \rightarrow R$ by

$$Cz = \int_0^1 z(x) dx,$$

then the controlled system (3.8) becomes

$$(3.9) \quad \begin{aligned} \frac{d}{dt} z_t &= Az_t + Bu(t), \\ y(t) &= Cz_t. \end{aligned}$$

(The operators B and C are bounded.) It is easily shown that the operator A has a pure point spectrum $\sigma(A) = \{\lambda_n \mid \lambda_n = 1 - (n\pi)^2, n = 0, 1, \dots\}$, and the normalized eigenvectors are given by

$$\{\phi_0 = 1 \text{ and } \phi_n = \sqrt{2} \cos(n\pi x), n = 1, 2, \dots\}.$$

Hence, $\{\phi_n \mid n = 0, 1, \dots\}$ is an orthonormal basis for the space Z and the semigroup $\{T(t); t \geq 0\}$ generated by A is given by

$$(3.10) \quad T(t)z = \sum_{n=0}^{\infty} e^{\lambda_n t} \phi_n(x) \langle \phi_n, z \rangle_Z.$$

If we choose to work with the mild solution, system (3.8) becomes

$$(3.11) \quad \begin{aligned} z(t) &= T(t)z_0 + \int_0^t T(t-s)Bu(s) ds, \\ y(t) &= Cz(t). \end{aligned}$$

Remark. The controlled system (3.11) is neither exactly controllable nor approximately controllable. (If we choose $b(x) = e^x$, the system will be approximately controllable.)

Now, let $Z_u = \text{span} \{\phi_0\}$ and $Z_s = \text{span} \{\phi_n, n = 1, 2, \dots\}$. If $z \in Z_u$, then z is a constant function and $Cz = \int_0^1 z \, dx = z$. Therefore (A_u, C_u) is trivially an observable pair on the space Z_u . Since the semigroup $\{T(t); t \geq 0\}$ is analytic, (A2) is satisfied. It is obvious that (A1) is satisfied. Now we show that (A3) is also satisfied.

Because Z is a Hilbert space we have $Z = Z^* = Z^{**}$. For any $z \in Z$

$$\begin{aligned} \langle z, Gu \rangle_Z &= \left\langle z, \int_0^{T_0} T(T_0 - s)Bu(s) \, ds \right\rangle_Z \\ &= \int_0^{T_0} \langle z, T(T_0 - s)Bu(s) \rangle_Z \, ds \\ &= \int_0^{T_0} \langle (T(T_0 - s)B)^*z, u(s) \rangle_R \, ds \\ &= \langle G^*z, u \rangle_W, \end{aligned}$$

i.e., $G^*z = (T(T_0 - s)B)^*z$. For any point $\gamma \in R$, (3.10) implies that

$$\begin{aligned} \langle T(T_0 - s)B\gamma, z \rangle_Z &= \left\langle \sum_{n=0}^{\infty} \phi_n e^{\lambda_n(T_0 - s)} \langle \phi_n, B\gamma \rangle, z \right\rangle_Z \\ &= \sum_{n=0}^{\infty} e^{\lambda_n(T_0 - s)} \langle \phi_n, B\gamma \rangle_Z \langle \phi_n, z \rangle_Z \\ &= \gamma \sum_{n=0}^{\infty} e^{\lambda_n(T_0 - s)} \langle \phi_n, 1 \rangle_Z \langle \phi_n, z \rangle_Z. \end{aligned}$$

Hence,

$$(T(T_0 - s)B)^*z = \sum_{n=0}^{\infty} e^{\lambda_n(T_0 - s)} \langle \phi_n, 1 \rangle_Z \langle \phi_n, z \rangle_Z.$$

If $z \in \text{Ker}(G^*)$, i.e.,

$$\sum_{n=0}^{\infty} e^{\lambda_n(T_0 - s)} \langle \phi_n, 1 \rangle_Z \langle \phi_n, z \rangle_Z = 0 \quad \text{for a.e. } s \in [0, T_0]$$

then

$$\langle \phi_n, 1 \rangle_Z \langle \phi_n, z \rangle_Z = 0 \quad \text{for all } n = 1, 2, \dots.$$

In particular, this must be true for $n = 0$; hence $\langle \phi_0, z \rangle_Z = 0$. This implies that

$$z \in \text{Ker}(\phi_0) = \bigcap_{z \in Z_u} \text{Ker}(z^{**});$$

hence by Theorem 2.2 Assumption (A3) is satisfied. We now summarize the main facts at our disposal.

(1) $\overline{R(T_0)} \supset Z_u$ for any number $T_0 > 0$. Note, there is no restriction on how small T_0 is allowed to be.

(2) Because $\dim Z_u = 1$, $(\exp(A_u\tau), C_u)$ is a trivially observable pair for any $\tau > 0$.

(3) Because $\dim Z_u = 1$, all the integers m_i , $i = 1, 2, \dots, q$ in Theorem 3.1 may be chosen equal to 1.

In the proof of Lemma 3.1.1, we can choose G to get $H = 0$, which implies that

$$T(\tau) + GC = \begin{bmatrix} 0 & GC_s \\ 0 & T_s(\tau) \end{bmatrix}.$$

Any point $z \in Z$ may be expressed in terms of the orthonormal eigenvectors, i.e., $z = \sum_{n=0}^{\infty} \alpha_n \phi_n$, $\alpha_n = \langle z, \phi_n \rangle_Z$, and $\|z\|^2 = \sum_{n=0}^{\infty} \alpha_n^2$.

$$(T(\tau) + GC)z = \begin{bmatrix} 0 & GC_s \\ 0 & T_s(\tau) \end{bmatrix} \begin{bmatrix} \alpha_0 \phi_0 \\ \sum_{n=1}^{\infty} \alpha_n \phi_n \end{bmatrix} = \begin{bmatrix} 0 \\ \sum_{n=1}^{\infty} \alpha_n e^{\lambda_n \tau} \phi_n \end{bmatrix}$$

implies that

$$\|(T(\tau) + GC)z\|_Z^2 = \sum_{n=1}^{\infty} \alpha_n^2 e^{\lambda_n 2\tau} \leq e^{2\lambda_1 \tau} \sum_{n=1}^{\infty} \alpha_n^2 \leq e^{2\lambda_1 \tau} \|z\|^2.$$

This last inequality implies that $\|T(\tau) + GC\|_Z^2 \leq e^{\lambda_1 \tau}$. Due to the fact that $\lambda_1 = 1 - \pi^2 < 0$, $e^{\lambda_1 \tau} < 1$ for any $\tau > 0$. In other words, the number τ^* in Theorem 3.1 may be taken to be arbitrarily small but positive, and the constraints imposed on the sampling time intervals in Theorem 3.1 are unnecessary for this example. Let us make the particular choice of $\tau^* = 0.08$, then $\|T(\tau_j) + G_j C\| \leq \frac{1}{4}$ for all $\tau_j \geq \tau^*$. If we choose $\hat{K}(s)$ such that

$$\left\| \int_0^{\tau_i} T(\tau_i - s) B \hat{K}(s) ds - G_i \right\| \leq \frac{1}{4} \quad \text{for } i = 1, 2, \dots, q,$$

then

$$\begin{aligned} \|z(t_{n+1})\| &= \left\| \prod_{j=1}^n \left(T(\tau_j) + \int_0^{\tau_j} T(\tau_j - s) B \hat{K}(s) ds C \right) z_0 \right\| \\ &\leq \prod_{j=1}^n \left\| \left(T(\tau_j) + G_j C \right) + \left(\int_0^{\tau_j} T(\tau_j - s) B \hat{K}(s) ds - G_j \right) C \right\| \|z_0\| \\ &\leq \left(\frac{1}{4} + \frac{1}{4} \right)^n \|z_0\| \\ &= \left(\frac{1}{2} \right)^n \|z_0\|. \end{aligned}$$

(For any $z \in Z$, $\|Cz\| = |\langle \phi_0, z \rangle| \leq \|z\|$; hence we know that $\|C\| \leq 1$. Note that we did not need to use Lemma 3.1.2 to obtain this result.)

At this point we make the following conclusion.

For any given sequence $0 < r_1 < r_2 < \dots < r_q < \infty$, there exists a feedback gain function $\hat{K} \in L^2(0, r_q; R)$ such that for any sampling time sequence $\{t_n\}_{n=0,1,\dots}$ which satisfies $\tau_n = t_n - t_{n-1} \in \{r_1, \dots, r_q\}$ for all $n = 1, 2, \dots$, the closed-loop system with the feedback control

$$u(t) = K(t)y(t_n),$$

$$K(t) = \hat{K}(t - t_n)$$

is stable in the sense that $\lim_{t \rightarrow \infty} \|z(t, z_0, u(t))\| = 0$ for any initial condition $z_0 \in Z$. (The function \hat{K} may be chosen to be piecewise constant, hence the control can be implemented exactly using a digital computer.)

In the case of finite-dimensional systems of the form (1.1) we can prove an interesting, stronger result which gives more freedom in the selection of the sample-time sequence. We just proved a result (Theorem 3.1) for infinite-dimensional systems which says that if we choose our sample-time sequence appropriately, then a stabilizing multirate sampled output feedback control exists. An admissible sample-time sequence allowed only a finite number of preordained intersample times. In addition when the sample-time sequence changed rate, or intersample times, the sample-time sequence was required to utilize this new intersample time for at least a certain number of consecutive samples. In the finite-dimensional case a result can be proven (Theorem 3.2) which shows that a much more general sample-time sequence may be used in designing a stabilizing multirate output feedback control. In loose terms, the intersample times must be uniformly bounded above and below and this interval can be represented as the union of closed subintervals with nonintersecting interiors which have the following significance. The intersample times are allowed to range within any of these closed subintervals at random, however, whenever the intersample-time sequence moves from one subinterval to another subinterval, then the intersample-time sequence is required to remain within this new subinterval for a certain number of consecutive samples.

THEOREM 3.2. *Assume a finite-dimensional system of the form (1.1) satisfies the following constraints:*

(i) (A, B) is controllable;

(ii) *There are positive reals m, M such that $(\exp(A\tau), C)$ is observable for all $\tau \in [m, M]$.*²

Then for this system there exists a finite set of integers n_1, n_2, \dots, n_k , a finite set of reals $0 < m = r_1 < r_2 < \dots < r_k \leq M$, and a function $\hat{K} \in L^2(0, M; \mathbb{R}^{p \times m})$ such that for any sampling sequence $\{t_n\}_{n=0,1,\dots}$ satisfying the following:

(i) $\tau_n \triangleq t_n - t_{n-1} \in [m, M]$ for all n ;

(ii) *If τ_k and τ_{k-1} are located in different intervals of the form $[r_i, r_{i+1})$ and τ_k is in interval $[r_i, r_{i+1})$, then $\{\tau_{k+j}\}_{j=0}^{n_i} \in [r_i, r_{i+1})$, the control*

$$\begin{aligned} u(t) &= K(t)y(t_n), & t_n \leq t < t_{n+1}, \\ K(t) &\triangleq \hat{K}(t - t_n) & \text{for } t \in [t_n, t_n + r_{n'}), \\ &\triangleq 0 & \text{for } t \in [t_n + r_n, t_{n+1}) \end{aligned}$$

where n' is defined by

$$r_{n'} \leq t_{n+1} - t_n < r_{n'+1}$$

achieves stability in the sense that

$$\lim_{t \rightarrow \infty} x(t, x_0, u) = 0$$

for all initial conditions $x_0 \in \mathbb{R}^n$.

Proof. Let $\mu \in (0, 1)$. For each point $d \in [m, M]$ the pair $(\exp(Ad), C)$ is observable; hence there exists a matrix $G_d \in \mathbb{R}^{n \times p}$ such that

$$(3.12) \quad \sigma(\exp(Ad) + G_d C) \subset \left\{ \lambda \in \mathbb{C} \mid |\lambda| < \frac{\mu}{2} \right\}$$

² The connection between the observability of the pair of matrices (A, C) and the observability of the related pair of matrices $(\exp(A\tau), C)$ for given $\tau > 0$ may be derived by using a result in [11] (stated in terms of controllability) and a duality argument.

and all eigenvalues of $(\exp(Ad) + G_d C)$ are distinct. Because this matrix has distinct eigenvalues there exists a nonsingular matrix Q_d , such that

$$Q_d^{-1}(\exp(Ad) + G_d C)Q_d = \Lambda_d$$

is a diagonal matrix and $\|\Lambda_d\| < \mu/2$. Let $Q_d^{-1}(\exp(As) + GC)Q_d \triangleq \Lambda'_d$. Then by continuity there exist $h_1(d) > 0$ and $\varepsilon(d) > 0$ such that³

$$(3.13) \quad \|\Lambda'_d\| < \mu/2$$

for all $s \in (d - h_1(d), d + h_1(d))$ and for any matrix G satisfying

$$\|G - G_d\| < \varepsilon(d).$$

By a similar continuity argument there exists $h_2(d) > 0$, such that

$$(3.14) \quad \|\exp(As) - I\| \leq \|G_d\|^{-1} \varepsilon(d)$$

for all $s \in (-h_2(d), +h_2(d))$. If we define $h(d)$ by $h(d) \triangleq \min(h_1(d), h_2(d))$, then for every $s \in (d - h(d), d + h(d))$, (3.13) holds for every $s \in (-h(d), +h(d))$, (3.14) holds.

The set of intervals $(d - h(d)/2, d + h(d)/2) | d \in [r_1, M]$ forms an open cover of the compact set $[r_1, M]$; hence there must exist a finite subcover which we denote by

$$\{(d_i - h_i/2, d_i + h_i/2)\}_{i=1}^k$$

where $h_i \triangleq h(d_i)$. Similarly, we shall denote G_{d_i} , Q_{d_i} , $\varepsilon(d_i)$, and Λ'_{d_i} by G_i , Q_i , ε_i , and Λ'_i , respectively. For each integer $i = 1, 2, \dots, k$,

$$(\exp(As) + GC) = Q_i \Lambda'_i(s) Q_i^{-1}$$

where the matrix $\Lambda'_i(s)$ satisfies $\|\Lambda'_i(s)\| < \mu/2$ for all $s \in I_i \triangleq (d_i - h_i/2, d_i + h_i/2)$ and for all G satisfying $\|G - G_i\| \leq \varepsilon_i$. This implies that

$$\begin{aligned} \|(\exp(As) + GC)^n\| &= \|Q_i \Lambda_i'^n Q_i^{-1}\| \\ &\leq \|Q_i\| \|Q_i^{-1}\| (\mu/2)^n. \end{aligned}$$

Hence there is an integer n_i such that

$$\|(\exp(As) + GC)^n\| \leq \mu^n$$

for all $n \geq n_i$, all $s \in (d_i - h_i/2, d_i + h_i/2)$ and all G with $\|G - G_i\| \leq \varepsilon_i$.

Without loss of generality, assume there is no open interval which is contained in any other interval in the chosen subcover. This means the points $\{d_i\}_{i=1,2,\dots,k}$ satisfy

$$0 < d_1 < d_2 < \dots < d_k \leq M.$$

For each integer $i = 1, 2, \dots, k$, we choose a point $r_{i+1} \in I_i \cap I_{i+1}$ such that $d_i < r_{i+1} < d_{i+1}$. These points are illustrated in Fig. 1.

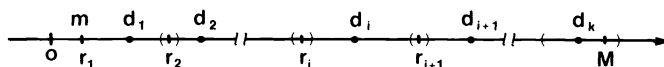


FIG. 1. The finite cover and the partition of the interval.

³ The matrix Λ'_d actually depends on s and in addition is not necessarily a diagonal matrix for all s for which it is defined.

Throughout the remainder of this paper every matrix norm will be assumed to be a spectral norm, unless stated otherwise.

Because (A, B) is controllable, for any matrix $Q \in R^{n \times m}$ and reals h_1 and h_2 there exists $K \in L^2(h_1, h_2; R^{p \times m})$ such that

$$\int_{h_1}^{h_2} \exp(A(h_2-s))BK(s) ds = Q.$$

In particular, for each integer $i = 1, 2, \dots, k$, there exists $K^i \in L^2(r_{i-1}, r_i; R^{p \times m})$ such that

$$\int_{r_{i-1}}^{r_i} \exp(A(r_i-s))BK^i(s) ds = G_i - \exp(A(r_i-r_{i-1}))G_{i-1}.$$

(We define G_0, r_0 , and t_0 by $G_0 = 0, r_0 = 0$, and $t_0 = 0$.) If we define \hat{K} by

$$\begin{aligned} \hat{K}(t) &= K^i(t) \quad \text{for } t \in (r_{i-1}, r_i], \\ &= 0 \quad \text{for } t \in (r_k, M], \end{aligned}$$

then we can prove the following technical lemma.

LEMMA 3.2.1. For each integer $i = 1, 2, \dots, k$

$$\int_0^{r_i} \exp(A(r_i-s))B\hat{K}(s) ds = G_i.$$

Proof. If $i = 0$, the conclusion holds by definition. Suppose that the conclusion is true for an arbitrary but fixed integer $i-1$; then

$$\begin{aligned} & \int_0^{r_i} \exp(A(r_i-s))B\hat{K}(s) ds \\ &= \int_0^{r_{i-1}} \exp(A(r_i-s))B\hat{K}(s) ds + \int_{r_{i-1}}^{r_i} \exp(A(r_i-s))B\hat{K}(s) ds \\ &= \exp(A(r_i-r_{i-1})) \int_0^{r_{i-1}} \exp(A(r_{i-1}-s))B\hat{K}(s) ds + G_i \\ & \quad - \exp(A(r_i-r_{i-1}))G_{i-1} \\ &= \exp(A(r_i-r_{i-1}))G_{i-1} + G_i - \exp(A(r_i-r_{i-1}))G_{i-1} \\ &= G_i \end{aligned}$$

and by the principle of induction the results holds for any integer $i = 1, 2, \dots, k$. \square

Suppose that we choose any sampling sequence $\{t_n\}_{n=0,1,\dots}$ satisfying $\tau_n \triangleq t_n - t_{n-1} \in [r_1, M]$ for all integers $n = 1, 2, \dots$ and the multirate output feedback control is given by

$$\begin{aligned} u(t) &= K(t)y(t_n), \quad t_n \leq t < t_{n+1}, \\ K(t) &\triangleq \hat{K}(t-t_n) \quad \text{for } t \in [t_n, t_n+r_{n'}), \\ &\triangleq 0 \quad \text{for } t \in [t_n+r_{n'}, t_{n+1}) \end{aligned}$$

where n' is defined by

$$r_{n'} \leq t_{n+1} - t_n < r_{n'+1},$$

then the state of the closed-loop system at the sampling time t_{n+1} can be expressed by

$$\begin{aligned}
 x(t_{n+1}) &= \exp(A t_{n+1}) x_0 + \int_0^{t_{n+1}} \exp(A(t_{n+1} - s)) B u(s) ds \\
 &= \exp(A(t_{n+1} - t_n)) \left[\exp(A t_n) x_0 + \int_0^{t_n} \exp(A(t_n - s)) B u(s) ds \right] \\
 &\quad + \int_{t_n}^{t_{n+1}} \exp(A(t_{n+1} - s)) B u(s) ds \\
 (3.15) \quad &= \left[\exp(A(t_{n+1} - t_n)) + \int_{t_n}^{t_{n+1}} \exp(A(t_{n+1} - s)) B K(s) ds C \right] x(t_n) \\
 &= \prod_{j=1}^n \left[\exp(A \tau_j) + \int_0^{\tau_j} \exp(A(\tau_j - s)) B \hat{K}(s) ds C \right] x_0.
 \end{aligned}$$

For each integer j , there is an integer i , such that $\tau_j \triangleq t_{j+1} - t_j \in [r_i, r_{i+1})$, i.e., $r_i < d_i < r_{i+1} < d_{i+1}$. The relative positions of these points is illustrated in Fig. 2. Consequently,

$$\begin{aligned}
 \int_0^{\tau_j} \exp(A(\tau_j - s)) B \hat{K}(s) ds &= \int_0^{r_i} \exp(A(\tau_j - s)) B \hat{K}(s) ds \\
 &= \exp(A(\tau_j - r_i)) \int_0^{r_i} \exp(A(r_i - s)) B \hat{K}(s) ds \\
 &= \exp(A(\tau_j - r_i)) G_i.
 \end{aligned}$$

Because $\tau_j - r_i < h_i$,

$$\begin{aligned}
 \left\| \int_0^{\tau_j} \exp(A(\tau_j - s)) B \hat{K}(s) ds - G_i \right\| &= \|(\exp(A(\tau_j - r_i)) - I) G_i\| \\
 &\leq \|\exp(A(\tau_j - r_i)) - I\| \|G_i\| \\
 &\leq \varepsilon_i.
 \end{aligned}$$

Since $\tau_j \in (d_i - (h_i/2), d_i + (h_i/2))$, the corresponding factor in the product (3.15) satisfies the inequality

$$\left\| \left(\exp(A \tau_j) + \int_0^{\tau_j} \exp(A(\tau_j - s)) B \hat{K}(s) ds C \right)^n \right\| \leq \mu^n$$

for all $n \geq n_i$. The elements of the set

$$\left\{ \left\| \prod_{j=1}^n \left(\exp(A \tau_j) + \int_0^{\tau_j} \exp(A(\tau_j - s)) B \hat{K}(s) ds C \right) \right\| \mid \tau_j \in [r_i, r_{i+1}) \right. \\
 \left. \text{for all } j = 1, 2, \dots, n, n < n_i, i = 1, 2, \dots, k \right\}$$

are uniformly bounded by virtue of the fact that the elements of the set $\{\tau_i, i = 1, 2, \dots\}$ are uniformly bounded. This means there exists a constant $M > 0$, independent of n , such that

$$\|x(t_{n+1})\| \leq M \|x_0\| \mu^m$$

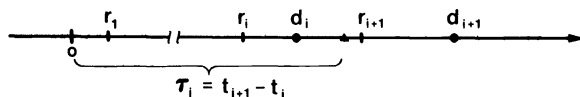


FIG. 2. Intersampling time and the partition of the interval.

where $m \rightarrow \infty$ as $n \rightarrow \infty$. This implies that

$$\lim_{n \rightarrow \infty} \|x(t_{n+1})\| = 0.$$

The proof that $\lim_{t \rightarrow \infty} \|x(t)\| = 0$ proceeds as in the end of the proof of Theorem 3.1. \square

We now present an algorithm for computing multirate sampled output feedback gains. In this algorithm we shall use the L^1 norm on R^n , i.e., $\|x\| = \sum_{i=1}^n |x_i|$ for any $x \in R^n$. For any element E of $R^{m \times n}$ we will define $\|E\|$ to be the maximum of the absolute values of all the elements of E . (There is no loss of generality in choosing these particular norms. The choice was made for reasons of simplicity and convenience.)

Algorithm for computing multirate feedback gains. Let m , M , and $\mu \in (0, 1)$ be fixed ($m < M$).

- Step 0.* Let $r_1 = m$, $k = 1$, choose μ' , μ'' such that $0 < \mu'' < \mu' < \mu$.
- Step 1.* Choose G_k such that $\sigma(e^{Ar_k} + G_k C) = \{\lambda_1, \dots, \lambda_n\}$, where the points λ_i are distinct and $|\lambda_i| < \mu''$.
- Step 2.* Determine the orthogonal matrix $Q_k \in R^{n \times n}$ such that $Q_k^{-1}(e^{Ar_k} + G_k C)Q_k = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$.
- Step 3.* Compute $\|G_k\|$, $\|Q_k^{-1}\|$ and $\|Q_k\|$.
- Step 4.* Choose $\varepsilon_k > 0$ and $h_k > 0$ so that
- (1) $\|e^{As} - I\| \leq \|G_k\|^{-1} \varepsilon_k$ for all $s \in [-h_k, h_k]$.
 - (2) $\|Q_k^{-1}(e^{As} + GC)Q_k\| \leq \mu'$ for all $s \in [r_k, r_k + h_k]$ and all G satisfying $\|G - G_k\| \leq \varepsilon_k$.
- Step 5.* Determine the smallest integer n_k such that $\|Q_k^{-1}\| \|Q_k\| (\mu')^{n_k} \leq \mu^{n_k}$.
- Step 6.* If $m + h_1 + h_2 + \dots + h_k \geq M$, then go to Step 7; else let $r_{k+1} = r_k + h_k$, increment k by 1 and go to Step 1.
- Step 7.* Determine the functions $K^i \in L^2(r_{i-1}, r_i; R^{p \times m})$ such that $\int_{r_{i-1}}^{r_i} e^{A(r_i-s)} B K^i(s) ds = G_i - e^{A(r_i-r_{i-1})} G_{i-1}$ for each integer $i = 1, 2, \dots, k$. The feedback gain function is given by

$$\begin{aligned} \hat{K}(t) &= K^i(t) \quad t \in (r_{i-1}, r_i], \\ &= 0, \quad t \in (r_k, M]. \end{aligned}$$

The particular feedback gain function computed by the algorithm depends on how computations are done at each step. This presents a great amount of latitude in designing an algorithm for a specific problem.

In Theorems 3.1 and 3.2, multirate sampled output feedback control was shown to provide a stabilizing control scheme which allows considerable freedom in the choice of sample times and yet admits a gain function K which is independent of the particular way in which sample times are chosen. Both theorems imply we can choose sample times randomly. In the infinite-dimensional case we can randomly progress from one sample-rate to another provided we stay with the new choice for a preordained number of successive samples. In the finite-dimensional case there is a finite set of bounded adjacent sampling intervals from which we can choose sample rates randomly, provided each time a choice is made which crosses into a new sampling interval a certain number of successive choices come from this interval. The gain function K in both cases can be computed by translations of a canonical gain function \hat{K} which means in practice we can store a representation of \hat{K} , on a bounded interval and use these stored values cyclicly. Another important aspect of this control scheme is, there is much latitude in our choice of the class of functions that the gain function comes from, as long as the class of functions is dense in L^2 . For example, we can choose a

gain function which is piecewise constant. This would be a natural choice for a digital computer implementation. On the other hand we might wish a control scheme which is very smooth in its action, in which case we could restrict the gain function to reside in the class of analytic functions which vanish at the endpoints of the interval of interest.

Example 3.3. In this section we will illustrate the computations involved in designing a multirate sampled output feedback control for a specific finite-dimensional control system of the form (1.1). The system we will consider is defined by

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = (1 \quad 0).$$

Let $G = (g_1, g_2)^T$; then

$$\exp(At) = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}, \quad \exp(At) + GC = \begin{bmatrix} g_1 + 1 & t \\ g_2 & 1 \end{bmatrix}$$

and $\det[\lambda I - \exp(At) - GC] = \lambda^2 - (g_1 + 2)\lambda + (g_1 + 1) - tg_2$.

For this example we have computed a multirate feedback gain by using the previous algorithm. The particular computations performed and choices made at each step are listed below. We start the procedure by assuming that $m = 1$, $M = 2$ and $\mu = 0.9$. (This example leads to algorithm Steps 2 and 3, which are trivial; hence they have been omitted.)

Step 0. $\mu' = 0.8$, $\mu'' = \frac{1}{3}$.

Step 1. $G_k = (-2, -8/9r_k)^T$, $\lambda_{1,2} = \pm \frac{1}{3}$.

Step 4. Condition (1) becomes $h_k \leq \|G_k\|^{-1} \varepsilon_k$.

Step 5. $n_k = 1$ for all k .

Step 6. The final value of $k = 14$.

Step 7. On each interval $[r_{i-1}, r_i)$, let

$$\begin{aligned} K^i(t) &= \alpha_i, & t \in [r_{i-1}, (r_{i-1} + r_i)/2), \\ &= \beta_i, & t \in [(r_{i-1} + r_i)/2, r_i). \end{aligned}$$

From the expression

$$\begin{aligned} \int_{r_{i-1}}^{r_i} \exp(A(r_i - s))BK^i(s) ds &= \int_{r_{i-1}}^{r_i} \begin{pmatrix} r_i - s \\ 1 \end{pmatrix} K^i(s) ds \triangleq \begin{pmatrix} g'_{i1} \\ g'_{i2} \end{pmatrix} \\ &= G_{i+1} - \exp(A(r_i - r_{i-1}))G_i = G_{i+1} - \begin{bmatrix} 1 & r_i - r_{i-1} \\ 0 & 1 \end{bmatrix} G_i \end{aligned}$$

we obtain the following formulas for α_i and β_i :

$$\begin{aligned} \alpha_i &= \frac{4}{(r_i - r_{i-1})^2} g'_{i1} - \frac{1}{(r_i - r_{i-1})} g'_{i2}, \\ \beta_i &= \frac{-4}{(r_i - r_{i-1})^2} g'_{i1} + \frac{3}{(r_i - r_{i-1})} g'_{i2}. \end{aligned}$$

For the particular choice of $\mu = 0.9$ for each integer i , we find the smallest integer n_i , such that

$$\|Q_i\| \|Q_i^{-1}\| (.8/.9)^{n_i} \leq 1.$$

For the problem under solution we obtain $k = 14$ and $n_i = 1$ for $i = 1, 2, 3, \dots, 14$. The numerical values for $\{r_i\}_{i=1}^{14}$, $\{\alpha_i\}_{i=1}^{14}$, $\{\beta_i\}_{i=1}^{14}$ are contained in Table 1.

TABLE 1
Numerical solution of Example 3.3.

i	r_i	α_i	β_i	i	r_i	α_i	β_i
1	1.00	-8.10	6.56	8	1.72	28.15	-27.69
2	1.09	21.92	-20.91	9	1.79	31.82	-31.40
3	1.22	21.69	-20.85	10	1.85	37.23	-36.85
4	1.34	21.98	-21.21	11	1.90	45.64	-45.30
5	1.45	22.62	-21.99	12	1.94	60.01	-59.71
6	1.55	23.81	-23.25	13	1.97	89.22	-89.00
7	1.64	25.59	-25.08	14	2.00	0.00	0.00

4. Directions for further research. An interesting area of further research is a generalization of the control technique outlined in this paper. We know that when a system of the form (1.1) is observable, then there exists a matrix G such that

$$\lim_{n \rightarrow \infty} \|(A + GC)^n\| = 0.$$

By using a continuity argument we can easily show that G can be chosen so that

$$\lim_{n \rightarrow \infty} \left\| \prod_{i=1}^n (A_i + GC) \right\| = 0$$

provided $\|A_i - A\| < \varepsilon$ for some $\varepsilon > 0$ (which depends on A and C). Example 3.1 shows that a sequence of matrices $\{E_i\}$ may have the property that

$$\lim_{n \rightarrow \infty} \left\| \prod_{i=1}^n E_i \right\|$$

does not exist when $\sigma(E_i) = \{0\}$ for all $i = 1, 2, \dots$. This does not prove it is impossible to choose G_i so that

$$\lim_{n \rightarrow \infty} \left\| \prod_{i=1}^n (A_i + G_i C) \right\| = 0$$

when (A_i, C) is observable for every integer i . This observation leads to the following conjecture.

If (A_i, C) is observable for each integer i , is it possible to choose $\{G_i\}$ such that

$$\lim_{n \rightarrow \infty} \left\| \prod_{i=1}^n (A_i + G_i C) \right\| = 0?$$

If the answer to this question is positive, then the result of Theorem 3.2 can be strengthened considerably.

Another important area of research is the integration of an objective function into the theory of multirate output feedback control. The theorems in this paper have proven that multirate output feedback will stabilize a wide variety of linear systems under mild hypotheses. Although these theorems prove the existence of stabilizing controls they say nothing about optimality. If we were to examine multirate output feedback control in the light of an objective function and random sampling corresponding to a finite-state Markov process, we would gain more insight and evaluate the practicality of this control technique.

Acknowledgment. The authors thank the referee for suggesting several improvements for the presentation of this work.

REFERENCES

- [1] A. B. CHAMMAS AND C. T. LEONDES, *On the design of linear time invariant systems by periodic output feedback. Part I, Discrete time pole assignment*, Internat. J. Control, 27 (1978), pp. 885-894.
- [2] ———, *On the design of linear time invariant systems by periodic output feedback. Part II, Output feedback controllability*, Internat. J. Control, 27 (1978), pp. 895-903.
- [3] ———, *Pole assignment by piecewise constant output feedback*, Internat. J. Control, 29 (1979), pp. 31-38.
- [4] ———, *On the finite time control of linear systems by piecewise constant output feedback*, Internat. J. Control, 30 (1979), pp. 227-234.
- [5] T.-J. TARN, J. R. ZAVGREN, JR., AND X. ZENG, *Stabilization of infinite-dimensional systems with periodic feedback gains and sampled output*, Automatica, 24 (1988), pp. 95-99.
- [6] X. ZENG, *Sampled output feedback stabilization of infinite-dimensional systems*, Ph.D. thesis, Washington University, St. Louis, MO, August 1985.
- [7] J. M. SCHUMACHER, *A direct approach to compensator design for distributed parameter systems*, SIAM J. Control Optim., 21 (1983), pp. 823-836.
- [8] T. KOBAYASHI, *Feedback stabilization of parabolic distributed parameter systems by piecewise constant output feedback*, SIAM J. Control Optim., 22 (1984), pp. 509-523.
- [9] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
- [10] A. TRIGGIANI, *On the stabilization problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383-402.
- [11] R. E. KALMAN, Y. C. HO, AND S. K. NARENDRA, *Controllability of linear dynamic systems*, Contrib. Differential Equations, 1 (1963), pp. 189-213.

SHAPE SENSITIVITY ANALYSIS VIA MIN MAX DIFFERENTIABILITY*

M. C. DELFOUR† AND J.-P. ZOLÉSIO‡

Abstract. The object of this paper is twofold. We introduce a new theorem on the differentiability of a Min Max with respect to a parameter and we show how such a theorem can be applied to compute the material derivative in shape sensitivity analysis problems. We consider the Min Max of a functional which is parametrized by t . We show that, under appropriate conditions, the derivative of the Min Max with respect to t is the Min Max with respect to the points solution of the Min Max problem of the derivative of the original functional with respect to t . To illustrate the use of this theorem, we apply it to the control of an elliptic equation with a nondifferentiable observation and to shape design problems.

Key words. shape sensitivity analysis, minimax, optimal design

AMS(MOS) subject classifications. 35R35, 49A29, 49A52

1. Introduction. Many problems in shape sensitivity analysis can be expressed as a Min Max of some Lagrangian functional which depends on the domain Ω . By introducing a velocity field of deformations V over Ω (cf. C  a [1]–[3], Zol  sio [1], [2]), a family of perturbations Ω_t , $t \geq 0$, of the domain Ω is obtained and the sensitivity analysis reduces to the study of the differentiability of a Min Max functional with respect to the parameter t for fixed velocity fields V and domains Ω .

The object of this paper is twofold. First we present a new theorem on the differentiability of the Min Max:

$$g(t) = \{\text{Min} [\text{Max } G(t, x, y): y \in \mathcal{B}]: x \in \mathcal{A}\}$$

of a functional $G(t, x, y)$ with respect to a real parameter $t \geq 0$ for some fixed subsets $\mathcal{A} \subset X$ and $\mathcal{B} \subset Y$ of two topological spaces X and Y . We show that, under appropriate hypotheses, the derivative of $g(t)$ with respect to t is the Min Max of the partial derivative $\partial_t G(t, x, y)$ with respect to all points $(x, y) \in \mathcal{A} \times \mathcal{B}$ which are solutions of the Min Max problem. The second objective is to show how such theorems can be applied to shape sensitivity analysis. Thus we give a precise mathematical justification to some results which are usually obtained formally in the literature. For instance, some interesting examples can be found in the book by Haug, Choi, and Komkov [1], Dems and Mroz [1], in the proceedings by Haug and C  a [1], and the recent paper on quick computations by C  a [1].

In order to motivate our approach we shall first consider a special and useful case in § 2 when $G(t, \cdot, \cdot)$ is a convex-concave functional with a unique saddle point (x_t^*, y_t^*) in $\mathcal{A} \times \mathcal{B}$ for each t . We show that under some reasonable hypotheses

$$\frac{dg(t)}{dt} = \partial_t G(t, x_t^*, y_t^*).$$

* Received by the editors November 4, 1986; accepted for publication (in revised form) July 15, 1987. This research was supported in part by Natural Sciences and Engineering Research Council of Canada operating grant A-8730 and a ‘‘Fonds pour la formation des chercheurs et l’aide    la recherche’’ grant from the ‘‘Minist  re de l’Education du Qu  bec.’’

† Centre de Recherches Math  matiques and D  partement de Math  matiques et Statistique, Universit   de Montr  al, Montr  al, Qu  bec, Canada H3C 3J7.

‡ Laboratoire de Physique Math  matique, Universit   des Sciences et Techniques du Languedoc, 34050-Montpellier C  dex, France.

In control theory this is a natural tool in the computation of the directional derivative of the cost function with respect to the control variable when the state is the unique solution of some partial differential equation. In fact this problem can also be expressed as the Min Max of an appropriate Lagrangian. In that context, the existence and uniqueness of the saddle point is equivalent to the well-posedness, existence and uniqueness of the solution to the associated adjoint problem. A very simple nonlinear example will be given in § 3 to illustrate this point.

In § 4 we give the main result where $G(t, \cdot, \cdot)$ is no longer assumed to be a convex-concave functional with saddle points. It is a slight generalization of the earlier result by Delfour and Zolésio [1], [2]. It also extends the work of Zolésio [3, Thm 1.1, p. 1458] on the differentiability of a Min or a Max with respect to a parameter in the shape sensitivity analysis context.

This theorem and its eventual generalizations have many interesting applications. To illustrate that point we describe the associated techniques for two examples. The first, in § 5, is a control or identification problem with a nondifferentiable observation which depends on the state which is the solution of an elliptic equation which itself depends on the control function u . The second example, in § 6, is a shape analysis problem where this technique makes it possible to completely bypass the problem of the existence and interpretation of the Eulerian or material derivative of the state. These two simple examples are given for the purpose of illustration. However, the techniques used here apply to the general linear case and some nonlinear situations (cf. § 7). A first version of the main theorem and its application to the examples of §§ 5 and 6 have been announced in Delfour and Zolésio [1], [2]. They generalize the earlier finite-dimensional result of Dem'yanov [1] for compact subsets \mathcal{A} and \mathcal{B} , G continuous in its arguments, and $\partial_t G(t, x, y)$ bounded on $[0, \tau] \times \mathcal{A} \times \mathcal{B}$, $\tau > 0$, and continuous with respect to t . Another interesting result has been obtained by Correa and Seeger [1] when the functional $G(t, x, y)$ has saddle points for t in $[0, \tau]$ for some $\tau > 0$. It contains as a special case the theorem given in § 2 and it is discussed in § 4 after the presentation of our main result.

Notation. R will denote the field of real numbers, R^+ the subset of positive or zero reals, and R^n ($n \geq 1$, an integer) the n -fold Cartesian product of R . The inner product and norm in R^n will be defined as

$$x \cdot y = \sum_{i=1, n} x_i y_i, \quad |x| = (x \cdot x)^{1/2}.$$

The dual operator of a continuous linear operator $A: X \rightarrow Y$ will be denoted by A^* . The identity matrix in R^n will be written I_d . The composition of two applications f and g will be denoted by $f \circ g$.

2. Derivative of a Min Max with a unique saddle point with respect to a parameter.

Let $\mathcal{A} \subset X$ and $\mathcal{B} \subset Y$ be two nonempty subsets of two topological spaces X and Y , let $\tau > 0$ be a real number, and let

$$(1) \quad t, x, y \rightarrow G(t, x, y): [0, \tau] \times \mathcal{A} \times \mathcal{B} \rightarrow R$$

be a functional which is differentiable with respect to t . Denote by $\partial_t G(t, x, y)$ its partial derivative with respect to t .

THEOREM 1. *Let \mathcal{A} , \mathcal{B} , and G be given as in (1) with G differentiable with respect to t for each $x \in \mathcal{A}$ and $y \in \mathcal{B}$. Assume that for each t in $[0, \tau]$, the functional $G(t, \cdot, \cdot)$ has a unique saddle point (x_t, y_t) and that the following set of hypotheses is verified.*

(HA) *There exist topologies τ_X on X and τ_Y on Y such that*

(i) the map

$$(2) \quad t \rightarrow (x_t, y_t) : [0, \tau[\rightarrow (X, \tau_X) \times (Y, \tau_Y)$$

is continuous (see also Remark 2.2);

(ii) $\forall y \in \mathcal{B}, (t, x) \rightarrow \partial_t G(t, x, y) : [0, \tau[\times (X, \tau_X) \rightarrow \mathbb{R}$
is lower semicontinuous;

(iii) $\forall x \in \mathcal{A}, (t, y) \rightarrow \partial_t G(t, x, y) : [0, \tau[\times (Y, \tau_Y) \rightarrow \mathbb{R}$
is upper semicontinuous.

Then the function $g(t) = G(t, x_t, y_t)$ is differentiable from the right and

$$(3) \quad dg(t) = \lim_{s \rightarrow 0^+} [g(t+s) - g(t)]/s = \partial_t G(t, x_t, y_t).$$

Proof. Write the saddle point conditions at $t+s$ and t for some $s > 0$:

$$(4) \quad G(t+s, x_{t+s}, y) \leq G(t+s, x_{t+s}, y_{t+s}) \leq G(t+s, x, y_{t+s})$$

for all x in \mathcal{A} and y in \mathcal{B} , and

$$(5) \quad -G(t, v, y_t) \leq -G(t, x_t, y_t) \leq -G(t, x_t, w)$$

for all v in \mathcal{A} and w in \mathcal{B} . Denote by ΔG the difference

$$G(t+s, x_{t+s}, y_{t+s}) - G(t, x_t, y_t).$$

Choose

$$v = x_{t+s}, \quad y = y_t, \quad x = x_t, \quad w = y_{t+s}$$

and add (4) and (5). We get

$$(6) \quad G(t+s, x_{t+s}, y_t) - G(t, x_{t+s}, y_t) \leq \Delta G \leq G(t+s, x_t, y_{t+s}) - G(t, x_t, y_{t+s}).$$

By the Mean Value Theorem, (6) can be rewritten as

$$(7) \quad \partial_t G(t + \theta_1 s, x_{t+s}, y_t) \leq \Delta G/s \leq \partial_t G(t + \theta_2 s, x_t, y_{t+s})$$

for some θ_1 and θ_2 in $]0, 1[$. By going to the limits as s goes to zero we get

$$(8) \quad \partial_t G(t, x_t, y_t) \leq \liminf \Delta G/s \leq \limsup \Delta G/s \leq \partial_t G(t, x_t, y_t)$$

by using hypothesis (HA). \square

Remark 2.1. It is important to emphasize that no differentiability of the map $t \rightarrow (x_t, y_t)$ is required to differentiate $g(t)$. So no implicit function theorem is necessary to obtain an equation for the derivatives of x_t and y_t with respect to t . The continuity of the saddle point with respect to t is sufficient.

Remark 2.2. When hypotheses (HA)(ii) and (HA)(iii) are both strengthened from lower and upper semicontinuity to continuity, then g is differentiable for all $t > 0$ in a neighborhood of $t = 0$:

$$(3') \quad \frac{dg(t)}{dt} = \lim_{s \rightarrow 0} [g(t+s) - g(t)]/s = \partial_t G(t, x_t, y_t), \quad t > 0.$$

This is readily seen if we take s positive or negative in the proof of Theorem 1.

Remark 2.3. In Theorem 1, the continuity hypothesis (HA)(i) can be modified in the following way: there exists $\tau > 0$ such that for all t_0 in $[0, \tau]$:

(i) There exists compact $K_0 \subset X \times Y$ (with respect to the $\tau_X \times \tau_Y$ -topology) such that for all $t_n \rightarrow t_0$, $t_n > t_0$, there exists a subsequence of $\{t_n\}$, still denoted $\{t_n\}$, such that for all n , $(x_{t_n}, y_{t_n}) \in K_0$;

(ii) For all $y \in \mathcal{B}$, $(t, x) \rightarrow G(t, x, y)$ is lower semicontinuous at (t_0, x_{t_0}) ;

(iii) For all $x \in \mathcal{A}$, $(t, y) \rightarrow G(t, x, y)$ is upper semicontinuous at (t_0, y_{t_0}) .

Under this hypothesis we prove Theorem 1 by using convergent subsequences in K_0 and the following lemma.

LEMMA 1. Assume that the above three hypotheses are verified. Let $t_n \rightarrow t_0$, $t_n > t_0$, and denote by (x^*, y^*) in $X \times Y$ a limit point of (x_{t_n}, y_{t_n}) . Then $(x^*, y^*) = (x_{t_0}, y_{t_0})$. \square

In order to make Theorem 1 a computational tool, it is necessary to obtain a characterization of the saddle point. This can obviously be done in several ways and independently. To illustrate that point we give a standard set of hypotheses under which the saddle point is easily characterized.

Assume that

- (HB) (i) For all t , $x \rightarrow G(t, x, y)$ is convex and Gâteaux differentiable;
 (ii) For all t , $y \rightarrow G(t, x, y)$ is concave and Gâteaux differentiable.

We quote the following proposition from Ekeland and Temam [1, Props. 1.6, 1.7, pp. 157–158].

PROPOSITION 1. Assume that \mathcal{A} and \mathcal{B} are convex subsets of two Banach spaces X and Y and that hypothesis (HB) is verified. Then for each t , (x_t, y_t) is the unique saddle point of $G(t, \cdot, \cdot)$ if and only if:

(HC) There exists a unique solution to the following system of inequalities:

$$(9) \quad \exists x_t \in \mathcal{A} \quad \forall x \in \mathcal{A}, \quad dG(t, x_t, y_t; x - x_t, 0) \geq 0,$$

$$(10) \quad \exists y_t \in \mathcal{B} \quad \forall y \in \mathcal{B}, \quad dG(t, x_t, y_t; 0, y - y_t) \leq 0$$

where when it exists

$$dG(t, x, y; v, w) = \lim_{s \rightarrow 0^+} [G(t, x + sv, y + sw) - G(t, x, y)]/s. \quad \square$$

So under the hypotheses of Proposition 1 and hypotheses (HB) and (HC), there exists a unique saddle point for each t and under hypothesis (HA) the derivative of $g(t)$ is given by (3).

Remark 2.4. When \mathcal{A} and \mathcal{B} are two linear subspaces of X and Y , respectively, then (9) and (10) are equivalent to

$$(9') \quad \exists x_t \in \mathcal{A} \quad \forall x \in \mathcal{A}, \quad dG(t, x_t, y_t; x, 0) = 0,$$

$$(10') \quad \exists y_t \in \mathcal{B} \quad \forall y \in \mathcal{B}, \quad dG(t, x_t, y_t; 0, y) = 0.$$

3. An immediate application to nonlinear control. We have seen at the end of § 2 that under hypotheses (HA), (HB)–(HC), the derivative of the function $g(t)$ is completely characterized by (2.3) provided that system (2.9)–(2.10) has a unique solution.

The above results are readily applicable to classes of nonlinear control problems. Equation (2.9) is usually the state equation which is independent of the y -variable. Equation (2.10) is the adjoint equation and hypothesis (HC) reduces to the well-posedness of that last equation.

So let U , X , and Y be three Banach spaces, U_{ad} a convex subset of U and for each u in U_{ad}

$$(1) \quad A_u: X \rightarrow Y'$$

a continuous linear mapping from X into the topological dual Y' of Y . Given f in $\text{Im}(A_u) \subset Y'$ and the convex, differentiable functional $F: X \rightarrow \mathbb{R}$, consider the problem of differentiating the functional

$$(2) \quad J(u) = F(x(u))$$

where we assume that, for each u in U , $x = x(u)$ is the unique solution in X of the equation

$$(3) \quad A_u x = f.$$

To recast this problem in our framework, introduce the Lagrangian functional

$$(4) \quad \underline{G}(u, x, y) = F(x) + \langle A_u x - f, y \rangle_Y$$

where $\langle \cdot, \cdot \rangle_Y$ denotes the duality pairing between Y' and Y . It is well known that the cost function can be rewritten:

$$(5) \quad J(u) = \text{Min}\{\text{Sup}[\underline{G}(u, x, y): y \in Y]: x \in X\}.$$

We can always formally introduce the adjoint problem:

$$(6) \quad \exists y(u) \in Y \quad \forall x' \in X, \quad dF(x(u); x') + \langle A_u x', y(u) \rangle_Y = 0.$$

Following § 2, assume that for all (x, y) in $X \times Y$, the mapping

$$(7) \quad x \rightarrow \langle A_u x, y \rangle_Y$$

has a Gâteaux derivative

$$(8) \quad \langle A'_{u,v} x, y \rangle_Y$$

in all admissible directions v in $U(u \in U_{\text{ad}}, t > 0 \text{ such that } u + tv \in U_{\text{ad}})$. Define

$$(9) \quad G(t, x, y) = \underline{G}(u + tv, x, y)$$

and assume that

$$(10) \quad \partial_t G(t, x, y) = \langle A'_{u+tv,v} x, y \rangle_Y$$

verifies hypothesis (HA) for the weak topologies of X and Y .

THEOREM 2. Assume that problems (3) and (6) are well posed (existence of a unique solution $(x(u), y(u)) \in X \times Y$ for each u in U_{ad}) and that the map

$$(11) \quad u \rightarrow (x(u), y(u)): U_{\text{ad}} \rightarrow (X\text{-weak}) \times (Y\text{-weak})$$

is continuous. Then the cost function $J(u)$ defined by (2) is Gâteaux differentiable and for any admissible direction v :

$$(12) \quad dJ(u; v) = \langle A'_{u,v} x(u), y(u) \rangle_Y. \quad \square$$

Consider the following illustrative example. Let Ω be a smooth bounded domain in R^N with boundary $\partial\Omega = \Gamma_0 \cup \Gamma_1$:

$$X = \{\varphi \in H^1(\Omega): \varphi = 0 \text{ on } \Gamma_0\}, \quad U = L^\infty(\Gamma_1), \quad U_{\text{ad}} = \{u \in U: \exists \alpha, u(x) \geq \alpha > 0 \text{ a.e.}\}.$$

To each u in U_{ad} , we associate the boundary value problem

$$(13) \quad \begin{aligned} -\Delta \varphi(u) &= 0 \quad \text{in } \mathcal{D}'(\Omega), & \varphi(u) &\in X, \\ \partial \varphi(u) / \partial n + u \varphi(u) &= g \quad \text{on } \Gamma_1 \end{aligned}$$

for a fixed g in $L^2(\Gamma_1)$.

We can associate with problem (13) the minimization over X of the energy functional

$$(14) \quad E(u, \varphi) = \frac{1}{2} \int_{\Omega} |\nabla \varphi|^2 dx + \int_{\Gamma_1} \left[\left(\frac{u}{2} \right) \varphi^2 - g \varphi \right] d\Gamma.$$

For each $x(u)$ define the cost function

$$(15) \quad J(u) = \frac{1}{2} \int_{\Gamma_1} \left[\frac{\partial \varphi(u)}{\partial n} \right]^2 d\Gamma = \frac{1}{2} \int_{\Gamma_1} [g - u\varphi(u)]^2 d\Gamma$$

and the Lagrangian functional $G: [0, \tau] \times X \times X \rightarrow \mathbb{R}$,

$$(16) \quad \begin{aligned} G(t, \varphi, \psi) &= F(\varphi) + dE(u + tv, \varphi; \psi) \\ &= \frac{1}{2} \int_{\Gamma_1} [g - (u + tv)\varphi]^2 d\Gamma + \int_{\Omega} \nabla \varphi \cdot \nabla \psi dx \\ &\quad + \int_{\Gamma_1} [(u + tv)\varphi - g]\psi d\Gamma. \end{aligned}$$

Apply Theorem 2 to get

$$(17) \quad \begin{aligned} dJ(u; v) &= \partial_t G(0, \varphi(u), \psi(u)) \\ &= \int_{\Gamma_1} [-g + u\varphi(u) + \psi(u)]\varphi(u)v d\Gamma \end{aligned}$$

where $\varphi(u)$ is the solution of (13) and $\psi = \psi(u) \in X$ is characterized by

$$(18) \quad dG(0, \varphi(u), \psi; \varphi, 0) = 0 \quad \forall \varphi \in X$$

or

$$(19) \quad \begin{aligned} &= \int_{\Gamma_1} -u[g - u\varphi(u)]\varphi d\Gamma + \int_{\Omega} \nabla \varphi \cdot \nabla \psi dx + \int_{\Gamma_1} u\varphi\psi d\Gamma \\ &= 0 \quad \forall \varphi \in X. \end{aligned}$$

But this is equivalent to

$$(20) \quad \begin{aligned} \Delta \psi &= 0 \quad \text{in } \mathcal{D}'(\Omega), \\ \partial \psi / \partial n + u\psi &= u\partial \varphi(u) / \partial n (= u(g - u\varphi) \in L^2(\Gamma_1)) \end{aligned}$$

which is a well-posed problem. To completely justify (17) it suffices to check the weak continuity of $t \rightarrow \varphi(u + tv)$ and $\psi(u + tv)$ which is easily done.

4. Derivative of a Min Max with respect to a parameter. Let $\mathcal{A} \subset X$ and $\mathcal{B} \subset Y$ be subsets of two topological spaces X and Y and let $\tau > 0$ be a real number. Given a map

$$G: [0, \tau] \times X \times Y \rightarrow \mathbb{R},$$

we consider the following functions:

$$(1) \quad H(t, x) = \text{Sup} \{G(t, x, y): y \in \mathcal{B}\}, \quad t \in [0, \tau], \quad x \in \mathcal{A},$$

$$(2) \quad g(t) = \text{Inf} \{H(t, x): x \in \mathcal{A}\}, \quad t \in [0, \tau].$$

As a result

$$(3) \quad g(t) = \text{Inf} \{\text{Sup} [G(t, x, y): y \in \mathcal{B}]: x \in \mathcal{A}\}.$$

We wish to show that under appropriate hypotheses the function g is differentiable at $t = 0$ from the right:

$$(4) \quad \lim_{t \rightarrow 0^+} (g(t) - g(0))/t \quad \text{exists.}$$

For $t \geq 0$ we shall need the set

$$(5) \quad A(t) = \{x \in \mathcal{A}: g(t) = H(t, x)\}$$

and for x in \mathcal{A} the set

$$(6) \quad B(t, x) = \{y \in \mathcal{B}: H(t, x) = G(t, x, y)\}.$$

In order to better see the role of each hypothesis in the final result, we proceed in a step-by-step fashion. We first introduce hypotheses to ensure that the Sup and Inf problems have solutions.

- (H1) There exists $\tau > 0$, for all t , $0 \leq t \leq \tau$:
- (i) $A(0) \neq \emptyset$, for all $x_0 \in A(0)$, $B(t, x_0) \neq \emptyset$,
 - (ii) $A(t) \neq \emptyset$, for all $x_t \in A(t)$, $B(0, x_t) \neq \emptyset$.

LEMMA 2. Under hypothesis (H1) we have the following estimates: for all t , $0 \leq t \leq \tau$

$$(7) \quad g(t) - g(0) \leq G(t, x_0, y^*) - G(0, x_0, y^*) \quad \forall x_0 \in A(0) \quad \forall y^* \in B(t, x_0)$$

and

$$(8) \quad g(t) - g(0) \geq G(t, x_t, z^*) - G(0, x_t, z^*) \quad \forall x_t \in A(t) \quad \forall z^* \in B(0, x_t).$$

Proof. The proof uses standard arguments and will be omitted. \square

- (9) In a second step we obtain upper and lower bounds on the differential quotient

$$\frac{(g(t) - g(0))}{t}.$$

We need the following additional hypothesis:

- (H2) (i) For all $x_0 \in A(0)$, the function

$$(10) \quad s \rightarrow G(s, x_0, y)$$

is differentiable in a neighborhood of $t=0$ for all y in

$$\cup \{B(t, x_0): 0 \leq t \leq \tau\};$$

- (ii) For all t , $0 \leq t \leq \tau$, for all $x_t \in A(t)$, the function

$$s \rightarrow G(s, x_t, y)$$

is differentiable in a neighborhood of $t=0$ for all y in

$$\cup \{B(0, x_t): 0 \leq t \leq \tau\}.$$

LEMMA 3. Under hypotheses (H1) and (H2), for each t , $0 < t \leq \tau$:

- (i) There exists θ_1 , $0 < \theta_1 < 1$, such that

$$(11) \quad \frac{(g(t) - g(0))}{t} \leq \partial_t G(\theta_1 t, x_0, y^*) \quad \forall x_0 \in A(0) \quad \forall y^* \in B(t, x_0);$$

- (ii) There exists θ_2 , $0 < \theta_2 < 1$, such that

$$(12) \quad \frac{(g(t) - g(0))}{t} \geq \partial_t G(\theta_2 t, x_t, z^*) \quad \forall x_t \in A(t) \quad \forall z^* \in B(0, x_t).$$

Proof. (i) For $x_0 \in A(0)$ and $y^* \in B(t, x_0)$ define

$$\lambda(s) = G(st, x_0, y^*).$$

By hypothesis (H2), λ is differentiable in a neighborhood of 0. So by Taylor's Theorem there exists $\theta_1 \in]0, 1[$ such that

$$\lambda(1) = \lambda(0) + \frac{d\lambda}{ds}(\theta_1)$$

and

$$G(t, x_0, y^*) - G(0, x_0, y^*) = t \partial_t G(\theta_1 t, x_0, y^*)$$

where $\partial_t G$ denotes the partial derivative of G with respect to the first argument. The proof of part (ii) is similar and will be omitted. \square

In the next step we go to the limit in (11) and (12) as t goes to 0. So we introduce

$$(13) \quad \bar{d}g(0) = \limsup_{t \rightarrow 0^+} \frac{(g(t) - g(0))}{t}, \quad dg(0) = \liminf_{t \rightarrow 0^+} \frac{(g(t) - g(0))}{t}.$$

They are the smallest upper and greatest lower bounds of the differential quotient in R . So there exist sequences $\{t_n\}$ and $\{t'_n\}$ of positive numbers in $]0, \tau]$ going to zero as n goes to $+\infty$ such that

$$(14) \quad \bar{d}g(0) = \lim_{n \rightarrow \infty} \frac{(g(t_n) - g(0))}{t_n},$$

$$(15) \quad dg(0) = \lim_{n \rightarrow \infty} \frac{(g(t'_n) - g(0))}{t'_n}.$$

We first consider the upper bound in (11). We use the following hypotheses of continuity.

- (H3) There exists a topology τ_Y on Y such that for all $x_0 \in A(0)$:
- (i) For all sequences $t_n \rightarrow 0$, $t_n > 0$, there exists $y_0 \in B(0, x_0)$ and a subsequence of $\{t_n\}$, still denoted $\{t_n\}$, such that for all n , there exists $y_n \in B(t_n, x_0)$, and $y_n \rightarrow y_0$ in the τ_Y -topology;
 - (ii) The map

$$(t, y) \rightarrow \partial_t G(t, x_0, y)$$

is upper semicontinuous in $\{0\} \times \cup \{B(t, x_0): 0 \leq t \leq \tau\}$ in the τ_Y -topology.

PROPOSITION 2. Under hypotheses (H1)(i), (H2)(i), and (H3)

$$(16) \quad \bar{d}g(0) \leq \inf_{x \in A(0)} \sup_{y \in B(0, x)} \partial_t G(0, x, y).$$

Proof. Fix x_0 in $A(0)$ and let $\{t_n\}$ be the sequence in (14). Then by hypothesis (H3)(i), there exists $y_0 \in B(0, x_0)$, there exists a subsequence of $\{t_n\}$, still denoted $\{t_n\}$, such that there exists $y_n \in B(t_n, x_0)$, for all n , and

$$y_n \rightarrow y_0 \text{ in } \tau_Y\text{-topology.}$$

From (11), using hypothesis (H3)(ii), we obtain

$$\bar{d}g(0) \leq \limsup_{n \rightarrow \infty} \partial_t G(\theta_1 t_n, x_0, y_n) \leq \partial_t G((0, x_0, y_0)).$$

As a result

$$\bar{d}g(0) \leq \sup \{\partial_t G(0, x_0, y): y \in B(0, x_0)\}.$$

The last estimate is true for all x_0 in $A(0)$. This is sufficient to establish (16). \square

Remark 4.1. (H3)(i) is the Kuratovsky hypothesis at $t=0$ for the set-valued function $t \Rightarrow B(t, x_0)$.

We now turn to the lower bound (12). We formulate the following hypotheses.

- (H4) There exist topologies τ_X on X and τ_Y on Y such that:
- (i) For all sequences $t_n \rightarrow 0$, $t_n > 0$, there exists $x_0 \in A(0)$, for all $y_0 \in B(0, x_0)$, there exists a subsequence of $\{t_n\}$, still denoted $\{t_n\}$, such that for all n , there exists $x_n \in A(t_n)$ and there exists $z_n \in B(0, x_n)$, such that $x_n \rightarrow x_0$ in the τ_X -topology and $z_n \rightarrow y_0$ in the τ_Y -topology;
 - (ii) The map $(t, x, y) \rightarrow \partial_t G(t, x, y)$ is lower semicontinuous in $\{0\} \times \{(x, y): x \in A(0), y \in B(0, x)\}$ in the $\tau_X \times \tau_Y$ -topology.

Remark 4.2. Hypothesis (H4)(i) is verified when the following two hypotheses are verified:

- (H4)(i₁) There exists a topology τ_X on X such that for all sequences $t_n \rightarrow 0$, $t_n > 0$, there exists $x_0 \in A(0)$, there exists a subsequence of $\{t_n\}$, still denoted $\{t_n\}$, and for all n , there exists $x_n \in A(t_n)$ such that $x_n \rightarrow x_0$ in the τ_X -topology.
- (H4)(i₂) There exists a topology τ_Y of Y for which the set-valued function $x \Rightarrow B(0, x)$ is lower semicontinuous on $A(0)$ in the sense of Aubin [1, Déf. 9.4, p. 121]: for all convergent sequences $x_n \rightarrow x_0$ in X and all z^* in $B(0, x_0)$, there exists a sequence $z_n^* \in B(0, x_n)$ such that $z_n^* \rightarrow z^*$ in the τ_Y -topology.

Hypothesis (H4(i₁)) is the Kuratovsky condition at $t = 0$ for the set-valued map $t \Rightarrow A(t)$.

We state the analogue of Proposition 2 in the other direction.

PROPOSITION 3. *Under hypotheses (H1)(ii), (H2)(ii), and (H4), we have*

$$(17) \quad dg(0) \geq \inf_{x \in A(0)} \sup_{y \in B(0, x)} \partial_t G(0, x, y).$$

Proof. Consider expression (15) and the converging sequence $t'_n \rightarrow 0^+$, $t'_n > 0$. By (H4)(i), there exist topologies τ_X on X and τ_Y on Y such that there exists $x_0 \in A(0)$, for all $y_0 \in B(0, x_0)$, there exists a subsequence of $\{t'_n\}$, still denoted $\{t'_n\}$, and for all n , there exists $x_n \in A(t'_n)$ and there exists $z_n \in B(0, x_n)$, such that

$$x_n \rightarrow x_0 \text{ in the } \tau_X\text{-topology and } z_n \rightarrow y_0 \text{ in the } \tau_Y\text{-topology.}$$

Now from (12)

$$[g(t'_n) - g(0)]/t'_n \geq \partial_t G(\theta_2 t'_n, x_n, z_n)$$

and, in view of (H4)(ii)

$$dg(0) \geq \liminf_{n \rightarrow \infty} \partial_t G(\theta_2 t_n, x_n, z_n) \geq \partial_t G(0, x_0, y_0)$$

for some $x_0 \in A(0)$ and all $y_0 \in B(0, x_0)$. Finally

$$dg(0) \geq \sup_{y_0 \in B(0, x_0)} \partial_t G(0, x_0, y_0) \geq \inf_{x_0 \in A(0)} \sup_{y_0 \in B(0, x_0)} \partial_t G(0, x_0, y_0). \quad \square$$

We now state our main result.

THEOREM 3. *Under hypotheses (H1)–(H4), we have*

$$(18) \quad dg(0) = \bar{d}g(0) = \inf_{x \in A(0)} \sup_{y \in B(0, x)} \partial_t G(0, x, y)$$

and the function g is differentiable at 0 from the right:

$$(19) \quad \lim_{t \rightarrow 0^+} \frac{(g(t) - g(0))}{t} \text{ exists.}$$

Proof. The proof follows from Propositions 2 and 3. \square

The following propositions give sets of sufficient conditions to verify (H3)(i) and (H4)(i₁). They are the conditions given in Delfour and Zolésio [1], [2] in their initial version of Theorem 3.

PROPOSITION 4. *If for each x_0 in $A(0)$:*

(a) *There exists a topology τ_Y on Y and a sequentially compact subset K_Y of Y such that for all sequences*

$$t_n \rightarrow 0, \quad t_n > 0, \quad B(t_n, x_0) \cap K_Y \neq \emptyset;$$

(b) *For all y in \mathcal{B} , the map $t \rightarrow G(t, x_0, y)$ is lower semicontinuous;*

(c) *The map $t, y \rightarrow G(t, x_0, y)$ is upper semicontinuous in $\{0\} \times \cup \{B(t, x_0): 0 \leq t \leq \tau\}$ in the τ_Y -topology.*

Then hypothesis (H3)(i) is verified.

Proof. Fix an arbitrary x_0 in $A(0)$ and let $t_n \rightarrow 0, t_n > 0$, be an arbitrary converging sequence. By (a) we can choose for each n

$$y_n \in B(t_n, x_0) \cap K_Y.$$

But since K_Y is sequentially compact, there exists a subsequence of $\{y_n\}$, still denoted $\{y_n\}$, such that

$$y_n \rightarrow y^* \quad \text{in the } \tau_Y\text{-topology.}$$

By definition of $B(t_n, x_0)$,

$$G(t_n, x_0, y_n) \geq G(t_n, x_0, y) \quad \forall y \in \mathcal{B}$$

and from (c)

$$G(0, x_0, y^*) \geq \limsup_{n \rightarrow \infty} G(t_n, x_0, y_n) \geq \limsup_{n \rightarrow \infty} G(t_n, x_0, y) \quad \forall y \in \mathcal{B}.$$

But from (b)

$$\limsup_{n \rightarrow \infty} G(t_n, x_0, y) \geq \liminf_{n \rightarrow \infty} G(t_n, x_0, y) \geq G(0, x_0, y) \quad \forall y \in \mathcal{B}.$$

By combining the two inequalities above we conclude that $y^* \in B(0, x_0)$. Thus hypothesis (H3)(i) is verified. \square

PROPOSITION 5. *Assume that the following hypotheses are verified:*

(a) *There exists a topology τ_X on X and a sequentially compact subset K of X such that $A(t) \cap K \neq \emptyset$, for all $t, 0 \leq t \leq \tau$;*

(b) *For all x in \mathcal{A} the map $t \rightarrow H(t, x)$ is upper semicontinuous at $t = 0$;*

(c) *The map $(t, x) \rightarrow H(t, x)$ is lower semicontinuous in $\{0\} \times \cup \{A(t): 0 \leq t \leq \tau\}$ in the τ_X -topology.*

Then hypothesis (H4)(i₁) is verified.

Proof. The proof is similar to that of the previous proposition. \square

Remark 4.3. In order to obtain the lower bound on $dg(0)$, we have used (H4)(i). This hypothesis is to be compared with its counterpart (H3(i)) for the upper bound. In the first we have for all $x_0 \in A(0)$, there exists $y_0 \in B(0, x_0)$ while in the second there exists $x_0 \in A(0)$, for all $y_0 \in B(0, x_0)$. If in (H4)(i) the statement is weakened to: there exists $x_0 \in A(0)$ and there exists $y_0 \in B(0, x_0)$, then the upper and lower bounds on the differential quotient will no longer coincide. This suggests the following weaker form of hypothesis (H4)(i).

(H4)(i)-weak There exist topologies τ_X on X and τ_Y on Y such that for all sequences $t_n \rightarrow 0, t_n > 0$, there exists $x_0 \in A(0)$, there exists $y_0 \in B(0, x_0)$ and there exists a subsequence of $\{t_n\}$, still denoted $\{t_n\}$, such that for all n ,

there exists $x_n \in A(t_n)$ and there exists $z_n \in B(0, x_n)$, such that $x_n \rightarrow x_0$ in the τ_X -topology and $z_n \rightarrow y_0$ in the τ_Y -topology.

THEOREM 4. *Under hypotheses (H1)–(H3), (H4)(i)-weak and (H4)(ii)*

$$(20) \quad \inf_{x \in A(0)} \inf_{y \in B(0, x)} \partial_t G(0, x, y) \leq dg(0),$$

$$(21) \quad \bar{d}g(0) \leq \inf_{x \in A(0)} \sup_{y \in B(0, x)} \partial_t G(0, x, y).$$

COROLLARY. *If, in addition to the hypotheses of Theorem 4, the set $B(0, x)$ is a singleton for each x in $A(0)$,*

$$(22) \quad \forall x \in A(0) \quad B(0, x) = \{y_0(x)\},$$

then g is differentiable at 0 from the right and

$$(23) \quad dg(0) = \inf \{\partial_t G(0, x, y_0(x)) : x \in A(0)\}.$$

Remark 4.4. The corollary can also be proved directly by two consecutive applications of the theorem on the differentiability of a Min.

Remark 4.5. A sufficient set of conditions to verify (H4(i)-weak) is given by (H4)(i₁) and a Kuratovsky condition on the set-valued map $x \Rightarrow B(0, x)$:

(H4)(i₁) There exists a topology τ_X on X such that for all sequences $t_n \rightarrow 0$, $t_n > 0$, there exists $x_0 \in A(0)$, there exists a subsequence of $\{t_n\}$, still denoted $\{t_n\}$, and for all n , there exists $x_n \in A(t_n)$ such that $x_n \rightarrow x_0$ in the τ_X -topology.

(H4)(i₂)-weak There exists a topology τ_Y on Y such that for all $x_n \rightarrow x_0$ in the τ_X -topology, there exists a subsequence of $\{x_n\}$, still denoted $\{x_n\}$, $z^* \in B(0, x_0)$, for all n , there exists $z_n \in B(0, x_n)$ such that $z_n \rightarrow z^*$ for the τ_Y -topology.

Remark 4.6. A set of sufficient conditions to verify (H4)(i₂)-weak is given in Delfour and Zolésio [1], [2]:

(a) Given a converging sequence $x_n \rightarrow x_0$ in X , there exists a subsequence of $\{x_n\}$, still denoted $\{x_n\}$, there exists $z^* \in Y$ and there exists $z_n \in B(0, x_n)$, for all n , such that $z_n \rightarrow z^*$ in the τ_Y -topology;

(b) For all z in \mathcal{B} , the map $x \rightarrow G(0, x, z)$ is lower semicontinuous on \mathcal{A} and the map $x, z \rightarrow G(0, x, z)$ is upper semicontinuous on $\mathcal{A} \times \mathcal{B}$.

In order to complete this section we quote a recent result by Correa and Seeger [1] which assumes the existence of a saddle point of the functional G . First introduce the function

$$(24) \quad h(t) = \sup \{ \inf [G(t, x, y) : x \in \mathcal{A}] : y \in \mathcal{B} \},$$

the associated sets

$$(25) \quad B(t) = \{y \in \mathcal{B} : h(t) = \inf [G(t, x, y) : x \in \mathcal{A}]\},$$

$$(26) \quad A(t, y) = \{x \in \mathcal{A} : \inf [G(t, x, y) : x \in \mathcal{A}] = G(t, x, y)\}, \quad y \in Y$$

and the set of saddle points

$$(27) \quad S(t) = \{(x_t, y_t) \in \mathcal{A} \times \mathcal{B} : g(t) = G(t, x_t, y_t) = h(t)\}.$$

Then the following lemma is immediate.

LEMMA 4. *If $S(t) \neq \emptyset$ for some $t \geq 0$, then*

$$(28) \quad S(t) = A(t) \times B(t), \quad A(t) \neq \emptyset, \quad B(t) \neq \emptyset,$$

$$(29) \quad \forall x_t \in A(t) \quad B(t, x_t) = B(t) \quad \text{and} \quad \forall y_t \in B(t) \quad A(t, y_t) = A(t). \quad \square$$

THEOREM 5 (Correa and Seeger [1]). Assume that there exists $\tau > 0$ such that the following hypotheses are verified:

- (HH1) $S(t) \neq \emptyset$, $0 \leq t \leq \tau$.
- (HH2) For all (x, y) in $\cup \{A(t): 0 \leq t \leq \tau\} \times \cup \{B(t): 0 \leq t \leq \tau\}$, the map $t \rightarrow G(t, x, y)$ is differentiable everywhere in $[0, \tau]$.
- (HH3) There exists a topology τ_X on X such that
- (i) For all $t_n \rightarrow 0$, $0 \leq t_n \leq \tau$, there exists $x_0 \in A(0)$, there exists a subsequence of $\{t_n\}$, still denoted $\{t_n\}$, and for all n , there exists $x_n \in A(t_n)$, such that $x_n \rightarrow x_0$ in the τ_X -topology;
 - (ii) For all $y \in \cup \{B(t): 0 \leq t \leq \tau\}$,

$$(t, x) \rightarrow \partial_t G(t, x, y)$$

is lower semicontinuous at $\{0\} \times A(0)$ for the τ_X -topology.

- (HH4) There exists a topology τ_Y on Y such that
- (i) For all $t_n \rightarrow 0$, $0 \leq t_n \leq \tau$, there exists $y_0 \in B(0)$, there exists a subsequence of $\{t_n\}$, still denoted $\{t_n\}$, and for all n , there exists $y_n \in B(t_n)$, such that $y_n \rightarrow y_0$ in the τ_Y -topology;
 - (ii) For all $x \in \cup \{A(t): 0 \leq t \leq \tau\}$,

$$(t, y) \rightarrow \partial_t G(t, x, y)$$

is upper semicontinuous at $\{0\} \times B(0)$ for the τ_Y -topology.

Then

$$(30) \quad dg(0) = \lim_{t \rightarrow 0^+} (g(t) - g(0))/t = \inf_{x \in A(0)} \sup_{y \in B(0)} \partial_t G(0, x, t)$$

$$(31) \quad = \sup_{x \in B(0)} \inf_{y \in A(0)} \partial_t G(0, x, y).$$

This theorem contains as a special case Theorem 1 in § 2 and can be proved by the same technique. The hypotheses are essentially those given in Correa and Seeger [1], except for the fact that we have used Kuratovsky's conditions (HH3)(i) and (HH4)(i) for the set-valued maps instead of the notion of *sequential semicontinuity* for a set-valued map $t \Rightarrow M(t)$ from $[0, \tau]$ to X :

There exists a topology τ_X on X such that for all sequences $t_n \rightarrow 0$, $t_n > 0$, there exists $x_0 \in M(0)$ and for all n , there exists $x_n \in M(t_n)$ such that $x_n \rightarrow x_0$ in the τ_X -topology.

5. Derivative of a nondifferentiable observation functional with respect to the control variable. Let Ω be a bounded domain in R^n with smooth boundary Γ , $f \in L^2(\Omega)$ and u be a function on the subset U_{ad} of $U = L^\infty(\Omega)$ defined as

$$(1) \quad U_{ad} = \{u \in L^\infty(\Omega): \exists \alpha > 0 \text{ such that } u(x) \geq \alpha \text{ a.e. in } \Omega\}.$$

Consider the solution $y = y(u)$ in $H_0^1(\Omega)$ of the variational problem

$$(2) \quad -\operatorname{div}(u \nabla y) = f \quad \text{in } \Omega, \quad y = 0 \quad \text{on } \Gamma.$$

Associate with u and y the cost function

$$(3) \quad J(u) = \int_{\Omega} |y - y_d| \, dx, \quad y_d \in L^1(\Omega).$$

We want to compute the derivative of $J(u)$ with respect to u subject to (2).

We consider the state equation (2) as a constraint and remove it by introducing a Min Sup. It is easy to check that

$$(4) \quad J(u) = \min_{\varphi \in H_0^1(\Omega)} \sup_{(\mu, \psi) \in M \times H_0^1(\Omega)} \left[\int_{\Omega} \mu(\varphi - y_d) dx + dE(u, \varphi; 0, \psi) \right]$$

where $dE(u, \varphi; 0, \psi)$ is the Gâteaux derivative of

$$(5) \quad E(u, \varphi) = \frac{1}{2} \int_{\Omega} [u|\nabla \varphi|^2 - 2f\varphi] dx$$

at (u, φ) in the direction $(0, \psi)$ and

$$(6) \quad M = \{\mu \in L^\infty(\Omega): |\mu(x)| \leq 1, \text{ a.e. in } \Omega\}.$$

In this form, it is not directly possible to apply Theorem 3 in § 4. It is necessary to introduce a perturbed functional indexed by a parameter $r > 0$ (which is not necessarily infinitesimally small):

$$(7) \quad -G_r(u, (\mu, \psi), \varphi) = \int_{\Omega} \mu(\varphi - y_d) dx + dE(u, \varphi; 0, \psi) + r\{E(u, \varphi) - e(u)\}$$

where

$$(8) \quad e(u) = \inf \{E(u, \varphi): \varphi \in H_0^1(\Omega)\}.$$

Define

$$(9) \quad J_r(u) = \min_{\varphi \in H_0^1(\Omega)} \sup_{(\mu, \psi) \in M \times H_0^1(\Omega)} -G_r(u, (\psi, \mu), \varphi)$$

and the dual quantity

$$(10) \quad J_r^*(u) = - \inf_{(\mu, \psi) \in M \times H_0^1(\Omega)} \sup_{\varphi \in H_0^1(\Omega)} G_r(u, (\psi, \mu), \varphi).$$

Theorem 3 will be applied to J_r^* with $r > 0$. Prior to doing this we show that G_r has saddle points and characterize the associated sets.

PROPOSITION 6. *For each u in U and r , $0 < r < 2$, the functional $G_r(u, \cdot, \cdot)$ has saddle points and*

$$(11) \quad J_r^*(u) = J_r(u) = - \min_{(\mu, \psi) \in M \times H_0^1(\Omega)} \max_{\varphi \in H_0^1(\Omega)} G_r(u, (\mu, \psi), \varphi).$$

Proof. (i) The first part of identity (11) follows from Ekeland and Temam [1, Prop. 2.4, p. 177] applied to the functional

$$(12) \quad F_r(u, \psi, \varphi) = \sup \{-G_r(u, (\mu, \psi), \varphi): \mu \in M\}$$

which is equal to

$$(13) \quad \int_{\Omega} \{|\varphi - y_d| dx + dE(u, \varphi; 0, \psi) + r[E(u, \varphi) - e(u)]\}.$$

It suffices to check the following two conditions:

$$(14) \quad \exists p \in H_0^1(\Omega), \text{ such that } \lim_{\|\varphi\| \rightarrow \infty} F_r(u, p, \varphi) = +\infty,$$

$$(15) \quad \lim_{\|p\| \rightarrow \infty} \inf_{\varphi \in H_0^1(\Omega)} F_r(u, p, \varphi) = -\infty.$$

The first condition is verified for $p = 0$. For the second condition, we fix p and choose $\varphi = -p$

$$\inf \{F_r(u, p, \varphi) : \varphi \in H_0^1(\Omega)\} \leq F_r(u, p, -p)$$

and show that the upper bound goes to $-\infty$ as $\|p\|$ goes to $+\infty$:

$$(16) \quad F_r(u, p, -p) = \int_{\Omega} \{|-p - y_d| - u|\nabla p|^2 - fp + r/2(u|\nabla p|^2 + 2fp)\} dx - re(u).$$

The L^2 -norm of ∇p goes to $+\infty$ since it is equivalent to the $H_0^1(\Omega)$ -norm. So for r , $0 < r < 2$, the right-hand side of (16) goes to $-\infty$ and (15) is verified. This shows the existence of a saddle point for $F_r(u, \cdot, \cdot)$:

$$(17) \quad \min_{\varphi} \sup_{\psi} F_r(u, \psi, \varphi) = \max_{\psi} \inf_{\varphi} F_r(u, \psi, \varphi).$$

(ii) The next step is to show that for a fixed p ,

$$(18) \quad \inf_{\varphi} \sup_{\mu \in M} -G_r(u, \mu, p, \varphi) = \max_{\mu \in M} \inf_{\varphi} -G_r(u, \mu, p, \varphi).$$

In view of the properties of $-G_r$ and the fact that M is bounded, this is a consequence of Remark 2.3 and Proposition 2.3 in Ekeland and Temam [1, p. 162]. If we combine (17) and (18)

$$\min_{\varphi} \sup_{(\mu, \psi)} -G_r = \max_{(\mu, \psi)} \inf_{\varphi} -G_r,$$

and by Proposition 1.2 in Ekeland and Temam [1, p. 155], $-G_r(u, (\cdot, \cdot), \cdot)$ has saddle points. In view of (10), this is sufficient to establish (11). \square

It is now important to note that for all $r \geq 0$

$$(19) \quad J_r(u) = J_0(u) = J(u).$$

We have shown that for $0 < r < 2$, $G_r(u, \cdot, \cdot)$ has saddle points and that

$$(20) \quad J_r(u) = J_r^*(u).$$

For $u \in U_{ad}$ and $v \in U = L^\infty(\Omega)$, there exists $\tau > 0$ small enough such that $u + \tau v \in U_{ad}$. Define for t in $[0, \tau]$

$$(21) \quad G(t, q, \varphi) = G_r(u + tv, q, \varphi)$$

for $q = (\mu, \psi) \in X = M \times H_0^1(\Omega)$ and $\varphi \in Y = H_0^1(\Omega)$. In view of the above proposition, the saddle points of $G(t, \cdot, \cdot)$ are completely characterized by the following set of equations (cf. Ekeland and Temam [1, Prop. 1.6, p. 157]):

$$(22) \quad -\operatorname{div} [(u + tv)\nabla y_t] = f \quad \text{in } \Omega, \quad y_t = 0 \quad \text{on } \Gamma,$$

$$(23) \quad -\operatorname{div} [(u + tv)\nabla p_t] + \mu_t = 0 \quad \text{in } \Omega, \quad p_t = 0 \quad \text{on } \Gamma,$$

$$(24) \quad \mu_t \in M_{dt} = \{\operatorname{sgn}(y_t - y_d) - \alpha \chi_{\Omega_{dt}} : \alpha \in M\}$$

where

$$(25) \quad \Omega_{dt} = \{x \in \Omega : y_t(x) = y_d(x)\}$$

is a measurable set.

Remark 5.1. For this special case where the functional F can be expressed as a Sup, we could have completely bypassed the technique with the term in r by noting that the system of equations (22)–(25) has solutions and applying Proposition 1.6 in Ekeland and Temam [1, p. 157] to show that they are saddle points of G_r for all $r \geq 0$.

We introduce the constants

$$(26) \quad \beta = \frac{1}{2} \|u\|_{L^\infty(\Omega)}, \quad \tau = \beta / \|v\|_{L^\infty(\Omega)}.$$

The sets \mathcal{A} , \mathcal{B} are

$$(27) \quad \mathcal{A} = M \times H_0^1(\Omega), \quad \mathcal{B} = H_0^1(\Omega)$$

and the sets $A(t)$, $0 \leq t \leq \tau$, and $B(s, q)$, $0 \leq s \leq \tau$, are characterized by the following lemma.

LEMMA 5. (i) Given $r \geq 0$, then for all t , $0 \leq t \leq \tau$,

$$(28) \quad A(t) = \{(\mu_t(\alpha), p_t(\alpha)) : \mu_t(\alpha) = \text{sgn}(y_t - y_d) - \alpha \chi_{\Omega_d}, \alpha \in M\}$$

where y_t is the solution of (22), $\mu_t(\alpha)$ is given by (24) and $p_t(\alpha)$ by (23) with $\mu_t = \mu_t(\alpha)$. For $r > 0$ and all $q = (\psi, \mu) \in H_0^1(\Omega) \times M$ and all s , $0 \leq s \leq \tau$,

$$(29) \quad B(s, q) = \{\varphi_s(q)\}$$

where $\varphi = \varphi_s(q)$ is the unique solution in $H_0^1(\Omega)$ of the variational equation

$$(30) \quad \int_{\Omega} [\mu \varphi + (u + sv) \nabla \varphi \cdot \nabla \psi] dx + r \int_{\Omega} [(u + sv) \nabla \varphi \cdot \nabla \varphi - f \varphi] dx = 0 \quad \forall \varphi \in H_0^1(\Omega).$$

(ii) Moreover for all $r \geq 0$, t , $0 \leq t \leq \tau$, and $q_t = (p_t, \mu_t) \in A(t)$,

$$B(t, q_t) = B(t) = \{y_t\}$$

where y_t is the solution of (22) which is independent of $q_t \in A(t)$. In particular for all $r > 0$, $S(t) = A(t) \times B(t)$ is the set of saddle points of $G_r(t, q, \varphi)$.

Proof. The proof follows from previous considerations and Lemma 4 in § 4. \square

We now apply Theorem 3 of § 4 to $J_r^*(u)$ with $r > 0$. For $0 \leq t \leq \tau$, $A(t) \neq \emptyset$ and (H1)(i) is verified. For each s in $[0, \tau]$, the set $B(s, q)$ reduces to a singleton. So by Lemma 5 it is nonempty and (H1)(ii) is verified. Hypothesis (H2) is obvious. For (H3)(i) we use Proposition 4. For (a) choose for τ_Y the weak topology on $Y = H_0^1(\Omega)$ and for K_Y the weakly compact ball of radius R since for all t in $[0, \tau]$ and $\{y_t\} = B(t)$

$$\begin{aligned} \beta \int_{\Omega} |\nabla y_t|^2 dx &\leq \int_{\Omega} (u + tv) |\nabla y_t|^2 dx = \int_{\Omega} f y_t dx \leq c \|f\|_{L^2} \|\nabla y_t\|_{L^2} \\ &\Rightarrow \|\nabla y_t\|_{L^2} \leq R = c/\beta \|f\|_{L^2} \quad \forall t \in [0, \tau]. \end{aligned}$$

Conditions (b) and (c) and (H3)(ii) are also obvious. For (H4)(i) we use Proposition 5. For (a) we choose for τ_X the weak topology on $X = M \times H_0^1(\Omega)$ and for K the ball of radius $R' = c'/\beta$ (c' as defined below). Indeed for all t in $[0, \tau]$

$$\begin{aligned} \beta \|\nabla p\|^2 &\leq \int_{\Omega} (u + tv) \nabla p \cdot \nabla p dx = - \int_{\Omega} [\text{sgn}(y_t - y_d) - \alpha \chi_{\Omega_d}] p dx \\ &\leq c \|p\| \leq c' \|\nabla p\| \\ &\Rightarrow \|\nabla p\| \leq R' = c'/\beta \quad \forall t \in [0, \tau]. \end{aligned}$$

For (b)

$$\partial_t G(s, q, \varphi) = - \int_{\Omega} v \nabla p \cdot \nabla \varphi dx + r [dE(u + sv, \varphi; v, 0) - dE(u + sv; v)].$$

From Zolésio [2, Thm. 1.1, p. 1458], it is known that

$$dE(u + sv; v) = dE(u + sv, y_s; v, 0)$$

where y_s is the solution of (22) with $t = s$. With τ_Y being the strong topology on $Y = H_0^1(\Omega)$, $\partial_t G$ is jointly continuous with respect to its arguments.

Conditions (b) and (c) are on the functional

$$(31) \quad H(t, q) = \sup_{\varphi \in H_0^1(\Omega)} G(t, q, \varphi)$$

where for $q = (\mu, \psi)$

$$(32) \quad G(t, q, \varphi) = - \left\{ \int_{\Omega} \mu(\varphi - y_d) dx + dE(u + tv, \varphi; 0, \psi) + r[E(u + tv, \varphi) - e(u + tv)] \right\}.$$

Since G is lower semicontinuous in the variables (t, q, φ) , the functional

$$(t, q) \rightarrow H(t, q)$$

is lower semicontinuous and (c) is verified. Condition (b) essentially requires that $t \rightarrow H(t, q)$ be continuous at 0. This follows from the continuity with respect to t of the minimizing element φ_t of $-G(t, q, \varphi)$ with respect to t . For (H4)(ii), the map

$$q \rightarrow B(0, q) = \{\varphi(q)\}$$

is single valued, affine and continuous with respect to q .

We summarize our results in the next proposition.

PROPOSITION 7. *For all u in U , v in $L^\infty(\Omega)$ and $r \geq 0$, there exists $\tau > 0$ such that hypotheses (H1)–(H4) on the functional $G(t, q, \varphi)$ in (32) be verified (recall that $q = (\mu, p) \in X = M \times H_0^1(\Omega)$ and that $\varphi \in Y = H_0^1(\Omega)$). For t in $[0, \tau]$, the sets $A(t)$ and $B(s, q)$ are given by (28) and (29).*

THEOREM 6. *For all u in U and v in $L^\infty(\Omega)$, the functional $J(u)$ is Gâteaux semidifferentiable at u in the direction v and*

$$(33) \quad dJ(u; v) = \lim_{t \rightarrow 0^+} [J(u + tv) - J(u)]/t = \sup \left\{ \int_{\Omega} v \nabla p(\alpha) \cdot \nabla y dx : \alpha \in M \right\}$$

where y and $p(\alpha)$ are the respective solutions of

$$(34) \quad -\operatorname{div}(u \nabla y) = f \quad \text{in } \Omega, \quad y = 0 \quad \text{on } \Gamma,$$

$$(35) \quad -\operatorname{div}(u \nabla p(\alpha)) + \operatorname{sgn}(y - y_d) - \alpha \chi_{\Omega_d} = 0 \quad \text{in } \Omega, \quad p(\alpha) = 0 \quad \text{on } \Gamma,$$

$$(36) \quad \Omega_d = \{x \in \Omega : y(x) = y_d(x)\}.$$

Proof. Recall that for $r \geq 0$ and $0 \leq t \leq \tau$

$$J_r(u + tv) = J(u + tv).$$

Computing the derivative of J is equivalent to computing the derivative of J_r for some fixed $r > 0$. The results of § 4 are now available. It is sufficient to note that the integral in (33) is $\partial_t G(0, (\mu(\alpha), p(\alpha)), y)$, where

$$\mu(\alpha) = \operatorname{sgn}(y - y_d) - \alpha \chi_{\Omega_d}, \quad \alpha \in M.$$

Expression (33) then follows from Proposition 7 and Theorem 3 in § 4. \square

Remark 5.2. Note that the map $\alpha \rightarrow p(\alpha)$ is affine and continuous. In (33) the Sup occurs at extremal points of M . By defining

$$(37) \quad M_d = \{\alpha \in L^\infty(\Omega) : \alpha(x) = \pm 1 \text{ in } \Omega_d \text{ and } \alpha(x) = 0 \text{ elsewhere}\},$$

we obtain

$$(38) \quad dJ(u; v) = \sup \left\{ \int_{\Omega} v \nabla p(\alpha) \cdot \nabla y dx : \alpha \in M_d \right\}.$$

Remark 5.3. Throughout our analysis, the parameter r is fixed but arbitrary and the saddle points of G_r are independent of r . Thus Theorem 6 could also have been obtained by applying Theorem 5 to $G(t, q, \varphi)$.

Remark 5.4. The interest behind this method for cost functionals of the form

$$(39) \quad J(u) = F(u, y(u))$$

is to justify the differentiation of $J(u)$ at u in the direction v without using the intermediate step of differentiating the state $y(u)$ at u in the direction v .

The above results formally extend to the class of linear variational problems. For instance, let

$$(40) \quad \underline{E}: U_{\text{ad}} \times X \rightarrow R, \quad \underline{E}(u, x) = \frac{1}{2}a(u, x, x) - L(u, x)$$

for some open set U_{ad} of a Banach space U , a Hilbert space X , a continuous symmetrical bilinear form $a(u, \cdot, \cdot)$, and a continuous linear form $L(u, \cdot)$. For each u in U_{ad} , define

$$(41) \quad \underline{e}(u) = \inf \{ \underline{E}(u, x) : x \in X \}$$

and assume that the minimizing element $y = y(u)$ in X is unique and completely characterized by the variational equation

$$(42) \quad d\underline{E}(u, y; 0, \psi) = 0 \quad \forall \psi \in X.$$

We can readily extend this method to cost functionals of the form

$$(43) \quad J(u) = \underline{F}(u, y(u))$$

for some map

$$(44) \quad \underline{F}: U_{\text{ad}} \times X \rightarrow R$$

where $y(u)$ is the solution of (42).

Then the associated functional \underline{G} is given by

$$(45) \quad -\underline{G}(u, \psi, x) = \underline{F}(u, x) + d\underline{E}(u, x; 0, \psi)$$

for $u \in U_{\text{ad}}$, $\psi \in X$ and $x \in X$. Assume that F is semidifferentiable at $y(u)$, that is,

$$(46) \quad \forall z \in Z, \quad d\underline{F}(u, y(u); z) = \lim_{t \rightarrow 0^+} [\underline{F}(u, y(u) + tz) - \underline{F}(u, y(u))]/t \quad \text{exists.}$$

Let $P(u)$ be the set of solutions $p = p(u)$ of the adjoint inequality

$$(47) \quad d\underline{F}(u, y(u); 0, \psi) + d^2 \underline{E}(u, y(u); 0, p; 0, \psi) \geq 0 \quad \forall \psi \in X$$

where $y(u)$ is the solution of (42).

THEOREM 7. Let E and F be as described above. Fix v in V and $\tau > 0$ such that for all t , $0 \leq t \leq \tau$, $u + tv \in U_{\text{ad}}$. Assume the following:

- (a) For all t , $0 \leq t \leq \tau$, $P(u + tv) \neq \emptyset$;
- (b) For all φ , $\psi \in X$, $t \rightarrow a(u + tv, \varphi, \psi)$ and $t \rightarrow L(u + tv, \psi)$ are continuously differentiable in $[0, \tau]$;
- (c) There exists $\alpha > 0$, for all t , $0 \leq t \leq \tau$, and for all φ , $a(u + tv, \varphi, \varphi) \geq \alpha \|\varphi\|^2$;
- (d) For all ψ , $(t, \varphi) \rightarrow d\underline{F}(u + tv, \varphi; 0, \psi)$ is upper semicontinuous at $(0, y(u))$;
- (e) There exists a neighborhood $W(u)$ of u , and there exists $c(u) > 0$, such that

$$(48) \quad \forall \psi \in X, \quad \forall w \in W(u), \quad d\underline{F}(w, y(w); 0, \psi) \leq c(u) \|\psi\|;$$

(f) For all φ , for all v and for all $t \in [0, \tau]$, the limit

$$(49) \quad d\underline{F}(u + tv, \varphi; v, 0) = \lim_{s \rightarrow 0^+} [\underline{F}(u + (t+s)v, \varphi) - \underline{F}(u + tv, \varphi)]/s$$

exists;

(g) For all v , $(t, \varphi) \rightarrow dF(u + tv, \varphi; v, 0)$ is lower semicontinuous at $(0, y_0)$;

(h) For all φ and for all v , $t \rightarrow dF(u + tv, \varphi; v, 0)$ is upper semicontinuous at $t = 0$.

Then $J(u)$ defined by (43) is semidifferentiable at u in the direction v and

$$(50) \quad dJ(u; v) = \sup \{dF(u, y(u); v, 0) + d^2E(u, y(u); 0, p; v, 0) : p \in P(u)\}.$$

The proof of this theorem will be given after a short discussion of the fundamental hypothesis (a).

Remark 5.5. If the map

$$(51) \quad \varphi \rightarrow dF(u, y; 0, \varphi)$$

is linear and continuous, the adjoint problem (47) is variational and $P(u)$ reduces to the usual solution of the associated variational problem.

When (54) is nonlinear we can use the augmented Lagrangian technique previously developed.

PROPOSITION 8. Assume the existence of a number r , $0 < r < 2$, such that

$$(52) \quad \begin{aligned} F(u, x) + \frac{r}{2} a(u, x, x) &\rightarrow +\infty, & \text{as } \|x\| \rightarrow \infty. \\ -F(u, -x) + (1 - r/2)a(u, x, x) &\rightarrow +\infty \end{aligned}$$

Then $P(u)$ is not empty. \square

Proof of Theorem 7. We show that the hypotheses of Theorem 5 are verified. From

(a) the sets of saddle points

$$S(t) = P(u + tv) \times \{y(u + tv)\} \neq \emptyset \quad \forall t \in [0, \tau]$$

and (HH1) is verified. Define

$$-G(t, \psi, \varphi) = F(u + tv, \varphi) + dE(u + tv, \varphi; 0, \psi).$$

From (b) and (f) for all t in $[0, \tau[$ and φ, ψ in X

$$-\partial_t G(t, \psi, \varphi) = dF(u + tv, \varphi; v, 0) + d^2E(u + tv, \varphi; 0, \psi; v, 0)$$

and (HH2) is verified. For each t in $[0, \tau]$, $y_t = y(u + tv) \in X$ is the unique solution of

$$a(u + tv, y_t, \psi) - L(u + tv, \psi) = 0 \quad \forall \psi \in X.$$

From the above equation and (b)

$$(53) \quad a(u + tv, y_t - y_0, \psi) + a(u + tv, y_0, \psi) - L(u + tv, \psi) = 0 \quad \forall \psi \in X.$$

But there exists $\theta \in]0, 1[$

$$\begin{aligned} a(u + tv, y_0, \psi) - L(u + tv, \psi) \\ = a(u, y_0, \psi) - L(u, \psi) + t[\partial_t a(u + \theta tv, y_0, \psi) - \partial_t L(u + \theta tv, \psi)] \end{aligned}$$

and for t small the term in the square bracket is bounded by $c\|\psi\|$ where $c > 0$ is a constant which is independent of t . So

$$a(u + tv, y_t - y_0, \psi) \leq tc\|\psi\|$$

and by (c) with $\psi = y_t - y_0$

$$\alpha \|y_t - y_0\|^2 \leq tc \|y_t - y_0\|.$$

So the map $t \rightarrow y_t$ is continuous at $t = 0$ in X -strong and (HH4)(i) is verified.

(HH4)(ii) reduces to the lower semicontinuity of the map

$$(t, \varphi) \rightarrow dF(u + tv, \varphi; v, 0) + \partial_t a(u + tv, \varphi, \psi) - \partial_t L(u + tv, \psi)$$

for each ψ in X . For the first term, it is true by (g). As for the other two terms it follows from (b) and the linearity of $\partial_t a$ with respect to φ . Similarly (HH3)(ii) follows from (h), (b) and the linearity of $\partial_t a(u + tv, y, \psi)$ with respect to ψ . As for (HH3)(i) we know that for all t in $[0, \tau]$ there exists at least one solution $p_t \in X$ to the variational inequality

$$(54) \quad dF(u + tv, y_t; 0, \psi) + d^2 E(u + tv, y_t; 0, p_t; 0, \psi) \geq 0 \quad \forall \psi \in X$$

where $y_t = y(u + tv)$ is the unique solution of (53). We first show that p_t is bounded in X . First (54) reduces to

$$(55) \quad dF(u + tv, y_t; 0, \psi) + a(u + tv, p_t, \psi) \geq 0 \quad \forall \psi \in X$$

and for $\psi = -p_t$

$$\alpha \|p_t\|^2 \leq dF(u + tv, y_t; 0, -p_t)$$

and by (e) for t small enough

$$\alpha \|p_t\|^2 \leq c(u) \|p_t\|.$$

So $\{p_t; 0 \leq t \leq \tau\}$ is bounded in X and there exists p^* and a sequence $t_n \rightarrow 0^+$ such that

$$p_n = p_{t_n} \rightarrow p^* \quad \text{in } X\text{-weak.}$$

Going back to (55)

$$dF(u + t_n v, y_n; 0, \psi) + a(u + t_n v, p_n, \psi) \geq 0 \quad \forall \psi \in X.$$

From (b)

$$a(u + t_n v, p_n, \psi) \rightarrow a(u, p^*, \psi)$$

and by (d)

$$\limsup_{n \rightarrow \infty} dF(u + t_n v, y_n; 0, \psi) \leq dF(u, y_0; 0, \psi)$$

since $y_n \rightarrow y_0$ in X -strong. As a result

$$dF(u, y_0; 0, \psi) + a(u, p^*, \psi) \geq 0 \quad \forall \psi \in X$$

and $p^* \in P(u)$. So (HH3)(i) is verified and Theorem 5 applies. \square

Remark 5.6. Theorem 7 can also be obtained by using for some $r > 0$ the Lagrangian

$$(56) \quad -G_r(u, \psi, \varphi) = F(u, \varphi) + dE(u, \varphi; 0, \psi) + r[E(u, \varphi) - \underline{e}(u)],$$

hypothesis (52) for $u + tv$ instead of $S(t) \neq \emptyset$ and applying Theorem 3 or 4 to

$$(57) \quad J^*(u) = - \inf_{\psi \in X} \sup_{\varphi \in X} G_r(u, \psi, \varphi).$$

In general it is not possible to verify hypothesis (H1) in Theorems 3 and 4 for

$$(58) \quad J(u) = \inf_{\varphi \in X} \sup_{\psi \in X} -G_r(u, \psi, \varphi).$$

Proof of Proposition 8. As in our illustrative example, we construct the augmented Lagrangian

$$-G_r(u, p, x) = -G(u, p, x) + r[E(u, x) - \underline{e}(u)].$$

It is easy to show that the functional $-G_r$ is convex and lower semicontinuous in x and concave and upper semicontinuous in p . From hypotheses (52),

$$-G_r(u, x, 0) \rightarrow +\infty \quad \text{as } \|x\| \rightarrow \infty,$$

$$\inf_{x \in X} -G_r(u, p, x) \leq G_r(u, p, -p) \rightarrow -\infty \quad \text{as } \|p\| \rightarrow \infty.$$

Again by Proposition 2.4 in Ekeland and Temam [1, p. 164], $\mathcal{G}_r(u, \cdot, \cdot)$ has saddle points in $X \times X$. They are completely characterized by the system

$$(59) \quad -d\mathcal{G}_r(u, p, y; 0, 0, \varphi - y) \geq 0 \quad \forall \varphi \in X,$$

$$(60) \quad -d\mathcal{G}_r(u, p, y; 0, \psi - p, 0) \leq 0 \quad \forall \psi \in X,$$

which is equivalent to (42) and (47). This is sufficient to establish that $P(u)$ is not empty. \square

6. Shape derivative of a functional: a simple example.

6.1. Shape optimization problem. Shape and structural optimal design is quite a broad field of activity. A good account of recent work can be found in the two volumes of the proceedings of the NATO Advanced Study Institute held in Iowa City (cf. Haug and C  a [1]) which contain an enormous amount of material covering the engineering and mathematical aspects of that class of problems. Relatively few books have been published on optimal design problems based on PDE models or in engineering terminology distributed parameter models. According to Haug and C  a [1], we find the book of Prager [1], the proceedings of the symposium held in Warsaw in 1973 edited by Sawczuk and Mroz [1], the books of Rozvany [1] in 1976, Haug and Arora [1], Banichuk [1] (original version in Russian, available in English), and the more recent books by Pironneau [1] and Haug, Choi, and Komkov [1]. The method presented here and in § 7 can be used to obtain a mathematical justification for some parts of the results formally obtained in C  a [1], Dems and Mroz [1], Haug and C  a [1], and Haug, Choi, and Komkov [1]. In this section we consider the shape sensitivity analysis problem. We do not cover variational inequalities. For this we refer to the recent papers by Sokolowski and Zol  sio [1], [2] and Sokolowski [1] and their bibliographies.

Consider the following simple example. Let Ω be a bounded open domain in R^n with a smooth boundary Γ . Let $y = y(\Omega)$ be the solution of the variational problem

$$(1) \quad \inf \{E(\Omega, \varphi) : \varphi \in H^1(\Omega)\}$$

where

$$(2) \quad E(\Omega, \varphi) = \frac{1}{2} \int_{\Omega} [|\nabla \varphi|^2 + |\varphi|^2 - 2f\varphi] \, dx$$

for some fixed function f in $H^1(R^n)$. We associate with y a cost function

$$(3) \quad J(\Omega) = F(\Omega, y(\Omega)).$$

For instance we can choose the standard cost function

$$(4) \quad F(\Omega, y) = \frac{1}{2} \int_{\Omega} (y - Y_d)^2 \, dx, \quad Y_d \in H^1(R^n).$$

6.2. The velocity field method. We briefly recall the notion of a shape derivative. Let $V(t, x)$, $t \geq 0$, $x \in R^n$, be a *velocity field of deformation*. Under the action of V , the points of Ω are transported onto a new domain $\Omega_t = T_t(\Omega)$, where the transformation $T_t : R^n \rightarrow R^n$ is generated by the solutions of the equation

$$(5) \quad (\partial/\partial t) T_t(x) = V(t, T_t(x)), \quad t \geq 0, \quad T_0(x) = x$$

(cf. Zol  sio [3]). Let y_t be the solution of problem (1) on the transformed domain Ω_t ,

$$(1_t) \quad \inf \{E(\Omega_t, \varphi) : \varphi \in H^1(\Omega_t)\}$$

and associate with y_t the cost function

$$(3_t) \quad J(\Omega_t) = F(\Omega_t, y_t).$$

Traditional methods involve the computation of the shape derivative (or partial derivative) Y' or the *Material derivative* \dot{y} . The shape derivative is defined as

$$Y'(x) = \lim_{t \rightarrow 0^+} [Y(t, x) - Y(0, x)]/t$$

where for some $\tau > 0$ $Y(t, x)$ is an appropriate extension of $y_t(x)$ to $[0, \tau] \times D$ for some fixed domain D containing all perturbations Ω_t of Ω , $0 \leq t \leq \tau$. The material derivative is defined as

$$\dot{y} = \lim_{t \rightarrow 0^+} [y^t - y]/t$$

in an appropriate function space on Ω , where y^t is the transported solution (from Ω_t to Ω)

$$y^t = y_t \circ T_t.$$

In classical examples Y' is the solution of a boundary value problem which depends on y and the normal component of the velocity field on the boundary Γ . However, in general, the material derivative is also the solution of a boundary value problem on Ω , but it depends on the velocity field in the whole domain. In fact the two derivatives are related through the formula

$$Y' = \dot{y} - \nabla y \cdot V$$

where V is the velocity field at 0, $x \mapsto V(0, x)$. In general Y' is "rougher" than the material derivative.

The next step consists in differentiating $J(\Omega_t)$ using the material derivative or Y' . Then an appropriate adjoint variable p is introduced to eliminate those derivatives and obtain a final expression which depends on Ω , y , p and V . The adjoint variable p is the solution of a boundary value problem which is dual to the corresponding boundary value problem for the material derivative or Y' .

The final expression can then be used for shape sensitivity analysis or as a necessary condition characterizing an eventual minimizing domain Ω^* .

6.3. The Inf Sup formulation of the perturbed problem. In general our objective is the minimization of the cost function J with respect to Ω . In particular we want to compute the shape derivative of J at Ω in the direction of the velocity field of deformations V . To do this we transform the problem (1_t)–(3_t) into an Inf Sup problem. This approach is widespread in the engineering and mathematical literature.

The solution of (1_t) is completely characterized by the variational equation

$$(6) \quad dE(\Omega_t, y_t; \varphi) = 0 \quad \forall \varphi \in H^1(\Omega_t)$$

where

$$(7) \quad dE(\Omega_t, \psi; \varphi) = \int_{\Omega_t} [\nabla \psi \cdot \nabla \varphi + \psi \varphi - f \varphi] dx, \quad \varphi, \psi \in H^1(\Omega_t).$$

Define for $r \geq 0$

$$(8) \quad -G_r(t, \varphi, p) = F(\Omega_t, \varphi) + dE(\Omega_t, \varphi; p) + r[E(\Omega_t, \varphi) - e(t)]$$

where

$$(9) \quad e(t) = \text{Min} \{E(\Omega_t, \varphi): \varphi \in H^1(\Omega_t)\}.$$

But

$$\text{Sup} \{-G_r(t, \varphi, p): p \in H^1(\Omega_t)\} = \begin{cases} F(\Omega_t, \varphi) & \text{if } \varphi \text{ is a solution of (6),} \\ +\infty & \text{otherwise.} \end{cases}$$

As a result for all $r \geq 0$

$$J(\Omega_t) = \text{Inf} \{\text{Sup} [-G_r(t, \varphi, p): p \in H^1(\Omega_t)]: \varphi \in H^1(\Omega_t)\}.$$

In this form the spaces depend on the parameter t and it is not readily possible to apply the theorems of § 6. However, for the Neuman problem it is possible to embed everything in $H^1(R^n)$. It is readily seen that

$$(10) \quad \text{Sup} \{-G_r(t, \varphi, p): p \in H^1(R^n)\} = \begin{cases} F(\Omega_t, \varphi) & \text{if } \varphi \text{ is a solution of (6),} \\ +\infty & \text{otherwise.} \end{cases}$$

As a result for all $r \geq 0$

$$(11) \quad J(\Omega_t) = \text{Inf} \{\text{Sup} [-G_r(t, \varphi, p): p \in H^1(R^n)]: \varphi \in H^1(R^n)\}.$$

The spaces involved are now fixed and independent of the parameter $t \geq 0$.

6.4. Perturbed dual functional J_r^* and existence of saddle points. Our objective is to show the existence of saddle points for $r > 0$ and use the results of § 4 together with identity (11).

For $r \geq 0$, define the functionals

$$(12) \quad J_r(\Omega_t) = -\text{Sup} \{\text{Inf} [G_r(t, \varphi, p): p \in H^1(R^n)]: \varphi \in H^1(R^n)\}$$

and

$$(13) \quad J_r^*(\Omega_t) = -\text{Inf} \{\text{Sup} [G_r(t, \varphi, p): \varphi \in H^1(R^n)]: p \in H^1(R^n)\}.$$

Recall that in view of (11)

$$(14) \quad J_r(\Omega_t) = J_0(\Omega_t) = J(\Omega_t) \quad \forall r \geq 0.$$

In general

$$(15) \quad J_r^*(\Omega_t) \leq J_r(\Omega_t)$$

since J_r^* is the dual functional associated with the perturbed functional G_r .

We have made the above construction in order to apply Theorem 3 to the dual problem for $r > 0$ (for $r = 0$ hypothesis (H1) would not be verified) or Theorem 5 for $r \geq 0$.

The functional G_r has a saddle point for all $r \geq 0$.

PROPOSITION 9. (i) *Given $\tau > 0$ small enough, then for all r , $0 \leq r$, and t , $0 \leq t \leq \tau$, $G_r(t, \cdot, \cdot)$ has saddle points (Y_r, P_r) in $H^1(R^n) \times H^1(R^n)$ and*

$$(16) \quad J_r^*(\Omega_t) = J_r(\Omega_t).$$

(ii) *The restriction of each saddle point to Ω_t*

$$(17) \quad (y'_t, p'_t) = (Y_r|_{\Omega_t}, P_r|_{\Omega_t})$$

coincides with the unique pair (y_t, p_t) solution of the system

$$(18) \quad dE(\Omega_t, y_t; \varphi) = 0 \quad \forall \varphi \in H^1(\Omega_t),$$

$$(19) \quad dF(\Omega_t, y_t; \psi) + d^2E(\Omega_t, y_t; p_t; \psi) = 0 \quad \forall \psi \in H^1(\Omega_t)$$

where

$$(20) \quad dF(\Omega_t, y_t; \psi) = \int_{\Omega_t} (y_t - Y_d) \psi \, dx,$$

$$(21) \quad d^2E(\Omega_t, y_t; p; \psi) = \int_{\Omega_t} [\nabla p_t \cdot \nabla \psi + p_t \psi] \, dx.$$

Proof. The conditions characterizing a saddle point (Y_r, P_r) of G_r are precisely (18)–(19). Both equations are elliptic with a unique solution in $H^1(\Omega_t)$ which is independent of $r \geq 0$. \square

6.5. Application of Theorem 3. The next step is the application of Theorem 3 from § 4 to the function $g(t) = J_r^*(\Omega_t)$. We first study the differentiability of the functional $G_r(t, \varphi, \psi)$ as defined by (8) with respect to t for all φ and ψ in the space $X = H^1(R^n)$:

$$(22) \quad \begin{aligned} -G_r(\Omega_t, \varphi, \psi) = & \frac{1}{2} \int_{\Omega_t} (\varphi - Y_d)^2 \, dx + \int_{\Omega_t} [\nabla \psi \cdot \nabla \varphi + \psi \varphi - f \varphi] \, dx \\ & + \frac{r}{2} \int_{\Omega_t} [|\nabla \varphi|^2 + |\varphi|^2 - 2f \varphi] \, dx. \end{aligned}$$

If ψ and φ were smoother (e.g., in $H^2(R^n)$) to make sure that the traces of $|\nabla \varphi|^2$ and $\nabla \psi \cdot \nabla \varphi$ exist, the standard expression for the derivative would be given by a boundary integral of the integrand in (22) (cf. Zolésio [3]). Unfortunately this is not the case. Expression (22) transported from Ω_t onto Ω is given by

$$\begin{aligned} & \int_{\Omega} A(t) \left[\nabla(\psi \circ T_t) + \frac{r}{2} (\nabla \varphi \circ T_t) \right] \cdot \nabla(\varphi \circ T_t) \, dx \\ & + \int_{\Omega} \left\{ \frac{1}{2} (\varphi \circ T_t - Y_d \circ T_t)^2 \right. \\ & \quad \left. + \left[(\psi \circ T_t - f \circ T_t) + \frac{r}{2} (\varphi \circ T_t - 2f \circ T_t) \right] (\varphi \circ T_t) \right\} J(t) \, dx \end{aligned}$$

where DT_t is the Jacobian matrix of the transformation T_t and

$$(23) \quad J(t) = \det(DT_t), \quad A(t) = J(t)((DT_t)^{-1})^*(DT_t)^{-1}$$

(* indicates the transposed matrix). Again to differentiate the above expression with respect to t would require that φ and ψ be in $H^2(R^n)$. To get around this difficulty we need the special technique which is described below and which does not seem to have any counterpart in control.

As before, given the smooth velocity field V , define the transformation

$$(24) \quad T_t = T_t(V) : R^n \rightarrow R^n$$

which transports R^n onto R^n , $\Omega_0 = \Omega$ onto Ω_t and $\Gamma_0 = \Gamma$ onto $\Gamma_t = \partial\Omega_t$. The space $H^1(\Omega_t)$ is transported in a similar way. As the functions φ and ψ fill in the whole space X , so do the functions $\psi \circ T_t^{-1}$ and $\varphi \circ T_t^{-1}$. As a result

$$(25) \quad g(t) = J_r(\Omega_t) = J_r^*(\Omega_t) = - \inf_{\varphi \in X} \sup_{\psi \in X} -G_r(t, \varphi, \psi)$$

where

$$(26) \quad G_r(t, \varphi, \psi) = G_r(t, \varphi \circ T_t^{-1}, \psi \circ T_t^{-1}).$$

By introducing the quantities

$$(27) \quad J(t) = \det(DT_t), \quad A(t) = J(t)((DT_t)^{-1})^*(DT_t)^{-1}$$

(DT_t is the Jacobian matrix of the transformation T_t) and the change of variable $x' = T_t(x)$, we obtain

$$(28) \quad -G_r(t, \varphi, \psi) = -G_0(t, \varphi, \psi) + r[E(\Omega, \varphi \circ T_t) - e(t)]$$

where

$$(29) \quad e(t) = \inf \{E(\Omega_t, \varphi) : \varphi \in H^1(R^n)\} = \inf \{E(\Omega, \varphi \circ T_t) : \varphi \in H^1(R^n)\}$$

and

$$(30) \quad -G_0(t, \varphi, \psi) = \int_{\Omega} [(A(t)\nabla\psi) \cdot \nabla\varphi + \psi\varphi J(t)] dx \\ + \int_{\Omega} \left[\frac{1}{2}(\varphi - Y_d \circ T_t)^2 - (f \circ T_t)\psi \right] J(t) dx,$$

$$(31) \quad E(\Omega, \varphi \circ T_t) = \int_{\Omega} \frac{1}{2} \{ (A(t)\nabla\varphi) \cdot \nabla\varphi + J(t)[\varphi\varphi - (f \circ T_t)\varphi] \} dx.$$

The derivative with respect to t can easily be obtained for every term in $G_r(t, \cdot, \cdot)$ except possibly $e(t)$. Fortunately we know from Zolésio [3, Thm. 1.1, p. 1458] that

$$(32) \quad d_t e(0) = \inf \{ \partial_t E(\Omega_0, \varphi) : \varphi \in X, E(\Omega_0, \varphi) = e(0) \} \\ = \partial_t E(\Omega_0, Y)$$

where Y is the solution of

$$E(\Omega_0, Y) = \inf \{ E(\Omega_0, \varphi) : \varphi \in H^1(R^n) \}$$

or equivalently

$$(33) \quad Y \in H^1(R^n), \quad dE(\Omega, Y; \varphi) = 0 \quad \forall \varphi \in H^1(R^n).$$

We finally obtain the following intermediate result prior to the application of Theorem 3 in § 4.

PROPOSITION 10. *For all r , $0 < r < 1$, t , $0 \leq t \leq \tau$, $G_r(t, \varphi, \psi)$ is differentiable with respect to t at $t=0$ and*

$$(34) \quad -\partial_t G_r(0, \varphi, \psi) = \int_{\Omega} [(A'(0)\nabla\psi) \cdot \nabla\varphi + \psi\varphi \operatorname{div} V(0)] dx \\ - \int_{\Omega} \left[\frac{1}{2}(\varphi - Y_d) \nabla Y_d + \psi \nabla f \right] \cdot V(0) dx \\ + \int_{\Omega} \left[\frac{1}{2}(\varphi - Y_d)^2 - f\psi \right] \operatorname{div} V(0) dx + r[\partial_t E(\Omega_0, \varphi) - \partial_t E(\Omega_0, Y)]$$

where

$$(35) \quad A'(0) = [\operatorname{div} V(0)]I_d - [DV(0) + (DV(0))^*].$$

(($DV(0))^*$ is the transposed matrix of $(DV(0))$). \square

All the hypotheses of Theorem 3 in § 4 are now verified.

THEOREM 8. For all r , $0 < r < 1$,

$$(36) \quad dJ_r(\Omega; V(0)) = -\partial_t \underline{G}_r(0, p, y)$$

where y and p are the solutions of

$$(37) \quad dE(\Omega, y; \varphi) = 0 \quad \forall \varphi \in H^1(\Omega)$$

and

$$(38) \quad dF(\Omega, y; \psi) + d^2E(\Omega, y; p; \psi) = 0 \quad \forall \psi \in H^1(\Omega). \quad \square$$

Remark 6.1. The derivative of J_r was obtained without our ever considering the problem of the differentiability of y .

Finally recall identity (14) to obtain the desired result.

THEOREM 9. (i) The function g is differentiable at $t = 0$ and

$$(39) \quad \begin{aligned} dJ(\Omega; V(0)) = & \int_{\Omega} \left[(A'(0) \nabla p) \cdot \nabla y - \left[\frac{1}{2} (y - Y_d) \nabla Y_d + p \nabla f \right] \cdot V(0) \right] dx \\ & + \int_{\Omega} \left[\frac{1}{2} (y - Y_d)^2 + py - fp \right] \operatorname{div} V(0) dx \end{aligned}$$

where y and p are the solutions of (37) and (38).

(ii) If, in addition, p and y belong to $H^{3/2+\rho}(\Omega)$, $\rho > 0$, then

$$(40) \quad dJ(\Omega; V(0)) = \int_{\Gamma} \left[\nabla y \cdot \nabla p + yp - fp + \frac{1}{2} (y - Y_d)^2 \right] V(0) \cdot n \, d\Gamma.$$

Proof. It suffices to note that for $\varphi = y$ the term which contains the r in identity (36) is identically zero. When y and p are sufficiently smooth, (39) is equivalent to the standard boundary integral formulation in shape optimization. To show that set $\varphi = \nabla p \circ V$ in (37) and $\psi = \nabla y \circ V$ in (38), add the resulting two equations and reorganize the terms. This yields an identity which shows that the right-hand sides of (39) and (40) are equal. \square

Remark 6.2. The simple example above contains several techniques which will turn out to be fundamental to the general theory. For instance the introduction of the functional

$$\underline{G}_r(t, \varphi, \psi) = G_r(t, \varphi \circ T_t^{-1}, \psi \circ T_t^{-1})$$

followed by the transport of the resulting expression from the domain Ω_t onto Ω makes it possible to keep the test functions in $H^1(\Omega)$ instead of going to the larger space $H^1(R^n)$. For instance this feature is extremely important in the homogeneous Dirichlet problem in $H_0^1(\Omega)$ where it would not be possible to substitute $H^1(R^n)$.

7. Shape derivative of a functional: The general method and other examples. In this last section we formalize the ideas and techniques introduced in § 6 and describe two other examples to further illustrate the applicability of Theorems 3 or 4 and our associated techniques.

The first goes over the discussion at the end of § 6 in Remark 6.2. Key details are provided to show how problems with Dirichlet boundary conditions can be handled. In fact the suggested construction could also have been used right from the beginning in § 6, but we preferred to do it in a different way to better emphasize its importance.

The second example shows that we can handle problems where the smoothness of the solution of the saddle point equations is minimal. Other techniques based on the Implicit Function Theorem would require more smoothness.

7.1. Dirichlet boundary condition. We go back to problem (1)–(4) in § 6 but with $H_0^1(\Omega)$ instead of $H^1(\Omega)$. Let $y = y(\Omega)$ in $H_0^1(\Omega)$ be the solution of the variational problem

$$(1) \quad \inf \{E(\Omega, \varphi) : \varphi \in H_0^1(\Omega)\}$$

where

$$(2) \quad E(\Omega, \varphi) = \frac{1}{2} \int_{\Omega} [|\nabla \varphi|^2 + |\varphi|^2 - 2f\varphi] dx$$

for some fixed function f in $H^1(R^n)$. This is the homogeneous Dirichlet problem. We associate with y a cost function

$$(3) \quad J(\Omega) = F(\Omega, y(\Omega)).$$

Again for simplicity we assume that it is of the form

$$(4) \quad F(\Omega, \varphi) = \frac{1}{2} \int_{\Omega} (\varphi - Y_d)^2 dx, \quad \varphi \in H_0^1(\Omega), \quad Y_d \in H^1(R^n).$$

Assume that V is a smooth vector field which transports Ω onto Ω_t , its boundary Γ onto Γ_t and the Sobolev space $H^1(\Omega)$ onto $H^1(\Omega_t)$ at time $t \geq 0$. As a result it also transports functions in $H_0^1(\Omega)$ onto functions in $H_0^1(\Omega_t)$ and

$$(5) \quad H_0^1(\Omega_t) = \{\varphi \circ T_t^{-1} : \varphi \in H_0^1(\Omega)\}.$$

Here we use techniques described at the end of § 6 in Remark 6.2. We introduce the new functional

$$(6) \quad \varphi \rightarrow \underline{E}(t, \varphi) = E(\Omega_t, \varphi \circ T_t^{-1}) : H_0^1(\Omega) \rightarrow R$$

and notice that

$$(7) \quad \inf_{\varphi \in H_0^1(\Omega)} \underline{E}(t, \varphi) = \inf_{\psi \in H_0^1(\Omega_t)} E(\Omega_t, \psi).$$

Denote by y' and y_t the minimizing unique solutions of $\underline{E}(t, \varphi)$ in $H_0^1(\Omega)$ and $E(\Omega_t, \psi)$ in $H_0^1(\Omega_t)$, respectively. Then in view of (5)

$$(8) \quad y_t = y' \circ T_t^{-1}.$$

The two formulations are equivalent, but the differentiation of $\underline{E}(t, \varphi)$ with respect to t does not require that the function φ be smoother than $H^1(\Omega)$ since

$$\underline{E}(t, \varphi) = \frac{1}{2} \int_{\Omega_t} [|\nabla(\varphi \circ T_t^{-1})|^2 + |\varphi \circ T_t^{-1}|^2 - 2f(\varphi \circ T_t^{-1})] dx$$

and after a change of variable

$$(9) \quad \underline{E}(t, \varphi) = \frac{1}{2} \int_{\Omega} \{(A(t)\nabla \varphi) \cdot \nabla \varphi + [|\varphi|^2 - 2(f \circ T_t)\varphi]J(t)\} dx,$$

where DT_t is the Jacobian matrix associated with the transformation T_t ,

$$(10) \quad J(t) = \det(DT_t), \quad A(t) = J(t)((DT_t)^{-1})^*(DT_t)^{-1}$$

and $*$ denotes the transposed matrix.

If we want to work with $\underline{E}(t, \Omega)$ and y' , we must also transform the functional F into a new functional

$$(11) \quad \varphi \rightarrow \underline{F}(t, \varphi) = F(\Omega_t, \varphi \circ T_t^{-1}): H_0^1(\Omega) \rightarrow \mathbb{R}.$$

As a result the cost function

$$(12) \quad J(\Omega_t) = F(\Omega_t, y_t) = F(\Omega_t, y' \circ T_t^{-1}) = \underline{F}(t, y').$$

Again the differentiability of $\underline{F}(t, \varphi)$ with respect to t does not require that the function φ be smoother than $H^1(\Omega)$:

$$\underline{F}(t, \varphi) = \frac{1}{2} \int_{\Omega_t} (\varphi \circ T_t^{-1} - Y_d)^2 dx$$

and after a change of variable

$$(13) \quad \underline{F}(t, \varphi) = \frac{1}{2} \int_{\Omega} (\varphi - Y_d \circ T_t)^2 J(t) dx.$$

Thus we are led to the construction of the functional

$$(14) \quad (\varphi, \psi) \rightarrow \underline{G}_r(t, \varphi, \psi) = G_r(\Omega_t, \varphi \circ T_t^{-1}, \psi \circ T_t^{-1}): H_0^1(\Omega) \times H_0^1(\Omega) \rightarrow \mathbb{R}$$

and the technique used in § 6.

We do not repeat the details here since the results are the same as those in Theorem 9 except that the functions y and p are the solutions of the variational equations

$$(15) \quad y \in H_0^1(\Omega), \quad dE(\Omega, y; \varphi) = 0 \quad \forall \varphi \in H_0^1(\Omega),$$

$$(16) \quad p \in H_0^1(\Omega), \quad dF(\Omega, y; \psi) + d^2E(\Omega, y; p, \psi) = 0 \quad \forall \psi \in H_0^1(\Omega).$$

So formally it suffices to substitute $H_0^1(\Omega)$ for $H^1(\Omega)$ in Theorem 9.

7.2. An example with less smoothness. In the two previous examples, the solutions (y, p) of the optimality system (42)–(47) or (15)–(16) are smoother than anticipated and belong to $H^2(\Omega)$. So it would be possible to argue that all the results can also be obtained by application of some form of the Implicit Function Theorem.

It is not difficult to slightly modify the example of § 6 to prevent this situation from happening. First change the functionals E and F to

$$(17) \quad E(\Omega, \varphi) = \frac{1}{2} \int_{\Omega} [|\nabla \varphi|^2 + |\varphi|^2 - 2f \cdot \nabla \varphi] dx, \quad \varphi \in H^1(\Omega)$$

where $f \in (H^1(\Omega))^n$

$$(18) \quad F(\Omega, \varphi) = \int_{\Omega} |\nabla \varphi| dx, \quad \varphi \in H^1(\Omega).$$

The minimization problem

$$(19) \quad e(\Omega) = \inf \{E(\Omega, \varphi): \varphi \in H^1(\Omega)\}$$

still has a unique solution y in $H^1(\Omega)$ which coincides with the solution of the boundary value problem

$$(20) \quad -\operatorname{div} \nabla y + y - \operatorname{div} f = 0 \quad \text{in } \Omega, \quad \left(\frac{\partial y}{\partial n} \right) = 0 \quad \text{on } \Gamma.$$

As in § 7.1 we introduce the new functionals

$$(21) \quad \underline{E}(t, \varphi) = E(\Omega_t, \varphi \circ T_t^{-1}), \quad \underline{F}(t, \varphi) = F(\Omega_t, \varphi \circ T_t^{-1}), \quad J(t) = J(\Omega_t)$$

and transport all the integrals from Ω_t to Ω . We are now back to the setup at the end of § 5, and Theorem 7 and Proposition 8 apply with $u = t$.

As a result

$$(22) \quad \begin{aligned} dJ(\Omega; V) &= (d/dt)J(t)|_{t=0} \\ &= \text{Sup} \{dF(0, y; 1, 0) + d^2E(0, y; 0, p; 1, 0) : p \in P(0)\} \end{aligned}$$

where y is the solution of

$$(23) \quad dE(0, y; 0, \psi) = 0 \quad \forall \psi \in H^1(\Omega)$$

and $P(0)$ is the set of solutions of the adjoint variational inequality

$$(24) \quad dF(0, y; 0, \psi) + d^2E(0, y; 0, p; 0, \psi) \geq 0 \quad \forall \psi \in H^1(\Omega).$$

Note that the set $P(0)$ is not empty since the hypotheses of Proposition 8 are verified. However, the elements of $P(0)$ belong to $H^1(\Omega)$ but again not much more. In fact (24) reduces to

$$(25) \quad \begin{aligned} \int_{\Omega_+} (\nabla y / |\nabla y|) \cdot \nabla \psi \, dx + \int_{\Omega_0} |\nabla \psi| \, dx + \int_{\Omega} [\nabla p \cdot \nabla \psi + p\psi] \, dx \geq 0, \\ p \in H^1(\Omega) \quad \forall \psi \in H^1(\Omega) \end{aligned}$$

where

$$(26) \quad \Omega_+ = \{x \in \Omega : \nabla y(x) \neq 0\}, \quad \Omega_0 = \{x \in \Omega : \nabla y(x) = 0\}.$$

It is readily seen that (25) has at least one solution since the following variational equation has a unique solution:

$$(27) \quad \int_{\Omega_+} (\nabla y / |\nabla y|) \cdot \nabla \psi \, dx + \int_{\Omega} [\nabla p \cdot \nabla \psi + p\psi] \, dx = 0, \quad p \in H^1(\Omega) \quad \forall \psi \in H^1(\Omega).$$

Of course, as in § 5, this problem can be solved by replacing the nondifferentiable cost function (18) by a Sup

$$F(\Omega, \varphi) = \text{Sup} \{F(\Omega, \mu, \varphi) : \varphi \in M\}$$

of the functional

$$F(\Omega, \mu, \varphi) = \int_{\Omega} \mu \cdot \nabla \varphi \, dx$$

over the weakly compact subset

$$M = \{\mu \in (L^2(\Omega))^n : |\mu(x)| \leq 1 \text{ a.e. in } \Omega\}$$

of $(L^2(\Omega))^n$ where $|\cdot|$ denotes the Euclidean norm in R^n . Instead of the variational inequality (25) we would obtain the set of variational equations

$$\begin{aligned} p &= p(\alpha) \in H^1(\Omega) \quad \forall \psi \in H^1(\Omega), \\ \int_{\Omega_+} (\nabla y / |\nabla y|) \cdot \nabla \psi \, dx + \int_{\Omega_0} \alpha \cdot \nabla \psi \, dx + \int_{\Omega} [\nabla p \cdot \nabla \psi + p\psi] \, dx &= 0 \end{aligned}$$

indexed by $\alpha \in M$ and the Sup in (22) would be taken over M :

$$dJ(\Omega; V) = \text{Sup} \{dF(0, \alpha, y; 1, 0, 0) + d^2E(0, y; 0, p(\alpha); 1, 0) : \alpha \in M\}$$

where

$$F(t, \alpha, \varphi) = F(t, \alpha \circ T_t^{-1}, \varphi \circ T_t^{-1}).$$

The proof is the same as the one given in § 5 for the nondifferentiable observation.

Note added in proof. The shape sensitivity analysis problem has also recently been studied by a penalization technique. The results apply to some classes of nonlinear problems and some problems governed by variational inequalities (see U. C. Delfour and J. P. Zolésio, Shape Sensitivity Analysis via a Penalization Method, Ann. Mat. Pura Appl., to appear).

REFERENCES

- J. P. AUBIN [1], *L'analyse non linéaire et ses motivations économiques*, Masson, Paris, New York, 1984.
- N. V. BANICHUK [1], *Optimization of the Shapes of Elastic Bodies*, Nauka, Moscow, 1980 (English translation 1984).
- J. CÉA [1], *Conception optimale ou identification de formes. Calcul rapide de la dérivée directionnelle de la fonction coût*, R.A.I.R.O., 20 (1986), pp. 371–402.
- [2], *Problems of shape optimal design*, in *Optimization of Distributed Parameter Structures*, Vol. II, E. J. Haug and J. Céa, eds., Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981, pp. 1005–1048.
- [3], *Numerical methods of shape optimal design*, in *Optimization of Distributed Parameter Structures*, Vol. II, E. J. Haug and J. Céa, eds., Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981, pp. 1049–1087.
- R. CORREA AND A. SEEGER [1], *Directional derivatives of a minimax function*, *Nonlinear Anal. Theory Methods and Appl.*, 9 (1985), pp. 13–22.
- M. C. DELFOUR AND J. P. ZOLÉSIO [1], *Dérivation d'un Min Max et application à la dérivation par rapport au contrôle d'une observation non différentiable de l'état*, *C. R. Acad. Sci. Paris Sér. I Math.*, 302 (1986), pp. 571–574.
- [2], *Differentiability of a Min Max and Application to Optimal Control and Design problems. Parts I and II.*, in *Control Problems for Systems described by Partial Differential Equations and Applications*, I. Lasiecka and R. Triggiani, eds., Springer-Verlag, New York, Berlin, 1987, pp. 204–219, pp. 220–229.
- K. DEMS AND Z. MROZ [1], *Variational approach by means of adjoint systems to structural optimization and sensitivity analysis, Part 2. Structure shape variations*, *Internat. J. Solids and Structures*, 20 (1984), pp. 527–552.
- V. F. DEM'YANOV [1], *Differentiability of a Maximin function. I*, *USSR Comput. Math. and Math. Phys.*, 8 (1968), pp. 1–15. (In English.) *Z. Vychisl. Mat. i Mat. Fiz.*, 8 (1968), pp. 1186–1195. (In Russian.)
- I. EKELAND AND R. TEMAM [1], *Analyse convexe et problèmes variationnels*, Dunod, Gauthiers-Villars, Paris, Brussels, Montreal, 1974.
- R. H. GALLAGHER AND O. C. ZIENKIEWICZ [1], *Optimum Structural Design Theory and Applications*, John Wiley, New York, 1972; Springer-Verlag, Berlin, New York, 1973.
- E. J. HAUG [1], *A review of distributed parameter structural optimization literature*, in *Optimization of Distributed Parameter Structures*, Vol. I, E. J. Haug and J. Céa, eds., Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981, pp. 3–74.
- E. J. HAUG AND J. S. ARORA [1], *Applied Optimal Design*, Wiley-Interscience, New York, 1979.
- E. J. HAUG AND J. CÉA, EDS. [1], *Optimization of Distributed Parameter Structures*, Vols. I and II, Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981.
- E. J. HAUG, K. K. CHOI AND V. KOMKOV [1], *Design Sensitivity Analysis of Structural Systems*, Academic Press, New York, 1986.
- A. MYSLINSKI AND J. SOKOŁOWSKI [1], *Nondifferentiable optimization problems for elliptic problems*, *SIAM J. Control Optim.*, 23 (1984), pp. 632–648.
- O. PIRONNEAU [1], *Optimal Design for Elliptic Systems*, Springer-Verlag, Berlin, New York, 1984.
- W. PRAGER [1], *Introduction to Structural Optimization, Courses and Lectures: International Centre for Mechanical Sciences*, Springer-Verlag, Vienna, 1974.
- G. I. N. ROZVANY [1], *Optimal Design of Flexural Systems*, Pergamon Press, New York, 1976.
- A. SAWCZUK AND Z. MROZ [1], *Optimization in Structural Design*, Springer-Verlag, Berlin, New York, 1975.
- J. SOKOŁOWSKI [1], *Optimal control in coefficients of boundary value problems with unilateral constraints*, *Bull. Polish Acad. Sci. Tech. Sci.*, 31 (1983), pp. 71–81.
- J. SOKOŁOWSKI AND J. P. ZOLÉSIO [1], *Dérivée par rapport au domaine de la solution d'un problème unilatéral*, *C. R. Acad. Sci. Paris Sér. I*, 301 (1985), pp. 103–106.
- [2], *Shape sensitivity analysis of an elasto-plastic torsion problem*, *Bull. Polish Acad. Sci. Tech. Sci.*, 33 (1985), pp. 579–586.
- J. P. ZOLÉSIO [1], *Identification de Domaine*, Thèse de doctorat d'état, Nice, France, 1979.
- [2], *Semi-derivatives of repeated eigenvalues*, in *Optimization of Distributed Parameter Structures*, Vol. II, E. J. Haug and J. Céa, eds., Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981, pp. 1457–1473.
- [3], *An optimal design procedure for optimal control support*, in *Convex Analysis and its Applications*, A. Auslender, ed., Springer-Verlag, Berlin, New York, 1977, pp. 207–219.

REALISATION AND APPROXIMATION OF LINEAR INFINITE-DIMENSIONAL SYSTEMS WITH ERROR BOUNDS*

KEITH GLOVER†, RUTH F. CURTAIN‡, AND JONATHAN R. PARTINGTON†

Abstract. The class of linear infinite-dimensional systems with finite-dimensional inputs and outputs whose impulse response h satisfies $h \in L_1 \cap L_2(0, \infty; \mathbb{C}^{p \times m})$ and induces a nuclear Hankel operator is said to be of nuclear type. For this class of systems it is shown that balanced or output normal realisations always exist and their truncations converge to the original system in various topologies. Furthermore, explicit L_∞ bounds on the transfer function errors, L_1 and L_2 bounds on the impulse response errors, and Hilbert-Schmidt and nuclear bounds on the Hankel operator errors are obtained. These truncations also generate an approximating sequence to the optimal Hankel-norm approximations to the original system, and various error bounds of these approximants are deduced.

Key words. infinite-dimensional system, linear infinite-dimensional system, Hankel operator, realisations, balanced realisations, optimal Hankel-norm approximations, Hankel operator

AMS(MOS) subject classifications. 93C25, 93B15, 41A65, 41A20

1. Introduction. Infinite-dimensional linear systems are quartets of linear operators (A, B, C, D) mapping between various infinite-dimensional linear vector spaces. Under suitable assumptions they lead to the existence of a nonrational transfer function, $G(s) = D + C(sI - A)^{-1}B$. In [15] Fuhrmann develops a realisation theory for linear systems for which the linear vector spaces are Hilbert spaces and the transfer functions have values in Hardy spaces. The results are reminiscent of the finite-dimensional case, but they are less complete. In any case, for a very large class of transfer functions there exists a restricted shift realisation ([15, p. 298]) and for the discrete-time case, a balanced realisation (Young [27]). Balanced realisations for continuous-time systems were treated by Curtain and Glover in [5].

While realisations are of theoretical interest, of more importance for the applications is the question of approximation of infinite-dimensional systems by finite-dimensional ones. For finite-dimensional systems it has been found that balanced realisations (Moore [21], Pernebo, and Silverman [23]) give good approximations when truncated (Enns [14], Glover [16]). In this paper we prove convergence of truncations of balanced realisations with error bounds for a class of linear infinite-dimensional continuous-time systems including the case when the impulse response $h \in L_1 \cap L_2(0, \infty; \mathbb{C}^{p \times m})$ and the induced Hankel operator is nuclear with distinct singular values. Moreover, we obtain approximations with error estimates to the optimal Hankel-norm approximations (cf. [1]). At the same time we show the existence of optimal Hankel-norm approximations to infinite-dimensional linear systems of nuclear type and we give a parametrization in terms of a linear fractional map. This is similar in spirit to the parametrization given by Ball and Ran in [3] for the special case of A, B, C bounded operators.

In § 2 we define systems of nuclear type and introduce the Hankel operator and prove some elementary properties of its singular values and Schmidt vectors. The output normal realisation derived in § 3 has the same sequence of truncated transfer

* Received by the editors February 3, 1986; accepted for publication (in revised form) September 18, 1987.

† Information Engineering Division, Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom.

‡ Department of Mathematics, University of Groningen, Groningen, the Netherlands.

functions as the balanced realisation found in [5], but it is more convenient to work with. In § 4 the convergence of the truncated realisations is proved: for nuclear-type real systems the truncations have impulse responses which converge in L_1 and transfer functions which converge in L_∞ . Using finite-dimensional results from [16], we find explicit bounds on the truncation errors in terms of the singular values in § 5. In § 6 the relationship between optimal Hankel-norm approximations to the infinite-dimensional system and its truncations are considered. This leads to an existence result to the optimal Hankel-norm approximation problem for infinite-dimensional systems of nuclear type (cf. [1]) and to a sequence of approximations with explicit error estimates. The estimates in §§ 5 and 6 are the natural analogues of the finite-dimensional estimates in [16], together with many that are also new for finite-dimensional systems. Finally, we conclude with an example in § 7 that can be completely analysed.

2. Hankel operators. We consider the class of linear, infinite-dimensional linear systems defined by the following input-output map:

$$(2.1) \quad y(t) = \int_0^\infty h(t-s)u(s) ds$$

where the outputs $y \in L_2(0, \infty; \mathbb{C}^p)$ and the inputs $u \in L_2(0, \infty; \mathbb{C}^m)$ and the impulse response satisfies

$$(2.2) \quad h \in L_1 \cap L_2(0, \infty; \mathbb{C}^{p \times m}).$$

Corresponding to (2.2) we have the Hankel operator $\Gamma: L_2(0, \infty; \mathbb{C}^m) \rightarrow L_2(0, \infty; \mathbb{C}^p)$ defined by

$$(2.3) \quad (\Gamma u)(t) = \int_0^\infty h(t+s)u(s) ds.$$

As in [1] and [20] the condition $h \in L_1 \cap L_2(0, \infty; \mathbb{C}^{p \times m})$ implies that Γ is a compact operator from the spaces $L_q(0, \infty; \mathbb{C}^m)$, $q \geq 1$ and $C^1(0, \infty; \mathbb{C}^m)$ to $L_q(0, \infty; \mathbb{C}^p)$ and $C^1(0, \infty; \mathbb{C}^p)$, where the latter has norm

$$(2.4) \quad \|f\| = \sup_{0 \leq s \leq \infty} |f(s)| + \int_0^\infty |\dot{f}(s)| ds.$$

Functions in C^1 are absolutely continuous, and their derivatives exist in the L_1 sense.

For the sake of completeness, we outline a proof of these assertions in Appendix 1 of this paper.

$\Gamma^*\Gamma$ is compact and positive on $L_2(0, \infty; \mathbb{C}^m)$ and so it has countably many positive eigenvalues $\sigma_1^2 \geq \sigma_2^2 \geq \dots \geq \sigma_i^2 \geq \dots \geq 0$. $\sigma_i \geq 0$ are the *singular values* of Γ . If v_i and w_i are the corresponding normalized eigenvectors of $\Gamma^*\Gamma$ and $\Gamma\Gamma^*$, respectively, then (v_i, w_i) , $i \geq 1$ are called the *Schmidt pairs* of Γ and therefore we have

$$(2.5) \quad \begin{aligned} \Gamma v_i &= \sigma_i w_i, \\ \Gamma^* w_i &= \sigma_i v_i, \end{aligned} \quad i = 1, 2, \dots$$

Now w_i is an eigenvector of $\Gamma\Gamma^*$ which is compact in the spaces $L_q(0, \infty; \mathbb{C}^p) \cap C^1(0, \infty; \mathbb{C}^p)$, $q \geq 1$. As in [1] it follows from results of Gohberg and Zambickii [28] that

$$(2.6) \quad w_i \in L_1(0, \infty; \mathbb{C}^p) \cap L_2(0, \infty; \mathbb{C}^p) \cap C^1(0, \infty; \mathbb{C}^p)$$

and similarly

$$(2.7) \quad v_i \in L_1(0, \infty; \mathbb{C}^m) \cap L_2(0, \infty; \mathbb{C}^m) \cap C^1(0, \infty; \mathbb{C}^m).$$

A more accessible source for the results of [28] is [29]. There the following results are derived independently.

Let B be a Banach space on which is defined a continuous scalar product (x, y) , which induces a new norm under which the completion of B is a Hilbert space H . Suppose that T is a bounded linear transformation on B such that $(x, Ty) = (Tx, y)$ for all x and y in B . Then

I: T is bounded with respect to the Hilbert norm.

II: The spectrum of T over H is a subset of the spectrum of T over B .

III: If λ belongs to the point spectrum of T over B , then λ belongs to the point spectrum of T over H and the nullspace of $T - \lambda I$ over B and H is the same.

For the application above we take B to be $L_1 \cap L_2 \cap C^1(0, \infty; \mathbb{C}^p)$ with the usual L_2 scalar product. The compactness of $T = \Gamma\Gamma^*$ guarantees that every nonzero point of the spectrum of T is actually in the point spectrum, and hence is an eigenvalue.

For simplicity of notation we shall subsequently suppress the $\mathbb{C}^m, \mathbb{C}^p$.

Hankel kernels may be expressed in terms of their Schmidt pairs and singular values.

LEMMA 2.1.

$$(2.8) \quad (\Gamma u)(t) = \int_0^\infty h(t+s)u(s) ds = \sum_{i=1}^\infty \sigma_i w_i(t) \int_0^\infty v_i^*(s)u(s) ds,$$

$$(2.9) \quad (\Gamma u)(t+s) = \int_0^\infty h(t+s+\alpha)u(\alpha) d\alpha \\ = \sum_{i=1}^\infty \sigma_i w_i(t) \int_0^\infty v_i^*(s+\alpha)u(\alpha) d\alpha \quad \text{in } L_2(dt) \text{ for all } s \geq 0,$$

$$(2.10) \quad h(t+\alpha) = \sum_{i=1}^\infty w_i(t)\sigma_i v_i^*(\alpha) \quad \text{in } L_2(d\alpha) \text{ for all } t \geq 0.$$

Proof. Equation (2.8) is just the polar decomposition for Γ [15], [20], [24].

For equation (2.9), we define for fixed s , the function

$$u_s(\beta) = \begin{cases} u(\beta-s), & \beta \geq s, \\ 0, & \beta < s. \end{cases}$$

Then we have

$$\int_0^\infty h(t+s+\alpha)u(\alpha) d\alpha = \int_0^\infty h(t+\beta)u_s(\beta) d\beta \\ = \sum_{i=1}^\infty \sigma_i w_i(t) \int_0^\infty v_i^*(\beta)u_s(\beta) d\beta \quad \text{from (2.8)} \\ = \sum_{i=1}^\infty \sigma_i w_i(t) \int_0^\infty v_i^*(s+\alpha)u(\alpha) d\alpha$$

and (2.9) is established.

For (2.10), since $h \in L_2(0, \infty)$ the shifted function $h_t(\cdot) = h(t+\cdot)$ and its rows h_t^k . Thus

$$h_t^k(\cdot) = \sum_{i=1}^\infty \langle h_t^k, v_i^* \rangle v_i^* + \sum_{i=1}^\infty \langle h_t^k, \tilde{v}_i^* \rangle \tilde{v}_i^*$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product in $L_2(0, \infty)$ and where \tilde{v}_i extends v_i to form a complete orthonormal basis for $L_2(0, \infty)$. From (2.5) we obtain

$$\langle h_i^k, v_i^* \rangle = \int_0^\infty h^k(t+s) v_i(s) ds = \sigma_i w_i^k(t)$$

and so for $t \geq 0$

$$h(t+s) = \sum_{i=1}^\infty w_i(t) \sigma_i v_i^*(s) + \sum_{i=1}^\infty A_i \tilde{v}_i^*(s).$$

Substituting $u = \tilde{v}_i$ in (2.8) shows that $A_i = 0$ for all i and (2.10) is established.

Of course relationships dual to (2.8)–(2.10) also hold for the adjoint Hankel operator Γ^* , but we shall need only the following:

$$(2.11) \quad h^*(t+\alpha) = \sum_{i=1}^\infty v_i(t) \sigma_i w_i^*(\alpha) \quad \text{in } L^2(d\alpha) \quad \text{for all } t \geq 0.$$

Since $\Gamma^* \Gamma$ and $\Gamma \Gamma^*$ are self-adjoint we have first

$$(2.12) \quad \langle v_i, v_j \rangle = \delta_{ij} = \langle w_i, w_j \rangle.$$

This equality together with the representations (2.10) and (2.11) for the square integrable functions h and h^* lead directly to the following useful inequalities:

$$(2.13) \quad \sum_{i=1}^\infty \sigma_i^2 |w_i(t)|^2 < \infty, \quad \sum_{i=1}^\infty \sigma_i^2 |v_i(t)|^2 < \infty, \quad t \geq 0.$$

Γ is Hilbert–Schmidt if and only if

$$(2.14) \quad \sum_{i=1}^\infty \sigma_i^2 < \infty$$

which is equivalent to the condition $t^{1/2} \lambda \in L_2(0, \infty; \mathbb{C}^{p \times m})$ [20]. However, in parts of §§ 4 and 5 we need to impose the stronger assumption that Γ is nuclear, that is

$$(2.15) \quad \sum_{i=1}^\infty \sigma_i < \infty.$$

We write $\|\Gamma\|_N$ to denote $\sum_{i=1}^\infty \sigma_i$, the nuclear norm of Γ .

We have the following relation holding between the nuclear norm of Γ and the L_1 norm of h .

THEOREM 2.1. $\|h\|_1 \leq 2\|\Gamma\|_N$.

Proof. Suppose first that h is continuous and of compact support $[0, M]$. For $\alpha > 0$ and $n = 0, 1, 2, \dots$ define $e_n^\alpha(t) = (1/\sqrt{\alpha}) u_n^\alpha \chi_{(n\alpha, (n+1)\alpha)}(t)$, where $u_n^\alpha \in \mathbb{C}^m$ satisfies $\|u_n^\alpha\| = 1$ and $\|h(2n\alpha) u_n^\alpha\| = \|h(2n\alpha)\|$. Also define $f_n^\alpha(t) = (1/\sqrt{\alpha}) y_n^\alpha \chi_{(n\alpha, (n+1)\alpha)}(t)$, where $y_n^\alpha \in \mathbb{C}^p$ satisfies $\|y_n^\alpha\| = 1$ and $\|(h(2n\alpha) u_n^\alpha, y_n^\alpha)\| = \|h(2n\alpha)\|$. Thus $e_0^\alpha, e_1^\alpha, e_2^\alpha, \dots$ forms an orthonormal sequence in $L_2(0, \infty; \mathbb{C}^m)$ and $f_0^\alpha, f_1^\alpha, f_2^\alpha, \dots$ forms one in $L_2(0, \infty; \mathbb{C}^p)$.

Consider

$$(\Gamma e_n^\alpha, f_n^\alpha) = \frac{1}{\alpha} \int_{n\alpha}^{(n+1)\alpha} \int_{n\alpha}^{(n+1)\alpha} (h(s+t) u_n^\alpha, y_n^\alpha) ds dt.$$

By the uniform continuity of h , given $\varepsilon > 0$ we have, for sufficiently small α , that

$$\|h(v_1) - h(v_2)\| < \frac{\varepsilon}{2M} \quad \text{if } |v_1 - v_2| < 2\alpha.$$

Writing $v = s + t$, $w = s - t$, we obtain

$$(\Gamma e_n^\alpha, f_n^\alpha) = \frac{1}{2\alpha} \int_{v=2n\alpha}^{2(n+1)\alpha} \int_{w=-\lambda(v)}^{\lambda(v)} (h(v) u_n^\alpha, y_n^\alpha) dv dw$$

where

$$\begin{aligned} \lambda(v) &= \min(v - 2n\alpha, 2(n+1)\alpha - v) \\ &= \frac{1}{2\alpha} \int_{2n\alpha}^{2(n+1)\alpha} 2\lambda(v) (h(v) u_n^\alpha, y_n^\alpha) dv \end{aligned}$$

so that

$$\left| (\Gamma e_n^\alpha, f_n^\alpha) - \frac{1}{2\alpha} \int_{2n\alpha}^{2(n+1)\alpha} 2\lambda(v) \|h(2n\alpha)\| dv \right| \leq \frac{1}{2\alpha} \int_{2n\alpha}^{2(n+1)\alpha} 2\lambda(v) \frac{\varepsilon}{2M} dv = \alpha \frac{\varepsilon}{2M},$$

since the mean value of $\lambda(v)$ is $\alpha/2$. Hence

$$|(\Gamma e_n^\alpha, f_n^\alpha) - \alpha \|h(2n\alpha)\|| \leq \alpha \frac{\varepsilon}{2M}$$

and so

$$\left| (\Gamma e_n^\alpha, f_n^\alpha) - \frac{1}{2} \int_{2n\alpha}^{2(n+1)\alpha} \|h(v)\| dv \right| \leq \alpha \frac{\varepsilon}{M},$$

using again the uniform continuity of h . Hence

$$\left| \sum_{n=0}^{\infty} (\Gamma e_n^\alpha, f_n^\alpha) - \frac{1}{2} \int_0^{\infty} \|h(v)\| dv \right| \leq \alpha \frac{\varepsilon}{M} \left(\frac{M}{\alpha} + 1 \right)$$

and so

$$\sum_{n=0}^{\infty} (\Gamma e_n^\alpha, f_n^\alpha) \rightarrow \frac{1}{2} \|h\|_1 \quad \text{as } \alpha \rightarrow 0.$$

(For each α the sum is finite, with at most $(M/\alpha + 1)$ nonzero terms.)

However, for any $h_0 \in L_1$ and $\alpha > 0$, with corresponding operator Γ_0 ,

$$|(\Gamma_0 e_n^\alpha, f_n^\alpha)| \leq \int_{2n\alpha}^{2(n+1)\alpha} \|h_0(v)\| dv,$$

since $\lambda(v) \leq \alpha$ for all v . Thus $\sum_{n=0}^{\infty} |(\Gamma_0 e_n^\alpha, f_n^\alpha)| \leq \|h_0\|_1$.

Given $\varepsilon > 0$, and arbitrary $h \in L_1$, we may write $h = h_1 + h_2$, where h_1 is continuous with compact support and $\|h_2\|_1 < \varepsilon/2$. Correspondingly, $\Gamma = \Gamma_1 + \Gamma_2$.

Let (e_n^α, f_n^α) be chosen as above, corresponding to h_1 . Then

$$\begin{aligned} \left| \sum_{n=0}^{\infty} |(\Gamma e_n^\alpha, f_n^\alpha)| - \frac{1}{2} \|h\|_1 \right| &\leq \sum_{n=0}^{\infty} |(\Gamma_2 e_n^\alpha, f_n^\alpha)| + \left| \sum_{n=0}^{\infty} |(\Gamma_1 e_n^\alpha, f_n^\alpha)| - \frac{1}{2} \|h\|_1 \right| \\ &\leq \|h_2\|_1 + \left| \sum_{n=0}^{\infty} |(\Gamma_1 e_n^\alpha, f_n^\alpha)| - \frac{1}{2} \|h_1\|_1 \right| + \frac{1}{2} \|h_2\|_1 \\ &< \varepsilon \quad \text{for sufficiently small } \alpha. \end{aligned}$$

It now follows from standard results, e.g., [33, p. 215], that

$$\|\Gamma\|_N \cong \frac{1}{2} \|h\|_1,$$

since, when $\Gamma: X \rightarrow Y$,

$$\|\Gamma\|_N = \sup \left\{ \sum_{n=1}^{\infty} |(\Gamma e_n, f_n)| : (e_n) \text{ orthonormal in } X, (f_n) \text{ orthonormal in } Y \right\}.$$

The above result was also proved by Gohberg and Doyle (see [13]) for the case when h has a rational Laplace transform.

There are a variety of conditions on h to ensure nuclearity (see Power [24], Curtain [9]). The class of systems satisfying (2.2) and (2.15) has particularly nice approximation properties and so we introduce the following definition.

DEFINITION 2.1. The infinite-dimensional linear system (2.1) is of *nuclear type* if it determines a bounded Hankel operator Γ whose singular values satisfy $\sum_1^{\infty} \sigma_j < \infty$.

The approximation of infinite-dimensional systems by finite-dimensional ones is aided by the results of Coifman and Rochberg [31] (see also [24]). These imply that there exist complex numbers ξ_i in the left half plane such that a Hankel operator Γ is nuclear if and only if its transfer function $G(s)$ can be expressed in the form

$$(2.16) \quad G(s) = \sum_{i=1}^{\infty} a_i (\operatorname{Re} \xi_i) (s - \xi_i)^{-1}$$

where $\sum |a_i| \leq C \|\Gamma\|_N$ (C an absolute constant). The series converges uniformly in $\operatorname{Re} s > 0$ and in the nuclear norm of the associated operators. This result will be of use to us later, but does have the following consequence.

COROLLARY 2.1. If $h(t)$ yields a nuclear Hankel operator Γ , then $h(t)$ is equal almost everywhere to a function $j(t)$ which is continuous on $(0, \infty)$ and satisfies $|j(t)| \leq M \|\Gamma\|_N / t$ for all $t > 0$, where M is an absolute constant.

Proof. Consider first the SISO case $p = m = 1$. Let

$$(2.17) \quad j(t) = \sum_{i=1}^{\infty} a_i (\operatorname{Re} \xi_i) e^{\xi_i t}$$

where (a_i) and (ξ_i) are as in (2.16). Since $\sum |a_i| \leq C \|\Gamma\|_N$,

$$\sup_{t \in [\delta, \infty)} |x e^{-tx}| = x e^{-\delta x}, \quad \sup_{x \geq 0} |x e^{-\delta x}| = \frac{1}{e\delta},$$

the series for $j(t)$ converges uniformly on every $[\delta, \infty)$ and $|j(t)| \leq c \|\Gamma\|_N / et$.

However, the right-hand side of (2.17) converges in L_1 norm to h . Hence $h(t) = j(t)$ almost everywhere.

In the MIMO case the argument proving continuity of j is still valid. The fact that M is independent of p and m follows because

$$\begin{aligned} |j(t)| &= \sup \{ |(j(t)e, f)| : e \in \mathbb{C}^m, f \in \mathbb{C}^p, \|e\| = \|f\| = 1 \} \\ &\leq M \sup_{e, f} \|\Gamma_{e, f}\|_N / t \end{aligned}$$

where $\Gamma_{e, f}$ is the Hankel operator with kernel $(j(t)e, f)$. Since $\|\Gamma_{e, f}\|_N \leq \|\Gamma\|_N$, the result follows.

In the sequel we restrict ourselves to the generic case where the singular values are distinct. Extensions to the case where σ_i has a multiplicity greater than one should be possible by introducing chains of Schmidt pairs as in [1], but it would add

complications to some of the proofs. Consequently we make the standing assumption that all singular values are distinct.

3. An output normal realisation. For approximation of finite-dimensional systems we often start with a balanced realisation (Moore [21] and Pernebo and Silverman [23]). For infinite-dimensional systems satisfying (2.2), Curtain and Glover have shown in [5] that *balanced* realisations also exist. However, for technical reasons we find it more convenient to work with a closely related realisation, called the *output normal* realisation.

First we give a time domain definition of a realisation for an infinite-dimensional system. This definition is less general than [15, p. 296] which covers impulse responses which are distributions. However a time domain definition is more appropriate for our purposes as our systems are specified in terms of an impulse response which is a function in $L_1(0, \infty; \mathbb{C}^{p \times m})$.

DEFINITION 3.1. A Hilbert space realisation of the Hankel operator Γ (as defined in (2.3)) is a triple $(C, T(t), B)$, where $T(t)$ is a C_0 -semigroup in the Hilbert space H (with infinitesimal generator A) and $B: \mathbb{C}^m \rightarrow \text{Dom}(A^*)^*$ and $C: \text{Dom}(A) \rightarrow \mathbb{C}^p$ are linear maps such that

(i) The *reliability operator* $\underline{B}: L_2(0, \infty; \mathbb{C}^m) \rightarrow H$ given by

$$(3.1) \quad \underline{B}u = \int_0^\infty T(t)Bu(t) dt$$

is bounded;

(ii) The *observability operator* $\underline{C}: H \rightarrow L_2(0, \infty; \mathbb{C}^p)$ given by

$$(3.2) \quad \underline{C}x = CT(t)x$$

is bounded;

(iii)

$$(3.3) \quad \Gamma = \underline{C}\underline{B}.$$

We remark that

$$(3.4) \quad \text{dom}(A) \subset H \subset \text{dom}(A^*)^*$$

and T restricts to a C_0 -semigroup on $\text{dom}(A)$ and has a unique extension to a C_0 -semigroup on $\text{dom}(A^*)^*$. This implies that (3.1) and (3.2) are well defined.

Furthermore the following controllability and observability gramians P and Q are well defined whenever the reachability and observability maps are: they are given by

$$(3.5) \quad P = \underline{B}\underline{B}^* = \int_0^\infty T(t)BB^*T^*(t) dt,$$

$$(3.6) \quad Q = \underline{C}^*\underline{C} = \int_0^\infty T^*(t)C^*CT(t) dt.$$

In [5], balanced realisations were considered, but here we are concerned with output normal realisations.

DEFINITION 3.2. $(C, T(t), B)$ is a *balanced* realisation for the Hankel operator Γ if (3.1)–(3.3) hold and the observability and controllability gramians are both equal to the same positive operator.

$(C, T(t), B)$ is an *output normal* realisation for Γ if (3.1)–(3.3) hold, the observability gramian equals the identity operator, and the controllability gramian is a positive operator.

Furthermore, it is often convenient to choose coordinates in which the above gramians have a diagonal form.

We now propose an explicit representation for an output normal realisation on the state l_2 , for the class of infinite-dimensional linear systems (2.1) satisfying (2.2). In fact it is a special case of the restricted translation realisation of [15, Thm. 6.5] using a suitable matrix representation. However, we have chosen to give a simpler self-contained derivation, which is directly related to the results on balanced realisations in [5], to which we appeal in § 4.

From (2.6) and (2.7) the following realisation is well defined:

$$(3.7) \quad T(t)_{ij} = \int_0^\infty w_i^*(s) w_j(t+s) ds,$$

$$(3.8) \quad C = [w_1(0), w_2(0), \dots, w_i(0), \dots],$$

$$(3.9) \quad B = [\sigma_1 v_1(0), \sigma_2 v_2(0), \dots, \sigma_i v_i(0), \dots]^*.$$

If we define Σ to be the diagonal operator

$$(3.10) \quad \Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_i, \dots),$$

then the balanced realisation given in [5] is precisely $(C\Sigma^{1/2}, \Sigma^{-1/2}T(t)\Sigma^{1/2}, \Sigma^{-1/2}B)$. We proceed to prove some results concerning $T(t)$.

LEMMA 3.1.

$$T(t)_{ij} = \frac{\sigma_i}{\sigma_j} \int_0^\infty v_i^*(t+s) v_j(s) ds.$$

Proof.

$$\begin{aligned} \sigma_j T(t)_{ij} &= \int_0^\infty w_i^*(s) \Gamma v_j(t+s) ds \quad \text{from (2.5)} \\ &= \int_0^\infty \int_0^\infty w_i^*(s) h(t+s+\alpha) v_j(\alpha) d\alpha ds \quad \text{from (2.3).} \end{aligned}$$

Formula (2.2) implies that we may interchange the order of integration. Thus

$$\begin{aligned} \sigma_j T(t)_{ij} &= \int_0^\infty (\Gamma^* w_i(t+\alpha))^* v_j(\alpha) d\alpha \\ &= \sigma_i \int_0^\infty v_i^*(t+\alpha) v_j(\alpha) d\alpha \quad \text{by (2.5).} \end{aligned}$$

LEMMA 3.2. $T(t)$ is a contraction on l_2 for each $t \geq 0$.

Proof. (a) Let l_2^0 denote the subspace of l_2 comprising all sequences of finite length. Then for $x \in l_2^0$ we calculate

$$\begin{aligned} \|T(t)x\|^2 &= \sum_{i=1}^\infty \left| \sum_{j=1}^N x_j \int_0^\infty w_i^*(s) w_j(t+s) ds \right|^2 \\ (3.11) \quad &= \sum_{i=1}^\infty \sum_{j=1}^N \sum_{k=1}^N x_j \bar{x}_k \int_0^\infty w_i^*(s) w_j(t+s) ds \int_0^\infty w_k^*(t+s) w_i(s) ds \\ &= \sum_{j=1}^N \sum_{k=1}^N x_j \bar{x}_k \langle w_k(t+\cdot), w_j(t+\cdot) \rangle \end{aligned}$$

since from (2.5) and (2.9) $w_j(t+\cdot)$ has an expansion in terms of the orthonormal series

$w_i(\cdot)$. Thus

$$\begin{aligned}\|T(t)x\|^2 &= \sum_{j=1}^N \sum_{k=1}^N x_j \bar{x}_k \left(\delta_{jk} - \int_0^t w_k^*(s) w_j(s) ds \right) \quad \text{by (2.13)} \\ &= \|x\|^2 - \int_0^t \left(\sum_{k=1}^N x_k w_k(s) \right)^* \left(\sum_{j=1}^N x_j w_j(s) \right) ds \\ &= \|x\|^2 - \int_0^t \left| \sum_{j=1}^N x_j w_j(s) \right|^2 ds \\ &\leq \|x\|^2.\end{aligned}$$

Now since l_2^0 is dense in l_2 , we can extend the inequality $\|T(t)x\| \leq \|x\|$ to all $x \in l_2$, which proves that $T(t)$ is a contraction.

LEMMA 3.3. $T(t)$ is a strongly continuous contraction semigroup on l_2 whose infinitesimal generator A satisfies

$$(3.12) \quad A_{ij} = \int_0^\infty w_i^*(s) \dot{w}_j(s) ds = \frac{\sigma_i}{\sigma_j} \int_0^\infty \dot{v}_i^*(s) v_j(s) ds.$$

Proof. (a) $T(0) = I$ follows from the fact that

$$\lim_{t \rightarrow 0} \int_0^\infty w_i^*(s) w_j(t+s) ds = \int_0^\infty w_i^*(s) w_j(s) ds = \delta_{ij}$$

since $w_i \in C^1(0, \infty) \cap L_2(0, \infty)$.

(b) For the semigroup property, we calculate

$$\begin{aligned}(T(t)T(s))_{ij} &= \sum_{k=1}^\infty T(t)_{ik} T(s)_{kj} \\ &= \sum_{k=1}^\infty \int_0^\infty w_i^*(\alpha) w_k(t+\alpha) d\alpha \int_0^\infty w_k^*(\beta) w_j(s+\beta) d\beta \quad \text{from (3.7)} \\ &= \sum_{k=1}^\infty \int_t^\infty w_i^*(\beta-t) w_k(\beta) d\beta \int_0^\infty w_k^*(\beta) w_j(s+\beta) d\beta \\ &= \int_t^\infty w_i^*(\beta-t) w_j(s+\beta) d\beta\end{aligned}$$

since $w_j(s+\cdot)$ can be expanded as a series in w_k

$$= \int_0^\infty w_i^*(\alpha) w_j(t+s+\alpha) d\alpha = T(t+s)_{ij} \quad \text{by (3.7)}.$$

(c) For strong continuity at the origin it suffices to show it for all $x \in l_2^0$, since by (a), $\|T(t)\| \leq 1$. Now for $x \in l_2^0$,

$$\begin{aligned}\|T(t)x - x\|^2 &= \|T(t)x\|^2 + \|x\|^2 - \langle T(t)x, x \rangle - \langle x, T(t)x \rangle \\ &= \sum_{j=1}^N \sum_{k=1}^N x_j \bar{x}_k \left[\int_0^\infty w_j^*(t+s) w_k(t+s) ds + \delta_{jk} \right. \\ &\quad \left. - \int_0^\infty w_k^*(s) w_j(t+s) ds - \int_0^\infty w_j^*(s) w_k(t+s) ds \right] \quad \text{by (3.11)} \\ &= \sum_{j=1}^N \sum_{k=1}^N x_j \bar{x}_k \left(\int_0^\infty [w_j^*(t+s) - w_j^*(s)] w_k(t+s) ds \right. \\ &\quad \left. + \int_0^\infty w_k^*(s) [w_j(s) - w_j(t+s)] ds \right) \quad \text{by (2.12)} \\ &\rightarrow 0 \quad \text{as } t \rightarrow 0 \quad \text{by (2.12)}\end{aligned}$$

and the Lebesgue dominated convergence theorem.

(d) Parts (a)–(c) show that $T(t)$ is a strongly continuous semigroup and so it has an infinitesimal generator A which satisfies [12]

$$Ax = \lim_{\delta \rightarrow 0} \frac{T(\delta)x - x}{\delta}.$$

Now since $w_i \in C^1(0, \infty) \cap L_2(0, \infty) \cap L_1(0, \infty)$, the integral in (3.12) is well defined and for all i, j

$$\lim_{\delta \rightarrow 0} \frac{1}{\delta} \int_0^\infty w_i^*(s)(w_j(\delta + s) - w_j(s)) ds = \int_0^\infty w_i^*(s)\dot{w}_j(s) ds.$$

So componentwise

$$(3.13) \quad A_{ij} = \lim_{\delta \rightarrow 0} \frac{1}{\delta} \left(\frac{T_{ij}(\delta) - \delta_{ij}}{\delta} \right) = \int_0^\infty w_i^*(s)\dot{w}_j(s) ds$$

and A satisfies (3.12) with a certain dense domain, which we do not specify.

We now examine the operators B and C .

LEMMA 3.4. B given by (3.9) is a bounded operator from \mathbb{C}^m to l_2 and

$$(3.14) \quad T(t)B = \{\sigma_i v_i^*(t)\}_{i=1}^\infty.$$

Furthermore, the reachability operator $\underline{B} \in \mathcal{L}(L_2(0, \infty), \Sigma^{-1}l_2)$ is given by

$$(3.15) \quad \underline{B}u = \{\sigma_i \langle v_i, u \rangle\}_{i=1}^\infty.$$

Proof. Equation (3.14): For $u \in \mathbb{C}^m$, we calculate

$$\begin{aligned} \|\underline{B}u\|^2 &= \sum_{i=1}^\infty |\sigma_i v_i^*(0)u|^2 \\ &\leq |u|^2 \sum_{i=1}^\infty \sigma_i^2 |v_i(0)|^2 \end{aligned}$$

and (2.13) shows that B is bounded. Now

$$\begin{aligned} (T(t)B)_i &= \sum_{k=1}^\infty \int_0^\infty w_i^*(s)w_k(t+s) ds \sigma_k v_k^*(0) \\ &= \int_0^\infty w_i^*(s)h(t+s) ds \quad \text{by (2.10)} \\ &= \sigma_i v_i^*(t) \quad \text{by (2.5).} \end{aligned}$$

Equation (3.15): For $u \in L_2(0, \infty)$ we have $\underline{B}u = \int_0^\infty T(s)\underline{B}u(s) ds$. Thus from (3.14), $(\underline{B}u)_i = \sigma_i \langle v_i, u \rangle$ and this proves (3.15). Finally, $\Sigma^{-1}\underline{B}u = \langle v_i, u \rangle$ and so

$$\|\Sigma^{-1}\underline{B}u\|^2 = \sum_{i=1}^\infty \langle v_i, u \rangle^2 \leq \|u\|^2.$$

LEMMA 3.5. C given by (3.8) is an A -bounded operator from l_2 to \mathbb{C}^p and the observability operator is a contraction from l_2 to $L_2(0, \infty; \mathbb{C}^p)$ and satisfies

$$(3.16) \quad \underline{C}x = \sum_{i=1}^\infty w_i(t)x_i.$$

Proof. (a) Since $w_i \in C^1(0, \infty)$, we have

$$-w_i^*(0)w_j(0) = \int_0^\infty (\dot{w}_i^*(s)w_j(s) + w_i^*(s)\dot{w}_j(s)) ds.$$

So for $x \in D(A)$, from (3.12)

$$\begin{aligned}\langle Ax, x \rangle + \langle x, Ax \rangle &= \sum_i \sum_j \bar{x}_i x_j \int_0^\infty (\dot{w}_i^*(s) w_j(s) + w_i^*(s) \dot{w}_j(s)) ds \\ &= - \sum_i \sum_j \bar{x}_i x_j w_i^*(0) w_j(0).\end{aligned}$$

Thus for $x \in D(A)$,

$$\begin{aligned}|Cx|^2 &= -\langle Ax, x \rangle - \langle x, Ax \rangle \\ &\leq 2\|Ax\| \|x\| \\ &\leq (\|Ax\| + \|x\|)^2.\end{aligned}$$

Hence $D(C) \supset D(A)$ and C is A -bounded.

(b) For $x \in l_2^0$ we have

$$\begin{aligned}\underline{C}x &= CT(t)x = \sum_{i=1}^\infty \sum_{j=1}^N w_i^*(0) \frac{\sigma_i}{\sigma_j} \int_0^\infty v_i^*(t+s) v_j(s) ds x_j \quad \text{by (3.11)} \\ &= \sum_{j=1}^N \frac{1}{\sigma_j} \int_0^\infty h(t+s) v_j(s) ds x_j \quad \text{by (2.9)} \\ &= \sum_{j=1}^N w_j(t) x_j \quad \text{by (2.5)}\end{aligned}$$

and

$$\|\underline{C}x\|^2 = \int_0^\infty \left\| \sum_{j=1}^N w_j(t) x_j \right\|^2 dt = \sum_{j=1}^N x_j^2 = \|x\|^2 \quad \text{by (2.12)}.$$

So \underline{C} is a contraction on l_2^0 and since l_2^0 is dense in l_2 we have proven (3.16) and that \underline{C} is a contraction on l_2 . \square

Finally we prove that (3.7)–(3.9) is the sought realisation.

THEOREM 3.1. *($C, T(t), B$) given by (3.7)–(3.9) is an approximately controllable, exactly observable output normal realisation for h on the state space l_2 .*

Proof. From Lemma 3.5, the observability operator \underline{C} is bounded from l_2 to $L_2(0, \infty; \mathbb{C}^p)$; since $B \in \mathcal{L}(\mathbb{C}^m, l_2)$, $\underline{C}B$ is a well-defined function in $L_2(0, \infty; \mathbb{C}^{p \times m})$ and from (3.16) and (3.9) we have

$$\begin{aligned}(3.17) \quad CT(t)B &= \underline{C}B = \sum_{i=1}^\infty w_i(t) \sigma_i v_i^*(0) \\ &= h(t) \quad \text{a.e. in } t \quad \text{by (2.11)}.\end{aligned}$$

Furthermore, the reachability and observability operators are bounded by Lemmas 3.4 and 3.5 and

$$\underline{C}Bu = CT(t) \int_0^\infty T(\tau) Bu(\tau) d\tau = \int_0^\infty h(t+\tau) u(\tau) d\tau = \Gamma u.$$

Hence $(C, T(t), B)$ realises Γ .

Now (3.15) shows that the controllability gramian $P = \Sigma^2$, since

$$(P)_{ij} = (\underline{B}\underline{B}^*)_{ij} = \sigma_i \langle v_i, \sigma_j v_j \rangle = \sigma_i^2 \delta_{ij}$$

and (3.16) shows that the observability gramian $Q = I$, since

$$Q_{ij} = (\underline{C}^* \underline{C})_{ij} = \langle w_i, w_j \rangle = \delta_{ij}.$$

Finally, $P = \Sigma^2 \geq 0$ shows that $(T(t), B)$ is approximately controllable and $Q = I > 0$ shows that $(C, T(t))$ is exactly observable [11].

COROLLARY 3.1. *The transfer function G of system (2.1) is given by*

$$(3.18) \quad G(s) = C(sI - A)^{-1}B.$$

Proof. By definition, the transfer function is given by [15]:

$$G(s) = \int_0^\infty e^{-st} h(t) dt = \int_0^\infty e^{-st} CT(t)B dt \quad \text{from (3.17)}$$

but C is unbounded and it is not so evident that (3.18) immediately follows. From Lemma 3.5, C is A -bounded and so it is a *bounded* operator on $D(A)$ with the graph norm. Thus for $x \in D(A)$ it follows that [20]

$$(3.19) \quad C(sI - A)^{-1}x = C \int_0^\infty e^{-st} T(t)x dt = \int_0^\infty e^{-st} CT(t)x dt.$$

Now $C(sI - A)^{-1}$ is bounded on l_2 as is C (Lemma 3.5) and so we may extend (3.19) to all $x \in l_2$. Letting $x = Bu$ proves (3.18).

4. Convergence of the truncated realisations. Truncations of (3.8), (3.9), and (3.12) yield finite-dimensional systems (C_n, A_n, B_n) given by

$$(4.1) \quad (A_n)_{ij} = \int_0^\infty w_i^*(s) \dot{w}_j(s) ds = \frac{\sigma_i}{\sigma_j} \int_0^\infty \dot{v}_i^*(s) v_j(s) ds, \quad i, j = 1, \dots, n,$$

$$(4.2) \quad C_n = [w_1(0), \dots, w_n(0)],$$

$$(4.3) \quad B_n = [\sigma_1 v_1(0), \dots, \sigma_n v_n(0)]^*.$$

In this section it will be shown that, as $n \rightarrow \infty$, $h_n(t) = C_n e^{A_n t} B_n$ converges to $h(t)$ and $G_n(s) = C_n(sI - A_n)^{-1} B_n$ converges to $G(s)$, the Laplace transform of $h(t)$. The proof involves using an intermediate sequence of rational approximants similar to the Coifman-Rochberg decomposition in (2.16).

Initially we consider system (2.1) under the same assumptions made in § 3, namely those given by (2.2). However, to obtain stronger convergence results we shall later make further assumptions, including the nuclear property (2.15).

LEMMA 4.1. *The truncated system (C_n, A_n, B_n) satisfies the Lyapunov equations*

$$(4.4) \quad A_n \Sigma_n^2 + \Sigma_n^2 A_n^* + B_n B_n^* = 0,$$

$$(4.5) \quad A_n + A_n^* + C_n^* C_n = 0$$

where $\Sigma_n = \text{diag}(\sigma_1, \dots, \sigma_n)$.

The singular values of (C_n, A_n, B_n) are σ_i , $i = 1, \dots, n$ and the controllability and observability gramians are $P_n = \Sigma_n^2$ and $Q_n = I_n$, respectively. A_n is asymptotically stable and

$$(4.6) \quad \|\exp(A_n t)\|_{\mathcal{L}(l_2)} \leq 1.$$

Proof.

$$\begin{aligned} (A_n \Sigma_n^2 + \Sigma_n^2 A_n^*)_{ij} &= \sigma_i \sigma_j \int_0^\infty \dot{v}_i^*(s) v_j(s) ds + \sigma_i \sigma_j \int_0^\infty v_i^*(s) \dot{v}_j(s) ds \\ &= -\sigma_i \sigma_j v_i^*(0) v_j(0) \quad \text{since } v_i \in C^1(0, \infty) \\ &= -(B_n B_n^*)_{ij}. \end{aligned}$$

So (4.4) is satisfied and similarly for (4.5). From (4.5) we may deduce for all $x \in l_2$

$$\langle A_n x, x \rangle + \langle A_n^* x, x \rangle = -|C_n x|^2$$

and so A_n and A_n^* are dissipative and A_n generates a contraction semigroup on l_2 ([4, p. 22]).

The rest of the proof uses results on truncations of the balanced realisation $(C\Sigma^{1/2}, \Sigma^{-1/2}T(t)\Sigma^{1/2}, \Sigma^{-1/2}B)$ from [5]. The balanced truncations are then $(\tilde{C}_n, \tilde{A}_n, \tilde{B}_n) = (C_n\Sigma_n^{1/2}, \Sigma_n^{-1/2}A_n\Sigma_n^{1/2}, \Sigma_n^{-1/2}B_n)$. Theorem 4.1 of [5] proves that \tilde{A}_n is asymptotically stable with gramians $\tilde{P}_n = \tilde{Q}_n = \Sigma_n$. So A_n is asymptotically stable with gramians $P_n = \Sigma_n^2$ and $Q_n = I_n$. $P_n Q_n = \Sigma_n^2 = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ and so (A_n, B_n, C_n) has singular values $\sigma_1, \dots, \sigma_n$ using Lemma 4.2 from [5]. \square

From now on we assume that the system is of nuclear type. As in § 2, the Coifman–Rochberg theorems give us a decomposition for the transfer function $G(s)$

$$(4.7) \quad G(s) = \sum_{i=1}^{\infty} a_i (\text{Re } \xi_i)(s - \xi_i)^{-1}.$$

Here the ξ_i are complex numbers in the left half-plane, and we have $\sum |a_i| < \infty$. Convergence occurs in the nuclear norm, which, by Theorem 2.1, implies that the impulse responses converge in L_1 , and the transfer functions converge in H_∞ , that is, uniformly over the right half-plane.

Given that h is in L_2 , our next result shows that convergence in L_2 can also be obtained. To simplify the notation, we shall write $\|G\|_N$ to denote $\|\Gamma\|_N$, where G is the transfer function corresponding to the operator Γ .

LEMMA 4.2. *There exists a sequence of finite-rank Hankel operators, corresponding to rational transfer functions F_m , $m = 1, 2, \dots$ such that $F_m \rightarrow G$ in nuclear norm and also in H_2 norm. Thus the impulse responses converge in both L_1 and L_2 norms.*

Proof. We write $G_m(s) = \sum_{i=1}^m a_i (\text{Re } \xi_i)(s - \xi_i)^{-1}$, and consider the Hankel operators X_m , $m = 1, 2, \dots$, associated with $F_m(s) = G_m(s)/(1 + s\varepsilon_m)$, where (ε_m) is a positive sequence to be determined.

H_2 convergence.

$$\begin{aligned} \|F_m - G\|_2 &\leq \left\| \frac{G_m - G}{1 + s\varepsilon_m} \right\|_2 + \left\| \frac{G}{1 + s\varepsilon_m} - G \right\|_2 \\ &\leq \|G_m - G\|_\infty / (2\varepsilon_m)^{1/2} + \left\| \frac{s\varepsilon_m}{1 + s\varepsilon_m} G \right\|_2. \end{aligned}$$

Thus if $\varepsilon_m \rightarrow 0$ are chosen such that $\|G_m - G\|_\infty / (2\varepsilon_m)^{1/2} \rightarrow 0$ (which is possible since $\|G_m - G\|_\infty \rightarrow 0$), both terms tend to zero (the second by dominated convergence).

Nuclear convergence. It is sufficient to prove the scalar case:

$$\begin{aligned} \|F_m - G\|_N &\leq \sum_{i=m+1}^{\infty} |a_i| \left\| \frac{\text{Re } \xi_i}{s - \xi_i} \right\|_N + \sum_{i=1}^m |a_i| \left\| \frac{\text{Re } \xi_i}{(s - \xi_i)} \frac{s\varepsilon_m}{1 + s\varepsilon_m} \right\|_N \\ &\leq 2 \sum_{i=m+1}^{\infty} |a_i| + \sum_{i=1}^m 2|a_i| \left\| \text{Re } \xi_i \frac{1}{(s - \xi_i)} \frac{s\varepsilon_m}{1 + s\varepsilon_m} \right\|_\infty \end{aligned}$$

since a rank k operator X satisfies $\|X\|_N \leq k\|X\|$

$$\leq 2 \sum_{i=n+1}^{\infty} |a_i| + \sum_{i=1}^n 2|a_i| \left\| \frac{\text{Re } \xi_i}{(s - \xi_i)} \frac{s\varepsilon_m}{1 + s\varepsilon_m} \right\|_\infty \quad \text{for all } n \leq m,$$

since in the second sum $\|\cdot\|_\infty \leq 1$. Thus, given any $\varepsilon > 0$, by choosing n sufficiently large, the first term can be made less than ε . Then choosing $m \geq n$ such that ε_m is sufficiently small, the second term can also be made less than ε since there are only n terms and each tends to zero. Hence $\|F_m - G\|_\infty \rightarrow 0$.

We remark that, in the case when h is real, we may also guarantee that its approximants are real. For since $G(s) = \overline{G(\bar{s})}$, we may replace (4.7) by the decomposition

$$(4.8) \quad G(s) = \frac{1}{2} \sum_{i=1}^{\infty} a_i \operatorname{Re}(\xi_i)(s - \xi_i)^{-1} + \bar{a}_i \operatorname{Re}(\bar{\xi}_i)(s - \bar{\xi}_i)^{-1}.$$

The partial sums of this modified decomposition may now be modified, as in the proof of Lemma 4.2, to obtain H_2 convergence.

We are now interested in the effects of truncating the realisations given in § 3. With (F_m) and G as in Lemma 4.2, let F_n^m and G_n denote the transfer functions corresponding to the n -dimensional truncations of the output-normal realizations analogous to (3.7)–(3.9), defined as in (4.1)–(4.3), of F_m , G , respectively.

Since $F_m \rightarrow G$ in nuclear norm (and hence in Hankel norm), and G has distinct singular values, the singular values and Schmidt vectors of the (F_m) also converge, as the next result shows.

LEMMA 4.3. *Let $(\sigma_i^{(m)})$, $(v_i^{(m)})$ and $(w_i^{(m)})$ be the singular values and Schmidt vectors of F_m . Then for some suitable normalization of the Schmidt vectors, we have, as $m \rightarrow \infty$*

- (i) $|\sigma_i^{(m)} - \sigma_i| \rightarrow 0$ for all i ;
- (ii) $\|v_i^{(m)} - v_i\|_2 \rightarrow 0$ and $\|w_i^{(m)} - w_i\|_2 \rightarrow 0$;
- (iii) $\|v_i^{(m)} - v_i\|_\infty \rightarrow 0$ and $\|w_i^{(m)} - w_i\|_\infty \rightarrow 0$.

Proof. Parts (i) and (ii) follow from standard operator-theoretic arguments, which are outlined in Appendix 2.

Now $w_i(t) = (1/\sigma_i) \int_0^\infty h(t+s)v_i(s) ds$ and $w_i^{(m)}(t) = (1/\sigma_i^{(m)}) \int_0^\infty h_m(t+s)v_i^{(m)}(s) ds$, where the (h_m) are the impulse responses corresponding to F_m . Since $\|h_m - h\|_2 \rightarrow 0$ and $\|v_i^{(m)} - v_i\|_2 \rightarrow 0$, we have that

$$\begin{aligned} |w_i^{(m)}(t) - w_i(t)| &\leq \|h\|_2 \left\| \frac{v_i}{\sigma_i} - \frac{v_i^{(m)}}{\sigma_i^{(m)}} \right\|_2 + \|h - h^{(m)}\|_2 \left\| \frac{v_i^{(m)}}{\sigma_i^{(m)}} \right\|_2 \\ &\rightarrow 0 \quad \text{uniformly in } t, \end{aligned}$$

and similarly $v_i^{(m)} \rightarrow v_i$ uniformly in t , for every i .

Examining (4.1)–(4.3) again, we see that the $C_n^{(m)}$ and $B_n^{(m)}$ corresponding to the truncations F_n^m do converge to C_n and B_n . Moreover, (4.4) and (4.5) imply that for $i \neq j$

$$(4.9) \quad A_{ij} = \frac{-\sigma_i^2 w_i^*(0) w_j(0) + \sigma_i \sigma_j v^*(0) v_j(0)}{\sigma_i^2 - \sigma_j^2}$$

and similarly for $A_{ij}^{(m)}$, and hence the off-diagonal elements of the $A_n^{(m)}$ also converge to those of A_n .

In the case when the system is real, (4.5) implies that the diagonal elements of the $A_n^{(m)}$ also converge to those of A_n , since $A_{ii}^{(m)} = -\frac{1}{2} w_i^*(0) w_i(0)$; however, in the complex case, (4.4) and (4.5) do not determine the imaginary part of $(A_n)_{ii}$ and we must impose stronger conditions on h to guarantee convergence.

LEMMA 4.4. *Suppose that $h \in L_1 \cap L_2$ is of nuclear type and either*

- (a) *h is purely real, or*
- (b) *h is complex, \dot{h} exists (in the sense that h is the integral of its derivative) and \dot{h} is the kernel of a bounded Hankel operator.*

Then the finite-dimensional truncations F_n^m converge to G_n in nuclear norm, H_2 and H_∞ as $m \rightarrow \infty$.

(Note also the standing assumption that the singular values of Γ are distinct.)

Proof. We verify first that the entries of the corresponding $A_n^{(m)}$, $B_n^{(m)}$, and $C_n^{(m)}$ do converge. In case (a) this follows from Lemma 4.3(iii) and the arguments above. In case (b) it remains to show that the diagonal entries

$$(4.10) \quad A_{ii}^{(m)} = \int_0^\infty w_i^{(m)*}(s) \dot{w}_i^{(m)}(s) ds$$

converge to A_{ii} .

Let Γ , Γ_m be the Hankel operators corresponding to h , h_m and Δ the operator corresponding to \dot{h} . By arguments similar to those in § 2, the eigenfunctions w_i , $w_i^{(m)}$ have derivatives in $L_1 \cap L_2$.

Consider the identity

$$(4.11) \quad \begin{aligned} \Gamma(v_i - v_i^{(m)}) + (\Gamma - \Gamma_m)(v_i^{(m)} - v_i) + (\Gamma - \Gamma_m)v_i &= \Gamma v_i - \Gamma_m v_i^{(m)} \\ &= (\sigma_i - \sigma_i^{(m)})w_i + \sigma_i^{(m)}(w_i - w_i^{(m)}). \end{aligned}$$

The derivative of each term is in L_2 and so we may consider its L_2 norm. Note that, if $u \in L_2 \cap L_\infty$

$$\begin{aligned} (\Gamma u)^*(s) &= \int_0^\infty \dot{h}(s+t)u(t) dt \\ &= -h(s)u(0) + \int_0^\infty h(s+t)\dot{u}(t) dt \end{aligned}$$

so that

$$(4.12) \quad \|(\Gamma u)^*\|_2 \leq \|h\|_2 \|u\|_\infty + \|\Gamma\| \|\dot{u}\|_2$$

and also

$$(4.13) \quad \|(\Gamma u)^*\|_2 \leq \|\Delta\| \|u\|_2.$$

Hence

$$\begin{aligned} \sigma_i^{(m)} \|\dot{w}_i - \dot{w}_i^{(m)}\|_2 &\leq \|(\Gamma(v_i - v_i^{(m)}))^*\|_2 + \|((\Gamma - \Gamma_m)(v_i^{(m)} - v_i))^*\|_2 \\ &\quad + \|((\Gamma - \Gamma_m)v_i)^*\|_2 + |\sigma_i - \sigma_i^{(m)}| \|\dot{w}_i\|_2 \\ &\leq \|\Delta\| \|v_i - v_i^{(m)}\|_2 + \|h - h_m\|_2 \|v_i - v_i^{(m)}\|_\infty + \|\Gamma - \Gamma_m\| \|\dot{v}_i - \dot{v}_i^{(m)}\|_2 \\ &\quad + \|h - h_m\|_2 \|v_i\|_\infty + \|\Gamma - \Gamma_m\| \|\dot{v}_i\|_2 \\ &\quad + |\sigma_i - \sigma_i^{(m)}| \|\dot{w}_i\|_2, \quad \text{using (4.13), (4.12), and (4.12) again.} \end{aligned}$$

Hence $\|\dot{w}_i - \dot{w}_i^{(m)}\|_2 \leq (\|\Gamma - \Gamma_m\| / \sigma_i^{(m)}) \|\dot{v}_i - \dot{v}_i^{(m)}\|_2 + \text{terms tending to zero}$. Similarly we can deduce that

$$\|\dot{v}_i - \dot{v}_i^{(m)}\|_2 \leq \frac{\|\Gamma^* - \Gamma_m^*\|}{\sigma_i^{(m)}} \|\dot{w}_i - \dot{w}_i^{(m)}\|_2 + o(1).$$

Hence $\|\dot{w}_i - \dot{w}_i^{(m)}\|_2 \rightarrow 0$ and $\|\dot{v}_i - \dot{v}_i^{(m)}\|_2 \rightarrow 0$ and thus $A_{ii}^{(m)} \rightarrow A_{ii}$ from (4.10) and Lemma 4.3.

Given that $A_n^{(m)} \rightarrow A_n$, $B_n^{(m)} \rightarrow B_n$ and $C_n^{(m)} \rightarrow C_n$ as $m \rightarrow \infty$ for each n , it is routine to verify that $F_n^m \rightarrow G_n$ in H_2 and H_∞ , and hence also in nuclear norm since $\|F_n^m - G_n\|_N \leq 2n\|F_n^m - G_n\|_\infty$.

LEMMA 4.5. *Under the hypotheses of Lemma 4.4, $\|G_n - G\|_\infty \rightarrow 0$.*

Proof. With the notation introduced earlier, and for $n \leq m$,

$$\|G_n - G\|_\infty \leq \|G_n - F_n^m\|_\infty + \|F_n^m - F_m\|_\infty + \|F_m - G\|_\infty.$$

Now, for all n ,

$$\|G_n - F_n^m\| \rightarrow 0 \quad \text{as } m \rightarrow \infty \quad \text{by Lemma 4.4.}$$

We have also that $\|F_n^m - F_m\|_\infty \leq 2 \sum_{i>n} \sigma_i^{(m)}$, from similar finite-dimensional results [16], since $F_n^m = C_n^{(m)}(sI - A_n^{(m)})B_n^{(m)}$ is an output-normal truncation of F_m , similar to a balanced truncation.

Suppose $\varepsilon > 0$ is given. From Theorem 2.1 there exists an m_1 such that for all $m > m_1$, $\|F_m - G\|_\infty < \varepsilon/3$. Moreover, there exists m_2 such that for $m > m_2$,

$$2 \sum_{i>n} \sigma_i^{(m)} \leq 2 \sum_{i=n+1}^{\infty} \sigma_i + \frac{\varepsilon}{6} < \frac{\varepsilon}{3}$$

for all $n \geq N$, say, by the nuclearity of Γ .

Finally, for any $n \geq N$, there exists an m_3 such that for $m > m_3$, $\|G_n - F_n^m\|_\infty < \varepsilon/3$, as above. Choosing such an $m > \max(m_1, m_2, m_3)$, we have $\|G_n - G\|_\infty < \varepsilon$. \square

We can now show that the Schmidt vectors of the truncations $\{v_i^n, w_i^n\}$ converge to the Schmidt vectors $\{v_i, w_i\}$ of the original Hankel operator Γ . From now on we shall be making the assumptions of § 2, together with those of Lemma 4.4.

Note that the Hankel operator Γ_n and its Schmidt vectors $\{v_i^n, w_i^n\}$ satisfy the relations

$$(4.14) \quad (\Gamma_n u)(t) = \int_0^\infty C_n \exp A_n(t+s) B_n u(s) ds,$$

$$(4.15) \quad \Gamma_n v_i^n = \sigma_i w_i^n, \quad \Gamma_n^* w_i^n = \sigma_i v_i^n,$$

$$(4.16) \quad h_n(t+\alpha) = C_n \exp A_n(t+\alpha) B_n = \sum_{i=1}^n w_i^n(t) \sigma_i v_i^n(\alpha)^*$$

where (4.16) now holds for all t and α because of the analyticity of h_n and the finite summation.

COROLLARY 4.1. $\|v_i^n - v_i\|_2 \rightarrow 0$ and $\|w_i^n - w_i\|_2 \rightarrow 0$ if suitably normalised.

Proof. Note that $\|\Gamma\| \leq \|G\|_\infty$ for any Hankel operator, since it acts on H_2 by multiplication, reflection, and projection. Thus, since $\|G_n - G\|_\infty \rightarrow 0$ and $\|\Gamma_n - \Gamma\| \leq \|G_n - G\|_\infty$, we have the result that the corresponding Hankel operators converge in norm to Γ . The result now follows as in Lemma 4.3, using standard arguments which are given in Appendix 2.

We can now prove convergence of Γ_n to Γ in the nuclear norm.

THEOREM 4.1. *If the hypotheses of Lemma 4.4 hold, then*

$$(a) \quad \|\Gamma_n - \Gamma\|_N \rightarrow 0;$$

$$(b) \quad \|h_n - h\|_1 \rightarrow 0;$$

$$(c) \quad \|h_n - h\|_2 \rightarrow 0.$$

Proof. (a) As in (4.16),

$$h_l(t+\alpha) - h_n(t+\alpha) = \sum_{i=1}^l \sigma_i w_i^l(t) v_i^l(\alpha)^* - \sum_{i=1}^n \sigma_i w_i^n(t) v_i^n(\alpha)^*.$$

So if $l > n$, writing $\Gamma_l - \Gamma_n$ as a sum of rank 1 operators gives

$$\|\Gamma_l - \Gamma_n\|_N \leq \sum_{i=1}^n \|w_i^l - w_i^n\|_2 \sigma_i \|v_i^l\|_2 + \sum_{i=1}^n \|w_i^n\|_2 \sigma_i \|v_i^l - v_i^n\|_2 + \sum_{i=n+1}^l \|w_i^l\|_2 \sigma_i \|v_i^l\|_2$$

$$\rightarrow 0 \quad \text{as } n, l \rightarrow \infty$$

since

$$w_i^l \rightarrow w_i \text{ in } L_2, \quad v_i^l \rightarrow v_i \text{ in } L_2, \quad \sum \sigma_i < \infty.$$

But since $\|G_n - G\|_\infty \rightarrow 0$ (Lemma 4.5), it follows that the limit of the Γ_n in nuclear norm is Γ .

(b) This follows from (a) and Theorem 2.1.

(c) Lemma 2.1 and (4.16) show that the following hold for $k < n$ and almost all $t \geq 0$:

$$h(t) - h_n(t) = \sum_{j=1}^k w_j(0) \sigma_j (v_j(t) - v_j^n(t))^* + \sum_{j>k} w_j(0) \sigma_j v_j(t)^* - \sum_{j=k+1}^n w_j(0) \sigma_j v_j^n(t)^*$$

since $w_j(0) = w_j^n(0)$ by (4.1)–(4.3) and (4.14)–(4.16).

Hence, since $\{v_j\}$ and $\{v_j^n\}$ are separately orthonormal sets in $L_2(0, \infty; \mathbb{C}^m)$ we have that

$$(4.17) \quad \|h - h_n\|_2 \leq \sum_{i=1}^k \sigma_i |w_i(0)| \|v_i - v_i^n\|_2 + 2 \left(\sum_{i>k} |w_i(0)|^2 \sigma_i^2 \right)^{1/2}.$$

Given any $\varepsilon > 0$, (2.13) shows that the second term can be made $< \varepsilon/2$ by choosing k sufficiently large. Corollary 4.1 then shows that the first term will be $< \varepsilon/2$ for all n sufficiently large.

A simple consequence of Corollary 4.1 gives convergence of the controllability operators, observability operators and semigroups to those of the infinite-dimensional system of § 3. Here we will identify (C_n, A_n, B_n) on \mathbb{C}^n with $(C\Pi_n, \Pi_n A \Pi_n, \Pi_n B)$ on l_2 , where Π_n is the orthogonal projection onto the first n coordinates in l_2 .

COROLLARY 4.2. $\|\underline{B} - \underline{B}_n\|_{HS} \rightarrow 0$ as $n \rightarrow \infty$, where the norm is the Hilbert–Schmidt norm in $\mathcal{L}(L_2(0, \infty; \mathbb{C}^p), l_2)$.

Proof. From (3.15) it follows that for $x \in l_2$,

$$\underline{B}^* x = \sum_{i=1}^{\infty} \sigma_i x_i v_i$$

and so calculating the Hilbert–Schmidt norm of \underline{B}^* , we obtain, with $\{e_i\}$ the usual orthonormal basis for l_2 ,

$$\begin{aligned} \|\underline{B}^* - \underline{B}_n^*\|_{HS}^2 &= \sum_{i=1}^{\infty} \|\underline{B}^* e_i - \underline{B}_n^* e_i\|^2 \\ &= \sum_{i=1}^n \sigma_i^2 \|v_i - v_i^n\|^2 + \sum_{n+1}^{\infty} \sigma_i^2 \|v_i\|^2 \\ &\rightarrow 0 \quad \text{as } n \rightarrow \infty \end{aligned}$$

by Corollary 4.1, since $\|v_i^n\| = 1$ and by the Hilbert–Schmidt assumption (2.14). $\|\underline{B}^* - \underline{B}_n^*\|_{HS} = \|\underline{B} - \underline{B}_n\|_{HS}$ completes the proof. \square

COROLLARY 4.3. $\|\underline{C}x - \underline{C}_n x\| \rightarrow 0$ as $n \rightarrow \infty$ for all $x \in l_2$, where the norms are in $L_2(0, \infty; \mathbb{C}^p)$.

Proof. Since from Lemma 3.5, $\|\underline{C}\| \leq 1$ and $\|\underline{C}_n\| \leq 1$, it suffices to prove the strong convergence on l_2^0 , a dense subspace of l_2 . Suppose that $x \in l_2^0$ has length N ; then for $n > N$ we calculate

$$\begin{aligned} \|(\underline{C} - \underline{C}_n)x\|^2 &= \int_0^\infty \left\| \sum_{i=1}^N (w_i^n(t) - w_i(t))x_i \right\|^2 dt \\ &\leq \|x\|^2 \sum_{i=1}^N \|w_i^n - w_i\|^2 \\ &\rightarrow 0 \quad \text{as } n \rightarrow \infty \quad \text{by Corollary 4.1.} \end{aligned}$$

Finally, the truncated semigroups converge strongly as $n \rightarrow \infty$.

COROLLARY 4.4. For all $x \in l_2$,

- (a) $\lim_{n \rightarrow \infty} \sup_{0 \leq t < \infty} \|\exp(A_n t)x - T(t)x\| = 0$,
- (b) $\lim_{n \rightarrow \infty} \sup_{0 \leq t < \infty} \|\exp(A_n t)B_n - T(t)B\| = 0$

where the norm is in $L(\mathbb{C}^m, l_2)$.

Proof. (a) This follows from (3.7) and Corollary 4.1, using the fact that $\exp(A_n t)$ is a contraction for each n (cf. Lemma 3.2).

(b) Since B and B_n have finite rank, $\|B - B_n\|_{l_2} \rightarrow 0$ as $n \rightarrow \infty$ and the result follows from (a). \square

5. Error bounds on the truncated realisations. It was shown in § 4 that the truncation to order n of the output normal realization of § 3, with impulse response $h_n(t)$, transfer function $G_n(s)$, and Hankel operator $\Gamma_n(s)$, satisfied $\|h - h_n\|_1 \rightarrow 0$, $\|h - h_n\|_2 \rightarrow 0$, $\|G - G_n\|_\infty \rightarrow 0$ and $\|\Gamma - \Gamma_n\|_\infty \rightarrow 0$ under the assumptions of Lemma 4.4 (principally that the system is of nuclear type). In this section explicit bounds on these truncation errors will be derived in terms of $\{\sigma_i\}$. The approach is first to calculate explicitly the L_2 error between the Schmidt vectors for G_n and \hat{G}_{n-1} (an optimal Hankel-norm approximation to G_n), and those for \hat{G}_{n-1} and G_{n-1} (Lemma 5.1). This new finite-dimensional calculation can then be exploited to give bounds on the L_1 and L_2 norms of $(h - h_n)$, which are also believed to be new for finite-dimensional systems. The bounds on $\|G - G_n\|_\infty$ are immediate from finite-dimensional results.

LEMMA 5.1. Let $G_n(s)$ be as defined in (4.1)–(4.3) and let $\hat{G}_{n-1}(s)$ be an optimal Hankel-norm approximation of degree $n-1$ to $G_n(s)$ and as obtained in [16, Lemma 9.1]. Let $\{v_i^n, w_i^n\}$ and $\{\hat{v}_i^{n-1}, \hat{w}_i^{n-1}\}$ be the Schmidt pairs for G_n and \hat{G}_{n-1} , respectively. Then

- (a) $\|w_i^n - \hat{w}_i^{n-1}\|_2 = \|v_i^n - \hat{v}_i^{n-1}\|_2 = \sqrt{2}\{1 - (1 - \sigma_n^2/\sigma_i^2)^{1/2}\}^{1/2} \leq \sqrt{2}\sigma_n/\sigma_i$,
- (b) $\|\hat{w}_i^{n-1} - w_i^{n-1}\|_2 = \|\hat{v}_i^{n-1} - v_i^{n-1}\|_2 = \sqrt{2}\{1 - (1 - \sigma_n^2/\sigma_i^2)^{1/2}\}^{1/2} \leq \sqrt{2}\sigma_n/\sigma_i$.

(Note that for $\sigma_n/\sigma_i \ll 1$ the exact expressions are approximately σ_n/σ_i , so that the above upper bounds will be conservative by a factor of $\sqrt{2}$.)

Proof. (a) The proof requires a detailed result from [16] and for convenience we will temporarily adopt the notation in [16]: i.e., let $G_n(s)$ have a balanced realization (A, B, C, D) and \hat{G}_{n-1} the realisation $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ given in [16, Thm. 6.3]. Further, let e_i^n be the i th standard unit vector in \mathbb{C}^n ; $\Sigma_1 = \text{diag}(\sigma_1, \dots, \sigma_{n-1})$, $\Sigma = \text{diag}(\Sigma_1, \sigma_n)$; $\Gamma = (\Sigma_1^2 - \sigma_n^2 I)$; $A_e = \text{diag}(A, \hat{A})$; $B_e^* = [B^*, \hat{B}^*]$.

Then Theorem 6.3 of [16] shows that

$$(5.1) \quad A_e P_e + P_e A_e^* + B_e B_e^* = 0$$

for

$$(5.2) \quad P_e = \begin{pmatrix} \Sigma_1 & 0 & I \\ 0 & \sigma_n & 0 \\ I & 0 & \Sigma_1 \hat{\Gamma}^{-1} \end{pmatrix}.$$

The Schmidt vectors will be given by

$$(5.3) \quad v_i^n = B^* e^{A^* t} \Sigma^{-1/2} e_i^n,$$

$$(5.4) \quad \hat{v}_i^{n-1} = \hat{B}^* e^{\hat{A}^* t} (\Sigma_1 \hat{\Gamma}^{-1})^{-1/2} e_i^{n-1},$$

$$(5.5) \quad v_i^n - \hat{v}_i^{n-1} = B_e^* e^{A_e^* t} \begin{bmatrix} \Sigma^{-1/2} & e_i^n \\ -(\Sigma_1 \hat{\Gamma}^{-1})^{-1/2} & e_i^{n-1} \end{bmatrix}.$$

Equation (5.1) implies that $P_e = \int_0^\infty e^{A_e^* t} B_e B_e^* e^{A_e^* t} dt$ and hence

$$\begin{aligned} \|v_i^n - \hat{v}_i^{n-1}\|_2^2 &= \begin{bmatrix} \Sigma^{-1/2} & e_i^n \\ -(\Sigma_1 \hat{\Gamma}^{-1})^{-1/2} & e_i^{n-1} \end{bmatrix}^* P_e \begin{bmatrix} \Sigma^{-1/2} & e_i^n \\ -(\Sigma_1 \hat{\Gamma}^{-1})^{-1/2} & e_i^{n-1} \end{bmatrix} \\ &= 2 - 2(e_i^{n-1})^* \Sigma_1^{-1/2} \hat{\Gamma}^{1/2} \Sigma_1^{-1/2} e_i^{n-1} \\ &= 2(1 - \sigma_i^{-1}(\sigma_i^2 - \sigma_n^2)^{1/2}). \end{aligned}$$

The result for $\|w_i^n - \hat{w}_i^{n-1}\|_2$ is obtained analogously and hence part (a) is proven.

(b) Note that $G_{n-1}(s)$ will have the realization (A_{11}, B_1, C_1, D) (the first $n-1$ rows and/or columns of (A, B, C, D)), since the output normal and balanced realisations are similar via a diagonal transformation. Hence

$$(5.6) \quad \begin{aligned} v_i^{n-1} &= B_1^* e^{A_{11}^* t} \Sigma_1^{-1/2} e_i^{n-1}, \\ \hat{v}_i^{n-1} - v_i^{n-1} &= [B_1^*, \hat{B}^*] \exp \begin{bmatrix} A_{11} & 0 \\ 0 & \hat{A} \end{bmatrix}^* \begin{bmatrix} -\Sigma_1^{-1/2} \\ (\Sigma_1 \hat{\Gamma}^{-1})^{-1/2} \end{bmatrix} e_i^{n-1}. \end{aligned}$$

Deleting the n th row and column from (5.1) we obtain

$$\begin{pmatrix} A_{11} & 0 \\ 0 & \hat{A} \end{pmatrix} \tilde{P} + \tilde{P} \begin{pmatrix} A_{11}^* & 0 \\ 0 & \hat{A}^* \end{pmatrix} + \begin{pmatrix} B_1 \\ \hat{B} \end{pmatrix} (B_1^*, \hat{B}^*) = 0 \quad \text{for } \tilde{P} = \begin{pmatrix} \Sigma_1 & I \\ I & \Sigma_1 \hat{\Gamma}^{-1} \end{pmatrix}.$$

Thus by the same reasoning as before,

$$\begin{aligned} \|\hat{v}_i^{n-1} - v_i^{n-1}\|_2^2 &= (e_i^{n-1})^* [-\Sigma_1^{-1/2}, (\Sigma_1 \hat{\Gamma}^{-1})^{-1/2}] \tilde{P} \begin{bmatrix} -\Sigma_2^{-1/2} \\ (\Sigma_1 \hat{\Gamma}^{-1})^{-1/2} \end{bmatrix} e_i^{n-1} \\ &= 2 - 2\sigma_i^{-1}(\sigma_i^2 - \sigma_n^2)^{1/2}. \end{aligned}$$

This, together with the analogous result for $\|\hat{w}_i^{n-1} - w_i^{n-1}\|_2$, completes the proof of part (b). \square

Parts (a) and (b) can now be combined to bound $\|v_i^n - v_i^{n-1}\|_2$ and then this is repeated one step at a time to bound $\|v_i^n - v_i^k\|_2$ as in the following corollary.

COROLLARY 5.1.

$$\begin{aligned} \|w_i^n - w_i^k\|_2, \|v_i^n - v_i^k\|_2 &\leq \sum_{j=k+1}^n 2\sqrt{2}\{1 - (\sigma_j^2/\sigma_i^2)^{1/2}\}^{1/2} \\ &\leq 2\sqrt{2}(\sigma_{k+1} + \sigma_{k+2} + \cdots + \sigma_n)/\sigma_i. \end{aligned}$$

The nuclearity assumption on Γ and Corollary 5.1 now imply that $\{v_i^n\}_{n=1}^\infty, \{w_i^n\}_{n=1}^\infty$ form Cauchy sequences in $L_2(0, \infty)$ and their unique limit points will be v_i and w_i , respectively, by Corollary 4.1.

Hence taking the limit as $n \rightarrow \infty$ in Corollary 5.1 gives the next corollary.

COROLLARY 5.2.

$$\begin{aligned} \|w_i - w_i^n\|_2, \|v_i - v_i^n\|_2 &\leq \sum_{j>n} 2\sqrt{2}\{1 - (1 - \sigma_j^2/\sigma_i^2)^{1/2}\}^{1/2} \\ &\leq 2\sqrt{2}(\sigma_{n+1} + \sigma_{n+2} + \cdots)/\sigma_i. \end{aligned}$$

Explicit bounds on the truncation errors in the transfer function and the impulse response can now be derived.

THEOREM 5.1. *Let (2.1) be real and of nuclear type with $G(s)$ the Laplace transform of $h(t)$. Define*

$$(5.7) \quad M_n = \sum_{j>n} \sigma_j$$

and let $h_n(t)$ and $G_n(s)$ be the impulse response and transfer function, respectively, of the truncated system given in (4.1)–(4.3). Then

$$\begin{aligned} (a) \quad &\|G - G_n\|_\infty \leq 2M_n; \\ (b) \quad &\|G - G_n\|_N \leq E_k = \min_{k \leq n} \{(4\sqrt{2}k - 1)M_n + 2M_k\} \end{aligned}$$

$$\leq 2(k\sigma_{k+1} + M_k)$$

for the largest k such that $\sigma_{k+1} \geq 2\sqrt{2}M_n$;

$$\begin{aligned} (c) \quad &\|h - h_n\|_1 \leq 2E_k; \\ (d) \quad &\|h - h_n\|_2 \leq \min_k \left\{ 2\sqrt{2}M_n \sum_{i=1}^k |w_i(0)| + 2 \left[\sum_{i>k} \sigma_i^2 |w_i(0)|^2 \right]^{1/2} \right\}. \end{aligned}$$

Proof. (a) Finite-dimensional results [16, Thm. 9.6] show that for $l > n$,

$$\|G_n - G_l\|_\infty \leq 2(\sigma_{n+1} + \cdots + \sigma_l) \leq 2M_n.$$

Hence taking the limit as $l \rightarrow \infty$ and using Lemma 4.5 gives the result.

(b) Let $l > n$. Then, as in the proof of Theorem 4.1,

$$h_l(t + \alpha) - h_n(t + \alpha) = \sum_{i=1}^n \sigma_i w_i^l(t) v_i^l(\alpha) - \sum_{i=1}^n \sigma_i w_i^n(t) v_i^n(\alpha).$$

So that if $k < n$, then

$$\begin{aligned} \|\Gamma_l - \Gamma_n\| &\leq \sum_{i=1}^k \sigma_i (\|w_i^l\|_2 \|v_i^l - v_i^n\|_2 + \|w_i^l - w_i^n\|_2 \|v_i^n\|_2) \\ &\quad + \sum_{i=k+1}^n \sigma_i \|w_i^l\|_2 \|v_i^l\|_2 + \sum_{i=k+1}^n \sigma_i \|w_i^n\|_2 \|v_i^n\|_2. \end{aligned}$$

Now by Corollary 5.1

$$(5.8) \quad \sigma_i \|w_i^l - w_i^n\|_2 \leq 2\sqrt{2}(M_n - M_l)$$

and

$$(5.9) \quad \sigma_i \|v_i^l - v_i^n\|_2 \leq 2\sqrt{2}(M_n - M_l)$$

and hence

$$\|\Gamma_l - \Gamma_n\|_N \leq 4k\sqrt{2}(M_n - M_l) + (M_k - M_l) + (M_k - M_n).$$

Taking the limit as $l \rightarrow \infty$ and using Theorem 4.1 gives the first upper bound and the second is immediate.

(c) Immediate from (b) and Theorem 2.1.

(d) The L_2 error bound is obtained from the proof of Theorem 4.1, (4.17) and substituting the bounds in (5.9). \square

6. Optimal Hankel-norm approximations. In this section we consider the minimization problem,

$$(6.1) \quad \inf_{X \in H_{-(k)}^{\infty, p \times m}} \|G - X\|_{\infty}$$

where G is the transfer function of the input-output system (2.1) of nuclear type (see Definition 2.1); $H_{-(k)}^{\infty, p \times m}$ is the space of $p \times m$ matrix-valued functions of a complex variable with a decomposition $X = \hat{G} + F$, where \hat{G} has MacMillan degree $\leq k$ with poles in $\text{Re}(s) < 0$ and $F \in H_{-}^{\infty, p \times m}$ is the Hardy space of functions analytic and bounded in the left half plane. For the rational functions in $H_{-}^{\infty, p \times m}$, we write $RH_{-}^{\infty, p \times m}$. These spaces can be regarded as being contained in $\mathcal{L}^{\infty, p \times m}$, the space of $p \times m$ matrix-valued functions of a complex variable with norm,

$$\|X\|_{\infty} = \sup_{-\infty < \omega < \infty} \bar{\sigma}(X(j\omega)).$$

The dimensions $p \times m$ may be omitted in the sequel.

Problem (6.1) is closely related to the Hankel-norm approximation problem

$$\inf_{\hat{G}} \|G - \hat{G}\|_H$$

where \hat{G} has MacMillan degree $\leq k$, and the Hankel-norm, $\|\cdot\|_H$, denotes the largest singular value of the corresponding Hankel operator.

Solutions to problem (6.1) showing that the minimum is σ_{k+1} have been obtained by Adamjan, Arov, and Krein [1], Ball and Helton [2], Glover [16], and Nikol'skii [22], for various classes of function G . For the case when G is rational and matrix-valued, Glover [16] gives explicit state-space algorithms characterizing all the solutions (see also Ball and Ran [3]). If G is not rational, Adamjan, Arov, and Krein [1] give a solution in the scalar case, but in the matrix case the solutions in [2] and [22] are not computationally explicit. Here we provide approximate algorithms for transfer functions of systems of nuclear type. This is done by relating solutions to problem (6.1) to the sequence of approximating problems

$$(6.2) \quad \inf_{X \in H_{-(k)}^{\infty, p \times m}} \|G_n - X\|_{\infty}$$

where G_n is the truncated transfer function of § 4. It will be shown that solutions to (6.2) are very close to those of (6.1), the difference being much smaller than can be deduced from only knowing that $G_n \rightarrow G$. Many of the error estimates are also believed to be new for the rational case.

6.1. Preliminaries. The solutions to (6.1) and (6.2) are characterized by a linear fractional map and the definition and some properties of such maps are first reviewed (see, for example, Redheffer [25], Helton [18], Doyle [13]).

DEFINITION 6.1. Let

$$J = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix}$$

and K be matrix valued functions defined on a subset of the complex plane (of dimensions $J_{11}: p \times m$; $J_{12}: p \times r$; $J_{21}: q \times m$; $J_{22}: q \times r$; $K: r \times q$); then the *linear fractional map* of K with coefficient matrix J , is the $p \times m$ matrix-valued function $\mathcal{F}(J, K)$ given by

$$\mathcal{F}(J, K) = J_{11} + J_{12}K(I - J_{22}K)^{-1}J_{21}$$

when all the terms are defined.

Note that if J is a transfer function, then $\mathcal{F}(J, K)$ will be the transfer function between its first set of inputs and outputs when the feedback K is connected from its second set of outputs and inputs (see Fig. 1). The linear fractional map takes K to $\mathcal{F}(J, K)$ for K such that $(I - J_{22}K)^{-1}$ exists, and some properties of \mathcal{F} for certain classes of J and K are now given.

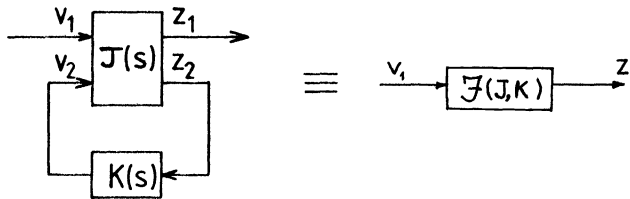


FIG. 1. Linear fractional map.

LEMMA 6.1 (Redheffer [25]). For $J, K \in \mathcal{L}^\infty$ (matrix valued) with $\|J\|_\infty \leq \sigma$, $\|K\|_\infty \leq \sigma^{-1}$, $\|J_{22}K\|_\infty < 1$ then $\mathcal{F}(J, K) \in \mathcal{L}^\infty$ and $\|\mathcal{F}(J, K)\|_\infty \leq \sigma$.

LEMMA 6.2. If $J \in RH_{-(k)}^{\infty, (p+q) \times (m+r)}$, $K \in RH_{-(l)}^{\infty, r \times q}$ have state space realisations

$$J = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} + \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} (sI - A)^{-1} (B_1 B_2),$$

$$K = \hat{D} + \hat{C}(sI - \hat{A})^{-1} \hat{B},$$

then

$$(a) \quad \mathcal{F}(J, K) = D_{11} + D_{12} \hat{D} L_1 D_{21} + (C_1 + D_{12} L_2 \hat{D} C_2, D_{12} L_2 \hat{C}) \\ \times (sI - \tilde{A})^{-1} \begin{bmatrix} B_1 + B_2 \hat{D} L_1 D_{21} \\ \hat{B} L_1 D_{21} \end{bmatrix}$$

where

$$L_1 = (I - D_{22} \hat{D})^{-1}, \quad L_2 = (I - \hat{D} D_{22})^{-1},$$

$$\tilde{A} = \begin{bmatrix} A + B_2 \hat{D} L_1 C_2, & B_2 L_2 \hat{C} \\ \hat{B} L_1 C_2, & \hat{A} + \hat{B} L_1 D_{22} \hat{C} \end{bmatrix};$$

(b) If $\|J_{22}K\|_\infty < 1$, then $\mathcal{F}(J, K) \in H_{-(k+l)}^\infty$.

Proof. Part (a) is obtained by considering the state space realization of the feedback system of Fig. 1 whose transfer function is $\mathcal{F}(J, K)$ (cf. Doyle [13]). Part (b) is proved by considering $\mathcal{F}(J, \alpha K)$ for $\alpha \in [0, 1]$ which will have the state-space realization in part (a) with \hat{D} and \hat{C} multiplied by α (denote the resulting \tilde{A} matrix by $\tilde{A}(\alpha)$). The

poles of $\mathcal{F}(J, \alpha K)$ will be the eigenvalues of $\tilde{A}(\alpha)$ in (a) and will be continuous functions of α . Since $J \in H_{-(k)}^\infty$ and $K \in H_{-(l)}^\infty$ at $\alpha = 0$ the matrix $\tilde{A}(\alpha)$ has $(k+l)$ eigenvalues in the open left half plane and none on the imaginary axis. Furthermore, it can be shown that as α increases the eigenvalues of \tilde{A} are $s \in \mathbb{C}$ such that $\det(I - \alpha J_{22}(s)K(s)) = 0$ together with a subset of the eigenvalues of A and \hat{A} . However, $\det(I - \alpha J_{22}(j(\omega)K(j\omega)) \neq 0$ for all $\omega \in \mathbb{R}$ and $\alpha \in [0, 1]$ since $\|\alpha J_{22}K\|_\infty < 1$, so that $s = j\omega$ is not an eigenvalue of \tilde{A} for any $\alpha \in [0, 1]$. Hence the eigenvalues of \tilde{A} do not cross the imaginary axis and hence $(k+l)$ of them remain in the left half plane which implies that $\mathcal{F}(J, K) \in H_{-(k+l)}^\infty$.

We now consider convergence of functions in $H_{-(k)}^\infty$.

LEMMA 6.3. $H_{-(k)}^{\infty, p \times m}$ is closed in $L^{\infty, p \times m}$.

Proof. Suppose that $X_n \rightarrow X$ in $L^{\infty, p \times m}$ with $X_n \in H_{-(k)}^{\infty, p \times m}$, $n = 1, 2, \dots$. Then Γ_{X_n} has rank at most k and $\Gamma_{X_n} \rightarrow \Gamma_X$ in operator norm, since $\|\Gamma_{X_n} - \Gamma_X\| \leq \|X_n - X\|_\infty$, so $\text{rank}(\Gamma_X) \leq k$. Thus $X \in H_{-(k)}^{\infty, p \times m}$. \square

6.2. Solutions to $\inf \|G_n - X\|_\infty$. We can now return to the minimization problem (6.1). All rational solutions to (6.2) are obtained from Theorem 8.7(2) and Remark 8.3 in [16], but an alternative interpretation using Theorem 7.2 in [16] and Lemmas 6.1 and 6.2 will be presented here. Theorem 7.2 in [16] gives a construction of $J^n \in H_{-(k)}^{\infty, (p \times m - 1) \times (p + m - 1)}$ which is an optimal \mathcal{L}^∞ approximation in that class to G_n augmented by zeros and such that the error

$$(6.3) \quad E_n \triangleq \begin{pmatrix} G_n - J_{11}^n & -J_{12}^n \\ -J_{21}^n & -J_{22}^n \end{pmatrix}$$

satisfies

$$(6.4) \quad E_n^* E_n = \sigma_{k+1}^2 I.$$

$X = J_{11}^n$ is clearly one solution to (6.2). However, for $p, m > 1$ a family of solutions can be obtained by letting

$$X = \mathcal{F}(J^n, K)$$

for $K \in H_{-(k)}^{\infty, (p-1) \times (m-1)}$, $\|K\|_\infty \leq \sigma_{k+1}^{-1}$. This follows since (i) $G_n - X = \mathcal{F}(E_n, -K)$ and by Lemma 6.1 and (6.4), $\|G_n - X\|_\infty \leq \sigma_{k+1}$; and (ii) $\mathcal{F}(J^n, K) \in H_{-(k)}^{\infty, p \times m}$ by Lemma 6.2(b).

In order to construct a state-space realization for J^n a $(m+p-1) \times (m+p-1)$ unitary matrix, U , is required such that (cf. [16], eq. (6.23))

$$(6.5) \quad [v_{k+1}^*(0), 0] = -[w_{k+1}^*(0), 0]U$$

where (v_i, w_i) are Schmidt pairs of Γ . A possible solution is to take

$$U = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & 0 \end{pmatrix}$$

where

$$(6.6) \quad U_{11} = -w_{k+1}(0)[w_{k+1}^*(0)w_{k+1}(0)]^{-1}v_{k+1}^*(0)$$

and the $(p-1)$ columns of U_{12} are orthonormal and orthogonal to $w_{k+1}(0)$, and the $(m-1)$ rows of U_{21} are orthonormal and orthogonal to $v_{k+1}^*(0)$. (This implies that $U^*U = I$ by noting that $|v_{k+1}(0)|^2 = |w_{k+1}(0)|^2$ follows from (3.12)). The state-space realization for J^n is now obtained from Theorem 6.3 in [16] as

$$(6.7) \quad J^n(s) = \hat{D}_n + \hat{C}_n(sI - \hat{A}_n)^{-1}\hat{B}_n = \begin{pmatrix} J_{11}^n & J_{12}^n \\ J_{21}^n & J_{22}^n \end{pmatrix}$$

where

$$(6.8) \quad \hat{A}_n = \hat{\Gamma}^{-1}(\sigma_{k+1}^2 A_{11}^* + \Sigma_1 A_{11} \Sigma_1 - \sigma_{k+1} C_1^* U_{11} B_1^*),$$

$$(6.9) \quad \hat{B}_n = [\hat{B}_1, \hat{B}_2] = \hat{\Gamma}^{-1}(\Sigma_1 B_1 + \sigma_{k+1} C_1^* U_{11}, \sigma_{k+1} C_1^* U_{12}),$$

$$(6.10) \quad \hat{C}_n = \begin{bmatrix} \hat{C}_1 \\ \hat{C}_2 \end{bmatrix} = \begin{bmatrix} C_1 \Sigma_1 + \sigma_{k+1} U_{11} B_1^* \\ \sigma_{k+1} U_{21} B_1^* \end{bmatrix},$$

$$(6.11) \quad \hat{D}_n = -\sigma_{k+1} U$$

and where ((4.1)–(4.5) transformed via $\Sigma_n^{-1/2}$ to balanced coordinates for compatibility with [16])

$$A_{11} = \Sigma_n^{-1/2} A_n \Sigma_n^{1/2} \quad \text{with } (k+1)\text{st row and column deleted,}$$

$$B_1 = \Sigma_n^{-1/2} B_n \quad \text{with } (k+1)\text{st row deleted,}$$

$$C_1 = C_n \Sigma_n^{1/2} \quad \text{with } (k+1)\text{st column deleted,}$$

$$\Sigma_1 = \Sigma_n \quad \text{with } (k+1)\text{st row and column deleted and slight abuse of notation,}$$

$$\hat{\Gamma} = \Sigma_1^2 - \sigma_{k+1}^2 I.$$

The following theorem now summarizes the result from [16]. (Note that this particular interpretation was not given in [16]; however the present derivation only shows that (6.12) is a family of solutions but not all solutions.)

THEOREM 6.1. *All rational solutions to problem (6.2) are characterized by*

$$(6.12) \quad X = \mathcal{F}(J^n, K)$$

where $K \in RH_-^{\infty, (p-1) \times (m-1)}$, and $\|K\|_\infty \leq \sigma_{k+1}^{-1}$, and J^n is given by (6.7)–(6.11). Further,

$$\|G_n - \mathcal{F}(J^n, K)\|_\infty = \sigma_{k+1}.$$

To verify that $\mathcal{F}(J^n, K)$ is well defined even when $\|K\|_\infty = \sigma_{k+1}^{-1}$, the following lemma is useful.

LEMMA 6.4. J_{22}^n given by (6.7)–(6.11) satisfies $\|J_{22}^n\|_\infty < \sigma_{k+1}$.

Proof. The construction ensures (6.3) and (6.4), which imply that

$$J_{12}^{n*} J_{12}^n + J_{22}^{n*} J_{22}^n = \sigma_{k+1}^2 I;$$

hence we just need to show that J_{12}^n is full rank for $\text{Re}(s) = 0$. From (6.7)–(6.11)

$$J_{12}^n = -\sigma_{k+1}(I - \hat{C}_1(sI - \hat{A}_n)^{-1} \hat{\Gamma}^{-1} C_1^*) U_{12}$$

and

$$\begin{aligned} \det(I - \hat{C}_1(sI - \hat{A}_n)^{-1} \hat{\Gamma}^{-1} C_1^*) &= \det(sI - \hat{A}_n - \hat{\Gamma}^{-1} C_1^* \hat{C}_1) / \det(sI - \hat{A}_n) \\ &= \det(sI + \hat{\Gamma}^{-1} A_{11}^* \hat{\Gamma}) / \det(sI - \hat{A}_n) \end{aligned}$$

$$(\text{from (6.8) and (6.10), and since } A_{11}^* \Sigma_1 + \Sigma_1 A_{11} + C_1 C_1^* = 0)$$

$$\neq 0 \quad \text{for } \text{Re}(s) \geq 0 \quad \text{since } \text{Re } \lambda_i(A_{11}) < 0$$

(by Thm. 4.1 in [5]).

By our construction of U , U_{21} has full column rank and this completes the proof.

6.3. Convergence of truncations. In order to study the properties of the truncations the following lemma is helpful (the proof is a minor variation on Lemma 9.5 in [16]).

LEMMA 6.5. Given (A, B, C) satisfying $AP + PA^* + BB^* = 0$, $A^*Q + QA + C^*C = 0$, for $P = \text{diag}(P_1, pI_r)$, $Q = \text{diag}(Q_1, qI_r)$ with P_1 and Q_1 diagonal and $(P_1Q_1 - pqI)$ invertible, then $pq \geq 0$ and

$$(6.13) \quad \|C(sI - A)^{-1}B - C_1(sI - A_{11})^{-1}B_1\|_\infty \leq 2\sqrt{pq}$$

where (A_{11}, B_1, C_1) is the truncated system obtained by deleting the final r rows and/or columns of (A, B, C) . \square

With these preliminaries established we will now consider the convergence of J^n .

THEOREM 6.2. Let (2.1) be of nuclear type with transfer function G and define

$$(6.14) \quad M_n = \sum_{i>n} \sigma_n.$$

If J^n is defined by (6.7)–(6.11), then

$$(6.15) \quad \begin{aligned} (a) \quad & \|J^n - J^{n-1}\|_\infty \leq 2\sigma_n, \\ (b) \quad & \text{there exists } J^\infty \in H_{-(k)}^{\infty, (p+m-1) \times (p+m-1)} \text{ such that} \end{aligned}$$

$$(6.16) \quad \|J^n - J^\infty\|_\infty \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

$$(6.17) \quad \|J^n - J^\infty\|_\infty \leq 2M_n.$$

Proof. Theorem 6.3 in [10] shows that

$$(6.18) \quad \hat{A}_n \Sigma_1 \hat{\Gamma}^{-1} + \Sigma_1 \hat{\Gamma}^{-1} \hat{A}_n^* + \hat{B}_n \hat{B}_n^* = 0,$$

$$(6.19) \quad \hat{A}_n^* \Sigma_1 \hat{\Gamma} + \Sigma_1 \hat{\Gamma} \hat{A}_n + \hat{C}_n^* \hat{C}_n = 0.$$

An examination of the equations for $(\hat{A}_n, \hat{B}_n, \hat{C}_n)$ shows that $(\hat{A}_{n-1}, \hat{B}_{n-1}, \hat{C}_{n-1})$ can be obtained by deleting the final row and/or column from $(\hat{A}_n, \hat{B}_n, \hat{C}_n)$. Hence Lemma 6.4 applies to give (6.15). This, together with the nuclearity assumption, shows that J^n is a Cauchy sequence in \mathcal{L}^∞ and by Lemma 6.3 there exists a $J^\infty \in H_{-(k)}^{\infty, (p+m-1) \times (p+m-1)}$ satisfying (6.16). Equation (6.17) follows from (6.15) and (6.16).

We can now show that the solution of problem (6.2) converges to the solution to problem (6.1).

THEOREM 6.3. For any $K \in RH_{-}^{\infty, (p-1) \times (m-1)}$ with $\|K\|_\infty \leq \sigma_{k+1}^{-1}$, and J^n defined above, we have

$$(6.20) \quad (a) \quad \|\mathcal{F}(J^n, K) - \mathcal{F}(J^{n-1}, K)\|_\infty \leq 2\sigma_n$$

and

$$(b) \quad \text{there exists } X_K \in H_{-(k)}^{\infty, p \times m} \text{ such that}$$

$$(6.21) \quad (i) \quad \|\mathcal{F}(J^n, K) - X_K\|_\infty \rightarrow 0 \quad \text{as } n \rightarrow \infty;$$

$$(6.22) \quad (ii) \quad \|\mathcal{F}(J^n, K) - X_K\|_\infty \leq 2M_n;$$

$$(6.23) \quad (iii) \quad \|G - X_K\|_\infty \leq \sigma_{k+1};$$

$$(6.24) \quad (iv) \quad \text{If in addition } \|K\|_\infty < \sigma_{k+1}^{-1} \text{ or } \|J_{22}^\infty\|_\infty < \sigma_{k+1}, \text{ then } X_K = \mathcal{F}(J^\infty, K).$$

(Note that $\mathcal{F}(J^n, K)$ solves problem (6.2) and X_K solves problem (6.1).)

Proof. (a) First consider $\mathcal{F}(J^n, K)$ when K is a constant matrix, such that $K^*K \leq \sigma_{k+1}^{-2}I$. Then a routine manipulation (a special case of Lemma 6.2(a)) gives us that

$$\mathcal{F}(J^n, K) = \tilde{D}_n + \tilde{C}_n(sI - \tilde{A}_n)^{-1} \tilde{B}_n$$

where

$$(6.25) \quad \tilde{A}_n = \hat{A}_n + \hat{B}_2 K \hat{C}_2 = \hat{\Gamma}^{-1}(\sigma_{k+1}^2 A_{11}^* + \Sigma_1 A_{11} \Sigma_1 - \sigma_{k+1} C_1^* V B_1^*),$$

$$(6.26) \quad \tilde{B}_n = \hat{B}_1 + \hat{B}_2 K \hat{D}_{21} = \hat{\Gamma}^{-1}(\Sigma_1 B_1 + \sigma_{k+1} C_1^* V),$$

$$(6.27) \quad \tilde{C}_n = \hat{C}_1 + \hat{D}_{12} K \hat{C}_2 = C_1 \Sigma_1 + \sigma_{k+1} V B_1^*,$$

$$(6.28) \quad \tilde{D}_n = \hat{D}_{11} + \hat{D}_{12} K \hat{D}_{21} = -\sigma_{k+1}(U_{11} - U_{12} \sigma_{k+1} K U_{21}) = -\sigma_{k+1} V.$$

Note that since U is unitary and $KK^* \leq \sigma_{k+1}^{-2} I$,

$$\begin{aligned} VV^* &= U_{11} U_{11}^* + U_{12} \sigma_{k+1}^2 K K^* U_{12}^* \\ &= I - U_{12}(I - \sigma_{k+1}^2 K K^*) U_{12}^* \\ &\leq I. \end{aligned}$$

Also (6.5) and (6.6) show that $v_{k+1}^*(0) = -w_{k+1}^*(0)V$. Hence by Corollary 7.3 in [16], (6.25)–(6.28) are the formulae for an optimal approximation corresponding to V . Now as in the proof of Corollary 7.3 in [16] we can augment this system into a square system satisfying equations analogous to (6.18) and (6.19). Hence Lemma 6.5 can be applied (as in the proof of Theorem 6.2), and (6.20) is proven for any constant matrix K with $\|K\|_\infty \leq \sigma_{k+1}^{-1}$. For K a rational function of s , (6.20), follows from the constant case by taking the constant equal to any boundary value $K(j\omega)$.

(b) Theorem 6.1 shows that $\mathcal{F}(J^n, K) \in H_{-(k)}^{\infty, p \times m}$ and (6.2) and Lemma 6.3 allow us to conclude that there exists an $X_K \in H_{-(k)}^{\infty, p \times m}$ satisfying (6.21), and (6.22) follows from (6.20). To prove (6.23) consider

$$\begin{aligned} \|G - X_K\|_\infty &= \|G - G_n + G_n - \mathcal{F}(J^n, K) + \mathcal{F}(J^n, K) - X_K\|_\infty \\ &\leq 2M_n + \sigma_{k+1} + 2M_n \rightarrow \sigma_{k+1} \quad \text{as } n \rightarrow \infty, \end{aligned}$$

by Theorem 5.1, (6.22), and Theorem 6.1.

Finally, if $\|KJ_{22}^\infty\|_\infty < \sigma_{k+1}$ then $\mathcal{F}(J^\infty, K)$ is a continuous function of J^∞ and K and

$$\lim_{n \rightarrow \infty} \mathcal{F}(J^n, K) = \mathcal{F}(J^\infty, K).$$

Note that (6.24) gives a linear fractional map describing a family of solutions to (6.1). The condition $\|J_{22}^n\|_\infty < \sigma_{k+1}$ is satisfied for all $n < \infty$ by Lemma 6.4.

6.4. Convergence of optimal Hankel-norm approximations. The optimal Hankel-norm approximants of G are given by \hat{G} of MacMillan degree $\leq k$, such that $\|G - \hat{G}\|_H$ is minimized. The solution to (6.1) induces solutions to the optimal Hankel-norm problem in that if $X \in H_{-(k)}^{\infty, p \times m}$ is such that

$$(6.29) \quad \|G - X\|_\infty = \sigma_{k+1}$$

with $X = \hat{G} + F$, \hat{G} of MacMillan degree k , and $F \in H_{-}^{\infty, p \times m}$, then by Nehari's Theorem [16, Thm. 6.1],

$$(6.30) \quad \|G - \hat{G}\|_H = \sigma_{k+1}.$$

We will again consider how solutions to (6.2) converge to solutions to (6.29).

THEOREM 6.4. *Suppose that G is the transfer function of a system (2.1) of nuclear type, and for any $K \in RH_{-}^{\infty, (p-1) \times (m-1)}$, $\|K\|_\infty \leq \sigma_{k+1}^{-1}$ and J^n defined by (6.7), decompose $\mathcal{F}(J^n, K)$ as*

$$(6.31) \quad \mathcal{F}(J^n, K) = \hat{G}^n + F^n$$

where \hat{G}^n has Macmillan degree k , $\hat{G}^n(\infty) = 0$ and $F^n \in RH_{-}^{\infty, p \times m}$.

The sequences \hat{G}^n and F^n have the following properties:

- (6.32) (a) $\|\hat{G}^n - \hat{G}^{n-1}\|_\infty \leq 4k\sigma_n$.
 (b) There exists a \hat{G}^∞ of Macmillan degree k such that $\|\hat{G}^n - \hat{G}^\infty\|_\infty \rightarrow 0$ as $n \rightarrow \infty$, and

$$(6.33) \quad \|\hat{G}^n - \hat{G}^\infty\|_\infty \leq 4kM_n,$$

$$(6.34) \quad \|G - \hat{G}^\infty\|_H = \sigma_{k+1}$$

and \hat{G}^∞ is an optimal Hankel-norm approximant for G .

- (c) There exists an $F^\infty \in H_-^{\infty, p \times m}$ such that $\|F^n - F^\infty\|_\infty \rightarrow 0$ as $n \rightarrow \infty$ and
- $$(6.35) \quad \|F^n - F^\infty\|_\infty \leq (4k+2)M_n.$$

(d) If K is a constant matrix and $G, G_n, \hat{G}^n, \hat{G}^\infty$ defined above have inverse transforms $h, h_n, \hat{h}^n, \hat{h}^\infty$, respectively, then

- (i) There exists a constant matrix D_0 such that

$$(6.36) \quad \|G - \hat{G}^\infty - D_0\|_\infty \leq M_k,$$

$$(6.37) \quad (ii) \quad \|h_n - \hat{h}_n\|_1 \leq 4k\sigma_{k+1} + 2(M_k - M_n),$$

$$(6.38) \quad (iii) \quad \|\hat{h}^n - \hat{h}^\infty\|_1 \leq 8kM_n,$$

$$(6.39) \quad (iv) \quad \|h - \hat{h}^\infty\|_1 \leq 4k\sigma_{k+1} + 2M_k,$$

$$(6.40) \quad (v) \quad \int_0^\infty t|h(t) - \hat{h}^\infty(t)|^2 dt \leq 2k\sigma_{k+1}^2 + \sum_{i>k} \sigma_i^2,$$

$$(6.41) \quad (vi) \quad \|G - \hat{G}^\infty\|_N \leq 2k\sigma_{k+1} + M_k.$$

Proof. (a) Expressions (6.20) and (6.31) give

$$(6.42) \quad \|\hat{G}^n + F^n - \hat{G}^{n-1} - F^{n-1}\|_\infty \leq 2\sigma_n$$

and hence by Nehari's Theorem,

$$(6.43) \quad \|\hat{G}^n - \hat{G}^{n-1}\|_H \leq 2\sigma_n.$$

Now since $\hat{G}^n - \hat{G}^{n-1}$ has MacMillan degree $\leq 2k$, Corollary 9.3 in [16] gives $\|\hat{G}^n - \hat{G}^{n-1}\|_\infty \leq 4k\sigma_n$ which is (6.32).

(b) Hence \hat{G}^n forms a Cauchy sequence in $\mathcal{L}^\infty \cap \{\text{the space of rational functions of MacMillan degree } \leq k\}$ and has a limit \hat{G}^∞ satisfying (6.33). Equation (6.34) follows from $\|G_n - \hat{G}^n\|_H = \sigma_{k+1}$ for all n .

(c) Now (6.42) and (6.33) imply $\|F^n - F^{n-1}\|_\infty \leq (4k+2)\sigma_n$ which proves (c).

(i) To prove (6.36) we observe that for all n there exists D^n such that $\|G_n - \hat{G}^n - D^n\|_\infty \leq \sigma_{k+1} + \dots + \sigma_n$ by Corollary 9.9 in [16], with $\|D^n\|_\infty \leq \sigma_{k+1} + \dots + \sigma_n$. The limit point of any convergent subsequence of $\{D^n\}$ can be used as a suitable D_0 .

(ii) Corollary 9.9 in [16] gives bounds for the singular values of the Hankel operator for the error system, $G_n - \hat{G}^n$ (of MacMillan degree $\leq n+k$). Hence using the bound in Theorem 2.1 we obtain (6.37).

(iii) Similarly since the MacMillan degree of $\hat{h}^n - \hat{h}^\infty$ is $\leq 2k$, Theorem 4.1 implies that

$$\|\hat{h}^n - \hat{h}^\infty\|_1 \leq 4k\|\hat{G}^n - \hat{G}^\infty\|_H \leq 4k \times 2M_n \quad \text{by (6.42)}$$

which gives (6.38).

- (iv) Expression (6.39) follows from (6.37), (6.38), and (4.23).
- (v) Expression (6.40) is the Hilbert–Schmidt norm of the error in the Hankel operators, and as in (ii) the result follows from [16, Cor. 9.9].
- (vi) This is similar to the proof of (6.37)–(6.39). \square

The L_1 error bound of (6.39) obtained for an optimal Hankel-norm approximation can be directly compared with the upper bound for truncated output-normal realizations given in Theorem 5.1(b); it is substantially smaller.

In [16, Thm. 9.2], a decomposition of a causal transfer function into the sum of causal all-pass terms scaled by σ_i is given for rational G . This is based on one-step-at-a-time optimal Hankel-norm approximations, and is now generalized to infinite-dimensional systems of nuclear type.

THEOREM 6.5. *Suppose G is the transfer function of a system (2.1) of nuclear type with $m = p$; then there exist $E_i(s)$ and a constant D_0 such that*

$$(6.44) \quad G(s) = D_0 + \sum_{i \geq 1} \sigma_i E_i(s)$$

where

- (i) $E_i(s)$ are all-pass and causal;
- (ii) For all $k \geq 1$, $\hat{G}_k = D_0 + \sum_{i=1}^k \sigma_i E_i$ has MacMillan degree $\leq k$.

Proof. The decomposition (6.44) for G_n was shown in [16] to be

$$G_n = D_0^n + \sum_{i=1}^n \sigma_i E_i^n$$

but it is not clear that E_i^n converge as $n \rightarrow \infty$. However, since D_0^n and E_i^n are all in compact sets, convergent subsequences can be selected, first from $\{D_0^n\}$, then from $\{E_i^n\}$, etc.

It is seen that whereas for finite-dimensional systems the decomposition of (6.44) was easy to calculate from a balanced realization, this is not the case for infinite-dimensional systems. \hat{G}_k does, however, provide a MacMillan degree k approximation to G satisfying $\|G - \hat{G}_k\|_\infty \leq M_k$.

7. Example. In this section realisation and approximation of the following impulse response is considered:

$$h(t) = \begin{cases} 1-t, & 0 \leq t < 1, \\ 0, & t \geq 1, \end{cases}$$

with transfer function

$$g(s) = \int_0^1 (1-t) e^{-st} dt = \frac{e^{-s} - 1 + s}{s^2}.$$

This system will be infinite-dimensional since $h(t)$ is not analytic and $g(s)$ contains e^{-s} . Note that $g(s)$ is an entire function and hence does not have a modal expansion. For the purpose of illustration this example is particularly attractive since closed form expressions can be obtained for most terms. First the Schmidt pairs $(v_i(t)$ and $w_i(t))$, and the singular values, σ_i , of the Hankel operator, Γ , are calculated. Then the output normal realization can be written and the errors in its truncations compared with the theoretical bounds. Finally, the method of § 6 for approximating optimal Hankel-norm reduced-order models is compared with the exact Hankel-norm solution.

7.1. The Schmidt pairs. Since the Hankel operator is real and symmetric, its Schmidt vectors will be its eigenfunctions and its singular values, $\sigma_i = \pm \lambda_i$ the eigenvalues.

TABLE 1

i	$2\alpha_i/\pi$	$\lambda_i = -(-1)^i/\alpha_i^2$
1	1.19373	0.284413
2	2.98835	-0.453834×10^{-1}
3	5.00049	0.162082×10^{-1}
4	6.99998	-0.827117×10^{-2}
5	9.00000	0.500351×10^{-2}
$i > 6$	$(2i-1)$	$-(-1)^i 4/\pi^2(2i-1)^2$

Values accurate to six significant figures.

Functions of this type will be considered in detail in Glover, Lam, and Partington [17] and the results for this particular function will be summarized here.

Let $\alpha > 0$ satisfy the transcendental equation

$$\tan \alpha/2 \tanh \alpha/2 = 1,$$

then it is shown that $\lambda = \alpha^{-2}$ is an eigenvalue with corresponding eigenfunction

$$v(t) = \begin{cases} \frac{\cosh \alpha(t-1/2)}{\cos(\alpha/2)} - \frac{\sin \alpha(t-1/2)}{\sin(\alpha/2)}, & 0 \leq t \leq 1, \\ 0, & t > 1. \end{cases}$$

Now let $\alpha > 0$ satisfy the equation

$$\tan \alpha/2 \tanh \alpha/2 = -1$$

then $\lambda = -\alpha^{-2}$ is an eigenvalue with corresponding eigenfunction

$$v(t) = \begin{cases} \frac{\sinh \alpha(t-1/2)}{\sinh(\alpha/2)} - \frac{\cos \alpha(t-1/2)}{\cos(\alpha/2)}, & 0 \leq t \leq 1, \\ 0, & t > 1. \end{cases}$$

Solving the equation $\tan(\alpha/2) = \pm \coth(\alpha/2)$ numerically gives the values for α and corresponding λ , which are given in Table 1. From Table 1 and $\sum_{i \geq 1} 1/(2i-1)^2 = \pi^2/8$ we obtain $\sum_{i \geq 1} \sigma_i = 0.380$ and hence Γ is indeed nuclear. The corresponding $w_i(t)$ functions will be

$$w_i(t) = (-1)^{i+1} v_i(t)$$

and further $v_i(0) = 2(-1)^{i+1}$, $w_i(0) = 2$.

7.2. State-space realization and its truncations. From the results of § 4 we can now write the state-space realization (A, B, C) , as follows:

$$(B)_i = |\lambda_i| v_i^*(0) = 2\lambda_i,$$

$$(C)_i = w_i(0) = 2,$$

$$(A)_{ii} = -\frac{1}{2} w_i^2(0) = -2,$$

$$(A)_{ij} = \frac{(B)_i (B)_j - \sigma_i^2 (C)_i (C)_j}{(\sigma_i^2 - \sigma_j^2)}$$

$$= \frac{4\lambda_i \lambda_j - 4\lambda_i^2}{\lambda_i^2 - \lambda_j^2} = \frac{-4\lambda_i}{\lambda_i + \lambda_j}.$$

Note that B is bounded from R^1 to l_2 , C is unbounded from l_2 to R^1 and A is unbounded from l_2 to l_2 .

The bounds of Theorem 5.1 on the truncation errors can be evaluated for large n using the accurate asymptotic formula $\sigma_n \rightarrow 1/\pi^2(n - \frac{1}{2})^2$. This gives us that $M_n \approx 1/\pi^2 n$, and for large n ,

$$\frac{1}{\pi^2(n + \frac{1}{2})^2} \leq \|G - G_n\|_\infty \leq 2/\pi^2 n.$$

Theorem 5.1(b) gives, after a small calculation,

$$\frac{1}{\pi^2(n + \frac{1}{2})^2} \leq \|h - h_n\|_1 \leq 53.8/(\pi^2 n^{0.5})$$

and Theorem 5.1(c) gives

$$\|h - h_n\|_2 \leq 6.37/(\pi^2 n^{0.6}).$$

(Note that the latter two bounds could be reduced by a factor of $\sqrt{2}$ using the remark in Lemma 5.1.) Although these upper bounds decrease quite slowly as n increases, their existence suggests that this approximation method will be well behaved. This is indeed confirmed in the following numerical calculations.

The impulse responses of the truncated realizations are plotted in Fig. 2(a) for orders 2, 5, and 20. The L_1 errors in these truncations are approximately

$$0.0033 \approx \sigma_6 \leq \|h - h_5\|_1 \approx 0.015 < 0.54,$$

$$0.016 \approx \sigma_3 \leq \|h - h_2\|_1 \approx 0.074 < 1.45$$

(upper bounds from Theorem 5.1(b)).

The Bode diagrams of the truncations are given in Fig. 2(b) and compared with that of $G(s)$. The L_∞ errors and the lower and upper bounds are as follows (see Theorem 5.1):

$$0.0033 \approx \sigma_6 < \|G - G_5\|_\infty \approx 0.0048 < 2M_5 \approx 0.04,$$

$$0.016 \approx \sigma_3 < \|G - G_2\|_\infty \approx 0.0245 < 2M_2 \approx 0.10.$$

It is seen that the approximation errors are much closer to the lower bound than the upper bounds.

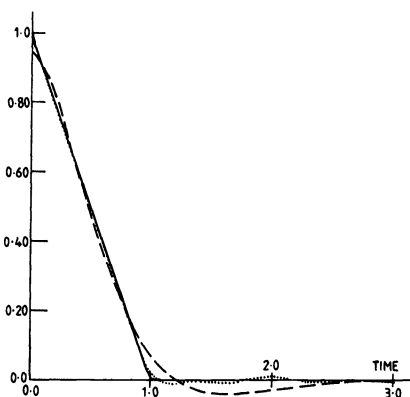


FIG. 2(a). Impulse responses.

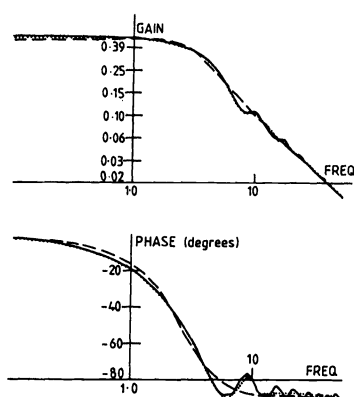


FIG. 2(b). Bode diagrams.

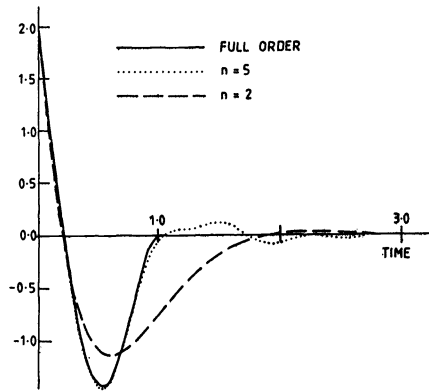


FIG. 2(c). Second Schmidt vectors.

The fact that the Schmidt vectors of the truncated realisations tend to the correct functions is illustrated in Fig. 2(c). Here the $v_i^n(t)$ is plotted for $i = 2$ and $n = 2, 5$, and 20. Upper bounds on the L_2 error were derived in Corollary 5.2.

7.3. Hankel-norm approximation. For this example the method of Adamjan, Arov, and Krein [1] can be performed explicitly. For this we need the Laplace transforms of $v_i(t)$ and $w_i(t)$:

$$\begin{aligned} V_i(s) &= \int_0^1 v_i(t) e^{-st} dt \\ &= \frac{-2}{(s^4 - \alpha_i^4)} \{(-s + \alpha_i \beta_i) s^2 + e^{-s} \alpha_i^2 (s + \alpha_i \beta_i)\} \quad \text{for } i \text{ odd} \\ &= \frac{-2}{(s^4 - \alpha_i^4)} \{(s - \alpha_i \beta_i^{-1}) s^2 + e^{-s} \alpha_i^2 (s + \alpha_i \beta_i^{-1})\} \quad \text{for } i \text{ even} \end{aligned}$$

where $\beta_i = \tanh(\alpha_i/2)$.

If $\hat{G}_k(s) + F(s)$ is the optimal L_∞ approximation to $G(s)$ with $F(s)$ anticausal, then the error is

$$G - \hat{G}_k - F = \lambda_{k+1} V_{k+1}(s) / V_{k+1}(-s)$$

and the optimal Hankel-norm approximation $\hat{G}_k(s)$ can be obtained by calculating the zeros of $V(-s)$ in the left half plane. These will be the k poles of \hat{G}_k and the corresponding residues found from

$$\begin{aligned} R_i &= \lim_{s \rightarrow p_i} \lambda_{k+1} (s - p_i) \frac{V_{k+1}(s)}{V_{k+1}(-s)} \\ &= -\lambda_{k+1} \frac{V_{k+1}(p_i)}{V_{k+1}^{(1)}(-p_i)}. \end{aligned}$$

For the purpose of illustration take $k = 1$; then the pole p satisfies

$$(-p - \alpha_2 \beta_2^{-1}) p^2 + e^{+p} \alpha_2^2 (-p + \alpha_2 \beta_2^{-1}) = 0.$$

TABLE 2

n	\hat{p}	\hat{R}	$\ \hat{G}_1 - \hat{G}_1''\ _\infty$	Upper bound $4M_n$
2	-2.768446	1.574761	0.0303	0.20
5	-2.427493	1.431449	0.005925	0.08
10	-2.450397	1.427323	0.001327	0.04
20	-2.446133	1.427402	0.001095	0.02

(Note this equation has a solution $p = -\alpha_2$ which is cancelled in $V(-s)$.)

$$\Rightarrow p = -2.436089, \quad R = 1.422083, \quad \hat{G}_1(s) = R/(s-p).$$

The procedure given in § 6 has also been carried out and has given the results in Table 2, where $\hat{G}_1^n = \hat{R}/(s - \hat{p})$ is the optimal Hankel-norm approximation to G_n of degree 1. This illustrates the convergence proven in Theorem 6.5.

Finally, we consider the error bounds given in Theorem 6.5 for the optimal Hankel-norm approximation of degree k , with impulse response \hat{h} and transfer function \hat{G} . These give

$$\begin{aligned} \|G - \hat{G} - D_0\|_\infty &\leq M_k \approx 1/(\pi^2 k), \\ \|h - \hat{h}\|_1 &\leq 6/(\pi^2 k), \\ \int_0^\infty t |h(t) - \hat{h}(t)|^2 dt &\leq 7/(3\pi^4 k^3) \approx \sum_{i>k} \sigma_i^2, \\ \|G - \hat{G}\|_N &\leq 3/(\pi^2 k). \end{aligned}$$

The second bound is substantially smaller than the corresponding one in § 7.2. The third bound is on the square of the Hilbert-Schmidt norm of the Hankel operator error and must be greater than $\sum_{i>k} \sigma_i^2$ for any rank k approximant. Hence for this example, although the Hankel norm is being minimized, the Hilbert-Schmidt norm is within a factor of $\sqrt{7}$ of an unachievable lower bound.

Similarly, the nuclear norm is within a factor of 3 of the lower bound $\sum_{i>k} \sigma_i$.

8. Conclusions. We have shown that output-normal and balanced realisations for real systems of nuclear type are attractive in that their truncations form a sequence of approximations whose impulse responses converge in L_1 and L_2 and whose transfer functions converge in L_∞ . Sufficient conditions are given for this convergence but in some cases weaker conditions may be sufficient. Further explicit bounds on the approximation errors have been derived in terms of the singular values $\{\sigma_i\}$ of the corresponding Hankel operator. It has also been shown that output-normal and balanced realisations give a suitable initial realisation from which to derive optimal Hankel-norm approximations. A family of solutions to this problem was derived via a linear fractional map of a ball in H^∞ , and a large number of error bounds were derived. Many of the error bounds in §§ 5 and 6 are believed to be new even for finite-dimensional systems. The error bounds can be used informally to compare the relative merits of the balanced realisation and Hankel-norm techniques. The Hankel norm scheme would appear to be generally superior in the $\|G - \hat{G}\|_H$, $\|G - \hat{G}\|_\infty$, $\|h - \hat{h}\|_1$, and $\|\Gamma - \hat{\Gamma}\|_{HS}$ norms, whereas the balanced realisation technique appears to be a good compromise between these norms and $\|h - \hat{h}\|_2$. Computational experience has indicated good performance for both methods on many examples.

These results can be applied to problems in robust control as follows. If the initial infinite-dimensional system can be modelled as a finite-dimensional totally unstable system in parallel with a system of nuclear type, then the methods of this paper will derive a rational approximation with a prescribed error bound. Robust design techniques for uncertain rational systems can then be applied. For example, the robust stabilisation problem has been solved by Curtain and Glover [7], [8]. Other applications are in the order reduction of filters and controllers. The major difficulty with the technique is in performing the initial polar decomposition of the Hankel operator. "Exact" algorithms are presently being developed for systems with dead time in [17], using similar techniques to those in § 7. For more general problems, however, it is likely that approximation techniques will be required as in, for example, [6] and [19].

Appendix 1. We show that $h \in L_1(0, \infty; \mathbb{C}^{p \times m})$ implies that Γ is a compact operator from $L_q(0, \infty; \mathbb{C}^m)$ to $L_q(0, \infty; \mathbb{C}^p)$ ($1 \leq q < \infty$) and also from $C^1(0, \infty; \mathbb{C}^m)$ to $C^1(0, \infty; \mathbb{C}^p)$.

Let $f \in L_q$ and $g \in L_r(1/q + 1/r = 1)$ satisfy $\|f\|_q, \|g\|_r \leq 1$. Then

$$\begin{aligned} \left| \int_0^\infty (\Gamma f)(t)^* g(t) dt \right| &= \left| \int_0^\infty \int_0^\infty (h(t+s))^* f(s) g(t) ds dt \right| \\ &\leq \int_{u=0}^\infty \int_{s=0}^u \|h(u)\| \|f(s)\| \|g(u-s)\| ds du \quad \text{letting } u = s+t \\ &\leq \int_{u=0}^\infty \|h(u)\| du \quad \text{since } \left| \int_{s=0}^\infty \|f(s)\| \|g(u-s)\| ds \right| \leq 1, \end{aligned}$$

by Holder's inequality.

It follows that $\|\Gamma f\|_q \leq \|h\|_1$, so that the operator norm of Γ is at most $\|h\|_1$.

For compactness in L_q we require that

- (i) $\lim_{x \rightarrow 0} \int_0^\infty |\Gamma f(x+y) - \Gamma f(y)|^p dy = 0$ uniformly over $\|f\|_q \leq 1$;
- (ii) $\lim_{A \rightarrow \infty} \int_A^\infty |\Gamma f(y)|^p dy = 0$ again uniformly over $\|f\|_q \leq 1$. (See Dunford and Schwarz [32, p. 298].)

To obtain (i) we note that, if $\|g\|_r \leq 1$

$$\begin{aligned} \int_0^\infty \|\Gamma f(x+t) - \Gamma f(t)\| \|g(t)\| dt &\leq \int_0^\infty \int_0^\infty \|h(s+t+x) - h(s+t)\| \|f(s)\| \|s(t)\| ds dt \\ &= \int_{u=0}^\infty \int_{s=0}^u \|h(u+x) - h(u)\| \|f(s)\| \|g(u-s)\| ds du \\ &\leq \int_{u=0}^\infty \|h(u+x) - h(u)\| du \rightarrow 0 \quad \text{as } x \rightarrow 0 \quad \text{since } h \in L_1. \end{aligned}$$

Similarly (ii) follows since

$$\begin{aligned} \int_{t=A}^\infty \int_{s=0}^\infty \|h(t+s)\| \|f(s)\| \|g(t)\| ds dt &= \int_{u=A}^\infty \int_{s=0}^u \|h(u)\| \|f(s)\| \|g(u-s)\| ds du \\ &\leq \int_{u=A}^\infty \|h(u)\| du \rightarrow 0 \quad \text{as } A \rightarrow \infty. \end{aligned}$$

To obtain compactness in C^1 , we may argue similarly using conditions in [32, p. 343], or alternatively as follows:

Write

$$\begin{aligned}
 j(t) &= \int_0^t h(s) ds, \quad \dot{j}(t) = h(t) \quad \text{a.e.}, \\
 (\Gamma \dot{f})(t) &= \frac{d}{dt} \int_0^\infty j(t+s)f(s) ds \quad \text{for } f \in C^1 \\
 &= \frac{d}{dt} \left\{ [j(t+s)f(s)]_0^\infty - \int_0^\infty j(t+s)\dot{f}(s) ds \right\} \\
 &= \frac{d}{dt} \left\{ -\dot{j}(t)f(0) - \int_0^\infty j(t+s)\dot{f}(s) ds \right\} \\
 &= -h(t)f(0) - (\Gamma \dot{f})(t);
 \end{aligned}$$

also

$$(\Gamma f)(0) = \int_0^\infty h(t)f(t) dt \leq \|h\|_1 \|f\|_\infty.$$

The map $D: C^1(0, \infty; \mathbb{C}^m) \rightarrow L_1(0, \infty; \mathbb{C}^m) \oplus \mathbb{C}^m$ defined by $Df = (\dot{f}, f(0))$ establishes an isomorphism of normed spaces. The induced map $E = D\Gamma D^{-1}: L_1(0, \infty; \mathbb{C}^m) \oplus \mathbb{C}^m \rightarrow L_1(0, \infty; \mathbb{C}^p) \oplus \mathbb{C}^p$ given by $E(g, x) = (-\langle \Gamma g \rangle - xh(t), \langle h, D^{-1}(g, x) \rangle)$ is compact by the above arguments in L_1 . Hence $\Gamma = D^{-1}ED$ is also compact.

It now follows, since $X = L_1 \cap L_2 \cap C^1$ is dense in each of L_1 , L_2 and C^1 and the operator $\Gamma: X \rightarrow X$ is compact, that all the singular vectors lie in X , and indeed in every L_q space ($1 \leq q < \infty$), as shown in § 2.

Appendix 2. We show here the result used in the proof of Lemma 4.3, namely that if (H_m) is a sequence of compact operators on a Hilbert space, such that $\|H_m - H\| \rightarrow 0$, and if H has distinct singular values (σ_i) and Schmidt vectors $\{v_i, w_i\}$, then the singular values of H_m , (σ_i^m) converge to σ_i and the Schmidt vectors $\{v_i^m, w_i^m\}$ suitably normalized converge in norm to $\{v_i, w_i\}$ for each i .

Proof. As given in [30, p. 30], $|\sigma_i^m - \sigma_i| \leq \|H_m - H\|$, and so the singular values converge.

We show first that the leading Schmidt vectors converge, that is, that $\|v_1^m - v_1\| \rightarrow 0$. Write $v_1^m = r_m v_1 + x_m$, where $(x_m, v_1) = 0$, so that $\|x_m\| = \sqrt{1 - |r_m|^2}$

$$\begin{aligned}
 \sigma_1^m &= \|H_m v_1^m\| \leq \|H v_1^m\| + \|H - H_m\| \\
 &= \|r_m \sigma_1 w_1 + H x_m\| + \|H - H_m\| \\
 &\leq (\sigma_1^2 |r_m|^2 + \sigma_2^2 (1 - |r_m|^2))^{1/2} + \|H - H_m\| \quad \text{since } (H x_m, w_1) = 0.
 \end{aligned}$$

Hence

$$\begin{aligned}
 \sigma_1^2 |r_m|^2 + \sigma_2^2 (1 - |r_m|^2) &\geq (\sigma_1^m - \|H - H_m\|)^2 \\
 &\geq (\sigma_1 - 2\|H - H_m\|)^2.
 \end{aligned}$$

Thus

$$|r_m|^2 \geq \frac{(\sigma_1 - 2\|H - H_m\|)^2 - \sigma_2^2}{\sigma_1^2 - \sigma_2^2} \rightarrow 1 \quad \text{as } m \rightarrow \infty.$$

Therefore, choosing the normalisation of v_1^m appropriately, we have that $\|v_1^m - v_1\| \rightarrow 0$, and since $\|H_m^* - H^*\| \rightarrow 0$, $\|w_1^m - w_1\| \rightarrow 0$ similarly.

The general result now follows by induction on i , since, if it is given for $i = 1, \dots, k-1$, then we observe that

$$H_m^k \rightarrow H^k \quad \text{where } H_m^k(x) = H_m(x) - \sum_{i=1}^{k-1} \sigma_i^m(x, v_i^m) w_i^m,$$

$$H^k(x) = H(x) - \sum_{i=1}^{k-1} \sigma_i(x, v_i) w_i.$$

The leading Schmidt vectors are now v_k^m and v_k and hence the result follows for $i = k$, using the arguments above.

The result extends easily to maps between different but isometric Hilbert spaces.

Acknowledgment. The assistance of J. Lam in computation of the numerical results presented in § 7.2 is gratefully acknowledged.

REFERENCES

- [1] V. M. ADAMJAN, D. Z. AROV, AND M. G. KREIN, *Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem*, Math. USSR-Sb., 15 (1971), pp. 31-73.
- [2] J. A. BALL AND J. W. HELTON, *A Beurling-Lax theorem for the Lie group $U(m, n)$ which contains most classical interpolation theory*, J. Operator Theory, 9 (1983), pp. 107-142.
- [3] J. A. BALL AND A. C. N. RAN, *Hankel-norm approximation of a rational matrix function in terms of its Realisation*, Mathematical Theory of Network Systems Conf., Stockholm, Sweden, June 1985, in Modelling, Identification and Robust Control, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, New York, 1986.
- [4] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite-Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sciences 8, Springer-Verlag, Berlin, New York, 1978.
- [5] R. F. CURTAIN AND K. GLOVER, *Balanced realisations for infinite dimensional systems*, in Proc. Workshop on Operator Theory and Systems, H. Bart, I. Gohberg, and M. A. Kaashoek, eds., Birkhäuser-Verlag, Basel, 1986. Amsterdam, the Netherlands, June 4-7, 1985.
- [6] R. F. CURTAIN, K. GLOVER, AND J. LAM, *Reduced order models for distributed systems based on optimal Hankel-norm approximations*, 5th Virginia Polytechnic Inst. and State Univ./American Inst. for Aeronautics and Astronautics Symp. on Dynamics and Control of Large Structures, L. Meirovitch, ed., Blacksburg, VA, June 12-14, 1985.
- [7] R. F. CURTAIN AND K. GLOVER, *Robust stabilization of infinite dimensional systems by finite dimensional controllers*, Systems Control Lett., (1986), pp. 41-47.
- [8] ———, *Robust stabilization of infinite dimensional systems by finite dimensional controllers: derivations and examples*, Mathematical Theory of Network and Systems Conf., Stockholm, Sweden, June 1985, in Modelling, Identification and Robust Control, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986.
- [9] R. F. CURTAIN, *Sufficient conditions for infinite rank Hankel operators to be nuclear*, J. Math. Control Inform., 2 (1985), pp. 171-180.
- [10] ———, *Balanced realisations for discrete-time infinite-dimensional systems*, in System Modelling and Optimization, Lecture Notes in Control and Inform. Sci., 84, A. Prékopa, J. Szelecsán, and B. Strazicky, eds., Springer-Verlag, Berlin, New York, 1986.
- [11] ———, *The linear quadratic control problem with fixed endpoint*, J. Optim. Theory Appl., 44 (1984), pp. 55-74.
- [12] E. B. DAVIES, *One-Parameter Semigroups*, Academic Press, London, 1980.
- [13] J. DOYLE, Lecture Notes, Office of Naval Research/Honeywell Workshop, Minneapolis, MN, 1984.
- [14] D. ENNS, *Model reduction for control system design*, Ph.D thesis, Dept. of Aeronautics and Astronautics, Stanford Univ., Stanford, CA, 1984.
- [15] P. A. FUHRMANN, *Linear Systems and Operators in Hilbert Space*, McGraw-Hill, New York, 1981.
- [16] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L_∞ error bounds*, Internat. J. Control, 39 (1984), pp. 1115-1193.
- [17] K. GLOVER, J. LAM AND J. PARTINGTON, *Balanced realisation and Hankel-norm approximation of systems involving delays*, Proc. IEEE Conf. on Decision and Control, Athens, Greece, December 1986.

- [18] J. W. HELTON, *Non-Euclidean functional analysis and electronics*, Bull. Amer. Math. Soc., 7 (1982), pp. 1–64.
- [19] E. A. JONCKHEERE AND L. M. SILVERMAN, *Singular value analysis of deformable systems*, in Rational Approximation in Systems Engineering, A. Bultheel and P. Dewilde, eds., Birkhäuser-Verlag, Basel, 1983.
- [20] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, 1976.
- [21] B. C. MOORE, *Principal component analysis in linear systems: controllability, observability and model reduction*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 17–32.
- [22] N. K. NIKOL'SKII, *Ha-plitz operators: a survey of some recent results*, in Operators and Function Theory, S. C. Power, ed., Reidel, Boston, 1985.
- [23] L. PERNEBO AND L. M. SILVERMAN, *Model reduction via balanced state space representation*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 382–387.
- [24] S. C. POWER, *Hankel Operators on Hilbert Space*, Pitman, Boston, 1982.
- [25] R. M. REDHEFFER, *On a certain linear fractional transformation*, J. Math. Phys., 39 (1960), pp. 269–286.
- [26] J. WEIDMANN, *Linear Operators in Hilbert Spaces*, Springer-Verlag, Berlin, New York, 1980.
- [27] N. Young, *Balanced Realisation in Infinite Dimensions*, Springer-Verlag, Berlin, New York, Heidelberg, 1980.
- [28] I. C. GOHBERG AND M. K. ZAMBICKII, *Normally solvable operators in spaces with two norms*, Bul. Akad. Sci. RSS Moldoven, 6 (1964), pp. 80–84. (In Russian.)
- [29] P. D. LAX, *Symmetrizable linear transformations*, Comm. Pure Appl. Math., 7 (1954), pp. 633–647.
- [30] I. C. GOHBERG AND M. G. KREIN, *Introduction to the Theory of Linear Nonselfadjoint Operators in Hilbert Space*, Trans. Math. Monographs, 18 (1969).
- [31] R. R. COIFMAN AND R. ROCHBERG, *Representation theorems for holomorphic and harmonic functions in L^p* , Asterisque, 77 (1980), pp. 11–66.
- [32] N. DUNFORD AND J. T. SCHWARZ, *Linear Operators. Part 1: General Theory*, Interscience, New York, 1957.
- [33] A. PIETSCH, *Operator Ideals*, North-Holland, Amsterdam, 1980.

THE LOCAL STRUCTURE OF TIME-OPTIMAL TRAJECTORIES IN DIMENSION THREE UNDER GENERIC CONDITIONS*

HEINZ SCHÄTTLER†

Abstract. We consider the problem of time-optimal control for systems of the form $\dot{x} = f(x) + g(x)u$, where f and g are smooth vector fields and admissible controls are measurable scalar functions u with values in $-1 \leq u \leq 1$. Under the assumption that f , g and $[f, g]$ are independent, and that also one of the triples $(g, [f, g], [f + g, [f, g]])$ or $(g, [f, g], [f - g, [f, g]])$ consists of independent vectors, we show that generically every point has a neighborhood U such that time-optimal trajectories that lie in U are concatenations of at most six bang and singular arcs. This implies that globally time-optimal trajectories are finite concatenations of bang and singular arcs with a bound on the number of switchings; in particular, time-optimal controls are piecewise smooth. Results of this type are relevant for the existence of a regular synthesis.

Key words. time-optimal control, nonlinear system, generic, regular synthesis, singular controls

AMS(MOS) subject classifications. 49B10, 93B10

1. Introduction. We study the problem of time-optimal control for a system

$$(1.1) \quad \Sigma: \quad \dot{x} = f(x) + g(x)u, \quad |u| \leq 1, \quad x \in \mathbb{R}^3$$

where f and g are smooth vector fields. Admissible controls are measurable scalar functions with values in $-1 \leq u \leq 1$ and a trajectory of the system corresponding to a control $u(\cdot)$ is an absolutely continuous curve $x(\cdot)$ such that $\dot{x} = f(x(t)) + g(x(t))u(t)$ holds almost everywhere.

The particular topic we are interested in is regularity properties of optimal trajectories, both local and global. For a system which is degenerate in the sense that all controls are time-optimal (i.e., time is a coordinate) obviously no such results can hold. In the class of smooth systems it is also easy to construct an example where an a priori given, but arbitrary, control u is the unique control which transfers a point q into a point p (see Sussmann [21]). This means that a selection of optimal trajectories with additional regularity properties may not be possible. Nevertheless, these are quite pathological cases, and it is natural to ask whether or not for a large class of systems optimal controls are more regular than just being measurable functions. The answer to this question has immediate connections with the concept of a "regular synthesis" (i.e., with sufficient conditions for optimality) as defined by Boltyansky [1] and later refined by Brunovsky [3] and Sussmann [14]. In order to explain this in more detail it is necessary to establish some terminology. To do so we review the necessary conditions of the Pontryagin Maximum Principle [9].

If $\Gamma = (x(\cdot), u(\cdot))$ is a time-optimal pair defined on an interval $[0, T]$, then there exist a constant $\lambda_0 \geq 0$ and an absolutely continuous curve $\lambda(\cdot): [0, T] \rightarrow \mathbb{R}^3$ (called the adjoint vector), which is not identically zero, such that

$$(1.2) \quad \dot{\lambda}(t)^T = -\lambda(t)^T (Df(x(t)) + u(t)Dg(x(t))),$$

$$(1.3) \quad \langle \lambda(t), g(x(t))u(t) \rangle = \min_{|v| \leq 1} \langle \lambda(t), g(x(t))v \rangle,$$

$$(1.4) \quad \langle \lambda(t), f(x(t)) + u(t)g(x(t)) \rangle + \lambda_0 = 0,$$

* Received by the editors August 11, 1986; accepted for publication (in revised form) October 2, 1987.

† Department of Systems Science and Mathematics, Campus Box 1040, Washington University, St. Louis, Missouri 63130.

almost everywhere on $[0, T]$. (We write vectors as columns and $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product on \mathbb{R}^3 ; Df and Dg are the Jacobian matrices of f and g .) Any pair Γ for which an adjoint vector λ and a constant $\lambda_0 \geq 0$ exist such that these conditions hold is called an *extremal*. The function $\phi_\Gamma(t) := \langle \lambda(t), g(x(t)) \rangle$ is called the *switching function*. On an interval where ϕ_Γ is positive (respectively, negative) the optimal control is constant and equal to -1 ($+1$) almost everywhere. We call trajectories which correspond to these constant controls *bang arcs*. A concatenation of bang arcs is a *bang-bang trajectory*. At times where the switching function vanishes the optimization condition (1.3) gives no information about the control. If, however, ϕ_Γ vanishes on an open interval I , then also all the derivatives of ϕ_Γ vanish on I and this may determine the control u on I . We have (as usual let $[f, g]$ be the Lie bracket of f and g)

$$\begin{aligned}\dot{\phi}_\Gamma(t) &= \langle \lambda(t), [f, g](x(t)) \rangle, \\ \ddot{\phi}_\Gamma(t) &= \langle \lambda(t), [f + ug, [f, g]] \rangle,\end{aligned}$$

and so, assuming $\langle \lambda, [g, [f, g]] \rangle$ does not vanish on I , we get

$$(1.5) \quad u(t) = -\frac{\langle \lambda(t), [f, [f, g]](x(t)) \rangle}{\langle \lambda(t), [g, [f, g]](x(t)) \rangle}.$$

A control of this type is called *singular* and the corresponding trajectory is a *singular arc*.

Therefore bang and singular arcs are the obvious candidates for time-optimal trajectories. It seems natural to conjecture—and many people seem to take it for granted—that apart from the degenerate case mentioned above optimal controls are concatenations of bang and singular arcs. However at the moment, little, if anything at all, is known to consolidate such a conjecture for arbitrary dimension n . If it is true, then in view of Brunovsky's results on regular synthesis a very important aspect is whether the number of switchings is bounded or not.

Brunovsky substantially relaxed the conditions given originally by Boltyansky [1] by applying the theory of subanalytic sets to the problem of synthesizing a globally optimal solution out of local information [3]. It also became clear that crucial for this approach is a certain degree of regularity of the optimal trajectories relating to a local finiteness property of subanalytic sets. Specifically, for an analytic system of the form (1.1) the crucial hypothesis is that time-optimal trajectories are finite concatenations of bang and singular arcs with a *bound on the number of switchings*. The precise meaning of this phrase is that for every compact set K and for every positive time T there exists an integer $N = N(K, T)$, with the property that whenever $\Gamma = (x(\cdot), u(\cdot))$ is a time-optimal trajectory which steers a point p_1 to a point p_2 in time $\leq T$ and lies in K , then $u(\cdot)$ has at most N switchings. If we consider the problem of steering points to a target point p time-optimally, then this implies that as long as an optimal trajectory is restricted to the compact set K consisting of all points which can be steered into p within a fixed time, then there exist an a priori bound on the number of switchings which depends only on K . If f and g are analytic vector fields, and if the singular control can be written in feedback form as an analytic function, then it follows from this that the problem of steering points to a target point p time-optimally has a piecewise analytic feedback solution [12, Thm. 6]. In particular, the corresponding value function V is piecewise analytic in the sense that there exists a stratification of the set of points which can be steered into p such that the restrictions of V to the strata are analytic functions [12, Thm. 5].

That regularity results of this type need not hold shows in Fuller's example [8], which can be viewed as a time-optimal control problem of the form (1.1) in dimension three. There we find optimal bang-bang trajectories with infinitely many switchings in arbitrarily small time. Recent work of Kupka [7] shows that this phenomenon occurs generically in sufficiently high dimension ($n \geq 13$).

Attempts to understand the mystery of the Fuller phenomenon as well as the work of Brunovsky again raised the interest in qualitative questions having to do with regularity properties of optimal trajectories. Note that the problem of describing time-optimal trajectories essentially is a local problem: the global structure can be obtained from the local structure at every point by a simple compactness argument. More precisely, if it is known that every point p has a neighborhood $U = U(p)$ such that time-optimal trajectories which lie in U are concatenations of at most $N = N(U)$ bang and singular arcs, then time-optimal trajectories are globally finite concatenations of bang and singular arcs with a bound on the number of switchings. (The argument will be carried out in § 2.) Therefore it suffices to study the local structure of time-optimal trajectories near reference points.

It is well known (cf., for instance, [16]) that for an analytic system the local behavior near p —in particular, the structure of time-optimal trajectories—is determined by the Lie algebra generated by f and g at p and all the Lie relations that hold at p , i.e., by all finite linear combinations of f , g and brackets of f and g that vanish at p . We call this the Lie bracket configuration. (For a precise definition see [16]; [15] contains a very readable explanation of the concept.) It is this connection with Lie algebra which makes the general n -dimensional case so difficult. Therefore attention has focused on low dimensions.

In a series of papers Sussmann [18]–[20] was able to solve the two-dimensional problem. For a generic smooth system in the plane he gave a classification of the local structure of time-optimal trajectories at every point. In particular, he showed that they are finite concatenations of bang and singular arcs with a bound on the number of switchings. For the class of analytic systems he extended his results to prove the existence of a regular synthesis for basically arbitrary analytic systems only subject to a mild “nonexplosion” condition. Such a precise description of time-optimal trajectories in higher dimensions seems unlikely, but it is natural to ask about the generic case. Is the Fuller phenomenon also generic in low dimensions? More generally, are infinite concatenations of bang and singular arcs possible in arbitrarily small time for a generic system? Can even more pathological situations occur? Since some of the crucial arguments in Sussmann's work on the plane are two-dimensional in nature, even the step to dimension three is nontrivial.

The first result in this direction is due to Bressan, who studied the problem of controlled stability (i.e., $f(p) = 0$) under generic conditions [2]. He proved that an equilibrium point p has a neighborhood U , such that any point $q \in U$ can be steered into p by a trajectory which is bang-bang with at most two switches or is at most a concatenation of a bang arc, followed by a singular arc and one more bang arc. It is clear that, if we want to get a global description of time-optimal trajectories, we must also consider the nonequilibrium case. Global structure results—such as time-optimal trajectories having a limited number of switchings—are equivalent to local results of this type at every point. It is not enough to know what happens near the equilibrium. Therefore the analysis of nonequilibria is equally important. In fact, generically equilibria will be an isolated phenomenon, whereas nonequilibria occur in an open dense set. Therefore the nonequilibrium case may even be considered more pressing to begin with.

To analyze the generic situation it is natural to begin with the least degenerate case and to proceed toward more degenerate situations. In dimension three the natural nondegeneracy assumption to make is that f , g and $[f, g]$ are independent. In [10] we presented a new technique to compute necessary conditions for optimality of bang-bang trajectories. This technique is based on Lie-algebraic computations. We applied it to the three-dimensional case with these vector fields as basis for our computations. We showed that under additional generic assumptions bang-bang trajectories with too many switchings (locally more than seven) are not time-optimal. In this paper we continue the analysis of time-optimal trajectories under this assumption by incorporating singular arcs. The aim is to restrict possible concatenations between bang and singular arcs and, in particular, to prove that time-optimal trajectories are in fact concatenations of bang and singular arcs with a bound on the number of switchings. Under the additional assumption that also one of the triples $(g, [f, g], [f+g, [f, g]])$ or $(g, [f, g], [f-g, [f, g]])$ consists of independent vectors this is true generically. Depending on the Lie bracket configuration we give a specific upper bound on the switching structure for time-optimal trajectories in each case.

A precise statement of our results will be given in § 2. We also show briefly how global structure results for optimal trajectories (including the case of controlled stability) follow from the local results we prove. Then we comment on how close our results are to a solution of the full generic problem. The proofs will then be given in §§ 3–6. In § 3 we analyze singular arcs and singular junctions (i.e., junctions where one arc is singular); in particular we will give new necessary conditions for optimality of concatenations of bang and singular arcs. In §§ 4 and 5 we then determine the local structure of time-optimal trajectories under generic conditions. We distinguish between the cases when optimal trajectories are bang-bang and when optimal singular arcs are possible. In § 6 we prove the genericity of our results.

2. Statement and interpretation of the results. We start with a convenient summarizing version of our results using standard genericity notions; then we give more detailed statements relating the local structure of time-optimal trajectories to the Lie bracket configuration.

We identify Σ with the C^∞ map $\Sigma = (f, g) : \mathbb{R}^3 \rightarrow \mathbb{R}^6$ and equip $C^\infty(\mathbb{R}^3, \mathbb{R}^6)$ with the Whitney C^∞ topology. (A definition of this topology is included in § 6.) Let \mathcal{B}_+ (respectively, \mathcal{B}_-) be the open subset of all systems Σ for which both triples $(f, g, [f, g])$ and $(g, [f, g], [f+g, [f, g]])$ (respectively, $(g, [f, g], [f-g, [f, g]])$) consist of independent vectors everywhere. Let $\mathcal{B} := \mathcal{B}_+ \cup \mathcal{B}_-$.

THEOREM 1. *For a generic system $\Sigma \in \mathcal{B}$ every point has a neighborhood U such that every time-optimal trajectory that lies in U is a concatenation of bang and singular arcs with at most six pieces.* \square

To state our results more precisely we need some notation. Our considerations will be local, i.e., they will hold only on a small neighborhood of some reference point p . We assume that f , g , and $[f, g]$ are independent there. If we let $X := f - g$ and $Y := f + g$, then we can write all higher order brackets as linear combinations of X , Y and $[X, Y]$. We write, using $\text{ad } X(Y) = [X, Y]$,

$$\begin{aligned}
 [X, [X, Y]] &= a_1 X + a_2 Y + a_3 [X, Y] = \alpha f + \cdots, \\
 [Y, [X, Y]] &= b_1 X + b_2 Y + b_3 [X, Y] = \beta f + \cdots, \\
 [X, [X, [X, Y]]] &= c_1 X + c_2 Y + c_3 [X, Y] = \gamma f + \cdots, \\
 (2.1) \quad [Y, [Y, [X, Y]]] &= e_1 X + e_2 Y + e_3 [X, Y] = \eta f + \cdots,
 \end{aligned}$$

$$\begin{aligned}
\text{ad}^4 X(Y) &= m_1 X + m_2 Y + m_3 [X, Y] = \mu f + \cdots, \\
\text{ad}^5 X(Y) &= k_1 X + k_2 Y + k_3 [X, Y] = \kappa f + \cdots, \\
-\text{ad}^4 Y(X) &= n_1 X + n_2 Y + n_3 [X, Y] = \nu f + \cdots, \\
-\text{ad}^5 Y(X) &= r_1 X + r_2 Y + r_3 [X, Y] = \rho f + \cdots, \\
[Y, [X, [X, [X, Y]]]] &= o_1 X + o_2 Y + o_3 [X, Y] = \omega f + \cdots.
\end{aligned}$$

The independence of g , $[f, g]$, and $[X, [X, Y]]$, respectively, $[Y, [X, Y]]$, is therefore equivalent to $\alpha \neq 0$, respectively, $\beta \neq 0$. We denote the Lie derivative of a function ϕ in direction of X by $L_X \phi$. Observe that

$$\begin{aligned}
L_X \alpha &= \gamma - a_3 \alpha, & L_X \gamma &= \mu - c_3 \alpha, \\
L_X \mu &= \kappa - m_3 \alpha, & L_Y \gamma &= \omega - c_3 \beta.
\end{aligned}$$

Finally, for a trajectory which is a concatenation of bang (X or Y) and singular arcs (S), we use the corresponding letter sequence to denote the trajectory. So an $XYSX$ -trajectory is a concatenation of an X -arc, followed by a Y -arc, then a singular arc and another X -arc. To simplify the notation we also allow this to stand for concatenations of at most these arcs, i.e., we allow for pieces to be absent. In this sense XSX is also of the form $XYSX$. Furthermore, we use B to denote a bang arc, i.e., an arc which is either X or Y . We can now state our results more precisely.

PROPOSITION 1. *Let $\beta(p) < 0$. There exists a neighborhood U of p such that time-optimal trajectories that lie in U have the following structure:*

- (0) $\alpha(p) \neq 0$: XYX if $\alpha(p) < 0$, BBB if $\alpha(p) > 0$;
- (i) $\alpha(p) = 0$, $\gamma(p) \neq 0$: $XYXY$ if $\gamma(p) > 0$, $YXYX$ if $\gamma(p) < 0$;
- (ii) $\alpha(p) = 0$, $\gamma(p) = 0$, $\mu(p) \neq 0$: $XYXYX$ if $\mu(p) < 0$, $YXYXY$ if $\mu(p) > 0$;
- (iii) $\alpha(p) = 0$, $\gamma(p) = 0$, $\mu(p) = 0$, $\kappa(p) \neq 0$: $XYXYXY$ if $\kappa(p) > 0$, $YXYXYX$ if $\kappa(p) < 0$.

PROPOSITION 2. *Let $\beta(p) > 0$. There exists a neighborhood U of p such that time-optimal trajectories that lie in U have the following structure:*

- (0) $\alpha(p) \neq 0$: YXY if $\alpha(p) > 0$, BSB if $\alpha(p) < 0$;
- (i) $\alpha(p) = 0$, $\gamma(p) \neq 0$: $BSXY$ if $\gamma(p) > 0$, $YXSB$ if $\gamma(p) < 0$;
- (ii) $\alpha(p) = 0$, $\gamma(p) = 0$, $\mu(p) \neq 0$: $YXSXY$ if $\mu(p) > 0$, $BSXSB$ if $\mu(p) < 0$;
- (iii) $\alpha(p) = 0$, $\gamma(p) = 0$, $\mu(p) = 0$, $\kappa(p) \neq 0$: $BSXSXY$ if $\kappa(p) > 0$, $YXSXSB$ if $\kappa(p) < 0$. \square

In § 6 we will prove that Propositions 1 and 2 contain all the generic cases in \mathcal{B} where β does not vanish. The remaining cases, i.e., those where α does not vanish, can easily be obtained from this using an input symmetry which interchanges X and Y (cf. [10, § 3.2]). Let Σ^* be the system where g is replaced by $-g$, i.e., with $X^* = Y$ and $Y^* = X$. Then

$$\alpha f + \cdots = [X, [X, Y]] = [Y^*, [Y^*, X^*]] = -\beta^* f + \cdots.$$

So we have $\beta^* = -\alpha$ and it follows similarly that $\alpha^* = -\beta$, $\gamma^* = -\eta$, $\eta^* = -\gamma$, $\mu^* = -\nu$, and $\kappa^* = -\rho$. Using this, we can translate all our results for the case $\beta(p) \neq 0$ into results for the case $\alpha(p) \neq 0$. For instance, if $\alpha(p) > 0$, $\beta(p) = 0$ and $\eta(p) > 0$, then $\beta^*(p) < 0$, $\alpha^*(p) = 0$ and $\gamma^*(p) < 0$. Therefore time-optimal Σ^* -trajectories are at most of the form $Y^*X^*Y^*X^*$ by Proposition 1. Hence time-optimal trajectories of Σ are of the form $XYXY$ near p . Analogously we get the following corollary.

COROLLARY 1. *Let $\alpha(p) > 0$. There exists a neighborhood U of p such that time-optimal trajectories that lie in U have the following structure:*

- (0) $\beta(p) \neq 0$: YXY if $\beta(p) > 0$, BBB if $\beta(p) < 0$;
- (i) $\beta(p) = 0$, $\eta(p) \neq 0$: $XYXY$ if $\eta(p) > 0$, $YXYX$ if $\eta(p) < 0$;
- (ii) $\beta(p) = 0$, $\eta(p) = 0$, $\nu(p) \neq 0$: $XYXYX$ if $\nu(p) < 0$, $YXYXY$ if $\nu(p) > 0$;
- (iii) $\beta(p) = 0$, $\eta(p) = 0$, $\nu(p) = 0$, $\rho(p) \neq 0$: $XYXYXY$ if $\rho(p) > 0$, $YXYXYX$ if $\rho(p) < 0$.

We will show below that singular arcs are the integral curves of the smooth vector field $S = (\alpha Y - \beta X)/(\alpha - \beta)$. So $S^* = S$ and we conclude the following from Proposition 2.

COROLLARY 2. *Let $\alpha(p) < 0$. There exists a neighborhood U of p such that time-optimal trajectories that lie in U have the following structure:*

- (0) $\beta(p) \neq 0$: XYX if $\beta(p) < 0$, BSB if $\beta(p) > 0$;
- (i) $\beta(p) = 0$, $\eta(p) \neq 0$: $BSYX$ if $\eta(p) < 0$, $XYSB$ if $\eta(p) > 0$;
- (ii) $\beta(p) = 0$, $\eta(p) = 0$, $\nu(p) \neq 0$: $XYSYX$ if $\nu(p) < 0$, $BSYSB$ if $\nu(p) > 0$;
- (iii) $\beta(p) = 0$, $\eta(p) = 0$, $\nu(p) = 0$, $\rho(p) \neq 0$: $BSYSYX$ if $\rho(p) < 0$, $XYSYSB$ if $\rho(p) > 0$.

Remark. The codimension-0 cases when α and β have the same sign are special instances of a general nonlinear bang-bang theorem of Sussmann [13]: if $g(p)$, $[f, g](p)$ and $[f, [f, g]](p)$ are independent, and if

$$[g, [f, g]] = ag + b[f, g] + c[f, [f, g]]$$

with $|c(p)| < 1$, then time-optimal trajectories are bang-bang with at most two switchings near p . (This follows from Lemma 3 of [13].) Under our standing assumption $f(p) \wedge g(p) \wedge [f, g](p) \neq 0$, $|c(p)| < 1$ is equivalent to $\alpha(p)\beta(p) > 0$. The other codimension-0 cases correspond to $|c(p)| > 1$ and the cases (i)–(iii) correspond to the degenerate situation when $|c(p)| = 1$. A bang-bang theorem is then still valid if $\alpha(p) > \beta(p)$ and α or β vanishes at p . Under generic conditions we have given a bound on the number of switchings.

These results give efficient upper bounds for the local switching structure under easy to verify assumptions. For several reasons, however, we will not answer the natural question whether our bounds are sharp. In the codimension-0 cases (0) it is intuitively clear and in fact easy to see (cf. [6]) that the bounds are sharp; for more degenerate cases such as the ones of codimension three this is not only *not* obvious, but also a nontrivial problem. From an aesthetical point of view it would be very satisfying to know that the bounds are sharp, but solutions to optimal control problems different from the standard textbook examples are extremely difficult and these degenerate cases do not fit into any easy category. More important, the issue of sharpness of the bounds is not really at hand here. In view of Brunovsky's results (see above) it is clear that the question whether the number of switchings is locally bounded or not is important, but the question of a precise bound is less significant. Furthermore, if we view the results as giving global structure results for time-optimal trajectories by analyzing the local behavior near every point, the issue of sharp bounds diminishes even more. To support this claim let us rephrase our results in the following theorem.

THEOREM 2. *Let Σ be a three-dimensional system which at every point satisfies one of the conditions of Propositions 1 or 2 or of Corollaries 1 or 2. Then time-optimal trajectories are finite concatenations of bang and singular arcs with a bound on the number of switchings.*

This follows from our results by a simple compactness argument. If K is a compact set in the state space, cover K by a finite number of neighborhoods U_{p_1}, \dots, U_{p_n} as

constructed in Propositions and Corollaries 1 and 2. Let L be a Lebesgue number for this covering. There exists a constant $C = C(K)$ such that if $x(\cdot)$ is any trajectory contained in K , then $\|x(t_1) - x(t_2)\| \leq C|t_1 - t_2|$ for all t_1, t_2 in the domain of $x(\cdot)$. Given T , choose M such that $ML \geq CT$. If $(x(\cdot), u(\cdot))$ is a time-optimal trajectory defined on $[0, t]$, $t \leq T$, and contained in K , then we can partition $[0, t]$ into at most M subintervals I_j such that $\|x(t_1) - x(t_2)\| \leq L$ for t_1, t_2 in I_j . Hence $x(\cdot)$ restricted to I_j lies entirely in one of the neighborhoods U_{p_i} and so it is a concatenation of at most six bang and singular arcs. Then $N = 6M$ gives a bound on the number of switchings. \square

Theorem 2 can readily be combined with Bressan's results to include the case of controlled stability.

COROLLARY 3. *Suppose Σ is as in Theorem 2 except for an equilibrium point p , $f(p) = 0$, where the triples $(g, [f, g], [f + g, [f, g]])$ and $(g, [f, g], [f - g, [f, g]])$ consist of independent vectors. Then time-optimal trajectories steering the system into p are finite concatenations of bang and singular arcs with a bound on the number of switchings which only depends on a compact set in which the trajectories lie.*

Proof. It was proved by Bressan [2] that there exists a neighborhood U of p such that time-optimal trajectories steering points in U to p are bang-bang with at most two switchings or are concatenations of a bang arc, followed by a singular arc and one more bang arc. Let R be the set of points that can be steered into p . Deleting an open neighborhood V of p with $\text{Clos } V \subseteq U$ from R , our analysis applies on $R \setminus V$ and trajectories steering points $p \in R$ into V are finite concatenations of bang and singular arcs with a bound on the number of switchings. Since the set of points which can be steered to p within a given time is compact, the corollary follows. \square

If Σ is an analytic system which has the properties mentioned in Corollary 3, then this implies the *existence of a piecewise analytic optimal feedback solution* for the problem of stabilizing the equilibrium time-optimally. This is an immediate corollary of Theorem 6 in [12]. (We shall show in § 3 that the singular control is given as $u = (\alpha + \beta)/(\alpha - \beta)$ where f, g and $[f, g]$ are independent, and that this is not an admissible control if α and β have the same sign. Near the equilibrium it is easy to see, as shown in [2], that the assumption of independence of $g, [f, g], [f \pm g, [f, g]]$ implies that singular controls can be written as analytic feedback as well. Therefore hypothesis (B3) of [12] holds; the nontrivial property is the one proved in Corollary 3.)

Let us close this discussion of our results with some brief remarks about the generic three-dimensional case. Our results cover a large portion of the generic cases, but there still is some work left to clear up the complete generic picture. We give affirmative answers to the key questions within our assumptions: time-optimal trajectories are finite concatenations of bang and singular arcs with a bound on the number of switchings, the Fuller phenomenon is not possible within the class \mathcal{A} of [10], more general, infinite concatenations of bang and singular arcs are not possible within the class \mathcal{B} considered in this paper. To answer these questions for a generic system in \mathbb{R}^3 , the major open task is to analyze nonequilibrium cases where $f(p)$, $g(p)$ and $[f, g](p)$ are dependent. Some of these cases are covered by Sussmann's nonlinear bang-bang theorem [13] (see the remark above). Recent work of Krener and Schättler [6] about the geometric structure of small-time reachable sets in low dimensions unifies Bressan's result with our results for the codimension-0 case, which give the same conclusion under different conditions and which were proved with very different techniques. They are in fact special instances of a more general result which holds only assuming the independence of $(g, [f, g], [f + g, [f, g]])$ and $(g, [f, g], [f - g, [f, g]])$ at p , but making no hypothesis about f . This technique, however, is restricted to nondegenerate (= codimension-0) cases and so many generic questions are still

unsettled and require further investigations. In particular, it is not known yet whether infinite concatenations of bang and singular arcs, the Fuller phenomenon, or an even more pathological behavior is possible for a generic three-dimensional system.

The local structure results do not simplify in the case of an analytic system. But the strong analyticity property has implications of a global nature. Basically, for an arbitrary analytic system the local structure result for the codimension-0 case holds away from an analytic set of positive codimension. But properties due to analyticity are the topic of a separate paper [11].

3. Singular arcs. Necessary conditions for optimality at singular junctions. Singular controls are uniquely determined due to our independence assumption on f , g , and $[f, g]$. If u is singular on an open interval I , we have $\phi_{1\cdot}(t) = \dot{\phi}_{1\cdot}(t) = \ddot{\phi}_{1\cdot}(t) = 0$ on I . So $\langle \lambda, g \rangle$ and $\langle \lambda, [f, g] \rangle$ vanish and

$$0 = \ddot{\phi}_{1\cdot}(t) = \frac{1}{4} \langle \lambda(t), [X + Y, [X, Y]](x(t)) - u(t) \cdot [X - Y, [X, Y]](x(t)) \rangle$$

is equivalent to

$$0 = \langle \lambda(t), \{(\alpha + \beta)(x(t)) - u(t) \cdot (\alpha - \beta)(x(t))\} \cdot f(x(t)) \rangle.$$

It follows from (1.4) that $\langle \lambda, f \rangle = -\lambda_0 \leq 0$. Since λ is nontrivial, $\langle \lambda, f \rangle$ cannot vanish and so λ_0 is positive for a singular extremal. Hence

$$(3.1) \quad u(t) = \frac{\alpha(x(t)) + \beta(x(t))}{\alpha(x(t)) - \beta(x(t))}.$$

LEMMA 1. *There does not pass a singular arc through a point q where $\alpha(q)\beta(q) > 0$. If $\alpha(q)\beta(q) < 0$, then there exists a unique singular extremal Γ_q passing through q . It is the integral curve of the smooth vector field*

$$(3.2) \quad S = \frac{\alpha Y - \beta X}{\alpha - \beta}$$

which is well defined near q .

Proof. If $\alpha(q)$ and $\beta(q)$ have the same sign, then (3.1) defines a number which is greater than one in absolute value, and hence not an admissible control. If α and β have opposite signs, then $S = f + (\alpha + \beta)/(\alpha - \beta)g = (\alpha Y - \beta X)/(\alpha - \beta)$ is well defined near q and the integral curves of S are singular arcs of Σ . Conversely, singular arcs are integral curves of S . \square

However, not all of these extremals Γ_q are time-optimal. They must also satisfy the Legendre–Clebsch condition [5]: If $u(\cdot)$ is a singular control of order k , which takes values in the interior of the control set, then a necessary condition for time optimality of $u(\cdot)$ is that

$$(-1)^k \frac{\partial}{\partial u} \frac{d^{2k}}{dt^{2k}} \frac{\partial}{\partial u} H(\lambda(t), x(t), u(t)) \leq 0$$

where H is the Hamiltonian of the problem. In our case $k=1$ and $H(\lambda, x, u) = \langle \lambda, f(x) + ug(x) \rangle$. So $H_u = \langle \lambda, g \rangle = \phi_{1\cdot}$, and hence $\ddot{H}_u = \langle \lambda, [f + ug, [f, g]] \rangle$. Therefore the Legendre–Clebsch condition states $\langle \lambda(t), [g, [f, g]](x(t)) \rangle \leq 0$. In our notation this is equivalent to

$$(3.3) \quad \alpha(x(t)) - \beta(x(t)) \leq 0 \quad \text{along } S.$$

COROLLARY. *If $\alpha(q) > 0$ and $\beta(q) < 0$, then Γ_q is not time-optimal near q .*

Our aim is to analyze the structure of time-optimal controls locally. One major issue is the question of optimality for finite concatenations of bang and singular arcs. Now we will develop necessary conditions for optimality of concatenations of this

type. Before we do this, we establish our terminology. Let $\Gamma: [0, T] \rightarrow \mathbb{R}^3$ be an extremal. We call $t \in (0, T)$ a *switching time* if $\phi_\Gamma(t) = 0$, but ϕ_Γ does not vanish identically on a neighborhood of t (i.e., times where the control is singular are not considered to be switching times). The point $q = x(t)$ is then called a *junction*; it is called a *singular junction* if also $\dot{\phi}_\Gamma(t) = 0$. This happens at concatenations of bang and singular arcs, but also at switching points which are limits of other switching points. Notice that λ_0 is positive for any extremal Γ that has a singular junction. We denote singular junctions by $*$, whereas a dot \cdot stands for any junction.

We will now examine necessary conditions for optimality at singular junctions and we start with well-known junction conditions. Suppose we have a $*X$ -junction at \tilde{q} at time \tilde{t} . Since $\phi_\Gamma(t) > 0$ for t shortly after \tilde{t} , we must have $\dot{\phi}_\Gamma(\tilde{t}+) \geq 0$. Now

$$\begin{aligned} \dot{\phi}_\Gamma(t) = & \alpha(x(t)) \langle \lambda(t), f(x(t)) \rangle + (a_2 - a_1)(x(t)) \langle \lambda(t), g(x(t)) \rangle \\ & + a_3(x(t)) \langle \lambda(t), [X, Y](x(t)) \rangle \end{aligned}$$

and so $\dot{\phi}_\Gamma(\tilde{t}+) = \alpha(\tilde{q}) \cdot (-\lambda_0)$. Since $\lambda_0 > 0$, we get $\alpha(\tilde{q}) \leq 0$. Similar considerations establish the following result.

LEMMA 2. *Necessary conditions for time optimality at singular junctions are*

- (i) α is nonpositive at singular junctions with X ;
- (ii) β is nonnegative at singular junctions with Y .

Additional necessary conditions have to hold if a singular junction is to be followed by another switching point. Assume for instance, we have a trajectory of the form $*X\cdot$. Call the two switching points p_0 and p_1 and let τ_1 be the time along the X -arc. Let $\tilde{\lambda}$ be the value of the adjoint vector corresponding to the singular junction. Then we have $\langle \tilde{\lambda}, g(p_0) \rangle = \langle \tilde{\lambda}, [f, g](p_0) \rangle = 0$, and if we transport the vector g from p_1 back to p_0 , we also have $\langle \tilde{\lambda}, e^{\tau_1 \text{ad} X} g(p_1) \rangle = 0$. Since $\tilde{\lambda} \neq 0$, these three vectors are dependent. (We say p_0 and p_1 are singular conjugate points (cf. [11]).) Analytically this yields

$$\begin{aligned} 0 &= g(p_0) \wedge [X, Y](p_0) \wedge e^{\tau_1 \text{ad} X} (Y - X)(p_1) \\ &= g(p_0) \wedge [X, Y](p_0) \wedge 2g(p_0) + \tau_1 [X, Y](p_0) + \frac{1}{2} \tau_1^2 [X, [X, Y]](p_0) \\ &\quad + \frac{1}{3!} \tau_1^3 \text{ad}^3 X(Y)(p_0) + \frac{1}{4!} \tau_1^4 \text{ad}^4 X(Y)(p_0) \\ &\quad + \frac{1}{5!} \tau_1^5 \text{ad}^5 X(Y)(p_0) + O(\tau_1^6). \end{aligned}$$

We now express the higher-order brackets as linear combinations of f , g , and $[X, Y]$. Since f , g , and $[X, Y]$ are independent, the coefficient at f of the last term must vanish. This gives the following *singular conjugate point relation*:

$$(3.4) \quad 0 = \alpha(p_0) + \frac{1}{3} \gamma(p_0) \tau_1 + \frac{1}{12} \mu(p_0) \tau_1^2 + \frac{1}{60} \kappa(p_0) \tau_1^3 + O(\tau_1^4).$$

For a $\cdot X*$ -concatenation only the time gets reversed and we have

$$(3.5) \quad 0 = \alpha(p_1) - \frac{1}{3} \gamma(p_1) \tau_1 + \frac{1}{12} \mu(p_1) \tau_1^2 - \frac{1}{60} \kappa(p_1) \tau_1^3 + O(\tau_1^4).$$

LEMMA 3. *If $\gamma(p) > 0$ (respectively, $\gamma(p) < 0$), then there exists a neighborhood U of p such that $\cdot X*$ - (respectively, $*X\cdot$ -) trajectories that lie in U are not time-optimal. In particular, no $*X*$ -junctions are time-optimal near p in case $\gamma(p) \neq 0$.*

Proof. Since f and g are independent, it is always possible to choose a neighborhood U of p so small that any trajectory of the system has to leave U within an a priori given time T . Therefore, if $\gamma(p) \neq 0$, we can assume that any value of $|\gamma|$ on U

dominates the time T along any trajectory that lies in U . Then we have for $\gamma(p) < 0$ that

$$\alpha(p_0) + \frac{1}{3}(\gamma(p_0) + O(\tau_1))\tau_1 \leq 0 + \frac{1}{6}\gamma(p)\tau_1 < 0$$

and so (3.4) cannot hold. Hence $*X$ -concatenations are not time-optimal near p in this case. Similarly (3.5) excludes $\cdot X$ -junctions if $\gamma(p) > 0$. \square

The conditions for $*Y$ - (and $\cdot Y$ -) concatenations follow by an analogous computation or simply by an application of the input symmetry which interchanges X and Y . If we call the time along Y τ_0 , then we have

$$(3.6) \quad 0 = \beta + \frac{1}{3}\eta\tau_0 + O(\tau_0^2) \quad (0 = \beta - \frac{1}{3}\eta\tau_0 + O(\tau_0^2))$$

where all the functions are evaluated at the singular junction. This implies the following lemma.

LEMMA 4. *If $\eta(p) > 0$ (respectively, $\eta(p) < 0$), then there exists a neighborhood U of p such that $*Y$ - (respectively, $\cdot Y$ -) trajectories that lie in U are not time-optimal. In particular, no $*Y$ -junctions are time-optimal near p if $\eta(p) \neq 0$.*

To deal with more degenerate cases—without loss of generality we may restrict ourselves to the case when $\gamma(p)$ vanishes—we first need some elementary observations about convexity properties of the switching function.

LEMMA 5. *There exists a neighborhood U of p with the following property: whenever Γ is an extremal steering a point $q_1 \in U$ to $q_2 \in U$ in U , and Γ is not bang-bang with at most one switching, then the following monotonicity relations hold.*

(i) *Along X -trajectories:*

$$(3.7) \quad \begin{aligned} \ddot{\phi}_\Gamma &\text{ has constant sign equal to } -\operatorname{sgn} \alpha(p) \text{ if } \alpha(p) \neq 0, \\ \phi_\Gamma^{(3)} &\text{ has constant sign equal to } -\operatorname{sgn} \gamma(p) \text{ if } \gamma(p) \neq 0, \\ \phi_\Gamma^{(4)} &\text{ has constant sign equal to } -\operatorname{sgn} \mu(p) \text{ if } \mu(p) \neq 0, \\ \phi_\Gamma^{(5)} &\text{ has constant sign equal to } -\operatorname{sgn} \kappa(p) \text{ if } \kappa(p) \neq 0. \end{aligned}$$

(ii) *Along Y -trajectories:*

$$(3.8) \quad \begin{aligned} \ddot{\phi}_\Gamma &\text{ has constant sign equal to } -\operatorname{sgn} \beta(p) \text{ if } \beta(p) \neq 0, \\ \phi_\Gamma^{(3)} &\text{ has constant sign equal to } -\operatorname{sgn} \eta(p) \text{ if } \eta(p) \neq 0. \end{aligned}$$

Proof. We choose coordinates (x_1, x_2, x_3) near p defined by

$$(x_1, x_2, x_3) \mapsto p \exp(-x_3[f, g]) \exp(x_2 f) \exp(x_1 g).$$

Here $\exp(\cdot Z)$ denotes the flow of a vector field Z and we let the diffeomorphisms act on the right. In these coordinates

$$f(x) = \begin{pmatrix} 0 \\ 1 \\ x_1 \end{pmatrix} + x_1 \cdot h(x), \quad g(x) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

where h is a smooth vector field near 0 with $h(0) = 0$. The adjoint equations are

$$\begin{aligned} \dot{\lambda}_1 &= -\lambda_3 - \sum_{j=1}^3 \lambda_j \left(h_j(x) + x_1 \frac{\partial h_j}{\partial x_1}(x) \right), \\ \dot{\lambda}_2 &= -x_1 \left(\sum_{j=1}^3 \lambda_j \frac{\partial h_j}{\partial x_2}(x) \right), \\ \dot{\lambda}_3 &= -x_1 \left(\sum_{j=1}^3 \lambda_j \frac{\partial h_j}{\partial x_3}(x) \right), \end{aligned}$$

and the switching function is just λ_1 . We normalize λ so that $\|\lambda(0)\|_1 = 1$. We want to show that the only interesting case is when we have approximately $\lambda_1(0) \approx 0$, $\lambda_2(0) \approx -1$ and $\lambda_3(0) \approx 0$.

We express all brackets of X and Y of orders 3, 4, 5, and 6 as linear combinations of f , g and $[X, Y]$ as in (2.1). On a small neighborhood V of p let M be an upper bound for the absolute values of all the coefficients that come up in this way. Choose ε , $0 < \varepsilon < \frac{1}{8}$, such that, whenever ϕ is a coefficient of f in such an expression with $\phi(p) \neq 0$, then we have $24M\varepsilon < |\phi(p)|$. We furthermore assume that $|\phi(q)| > \frac{2}{3}|\phi(p)|$ for all $q \in V$.

We first show that, if we choose U small enough, then any extremal in U for which $|\lambda_2(0)| \leq 1 - \varepsilon$ is bang-bang with at most one switching. Since f and g are independent, it suffices to show that there exists a small time T such that any extremal of this type with total time $\leq T$ is bang-bang with at most one switching. Gronwall's inequality gives an a priori bound $\|\lambda(t) - \lambda(0)\|_\infty \leq CT$ where the constant C depends only on V . Near p we can also assume that the components of h and the coordinates x_i are of order ε . Then we have for sufficiently small T

$$|\dot{\lambda}_1(t)| \leq |\lambda_3(0)| + C\varepsilon, \quad |\dot{\lambda}_i(t)| \leq C\varepsilon, \quad i = 2, 3.$$

Note that $|\lambda_1(0)| + |\lambda_3(0)| > \varepsilon$. If $|\lambda_3(0)| < \varepsilon/4$, then we get

$$|\lambda_1(t)| \geq |\lambda_1(0)| - |\lambda_1(t) - \lambda_1(0)| \geq \frac{\varepsilon}{2} - CT \geq \frac{\varepsilon}{4}$$

for small T , so u is constant. If $|\lambda_3(0)| \geq \varepsilon/4$, then

$$\begin{aligned} |\dot{\lambda}_1(t)| &\geq |\lambda_3(t)| - \sum_{j=1}^3 \left| \lambda_j(t) \left(h_j(x(t)) + x_1(t) \frac{\partial h_j}{\partial x_1}(x(t)) \right) \right| \\ &\geq |\lambda_3(0)| - |\lambda_3(t) - \lambda_3(0)| - \frac{\varepsilon}{8} \geq \frac{\varepsilon}{8} - CT \geq \frac{\varepsilon}{16} \end{aligned}$$

for small T . So in this case the derivative of the switching function has constant sign, i.e., u is bang-bang with at most one switching. This proves our assertion above.

We now consider extremals which satisfy $|\lambda_2(0)| > 1 - \varepsilon$. By shrinking U and T further, if necessary, we can assume that for any such multiplier the following inequalities hold for $t \in [0, T]$ and $q \in U$:

$$|\langle \lambda(t), g(q) \rangle| \leq 2\varepsilon, \quad |\langle \lambda(t), f(q) \rangle| \geq \frac{3}{4}, \quad |\langle \lambda(t), [f, g](q) \rangle| \leq 2\varepsilon.$$

If $\lambda_0 = 0$, then we have $\langle \lambda(t), f(x(t)) \rangle = 0$ at switching times. Therefore extremals with $\lambda_0 = 0$ correspond to constant controls. For a positive λ_0 we get from (1.4) $\langle \lambda(t), f(q) \rangle \leq -\frac{3}{4}$. We can now prove the monotonicity relations (3.7) and (3.8). For instance, assume $\beta(p) \neq 0$. Then we have along Y -trajectories

$$\begin{aligned} 2\ddot{\phi}_1(t) &= \langle \lambda(t), [Y, [X, Y]](x(t)) \rangle \\ &= \beta(x(t)) \langle \lambda(t), f(x(t)) \rangle + (b_2 - b_1)(x(t)) \cdot \langle \lambda(t), g(x(t)) \rangle \\ &\quad + b_3(x(t)) \cdot \langle \lambda(t), [X, Y](x(t)) \rangle \end{aligned}$$

and so by construction

$$\begin{aligned} 2|\ddot{\phi}_\Gamma(t)| &> \frac{4}{3}|\beta(p)|^{\frac{3}{4}} - 8M\varepsilon - 4M\varepsilon \\ &= |\beta(p)| - 12M\varepsilon \\ &> \frac{1}{2}|\beta(p)| > 0. \end{aligned}$$

Hence $\ddot{\phi}_\Gamma$ has constant sign. Since $\langle \lambda, f \rangle < 0$ it follows that the sign is $-\text{sgn } \beta(p)$. This proves the first relation in (ii). The other relations follow exactly like this one. \square

We now return to the analysis of singular junctions in more degenerate cases. The following lemmas are elementary, but useful observations.

LEMMA 6. *Suppose $\gamma(p) = 0$ and $\mu(p) > 0$. There exists a neighborhood U of p such that $*X*$ -trajectories that lie in U are not time-optimal. \square*

Proof. We choose U so small that μ is positive on U and such that the monotonicity relations (3.7) hold. So $\phi_\Gamma^{(3)}$ is decreasing along X -trajectories. Let $[t_1, t_2]$ be the interval where X is used, with singular junctions at t_1 and t_2 . We have $\phi_\Gamma(t_1+) \geq 0$ and $\phi_\Gamma(t_2-) \leq 0$; ϕ_Γ has a maximum on $[t_1, t_2]$, say at $\tilde{t} \in (t_1, t_2)$, and there $\ddot{\phi}_\Gamma(\tilde{t}) \leq 0$. So ϕ_Γ cannot increase everywhere in (t_1, \tilde{t}) and it cannot decrease everywhere in (\tilde{t}, t_2) . Hence there exist a time less than \tilde{t} where $\ddot{\phi}_\Gamma$ is negative and a time greater than \tilde{t} where $\ddot{\phi}_\Gamma$ is positive. But $\ddot{\phi}_\Gamma$ decreases on (t_1, t_2) . This is a contradiction.

LEMMA 7. *Suppose $\gamma(p) = 0$ and $\mu(p) < 0$. A necessary condition for time optimality of a $*X*$ - (respectively, $*X\cdot$ -) trajectory near p is that γ be negative (respectively positive) at the singular junction. In particular, if γ decreases along a singular arc, then $*XSX\cdot$ -trajectories are not time-optimal.*

Proof. We choose U so small that $\phi_\Gamma^{(4)}$ is positive along X -trajectories that lie in U . Without loss of generality we consider a $*X*$ -junction. Let $[t_1, t_2]$ be the X -interval and let $\tilde{t} \in (t_1, t_2)$ be a time where the switching function has a maximum on $[t_1, t_2]$. Then we have $\ddot{\phi}_\Gamma(\tilde{t}) \leq 0$ and $\ddot{\phi}_\Gamma(t_2-) \geq 0$. Since ϕ_Γ does not vanish identically on $[t_1, t_2]$, $\phi_\Gamma^{(3)}$ is positive somewhere in (\tilde{t}, t_2) , and since $\phi_\Gamma^{(3)}$ increases, $\phi_\Gamma^{(3)}(t_2-) > 0$ follows. But $\phi_\Gamma^{(3)}(t_2-) = -\lambda_0 \cdot \gamma(x(t_2))$ and λ_0 is positive. Hence $\gamma(x(t_2)) < 0$ follows. Analogously, $\gamma(x(t_1)) > 0$ follows in case of a $*X\cdot$ -concatenation. The last observation is a consequence of these conditions. \square

LEMMA 8. *Suppose $\gamma(p) = 0$, $\mu(p) = 0$ and $\kappa(p) > 0$. There exists a neighborhood U of p such that any time-optimal $*X*$ -trajectory (with switching points p_1 and p_2) that lies in U has to satisfy $\gamma(p_1) > 0$ and $\mu(p_1) < 0$.*

Proof. We choose U so small that $\phi_\Gamma^{(5)}$ is negative on U along X -trajectories. If we call the X_1 -interval $[t_1, t_2]$, then it follows that $\phi_\Gamma^{(3)}$ is concave over (t_1, t_2) . As above, we have as necessary conditions $\ddot{\phi}_\Gamma(t_1+) \geq 0$, $\ddot{\phi}_\Gamma(t_2-) \geq 0$ and $\ddot{\phi}_\Gamma(\tilde{t}) \leq 0$, where \tilde{t} is a time where ϕ_Γ has a maximum over $[t_1, t_2]$. This implies that $\phi_\Gamma^{(3)}$ is negative somewhere in (t_1, \tilde{t}) and positive somewhere in (\tilde{t}, t_2) . Hence we have $\phi_\Gamma^{(3)}(t_1+) < 0$ and $\phi_\Gamma^{(4)}(t_1+) > 0$. Now $\phi_\Gamma^{(3)}(t_1+) = -\lambda_0 \gamma(p_1)$ and $\phi_\Gamma^{(4)}(t_1+) = -\lambda_0 \cdot \mu(p_1)$. Since $\lambda_0 > 0$, the lemma follows. \square

4. The bang-bang cases. We now start to analyze the local structure of time-optimal trajectories for a generic system in \mathcal{B} . As shown in § 2, we can restrict ourselves to the case when $\beta(p) \neq 0$. Here we consider those systems for which time-optimal controls are bang-bang. These are the cases where $\beta(p)$ is negative.

It is clear that singular arcs are not optimal near p if $\beta(p) < 0$. For, if also $\alpha(p) < 0$, then no singular arcs exist near p , and if $\alpha(p) > 0$, then the Legendre–Clebsch condition is violated. For bang-bang trajectories we proved the following result in [10] (cf. Propositions 5.1 and 5.2).

PROPOSITION 3. *Let $\beta(p) < 0$. There exists a neighborhood U of p such that time-optimal bang-bang trajectories that lie in U are at most concatenations of the following forms:*

- (0) XYX if $\alpha(p) < 0$, BBB if $\alpha(p) > 0$;
- (i) $XYXY$ if $\alpha(p) = 0$, $\gamma(p) > 0$, $YXYX$ if $\alpha(p) = 0$, $\gamma(p) < 0$;
- (ii) $XYXYX$ if $\alpha(p) = 0$, $\gamma(p) = 0$, $\mu(p) < 0$, $YXYXY$ if $\alpha(p) = 0$, $\gamma(p) = 0$, $\mu(p) < 0$;
- (iii) $XYXYXY$ if $\alpha(p) = 0$, $\gamma(p) = 0$, $\mu(p) = 0$, $\kappa(p) > 0$, $YXYXYX$ if $\alpha(p) = 0$, $\gamma(p) = 0$, $\mu(p) = 0$, $\kappa(p) < 0$. \square

Remark. Actually, we proved a slightly more general result than this one in the sense that we excluded the optimality of concatenations where the first or last switchings did not need to be bang-bang, but could be arbitrary junctions. For instance, we excluded the optimality of $\cdot XYX \cdot$ -concatenations in the case $\alpha(p) = 0$, $\gamma(p) = 0$ and $\mu(p) < 0$. For bang-bang trajectories this yields the result above; below we will also use the more general version.

Proposition 1 will follow immediately from this result once we know that time-optimal trajectories are bang-bang. There is, however, a minor technical problem here which is due to the fact that an optimal control is only known to be measurable. In particular, we do not know a priori that optimal controls are concatenations of bang and singular controls. We take care of this now.

We call a switching time t which is a limit of switching times t_n , $n \in \mathbb{N}$, $t_n \neq t$, a *nonisolated switching time*. For a trajectory Γ we denote the set of its nonisolated switching times by $N = N_\Gamma$. Basic properties of N are:

- (i) If $t \in N$, then $\phi_\Gamma(t) = 0$ and $\dot{\phi}_\Gamma(t) = 0$;
- (ii) N is closed and nowhere dense.

For all the cases considered in Proposition 1 we know in addition that a bang-bang trajectory with more than five switchings is not time-optimal. This implies that no point in N can be isolated if Γ is time-optimal. Therefore, if N is not empty, then N is a perfect set (i.e., it is closed and every point is a limitpoint of points in N). If N is empty, then Γ is a finite concatenation of bang and singular arcs, and so is a bang-bang trajectory with finitely many switchings in our case. Hence Proposition 1 will follow from Proposition 3 and the following lemma.

LEMMA 9. *Let U be a sufficiently small neighborhood of p and let Γ be a time-optimal trajectory that lies in U . Then, in all the cases considered in Proposition 1, we have $N = N_\Gamma = \emptyset$.*

Proof. We assume Γ is a time-optimal trajectory defined over an interval $[0, T]$ and that N is not empty. Then there exist times $t_1, t_2 \in N$ —without loss of generality we may assume $0 < t_1$ and $t_2 < T$ —such that Γ is bang-bang over (t_1, t_2) . At t_1 and t_2 we have singular junctions. Since $\beta(p) \neq 0$, the singular conjugate point relation (3.6) cannot hold for small time T , i.e., on a sufficiently small neighborhood of p . Equivalently, we could say that ϕ_Γ is strictly convex along Y -trajectories by (3.8) and so at ends of Y -arcs a switching to X must occur. Similarly, no singular junctions with X are possible near p if $\alpha(p) \neq 0$ by (3.4) and (3.5). So N can only be nonempty if $\alpha(p) = 0$.

We first consider the case when the bang-bang trajectory over (t_1, t_2) contains a Y -arc. Since at the endpoints a switching to X occurs, Γ contains then both a $\cdot XYX \cdot$ - and a $\cdot XYX \cdot$ -concatenation over $[t_1, t_2]$. We claim that this part of Γ is not time-optimal. In the cases $\gamma(p) \neq 0$ and $\gamma(p) = 0$, $\mu(p) < 0$ this follows from Proposition 3, (i) and (ii), and the subsequent remark. To treat the other cases we have to recall a couple of necessary conditions for optimality of bang-bang trajectories which we proved in [10]: we consider a $\cdot XYX \cdot$ -concatenation with switching points p_0, p_1, p_2 ,

p_3 and times τ_1, τ_2, τ_3 along X, Y , and X , respectively. A necessary condition for time optimality of the $\cdot XYX$ -trajectory with first switching point p_0 is that

$$(4.1) \quad 0 \leq \frac{1}{3}\beta(p)\tau_2 + \frac{1}{3}\tau_1^2(1 + O(1))\{-\gamma(p_0) - \frac{3}{4}\tilde{\mu}(p_0)\tau_1 - \frac{3}{10}\tilde{\kappa}(p_0)\tau_1^2\}$$

where $O(1)$ stands for times which are at least linear in the switching times τ_1 and τ_2 , and $\tilde{\phi}$ stands for expressions of the form $\phi(1 + O(1))$ (cf. [10, (5.4)] and note that the specific form of the first junction is irrelevant for the condition obtained there. It holds equally for $YXYX$ - and for $*XYX$ -concatenations). Furthermore, a necessary condition for time optimality of a $\cdot XYX$ -concatenation is that

$$(4.2) \quad 0 \leq \mu(p_2) + \frac{1}{5}\tilde{\kappa}(p_2)(\tau_3 - \tau_1) + O(\tau_1^2, \tau_3^2)$$

(cf. [10, (5.5)]).

We now assume that $\gamma(p) = 0$ and $\mu(p) > 0$. Since there is a singular junction at p_0 , in addition we have (3.4) as a necessary condition and since $\alpha(p_0) \leq 0$ (Lemma 2) we get

$$0 \leq \gamma(p_0) + \frac{1}{4}\mu(p_0)\tau_1 + \frac{1}{20}\tilde{\kappa}(p_0)\tau_1^2.$$

We use this in (4.1) to conclude that

$$0 \leq \frac{1}{3}\beta(p)\tau_2 + \frac{1}{3}\tau_1^2(1 + O(1))\{-\frac{1}{2}\tilde{\mu}(p_0)\tau_1 - \frac{1}{4}\tilde{\kappa}(p_0)\tau_1^2\}.$$

Next we transport the expression in $\{\cdot \cdot \cdot\}$ from p_0 to p_2 by Taylor's theorem. The terms with a factor τ_2 that come up in this process are all dominated by $\beta(p)\tau_2 < 0$ and so we obtain as necessary condition for optimality that

$$(4.3) \quad \mu(p_2) - \frac{1}{2}\tilde{\kappa}(p_2)\tau_1 \leq 0.$$

It is clear from this that $*XYX$ -concatenations are not time-optimal near p if $\mu(p) > 0$. If $\gamma(p) = 0$ and $\mu(p) = 0$, we combine (4.2) and (4.3) to get the following necessary condition for a $*XYX$ -concatenation:

$$0 \leq \frac{1}{5}\tilde{\kappa}(p_2)(\tau_3 - \tau_1) + \frac{1}{2}\tilde{\kappa}(p_2)\tau_1 = \frac{1}{10}\tilde{\kappa}(p_2)(3\tau_1 + 2\tau_3).$$

This excludes $*XYX$ -trajectories near p if $\kappa(p) < 0$. In the case $\kappa(p) > 0$ a time reversal argument shows that $\cdot XYX$ -concatenations cannot be time-optimal near p : Time reversal is the input symmetry which changes X into $-X$ and Y into $-Y$. The functions α, β, μ are fixed under this operation, but γ and κ change into $-\gamma$ and $-\kappa$. Therefore, if $*XYX$ -concatenations are not time-optimal in the case $\kappa(p) > 0$, it follows that $\cdot XYX$ -trajectories are not optimal if $\kappa(p) < 0$ (cf. [10, § 3.2]).

This proves that Γ does not contain a Y -arc over (t_1, t_2) . So we can assume that Γ is an X -trajectory on (t_1, t_2) . Concatenations of the type $*X*$ are not time-optimal if $\gamma(p) \neq 0$ (Lemma 3) or if $\gamma(p) = 0$ and $\mu(p) > 0$ (Lemma 6). In the remaining cases— $\mu(p) < 0$ and $\mu(p) = 0, \kappa(p) \neq 0$ —we cannot quite exclude $*X*$ -concatenations. But since N is a perfect set, there are $*X*$ -trajectories arbitrarily close to each other. We will show that such a structure cannot be time-optimal.

If $\mu(p) < 0$, we have by Lemma 7, as necessary conditions for time optimality of a $*X*$ -concatenation, $\gamma(p_1) > 0$ and $\gamma(p_2) < 0$, where p_1 and p_2 are the consecutive switching points. Since N is a perfect set, there exist times $t^* \in N$ arbitrarily close to t_2 ; in particular, we can choose $t^* - t_2$ so small that $\gamma(x(t))$ is negative on $[t_2, t^*]$. Since N does not contain an interval, there exist times $\tilde{t}_1, \tilde{t}_2 \in N, (\tilde{t}_1, \tilde{t}_2) \subseteq [t_2, t^*]$ such that Γ is bang-bang on $(\tilde{t}_1, \tilde{t}_2)$. Without loss of generality we may assume that this is another $*X*$ -concatenation since the optimality of other concatenations has been

excluded already. But then $\gamma(x(t_1)) < 0$, contradicting the necessary condition $\gamma(x(t_1)) > 0$ of Lemma 7.

In the case $\mu(p) = 0$, $\kappa(p) \neq 0$ we can restrict ourselves to the subcase $\kappa(p) > 0$ since the result will then follow for $\kappa(p) < 0$ as well by time reversal. By Lemma 8 we have as necessary conditions $\gamma(p_1) > 0$ and $\mu(p_1) < 0$. As above, we conclude that there exists another $*X*$ -concatenation over an interval $(\tilde{t}_1, \tilde{t}_2)$ arbitrarily close to the left of t_1 . By choosing $t_1 - \tilde{t}_1$ small enough we get $\gamma(x(t)) > 0$ and $\mu(x(t)) < 0$ on $[\tilde{t}_1, t_1]$. So we have $\phi_1^{(3)}(t_1+) < 0$ and $\phi_1^{(3)}(t_2-) < 0$. Also, $\phi_1^{(3)}$ is positive somewhere in $(\tilde{t}_1, \tilde{t}_2)$ and so $\phi_1^{(4)}$ has a zero in $(\tilde{t}_1, \tilde{t}_2)$. The monotonicity relation (3.7) implies that $\phi_1^{(4)}$ is strictly decreasing along X -trajectories— $\kappa(p) > 0$ —and so $\phi_1^{(4)}(t_2-) < 0$. This contradicts $\mu(\tilde{t}_2) < 0$.

So in all the cases considered in Proposition 1 we have contradicted $N \neq \emptyset$. This proves Proposition 1. \square

5. The singular cases. We now consider the cases when optimal singular arcs can exist and when bang-bang trajectories occur only in a trivial way. This happens for $\beta(p) > 0$. On a sufficiently small neighborhood of p the monotonicity relations (3.7) and (3.8) hold; in particular it follows that the switching function is concave (and negative) along Y -trajectories. Therefore Y -trajectories necessarily lie at the ends of the interval $[0, T]$ over which the extremal is defined. Here now the problem is to control concatenations of X and singular arcs.

In the case when $\alpha(p) \neq 0$ nonisolated switching times cannot exist since the singular conjugate point relations would be violated otherwise. If $\alpha(p) > 0$, then singular controls do not exist near p and so time-optimal trajectories can only be of the form YXY . If $\alpha(p) < 0$, then the switching function is in addition convex (and positive) along X -trajectories; so they also lie at the ends of the interval $[0, T]$. Now singular arcs can be optimal and therefore this allows for time-optimal trajectories which are of the form BSB . This proves case (0) in Proposition 2.

PROPOSITION 4. *Suppose $\alpha(p) = 0$. There exists a neighborhood U of p such that the following concatenations are not time-optimal in U in the cases specified:*

- (i) $\cdot X \cdot$ if $\gamma(p) > 0$, $*X \cdot$ if $\gamma(p) < 0$;
- (ii) $*X*$ if $\gamma(p) = 0$, $\mu(p) > 0$;
- (iii) $\cdot XSX \cdot$ if $\gamma(p) = 0$, $\mu(p) < 0$;
- (iv) $\cdot XSX*$ if $\gamma(p) = 0$, $\mu(p) = 0$, $\kappa(p) > 0$, $*XSX \cdot$ if $\gamma(p) = 0$, $\mu(p) = 0$, $\kappa(p) < 0$.

Proof. Statements (i) and (ii) were proved in Lemmas 3 and 6 and we just listed them for convenience. Also, (iii) will follow from Lemma 7 once we show that γ decreases along singular arcs. This is so since $S = X + (\alpha/\alpha - \beta)(Y - X)$, and therefore $L_S\gamma(p) = L_X\gamma(p) = \mu(p) < 0$. Hence $L_S\gamma$ is negative on a small neighborhood of p .

A new result is (iv). We may assume $\kappa(p) > 0$ since the result for the case $\kappa(p) < 0$ follows then by time reversal. By (3.7), $\phi_1^{(5)}$ is then negative on intervals where X is used. First let us establish notation. We consider a $\cdot XSX*$ -concatenation and call the switching times $t_0 < t_1 < t_2 < t_3$ and the switching points p_0, p_1, p_2, p_3 . Lemma 8 implies as necessary conditions $\gamma(p_2) > 0$ and $\mu(p_2) < 0$. We claim that it suffices to show that γ decreases along the singular arc between p_1 and p_2 , since, if this holds, we have $\gamma(p_1) > 0$, and therefore the singular conjugate point relation (3.5),

$$0 = \alpha(p_1) - \frac{1}{3}\gamma(p_1)\tau_1 + \frac{1}{12}\mu(p_1)\tau_1^2 - \frac{1}{60}\kappa(p_1)\tau_1^3 + O(\tau_1^4),$$

and Lemma 2 imply $\mu(p_1) > 0$. Also $L_S\mu(p) = L_X\mu(p) = \kappa(p) > 0$, and so μ increases along singular arcs near p . So we get $\mu(p_2) > 0$ as well and this contradicts the necessary condition $\mu(p_2) < 0$.

So it only remains to show that γ does indeed decrease along S over $[0, T]$. Note that

$$L_S \gamma = \frac{\alpha}{\alpha - \beta} L_Y \gamma - \frac{\beta}{\alpha - \beta} L_X \gamma = \frac{\alpha\omega - \beta\mu}{\alpha - \beta}.$$

Since $\alpha(p) = 0$, $\gamma(p) = 0$, and $\mu(p) = 0$, we have $L_S \alpha(p) = 0$ and so

$$L_S^2 \gamma(p) = L_S \mu(p) = L_X \mu(p) = \kappa(p) > 0.$$

Therefore γ is convex along singular arcs near p . In particular, $L_S \gamma(x(t))$ increases on $[t_1, t_2]$ and so it suffices to show that $L_S \gamma(p_2) < 0$. Since $\alpha - \beta$ is negative near p , we have to show that $\alpha(p_2)\omega(p_2) - \beta(p_2)\mu(p_2) > 0$.

If $\omega(p_2) \leq 0$, this follows from the necessary conditions $\alpha(p_2) \leq 0$ (Lemma 2) and $\mu(p_2) < 0$ (Lemma 8). The case $\omega(p_2) > 0$ needs more care. We use the singular conjugate point relation along the SX^* -concatenation (cf. (3.4)) to say

$$(5.1) \quad 0 = \alpha(p_2) + \frac{1}{3}\gamma(p_2)\tau + \frac{1}{12}\mu(p_2)\tau^2 + \frac{1}{60}\kappa(p_2)\tau^3 + O(\tau^4)$$

where $\tau = t_3 - t_2$. Since the junctions are singular at both ends we have a second singular conjugate point relation, namely also the vectors $g(p_2)$, $[X, Y](p_2)$ and $e^{\tau \text{ad}^X}([X, Y](p_3))$ are dependent (since $\dot{\phi}_1(t_3) = 0$ as well). Analytically this yields

$$\begin{aligned} 0 &= g(p_2) \wedge [X, Y](p_2) \wedge e^{\tau \text{ad}^X}([X, Y](p_3)) \\ &= g(p_2) \wedge [X, Y](p_2) \wedge \tau[X, [X, Y]](p_2) + \frac{1}{2}\tau^2 \text{ad}^3 X(Y)(p_2) \\ &\quad + \frac{1}{6}\tau^3 \text{ad}^4 X(Y)(p_2) + \frac{1}{24}\tau^4 \text{ad}^5 X(Y)(p_2) + O(\tau^5). \end{aligned}$$

If we express the higher-order brackets as linear combinations of f , g and $[X, Y]$, the coefficient at f must vanish. This yields

$$(5.2) \quad 0 = \alpha(p_2) + \frac{1}{2}\gamma(p_2)\tau + \frac{1}{6}\mu(p_2)\tau^2 + \frac{1}{24}\kappa(p_2)\tau^3 + O(\tau^4).$$

Now subtract twice (5.2) from three times (5.1) to eliminate γ . This gives

$$\alpha(p_2) = \frac{1}{12}\mu(p_2)\tau^2 + \frac{1}{30}\kappa(p_2)\tau^3 + O(\tau^4)$$

and so we have in small time $\alpha(p_2) \cong (1/12)\mu(p_2)\tau^2$. Hence, for $\omega(p_2) > 0$, we get

$$\alpha(p_2)\omega(p_2) - \beta(p_2)\mu(p_2) \cong \underbrace{\mu(p_2)}_{<0} \underbrace{(-\beta(p_2) + \frac{1}{12}\omega(p_2)\tau^2)}_{<0 \text{ for small } \tau} > 0.$$

This proves that γ decreases along S on $[t_1, t_2]$. \square

Proposition 2 will follow immediately from Proposition 4 once we know that time-optimal trajectories are concatenations of bang and singular arcs. We now prove this.

LEMMA 10. *Let U be a sufficiently small neighborhood of p and let Γ be a time-optimal trajectory that lies in U . Then, in all the cases considered in Proposition 2 we have that $N = N_\Gamma = \emptyset$.*

Proof. We may assume $\alpha(p) = 0$. Again we prove Lemma 10 by contradiction, i.e., we assume that $N_\Gamma \neq \emptyset$.

A consequence of Proposition 4 is that N is a perfect set—it has no isolated points. We first show that there exist two SXS^* -concatenations (where the singular arcs may be absent) arbitrarily close to each other. By omitting finitely many bang and singular arcs we may assume $O \in N$ and $T \in N$. Then no Y -arcs occur in Γ and since ϕ_1 does not vanish identically, it is positive somewhere, say at \tilde{t} . Then there exists an interval (t_1, t_2) , $t_1 \in N$, $t_2 \in N$, $\tilde{t} \in (t_1, t_2)$ such that Γ is a finite concatenation of X and

singular arcs on (t_1, t_2) , at least one of them being an X -arc. The optimality of an $*XSX*$ -concatenation is ruled out by Proposition 4, and so we may assume that Γ is of the form $*XSX*$ on (t_1, t_2) , allowing for the possibility that the singular arcs be absent. Furthermore, since N is a perfect set, t_2 is a limit point of times in N from above and so for every $\varepsilon > 0$ there exists a $\bar{t} \in N$, $0 < \bar{t} - t_2 < \varepsilon$. Now repeat the same construction over the interval (t_2, \bar{t}) . This establishes our claim. We call the second interval (t_3, t_4) .

In the cases $\gamma(p) \neq 0$ and $\gamma(p) = 0, \mu(p) > 0$ this already contradicts the optimality of Γ . For the other cases we can basically mimic the arguments used to prove (iii) and (iv) of Proposition 4 since it is possible to choose t_2 and t_3 arbitrarily close to each other. First, let us establish notation: let $(\tilde{t}_1, \tilde{t}_2)$ and $(\tilde{t}_3, \tilde{t}_4)$ be the X -intervals in (t_1, t_2) and (t_3, t_4) , respectively. If, for instance, there is no singular arc before the first X -trajectory, then $t_1 = \tilde{t}_1$; we call the corresponding switching points p_i and \tilde{p}_i , $i = 1, 2, 3, 4$ (see Fig. 1).

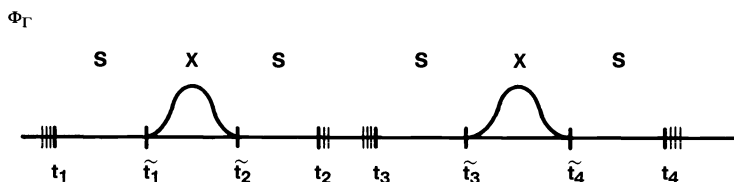


FIG. 1

Now consider the case $\gamma(p) = 0$ and $\mu(p) < 0$. By Lemma 7 we have as necessary conditions for optimality $\gamma(\tilde{p}_2) < 0$ and $\gamma(\tilde{p}_3) > 0$. In this case γ decreases along singular arcs near p , and so we have also $\gamma(p_2) < 0$ and $\gamma(p_3) > 0$. This cannot be since it is possible to choose t_3 arbitrarily close to t_2 . This is a contradiction.

Now suppose $\gamma(p) = 0$, $\mu(p) = 0$, and without loss of generality, assume that $\kappa(p) > 0$. Then by Lemma 8 we have as necessary conditions $\gamma(\tilde{p}_1) > 0$, $\mu(\tilde{p}_1) < 0$ and $\gamma(\tilde{p}_3) > 0$, $\mu(\tilde{p}_3) < 0$. We will be done if we can show that $\gamma(\tilde{p}_2) > 0$, since this will lead to the same contradiction as in the proof of Proposition 4(iv): The singular conjugate point relation at p_2 implies $\mu(\tilde{p}_2) > 0$ and since μ increases along singular arcs, and since we can choose t_3 arbitrarily close to t_2 , we then get $\mu(\tilde{p}_3) > 0$ as well, contradicting $\mu(\tilde{p}_3) < 0$. To show that $\gamma(\tilde{p}_2) > 0$, we first recall the following formula from the proof of Proposition 4:

$$L_S \gamma(x(\tilde{t}_3)) = \frac{\alpha\omega - \beta\mu}{\alpha - \beta} \Big|_{\tilde{p}_3} < 0.$$

Since γ is convex along S , we get $L_S \gamma(p_3) < 0$ and if we choose t_3 sufficiently close to t_2 also $L_S \gamma(p_2) < 0$. This implies that γ decreases on the intervals (\tilde{t}_2, t_2) and (t_3, \tilde{t}_3) . Therefore $\gamma(p_3) > 0$ and for t_3 sufficiently close to t_2 also $\gamma(p_2) > 0$. This then implies $\gamma(\tilde{p}_2) > 0$. This concludes the proof of Proposition 2. \square

6. The genericity of the conditions. In this section we show that the conditions of Propositions 1 and 2 and of Corollaries 1 and 2 are generic within the class \mathcal{B} . This is an almost immediate consequence of Thom's Transversality Theorem.

For the convenience of the reader we briefly recall some basic concepts and results from differential topology, following [4] in our exposition. In particular, all the proofs can be found there. Let M and N be smooth manifolds; let TM and TN be the tangent bundles and let $T_p M$ be the tangent space at a point $p \in M$. We denote the set of all

smooth mappings from M to N by $C^\infty(M, N)$. If $p \in M$ and f and g are smooth mappings with $f(p) = g(p) = q$, we say f has first order contact with g at p if $(Df)(p) = (Dg)(p)$ as maps from $T_p M$ to $T_q N$; f has k th order contact with g at p , $f \sim_k g$, if $Df: T_p M \rightarrow T_q N$ has $(k-1)$ st order contact with Dg at every point in $T_p M$. Let $J^k(M, N)_{(p,q)}$ denote the set of equivalence classes under " \sim_k at p " of mappings $f \in C^\infty(M, N)$ with $f(p) = q$ and let $J^k(M, N) := \bigcup_{(p,q) \in M \times N} J^k(M, N)_{(p,q)}$ be the disjoint union. Elements in $J^k(M, N)$ are called k -jets of mappings, p is the source and q is the target. It can be shown that $J^k(M, N)$ is itself a manifold and that it has a fiber bundle structure; $J^k(M, N)$ is called the k -jet bundle. For every $f \in C^\infty(M, N)$ there is a canonically defined mapping $j^k f: M \rightarrow J^k(M, N)$ which assigns to every $p \in M$ as image $j^k f(p)$ the equivalence class of f in $J^k(M, N)_{(p, f(p))}$. This map is called the k -jet of f . It is an invariant way to describe the Taylor expansion of f at p up to order k . If $f \in C^\infty(M, N)$, then $j^k f: M \rightarrow J^k(M, N)$ is smooth. Finally, let $j^k: C^\infty(M, N) \rightarrow C^\infty(M, J^k(M, N))$ be the map which assigns to $f \in C^\infty(M, N)$ its k -jet.

We now define the Whitney C^∞ topology. For $U \subseteq J^k(M, N)$ let $M(U) := \{f \in C^\infty(M, N) : j^k f(M) \subseteq U\}$. The family of sets $\{M(U)\}$, where U is an open subset of $J^k(M, N)$, form a basis for a topology on $C^\infty(M, N)$, called the Whitney C^k topology. Let O_k be the set of open subsets of $C^\infty(M, N)$ in the Whitney C^k topology. The Whitney C^∞ topology on $C^\infty(M, N)$ is the topology whose basis is $O = \bigcup_{k \in \mathbb{N}} O_k$. This is a well-defined basis since $O_k \subseteq O_l$ for $k \leq l$. It can be shown that $C^\infty(M, N)$ is a Baire space in the Whitney C^∞ topology. Therefore a countable intersection of open dense subsets (i.e., a residual subset) is dense. We also note that the map j^k is continuous in the Whitney C^∞ topology.

Let $f \in C^\infty(M, N)$, let W be an embedded submanifold of N , and let $x \in M$. We say f intersects W transversely at x if either $f(x) \notin W$ or $f(x) \in W$ and $T_{f(x)} N = T_{f(x)} W + (Df)(x)(T_x M)$; f intersects W transversely— $f \pitchfork W$ —if f intersects W transversely at every $x \in M$. Notice that, if the codimension of W is greater than the dimension of M , then $f \pitchfork W$ means $f(M) \cap W = \emptyset$.

THOM'S TRANSVERSALITY THEOREM. *Let M and N be smooth manifolds and let W be an embedded submanifold of $J^k(M, N)$. Then the set $T_W = \{f \in C^\infty(M, N) : j^k f \pitchfork W\}$ is a residual subset of $C^\infty(M, N)$ in the Whitney C^∞ topology; in particular, it is dense in $C^\infty(M, N)$. If W is closed, then T_W is also open.*

We now apply this to our situation. Recall that we identify Σ with the C^∞ map $\Sigma = (f, g): \mathbb{R}^3 \rightarrow \mathbb{R}^6$ and that we equip $C^\infty(\mathbb{R}^3, \mathbb{R}^6)$ with the Whitney C^∞ topology. Let $J^k := J^k(\mathbb{R}^3, \mathbb{R}^6)$. The sets

$$B_+ := \{p \in J^2 : f(x) \wedge g(x) \wedge [f, g](x) \neq 0, g(x) \wedge [f, g](x) \wedge [X, [f, g]](x) \neq 0, \\ \text{where } x \text{ is the source of } P\}$$

and

$$B_- := \{p \in J^2 : f(x) \wedge g(x) \wedge [f, g](x), g(x) \wedge [f, g](x) \wedge [Y, [f, g]](x), \\ \text{where } x \text{ is the source of } P\}$$

are open in J^2 . Therefore $M(B_+)$ and $M(B_-)$ are open subsets of $C^\infty(\mathbb{R}^3, \mathbb{R}^6)$ in the Whitney C^2 (and hence in the C^∞) topology. Observe that these sets consist of all the systems Σ for which f , g and $[f, g]$ are independent everywhere and α , respectively, β does not vanish on \mathbb{R}^3 , i.e., these sets are the sets \mathcal{B}_+ respectively \mathcal{B}_- . We will show that the systems $\Sigma \in \mathcal{B} = \mathcal{B}_+ \cup \mathcal{B}_-$ which satisfy at every point at least one of the conditions of Propositions 1 and 2 and Corollaries 1 and 2 contain an open and dense subset of \mathcal{B} , and hence are generic. This will prove our theorem.

Let $\tilde{J}^k := \{p \in J^k : f(x) \wedge g(x) \wedge [f, g](x) \neq 0, \text{ where } x \text{ is the source of } p\}$. Define

$$W_+ := \{P \in \tilde{J}^5 : g(x) \wedge [f, g](x) \wedge \text{ad}^r X(Y)(x) = 0, r = 2, 3, 4, 5 \text{ } x = \text{source of } P\},$$

$$W_- := \{P \in \tilde{J}^5 : g(x) \wedge [f, g](x) \wedge \text{ad}^r Y(X)(x) = 0, r = 2, 3, 4, 5 \text{ } x = \text{source of } P\}.$$

LEMMA. W_+ and W_- are closed embedded submanifolds of codimension 4 in \tilde{J}^5 .

Proof. The functions defining the sets W_{\pm} are smooth and so it is clear that the sets W_{\pm} are closed. To prove that they are embedded submanifolds of codimension 4, it suffices to show that in each case the defining map $\phi : \tilde{J}^5 \rightarrow \mathbb{R}^4$ has maximal rank everywhere. To see this we compute some special partial derivatives of the functions $\phi_{r-1}(P) := f(x) \wedge g(x) \wedge \text{ad}^r X(Y)(x)$, x the source of P , in local coordinates. (We only do it for W_+ .) Note that the i th component of $[f, g]$ is

$$[f, g]_i = \sum_{j=1}^3 \frac{\partial g_i}{\partial x_j} f_j - \frac{\partial f_i}{\partial x_j} g_j,$$

$$[f - g, [f, g]]_i = \sum_{k=1}^3 \left(\sum_{j=1}^3 \frac{\partial^2 g_i}{\partial x_k \partial x_j} f_j (f_k - g_k) + \text{derivative terms of order one or second order derivatives of } f \right).$$

Therefore, if e_i denotes the i th unit vector in \mathbb{R}^3 , we have

$$\frac{\partial \phi_1}{\partial (\partial^2 g_i / \partial x_r \partial x_s)} = (f_r(f_s - g_s) + f_s(f_r - g_r))(g(x) \wedge [f, g](x) \wedge e_i).$$

We claim that not all of these partial derivatives vanish. Since $g(x)$ and $[f, g](x)$ are independent, $g(x) \wedge [f, g](x) \wedge e_i \neq 0$ for at least one i . So, if the partials are all zero, we must have

$$f_r(f_s - g_s) + f_s(f_r - g_r) = 0 \quad \text{for all } r, s \in \{1, 2, 3\}.$$

In particular, we get for $r = s$ that $f_r(f_r - g_r) = 0$ for all f . If $f_r \neq 0$, this implies $f_r = g_r$. For $f_r = 0$ we have $f_s g_r = 0$ for all s . Since $f(x) \neq 0$ there exists an s such that $f_s \neq 0$, and hence $g_r = 0$. Therefore we have $f(x) = g(x)$, contradicting the independence of f and g . This shows that not all partial derivatives of ϕ_1 with respect to second-order partials of g vanish. Since ϕ_1 depends only on the 2-jet of Σ , it is clear that all partial derivatives of ϕ_1 with respect to higher-order partials of g do vanish. Analogously it follows for $r = 3, 4, 5$ that not all the partial derivatives of ϕ_{r-1} with respect to r th order partials of g vanish, whereas all partial derivatives with respect to higher order partials of g vanish trivially since ϕ_{r-1} only depends on the r -jet of Σ . This shows that the gradients of ϕ_1, ϕ_2, ϕ_3 , and ϕ_4 are independent everywhere in \tilde{J}^5 . \square

Let $T_W := \{\Sigma \in \mathcal{B} : j^5 \Sigma \not\subset W\}$, $W = W_+$, or $W = W_-$. By Thom's Transversality Theorem the set $\{\Sigma \in C^\infty(\mathbb{R}^3, \mathbb{R}^6) : j^5 \Sigma \not\subset W\}$ is dense in $C^\infty(\mathbb{R}^3, \mathbb{R}^6)$ and so T_W is dense in \mathcal{B} . Also, since W is a closed embedded submanifold of \tilde{J}^5 , the set $\{j^5 \Sigma \in C^\infty(\mathbb{R}^3, \tilde{J}^5) : j^5 \Sigma \not\subset W\}$ is open in $C^\infty(\mathbb{R}^3, \tilde{J}^5)$. Furthermore, \tilde{J}^5 is open in J^5 and the map j^5 is continuous in the Whitney C^∞ topology. Therefore T_W is open and dense in \mathcal{B} , then so is $T_{W_+} \cap T_{W_-}$. And for $\Sigma \in T_{W_+} \cap T_{W_-}$ at every point at least one of the conditions of Propositions 1 and 2 and Corollaries 1 and 2 holds. For, if $\beta(p) \neq 0$, then not all of α, γ, μ , and κ can vanish at p , since $j^5 \Sigma(p) \not\subset W_-$ means $j^5 \Sigma(p) \notin W_-$. So one of the conditions of Proposition 1 or 2 holds. Analogously, if $\alpha(p) \neq 0$, then one of the conditions of Corollary 1 or 2 is satisfied since $j^5 \Sigma(p) \notin W_+$. This proves our result.

Acknowledgment. The author gratefully acknowledges the constructive remarks of an anonymous referee which greatly improved the presentation of the results in this paper in the first two sections.

REFERENCES

- [1] V. G. BOLTYANSKY, *Sufficient conditions for optimality and the justification of the dynamic programming method*, SIAM J. Control Optim., 4 (1966), pp. 326–361.
- [2] A. BRESSAN, *The generic local time-optimal stabilizing controls in dimension 3*, SIAM J. Control Optim., 24 (1986), pp. 170–190.
- [3] P. BRUNOVSKY, *Existence of a regular synthesis for general control problems*, J. Differential Equations, 38 (1980), pp. 317–343.
- [4] M. GOLUBITSKY AND V. GUILLEMIN, *Stable Mappings and Their Singularities*, Springer-Verlag, New York, 1973.
- [5] A. J. KRENER, *The high-order maximum principle and its application to singular extremals*, SIAM J. Control Optim., 15 (1977), pp. 256–293.
- [6] A. J. KRENER AND H. SCHÄTTLER, *The structure of small-time reachable sets in low dimensions*, to appear.
- [7] I. A. K. KUPKA, *The ubiquity of the Fuller-phenomenon*, Report 52, Institut Fourier, Grenoble, France, 1986.
- [8] C. MARCHAL, *Chattering arcs and chattering controls*, J. Optim. Theory Appl., 11 (1973), pp. 441–468.
- [9] L. S. PONTRYAGIN, V. G. BOLTYANSKY, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *Mathematical Theory of Optimal Processes*, Wiley-Interscience, New York, 1962.
- [10] H. SCHÄTTLER, *On the local structure of time-optimal bang-bang trajectories in R^3* , SIAM J. Control Optim., 26 (1988), pp. 186–204.
- [11] ———, *Regularity properties of time-optimal trajectories of an analytic single-input control-linear system in dimension 3*, J. Optim. Theory Appl., to appear.
- [12] H. J. SUSSMANN, *Analytic stratifications and control theory*, Proc. ICM, Helsinki, 1978, pp. 865–871.
- [13] ———, *A bang-bang theorem with a bound on the number of switchings*, SIAM J. Control Optim., 17 (1979), pp. 629–651.
- [14] ———, *Lie brackets, real analyticity, and geometric control*, in Differential Geometric Control, R. Brockett, R. Milman, and H. Sussmann, eds., Progress in Mathematics, Vol. 27, Birkhauser, Boston, 1979.
- [15] ———, *Time optimal control in the plane*, in Feedback Control of Linear and Nonlinear Systems, Lecture Notes in Control and Information Sciences, Vol. 39, Springer-Verlag, Berlin, 1985, pp. 244–260.
- [16] ———, *Lie brackets and real analyticity in control theory*, in Mathematical Control Theory, Banach Center Publications, Vol. 14, Warsaw, 1985, pp. 515–542.
- [17] ———, *Envelopes, conjugate points, and optimal bang-bang extremals*, in Proc. 1985 Paris Conference on Nonlinear Systems, M. Fliess and M. Hazewinkel, eds., Reidel, Boston, 1986.
- [18] ———, *The structure of time-optimal trajectories for single-input systems in the plane: the C nonsingular case*, SIAM J. Control Optim., 25 (1987), pp. 433–465.
- [19] ———, *The structure of time-optimal trajectories for single-input systems in the plane: the general real analytic case*, SIAM J. Control Optim., 25 (1987), pp. 868–904.
- [20] ———, *Regular synthesis for time-optimal control of single-input real analytic systems in the plane*, SIAM J. Control Optim., 25 (1987), pp. 1145–1162.
- [21] ———, *Remarks on real analyticity and the regularity of optimal controls*, to appear.

AN APPROACH TO SIMULTANEOUS SYSTEM DESIGN. PART II: NONSWITCHING GAIN AND DYNAMIC FEEDBACK COMPENSATION BY ALGEBRAIC GEOMETRIC METHODS*

BIJOY K. GHOSH†

Abstract. This paper studies structured uncertainty problems in feedback system design, considers a compact parameterization of the space of linear dynamical systems and introduces “base points” and “critical points” as two algebraic-geometric objects that have significance in sensitivity and robustness studies, respectively. Using the Nevanlinna-Pick interpolation theory, the author obtains a necessary and sufficient condition for simultaneous stabilization of a structured one-parameter family of plants. A recent result due to Kharitonov, on the simultaneous stability of a parameterized family of polynomials, leads to a sufficiency condition for simultaneous stabilization of a structured multiparameter family of plants. Furthermore the author considers “simultaneous pole placement” of an r -tuple of plants as a means to arbitrarily tune the natural frequencies of a multimode linear dynamical system. The concept of “nondegenerate” and “twisted” r -tuples of plants is introduced as the pole placement problem is studied via Schubert enumerative geometry as an intersection problem on the associated Grassmannian. Various other design problems, viz., the strong stabilization problem and the dead beat control problem, are also considered.

Key words. simultaneous stabilization, pole assignment, base point, critical point, interpolation, nondegenerate, twisted

AMS(MOS) subject classifications. 93, 14

1. Introduction. In the last decade, control theorists have witnessed significant progress in multi-input multi-output system design. The central issue is the classification of plants or families of plants that admit a robust, nonswitching, dynamic output feedback compensator and satisfy a specific set of design constraints (viz. sensitivity minimization, stabilization, pole assignment, etc.). The basic problem without any additional design restriction is the robust stabilization problem described as follows.

Problem 1.1 (Nonswitching compensator problem). Given a family G of $p \times m$ real, linear dynamical systems, does there exist a nonswitching $m \times p$ real compensator $K(s)$ such that the closed-loop systems $G(s)[I + K(s)G(s)]^{-1}$ have poles only in the open left half of the complex plane for every $G(s) \in G$?

The above problem is important in the design of a compensator for a dynamical system whose parameters are uncertain and where Λ , the region of uncertainty, is known. This would indeed be the case if the parameters of the system were poorly identifiable. In this situation it is important to ascertain the existence of a compensator which is robust with respect to the parameter uncertainty and which stabilizes the family $(G_\lambda(s), \lambda \in \Lambda)$ of plants. Likewise, the above problem would arise if the dynamical system has a continuum of operating points (for example, the altitude of an aircraft or the r.p.m. of an induction motor). The parameters of the system may vary depending upon the choice of the operating points and the objective of the design is to synthesize a compensator robust with respect to the parameter variation. As has been originally pointed out by Sacks and Murray [37], this problem also arises if a control system $G(s)$ operates in many failed modes, all within a given family G , and the design

* Received by the editors September 22, 1986; accepted for publication (in revised form) July 27, 1987. This research was partially supported by the National Aeronautics and Space Administration under grant NSG-2265 while the author was at Harvard University, Cambridge, Massachusetts, and by the National Science Foundation under grant ECS-8414220. This paper is part of the author's Ph.D. thesis at Harvard University.

† Department of Systems Science and Mathematics, Washington University, Saint Louis, Missouri 63130.

objective is to continue to maintain stability even if a component of the control system $G(s)$ should fail. For a detailed discussion concerning many other illustrative examples of the above design problem, we refer the reader to the monograph edited by Ackermann [1], and many other references, e.g., [8]–[10].

We now detail the contribution of this paper. In § 2 we pose Problem 1.1 as a classification or parameterization problem and propose an algebraic geometric framework which enables us to characterize some of the design aspects of the stabilization and pole assignment problem that have not yet been recognized. In particular we compactify the space of dynamic compensators, study the continuity of the pole placement map, and show the existence of “base points.” Finally we argue that the base points are significant, and that, in fact, around these points the location of the closed-loop poles is extremely sensitive with respect to the plant and the compensator parameters. We also study the asymptotic behavior of the closed-loop system as the plant/compensator parameters tend to the base points. We also show the existence of critical points, which are those infinite points in a neighborhood of which the stability of the closed-loop system is not robust with any margin of stability. The principal message of § 2 is therefore to avoid the base points and the critical points in any feedback system design.

In § 3 we describe explicit system design techniques for robust stabilization of a structured family of plants simultaneously. We consider the one-parameter family of plants and, using Nevanlinna–Pick interpolation methods [35], [36], obtain a necessary and sufficient condition for simultaneous stabilization of the family. We also obtain a bound on the degree of the compensator if it exists. Thus whereas the conjecture that “the set of pairs of simultaneously stabilizable plants of bounded degree has simultaneously stabilizing compensators of bounded degree” is false [13], in this paper we show the existence of a suitable class of one-parameter families of simultaneously stabilizable plants of bounded degree that admits simultaneously stabilizing compensators of bounded degree. In other words, the bound on the degree of the compensator can be a priori computed as a function of the degree of the plants to be compensated. In § 3 we also consider the multiparameter family of plants and obtain a sufficiency condition for simultaneous stabilization of such a family. Finally we consider robust simultaneous stabilization problems, wherein we study the extent to which a structured family of simultaneously stabilizable plants can be perturbed by an unstructured perturbation.

In §§ 4 and 5 we describe explicit system design techniques for pole assignment. In § 4 we consider a general ansatz for pole placement as a Schubert intersection problem [19] and derive explicit bounds on the degree of the compensator for generic simultaneous pole placement of proper and strictly proper plants. We also obtain similar bounds for the strong stabilization problem, wherein the dynamic compensator is assumed to be stable. In § 5 we analyze a simultaneous pole assignment problem via gain feedback. We also pose and analyze a classically well-known “dead beat control” problem.

Section 6 concludes this paper with a summary and discussion on future research directions.

2. A parameterization of the space of linear dynamical systems. In this section we consider the set of $m \times p$ proper or strictly proper transfer functions of McMillan degree q denoted, respectively, $S_{p,m}^q$ and $\Sigma_{p,m}^q$. In order to study degenerating families of systems and asymptotic properties of systems under deformation, it is useful to compactify the spaces $\Sigma_{p,m}^q$ and $S_{p,m}^q$ and obtain the spaces $\overline{\Sigma}_{p,m}^q$ and $\overline{S}_{p,m}^q$, respectively.

This construction enables us to study the behavior of a sequence of feedback control systems, in particular the closed-loop poles, as an asymptotic property of a sequence of compensators of a given McMillan degree q . Clearly the above study is useful in system identification, adaptive control and in high gain feedback control system design. It also enables us to study variations in the plant and in the compensator parameters simultaneously arising possibly as a result of parameter changes or structured uncertainty. (For some further details and previous work on families of systems, see [20]–[22], [25], and [32].) The main point that we wish to illustrate in this section is that in every feedback system design problem which involves a family of plants and a family of compensators, there exists a set of points called “base points” in the plant/compensator space which needs to be avoided. Otherwise near the base point the corresponding closed-loop poles are sensitive with respect to changes in the plant and the compensator parameters. Likewise we show that there exists a set of points called “critical points” in the compactified plant/compensator space which cannot be robustly stabilized by a nonswitching regulator.

2.1. A compactification of the space of dynamic compensators. Using a construction due to Hermann and Martin [26], compactification of $S_{p,m}^q$ has been developed by Brockett and Byrnes [3] for $q=0$ and more recently by Byrnes [4] for $q \geq 0$. Although it is repetitive, we discuss the main results for the sake of completeness. (For a detailed description, see [12].)

To begin with, consider $S_{p,m}^0$ of $m \times p$ gain matrices $-K$. If we consider $-K$ to be a feedback gain matrix, it defines a relation between the m input u and p output y as follows:

$$(2.1.1) \quad [I_m \quad K][u^T \quad y^T]^T = 0.$$

The matrix $[I_m \quad K]$ is an $m \times (m+p)$ matrix with linearly independent rows. In fact, multiplying (2.1.1) by a nonsingular $m \times m$ matrix does not change the relation between u and y . Thus the matrix $[I_m \quad K]$ in fact defines an m -plane in \mathbb{R}^{m+p} via its row span, and is therefore a point in Grass $(m, m+p)$. (See [19] for details about a Grassmannian.)

More generally, consider $S_{p,m}^q$ of $m \times p$ proper transfer functions $K(s)$ of McMillan degree q . If we consider $-K(s)$ to be a feedback compensator, it defines a relation between $u(s)$ and $y(s)$ as follows:

$$(2.1.2) \quad [I_m \quad K(s)][u(s)^T \quad y(s)^T]^T = 0.$$

Let us now consider the left coprime factorization of $K(s)$ given by

$$(2.1.3) \quad K(s) = \begin{bmatrix} k(s) & 0 \\ 0 & k(s) \end{bmatrix}^{-1} \begin{bmatrix} k_{11}(s) & k_{1p}(s) \\ k_{m1}(s) & k_{mp}(s) \end{bmatrix}$$

where $k(s)$, $k_{ij}(s)$, $i = 1, \dots, m$; $j = 1, \dots, p$ are polynomials in s of degree q . Let us write

$$(2.1.4) \quad k(s) \triangleq k^{(0)} + \dots + k^{(q)} s^q,$$

$$(2.1.5) \quad k_{ij}(s) \triangleq k_{ij}^{(0)} + \dots + k_{ij}^{(q)} s^q$$

and define

$$(2.1.6) \quad \underline{k} \triangleq [k^{(0)} \dots k^{(q)}],$$

$$(2.1.7) \quad \underline{k}_{ij} \triangleq [k_{ij}^{(0)} \dots k_{ij}^{(q)}],$$

$$(2.1.8) \quad \underline{u}_i \triangleq [u_i(s), su_i(s), \dots, s^q u_i(s)]^T, \quad i = 1, \dots, m,$$

$$(2.1.9) \quad \underline{y}_j \triangleq [y_j(s), sy_j(s), \dots, s^q y_j(s)]^T, \quad j = 1, \dots, p.$$

We now rewrite (2.1.2) as

$$(2.1.10) \quad \begin{bmatrix} \underline{k} & 0 & \underline{k_{11}} & \cdots & \underline{k_{1p}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \underline{k} & \underline{k_{m1}} & \cdots & \underline{k_{mp}} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_m \\ y_1 \\ \vdots \\ y_p \end{bmatrix} = 0.$$

The associated matrix in (2.1.10) defines (via its row span) an m -plane in $\mathbb{R}^{(q+1)(m+p)}$ and is therefore a point in $\text{Grass}(m, (q+1)(m+p))$. Thus we have

$$(2.1.11) \quad S_{p,m}^q \subset \text{Grass}(m, (q+1)(m+p)).$$

Dually, by considering the right coprime factorization of $K(s)$ in (2.1.3), we also have

$$(2.1.12) \quad S_{p,m}^q \subset \text{Grass}(p, (q+1)(m+p)).$$

The closure $\overline{S_{p,m}^q}$ of $S_{p,m}^q$ inside either $\text{Grass}(m, (q+1)(m+p))$ or $\text{Grass}(p, (q+1)(m+p))$ is compact and is the compactification obtained by Byrnes [4] for $q \geq 0$ and by Brockett and Byrnes [3] for $q = 0$.

DEFINITION 2.1.1. The set of points

$$[\overline{S_{p,m}^q} - S_{p,m}^q]$$

is defined to be the “infinite points” in the above compactification.

Note that the infinite points are the set of limit points which needs to be added to $S_{p,m}^q$ in order to obtain a compact set.

The following example taken from [12] explicitly describes $\overline{S_{2,2}^1}$ in $\text{Grass}(2, 8)$.

Example 2.1.2. Consider the set of two-input and two-output proper systems of degree 1 given by

$$(2.1.13) \quad G(s) = \begin{bmatrix} (a_7s + a_1)/(a_6s + a_5) & (a_8s + a_2)/(a_6s + a_5) \\ (a_9s + a_3)/(a_6s + a_5) & (a_{10}s + a_4)/(a_6s + a_5) \end{bmatrix}$$

where

$$(2.1.14) \quad (a_7a_5 - a_1a_6)(a_8a_5 - a_2a_6) = (a_9a_5 - a_3a_6)(a_{10}a_5 - a_4a_6).$$

The transfer function (2.1.13) defines the point

$$(2.1.15) \quad \text{row span of} \begin{bmatrix} a_5 & a_6 & 0 & 0 & a_1 & a_7 & a_2 & a_8 \\ 0 & 0 & a_5 & a_6 & a_3 & a_9 & a_4 & a_{10} \end{bmatrix}$$

in $\text{Grass}(2, 8)$. Not every element in $\text{Grass}(2, 8)$, however, corresponds to an element in $S_{2,2}^1$ or in $\overline{S_{2,2}^1}$. To see that, consider the point p defined by

$$(2.1.16) \quad p \triangleq \text{row span of} \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \\ y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 & y_8 \end{bmatrix}$$

in $\text{Grass}(2, 8)$. In order that $p \in S_{2,2}^1$ it should satisfy

$$(2.1.17) \quad \eta_{12} = \eta_{34} = 0,$$

$$(2.1.18) \quad \eta_{14} = \eta_{23},$$

$$(2.1.19) \quad \eta_{13}\eta_{68} + \eta_{24}\eta_{57} + \eta_{23}[\eta_{58} + \eta_{67}] = 0,$$

$$(2.1.20) \quad \eta_{24} \neq 0,$$

and

$$(2.1.21) \quad (\eta_{24}\eta_{54} \neq \eta_{14}\eta_{64} \text{ or } \eta_{24}\eta_{74} \neq \eta_{14}\eta_{84} \text{ or } \eta_{24}\eta_{25} \neq \eta_{14}\eta_{26} \text{ or } \eta_{24}\eta_{27} \neq \eta_{14}\eta_{28})$$

where $\eta_{ij} = x_i y_j - x_j y_i$ is the set of Plucker coordinates (see [7]) via which Grass $(2, 8)$ can be imbedded in \mathbb{RP}^{27} . Note that the infinite points in $\overline{S}_{2,2}^1$ are described by (2.1.17)–(2.1.19) and

$$(2.1.22) \quad \eta_{24} = 0$$

or

$$(2.1.23) \quad \frac{\eta_{24}}{\eta_{14}} = \frac{\eta_{64}}{\eta_{54}} = \frac{\eta_{84}}{\eta_{74}} = \frac{\eta_{26}}{\eta_{25}} = \frac{\eta_{28}}{\eta_{27}}.$$

We remark that (2.1.17)–(2.1.19) and (2.1.22) contain the set of improper transfer functions in $\overline{S}_{2,2}^1$, and that (2.1.17)–(2.1.19) and (2.1.23) describe the set of transfer functions of degree 0 in $\overline{S}_{2,2}^1$.

By analogous technique we can compactify $\Sigma_{p,m}^q$, the set of $m \times p$ strictly proper transfer functions of degree q as a point in Grass $(m, (q+1)m + qp)$ or Grass $(p, (q+1)p + qm)$. Let us now consider the following example.

Example 2.1.3. Consider the set $\Sigma_{2,2}^1$ given by (2.1.13), (2.1.14) with $a_7 = a_8 = a_9 = a_{10} = 0$. Every transfer function in $\Sigma_{2,2}^1$ defines a point in Grass $(2, 6)$ given by

$$(2.1.24) \quad \text{row span of } \begin{bmatrix} a_5 & a_6 & 0 & 0 & a_1 & a_2 \\ 0 & 0 & a_5 & a_6 & a_3 & a_4 \end{bmatrix}.$$

In fact if p is a point in Grass $(2, 6)$ defined by

$$(2.1.25) \quad \text{row span of } \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ y_1 & y_2 & y_3 & y_4 & y_5 & y_6 \end{bmatrix},$$

then $p \in \Sigma_{2,2}^1$ if and only if

$$(2.1.26) \quad \eta_{12} = \eta_{34} = 0, \quad \eta_{14} = \eta_{23}, \quad \eta_{56} = 0, \quad \eta_{24} \neq 0.$$

It follows that $\overline{\Sigma}_{2,2}^1 - \Sigma_{2,2}^1$ is described by

$$(2.1.27) \quad \eta_{12} = \eta_{34} = \eta_{56} = \eta_{24} = 0, \quad \eta_{14} = \eta_{23} = 0.$$

2.2. Continuity of the pole placement map and the base locus condition. Classically, root locus [11] was studied by considering a single-input single-output plant together with a scalar gain feedback, and a plot of the closed-loop poles as a function of the feedback gain was obtained. Since root locus is continuous with respect to the scalar gain, the closed-loop poles asymptotically and continuously approach the open-loop zeros as the gain tends to infinity. For the multi-input multi-output case the generalization is not immediate. However, as we shall see in this section, the asymptotic locus of the closed-loop poles as a function of the compensator parameters is not necessarily continuous. Furthermore, as we have seen in § 2.1, infinity is not a single point and therefore there is more than one way to tend to infinity leading to distinct closed-loop behavior of the dynamical system. These are now detailed as follows.

Let $G(s)$ be a proper $p \times m$ plant of McMillan degree n and let $K(s)$ be a proper $m \times p$ compensator of McMillan degree q . Let us represent them in their coprime fraction representation as follows:

$$(2.2.1) \quad G(s) = N_p(s)D_p(s)^{-1},$$

$$(2.2.2) \quad K(s) = D_c(s)^{-1}N_c(s)$$

where $N_p(s)$, $D_p(s)$, $N_c(s)$, $D_c(s)$ are matrices of appropriate orders with elements in the ring H of stable proper rational functions. It is well known that $K(s)$ places the poles of the plant $G(s)$ at $s = s_1, s_2, \dots, s_{n+q}$ provided

$$(2.2.3) \quad \det [N_c(s)N_p(s) + D_c(s)D_p(s)](s_i) = 0$$

for $i = 1, \dots, n+q$. Let

$$(2.2.4) \quad c_0 + c_1s + \dots + c_{n+q}s^{n+q} \triangleq p(s)$$

be a polynomial which vanishes precisely where $\det [N_cN_p + D_cD_p]$ vanishes. The polynomial $p(s)$ is called the characteristic polynomial. For a fixed $G(s)$, (2.2.3) defines a mapping between the compensator parameters and the coefficients c_i of the characteristic polynomial $p(s)$ in (2.2.4). We therefore define the pole placement map

$$(2.2.5) \quad \chi: \overline{S_{p,m}^q} - B \rightarrow \mathbb{RP}^{n+q}$$

defined by

$$(2.2.6) \quad \chi(K(s)) = [c_0, \dots, c_{n+q}]$$

where the right-hand side of (2.2.6) is written in homogeneous coordinates. The set B is called the base locus (see [19], [34]) and is defined to be the set of points in $\overline{S_{p,m}^q}$, where χ does not have a continuous extension. It is a well-known fact (see [34, p. 98]) that the base locus B is given by

$$(2.2.7) \quad B \triangleq \bigcap_H \overline{\chi^{-1}(H)}$$

where H is a hyperplane in \mathbb{RP}^{n+q} and $\overline{\chi^{-1}(H)}$ is the closure of $\chi^{-1}(H)$ inside $\overline{S_{p,m}^q}$. The significance of the base locus is made clear in the following proposition.

PROPOSITION 2.2.1. *The base locus B is equivalently described as*

$$(2.2.8) \quad B = \{K(s) \mid \det [N_c(s)N_p(s) + D_c(s)D_p(s)] \equiv 0\}.$$

Proof. Every $s_i \in \mathbb{R}$ defines a hyperplane H_i in \mathbb{RP}^{n+q} , viz., the hyperplane orthogonal to the vector $[1 \ s_i \ s_i^2 \ \dots \ s_i^{n+q}]$. It follows that

$$(2.2.9) \quad \overline{\chi^{-1}(H_i)} = \{K(s) \mid \det [N_cN_p + D_cD_p](s_i) = 0\}.$$

Let s_1, \dots, s_{n+q+1} be a set of distinct real numbers. It follows that

$$(2.2.10) \quad \bigcap_{i=1}^{n+q+1} \overline{\chi^{-1}(H_i)} = \{K(s) \mid \det [N_cN_p + D_cD_p](s) \equiv 0\},$$

since the characteristic polynomial of the closed-loop transfer function is of degree $n+q$. Finally since

$$(2.2.11) \quad B \subset \bigcap_{i=1}^{n+q+1} \overline{\chi^{-1}(H_i)},$$

we have

$$(2.2.12) \quad B \subset \{K(s) \mid \det [N_c(s)N_p(s) + D_c(s)D_p(s)] \equiv 0\}.$$

Conversely, let us assume that

$$(2.2.13) \quad K_1(s) \in \{K(s) \mid \det [N_c(s)N_p(s) + D_c(s)D_p(s)] \equiv 0\}.$$

Clearly $K_1(s)$ is a compensator for which the associated characteristic polynomial $p(s)$ is identically zero, i.e., $c_0 = c_1 = \dots = c_{n+q} = 0$. Hence χ cannot be defined at $K_1(s)$. By choosing two different sequences of compensators approaching $K_1(s)$, we can show

that χ does not even have a continuous extension to a mapping which is defined at $K_1(s)$. Otherwise χ has to be multivalued at $K_1(s)$, which is absurd. Therefore

$$(2.2.14) \quad \{K(s) | \det [N_c(s)N_p(s) + D_c(s)D_p(s)] \equiv 0\} \subset B. \quad \square$$

In the following example we characterize the base locus for a particular compensation problem.

Example 2.2.2. Consider a single-input single-output plant $g(s)$ of degree 1 given by $(s+1)/(s+2)$. Consider a proper compensator $k(s)$ of degree 1 given by $(k_1s-2)/(s+k_2)$. The compensator is parameterized by two parameters k_1 and k_2 . The closed-loop transfer function is given by

$$(2.2.15) \quad \frac{(s+1)(s+k_2)}{(k_1+1)s^2 + (k_1+k_2)s + (2k_2-2)}.$$

Using Proposition 2.2.1 we infer that the base locus B is the point set given by

$$(2.2.16) \quad B \triangleq \{k(s) | k_1 = -1, k_2 = 1\}.$$

The pole placement map χ is described as

$$(2.2.17) \quad \chi: S_{1,1}^1 - B \rightarrow \mathbb{RP}^2,$$

$$(2.2.18) \quad \chi(k(s)) = \{k_1 + 1, k_1 + k_2, 2k_2 - 2\}.$$

Clearly χ is continuous everywhere except at $k(s) = -(s+2)/(s+1)$ where it is not even defined. To show that χ cannot be extended to a function χ_1 continuous at $-(s+2)/(s+1)$, we assume the parameterized family of compensators F_t given by

$$(2.2.19) \quad F_t \triangleq \{k(s) | (k_1 + 1) = t(k_2 - 1)\}$$

where $t \in \mathbb{R}$. Consider a sequence of compensators $h_1^{(i)}(s), h_2^{(i)}(s), \dots$ in F_t approaching the compensator $-(s+2)/(s+1)$. It is easy to see that

$$(2.2.20) \quad \chi(h_i^{(i)}(s)) = [t \quad (t+1) \quad 2]$$

independent of i but dependent on t . Note that the closed-loop system is stable if $t > 0$ and unstable otherwise. This indicates that when the compensator parameters k_1, k_2 approach arbitrarily close to the point $(-1, 1)$, possibly along two different lines in the (k_1, k_2) plane, the closed-loop system can be stable in one case and unstable in the other. Thus stability is not a robust property in the neighborhood of $(-1, 1)$ primarily because the closed-loop poles are extremely sensitive with respect to the parameter t . Finally the asymptotic locus of the roots of the characteristic polynomial is not continuous with respect to the parameters k_1, k_2 at the point $(-1, 1)$. Hence χ does not have a continuous extension to a function χ_1 defined at the point $(-1, 1)$.

2.3. Degeneration of a family of closed-loop systems. In the last section we derived the base locus condition assuming that the plant is fixed and the compensator parameters are allowed to vary. In this section we describe a general theory, wherein both the plant and the compensator parameters are allowed to vary. Let us consider the set $\overline{S_{m,p}^n}$ of $p \times m$ plants of degree n and the set $\overline{S_{p,m}^q}$ of $m \times p$ compensators of degree q suitably compactified. As an extension of the pole placement map (2.2.5) we define

$$(2.3.1) \quad \psi: \overline{S_{m,p}^n} \times \overline{S_{p,m}^q} - \tilde{B} \rightarrow \mathbb{RP}^{n+q},$$

$$(2.3.2) \quad \psi(G(s), K(s)) = [c_0, \dots, c_{n+q}]$$

where $G(s)$, $K(s)$, c_i , $i = 0, \dots, n+q$ are described via (2.2.1)–(2.2.4). Similar to the definition of B , we define \tilde{B} to be the base locus of the map ψ . We now consider the following proposition.

PROPOSITION 2.3.1. *The base locus \tilde{B} is described as*

$$(2.3.3) \quad \tilde{B} = \{(G(s), K(s)) \mid \underline{\sigma}[N_c(s)N_p(s) + D_c(s)D_p(s)] = 0 \ \forall s \in \mathbb{C}\} \\ \text{where } \underline{\sigma}(\cdot) \text{ denotes the minimum singular value} \}.$$

Proof. The proof follows from Proposition 2.2.1 since

$$(2.3.4) \quad \underline{\sigma}[N_c(s)N_p(s) + D_c(s)D_p(s)] = 0 \quad \text{for all } s \in \mathbb{C},$$

$$(2.3.5) \quad \Leftrightarrow \det[N_c(s)N_p(s) + D_c(s)D_p(s)] \equiv 0. \quad \square$$

We now consider the following illustrative example from the literature [33, Ex. 2.1], wherein the compensator is fixed and the plant parameters degenerate into the base locus.

Example 2.3.2. Assume $b_{12} \neq 0$. Consider the following state-space system:

$$(2.3.6) \quad \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 & b_{12} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix},$$

$$(2.3.7) \quad [y_1^T \ y_2^T]^T = [x_1^T \ x_2^T]^T$$

together with the feedback compensator

$$(2.3.8) \quad [u_1^T \ u_2^T]^T = -[y_1^T \ y_2^T]^T.$$

The closed-loop system above is stable and has a pair of poles at $-2, -2$. Let us now perturb (2.3.6) and consider

$$(2.3.9) \quad \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 & b_{12} \\ 5/b_{12} & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

The perturbed system (2.3.9), (2.3.7), (2.3.8) is an unstable system with closed-loop poles at the zeros of $(s+2)^2 - 5$. Note that the stability of the closed-loop system and the instability of the perturbed closed-loop system are independent of b_{12} . This fact is quite disturbing because, for arbitrary large b_{12} , the matrix components of the two systems are “close.”

From the viewpoint of the parameterization proposed in § 2.1, it may be trivially checked that the unperturbed system defines a point in Grass $(2, 8)$ given by

$$(2.3.10) \quad \text{row span of } \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & b_{12} & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

which degenerates to the point

$$(2.3.11) \quad \text{row span of } \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

as $b_{12} \rightarrow \infty$. The perturbed system, on the other hand, defines the point

$$(2.3.12) \quad \text{row span of } \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & b_{12} & 0 \\ 0 & 0 & b_{12} & b_{12} & 5 & 0 & 1 & 0 \end{bmatrix}$$

which also degenerates to the point (2.3.11) as $b_{12} \rightarrow \infty$. Thus the two closed-loop systems under consideration are “close” as points on Grass $(2, 8)$.

It may be trivially checked that the point (2.3.11) is on the base locus of the map

$$(2.3.13) \quad \psi: \overline{S_{2,2}^2} \times \overline{S_{2,2}^0} - \tilde{B} \rightarrow \mathbb{RP}^2.$$

Thus, following the argument presented in Example 2.2.2, we conclude that stability is not a robust property around the point (2.3.11). This may also be viewed as an algebraic-geometric interpretation of the conclusion presented in Example 2.1 of [33].

Remark 2.3.3. Proposition 2.3.1 serves to justify the comment on page 78 of [33]: “A multivariable system will not be robust . . . if its return difference transfer function matrix . . . is nearly singular at some frequency . . .” In fact, a near singular transfer function matrix is close to the base locus.

In multivariable feedback system design, a good design strategy is to avoid the base locus. In particular, around the base locus, the closed-loop poles are sensitive with respect to the plant and the compensator parameters. (For an original reference on the application of base locus in system theory, see Byrnes [45].)

2.4. Robust system design and the critical point condition. Let $\overline{S_{p,m}^q}$ be as defined in § 2.1. For the purpose of robust system design, in this section we pose the following problem.

Problem 2.4.1. Parameterize the set $\Omega_{p,m}^q$ of points in $\overline{S_{p,m}^q}$ with the property that for every $\sigma \in \Omega_{p,m}^q$ there exists an open neighborhood $N(\sigma)$ of σ in $\overline{S_{p,m}^q}$ such that every plant in $S_{p,m}^q \cap N(\sigma)$ is simultaneously stabilizable.

Analysis of the above problem is clearly important in the study of “system degeneration” and “high gain” compensation. (For motivational remarks and other details pertaining to the case $m = p = 1$, see [6].)

DEFINITION 2.4.2. The set of points

$$(2.4.1) \quad \overline{S_{p,m}^q} - \Omega_{p,m}^q$$

is defined as the “critical points” in the compactified $\overline{S_{p,m}^q}$.

We now state the following simple proposition.

PROPOSITION 2.4.3. Every critical point is an infinite point in $\overline{S_{p,m}^q}$.

Proof. The finite points in $\overline{S_{p,m}^q}$ are points in $S_{p,m}^q$. Moreover, every point in $S_{p,m}^q$ has an open neighborhood in $S_{p,m}^q$ that can be simultaneously stabilized by a dynamic compensator. Therefore no point in $S_{p,m}^q$ is a critical point. \square

The set of critical points are now explicitly characterized as follows.

THEOREM 2.4.4. The points

$$(2.4.2) \quad \text{row span of } \left[\begin{array}{cc|ccc} \underline{k} & 0 & \underline{k}_{11} & \underline{k}_{12} & \cdots & \underline{k}_{1p} \\ & \ddots & \vdots & \vdots & & \vdots \\ 0 & \underline{k} & \underline{k}_{m1} & \underline{k}_{m2} & \cdots & \underline{k}_{mp} \end{array} \right]$$

belonging to $\overline{S_{p,m}^q}$ are critical points of $\overline{S_{p,m}^q}$ if and only if there exists s^* in the closed right half of the complex plane such that the row rank of the matrix $M(s)$ given by

$$(2.4.3) \quad M(s) \triangleq \left[\begin{array}{cc|ccc} k(s) & 0 & k_{11}(s) & \cdots & k_{1p}(s) \\ & \ddots & \vdots & & \vdots \\ 0 & k(s) & k_{m1}(s) & \cdots & k_{mp}(s) \end{array} \right]$$

at $s = s^*$ is $< m$.

Remark 2.4.5. The vectors \underline{k} , \underline{k}_{ij} and the polynomials $k(s)$, $k_{ij}(s)$, $i = 1, \dots, m$; $j = 1, \dots, p$ are defined in (2.1.4)–(2.1.7).

Proof of Theorem 2.4.4. The sufficiency is clear, since the closed-loop characteristic polynomial

$$(2.4.4) \quad \det \left[\begin{array}{ccc|ccc} k(s) & & 0 & k_{11}(s) & \cdots & k_{1p}(s) \\ & \ddots & & \vdots & & \vdots \\ 0 & & k(s) & k_{m1}(s) & \cdots & k_{mp}(s) \\ \hline & & K(s) & & & I \end{array} \right]$$

corresponding to any $p \times m$ compensator transfer function $K(s)$ always vanishes at s^* and therefore (2.4.2) is indeed a critical point.

In order to show “necessity” assume that the row rank of the matrix $M(s)$ in (2.4.3) is m for all $s \in \mathbb{C}$. Then clearly $M(s)$ is right invertible, i.e., there exists an $(m+p) \times m$ matrix $N(s)$ such that $M(s)N(s) = \Delta(s)$, where $\det \Delta(s)$ vanishes only in \mathbb{C}_s , the open left half of the complex plane. We partition $N(s)$ as $[N_1(s)^T \ N_2(s)^T]^T$, where $N_1(s)$ is $m \times m$ and where we choose $N(s)$ such that $\det N_1(s) \neq 0$ and $N_1^{-1}N_2$ is proper. Thus we have a transfer function $N_1^{-1}N_2$ of a compensator which stabilizes the transfer function corresponding to the plant (2.4.2). Moreover, a sufficiently small neighborhood of (2.4.2) in $S_{p,m}^q$ is stabilized by the compensator. It follows that (2.4.2) is not a critical point. The only case left is that where the rank of the matrix $M(s)$ in (2.4.3) is $< m$ at some points in \mathbb{C}_s , the open left half of the complex plane. This fact, however, implies that $M(s)$ can be factored as

$$(2.4.5) \quad M(s) = \Delta_1(s)M_1(s)$$

where $\det \Delta_1(s)$ vanishes at $s = s_1$ and where $M_1(s)$ has either rank m at all $s \in \mathbb{C}_s$ or rank $< m$ for some $s = s_2$ in \mathbb{C}_s . Proceeding as above it follows that $M(s)$ can be factored as

$$(2.4.6) \quad M(s) = \Delta^*(s)M^*(s)$$

where $\det \Delta^*(s)$ vanishes only in \mathbb{C}_s and where $M^*(s)$ is of rank m at all $s \in \mathbb{C}$. By obtaining a right inversion of $M^*(s)$ and considering the reasoning as before it follows that (2.4.2) is not a critical point. \square

Example 2.4.6. If $q = m = p = 1$, the space $S_{1,1}^1$ contains proper transfer functions

$$(2.4.7) \quad q(s) = \frac{a_1s + a_0}{s + b_0}, \quad a_0 \neq a_1b_0$$

of degree 1, that can be parameterized as points in \mathbb{RP}^3 via the homogeneous coordinates $[1 \ b_0 \ a_1 \ a_0]$. The set of infinite points $\overline{S_{1,1}^1} - S_{1,1}^1$ in \mathbb{RP}^3 is described as

$$(2.4.8) \quad \{[h_0, h_1, h_2, h_3]: h_0h_3 = h_1h_2\}$$

and the set of critical points are described as

$$(2.4.9) \quad \{[h_0, h_1, h_2, h_3]: h_0h_3 = h_1h_2 \text{ and } h_0h_1 \leq 0 \text{ and } h_2h_3 \leq 0\}.$$

Remark 2.4.7. The significance of a critical point may now be emphasized. By definition, if ξ is a critical point in $\overline{S_{p,m}^q}$, then there exists a sequence of points $\xi_i, i = 1, 2, \dots$, in $S_{p,m}^q$ which tends to ξ in the limit and with the property that the family of plants $\{\xi_i\}$ is simultaneously unstabilizable by a dynamic compensator. Since critical points are those infinite points in $\overline{S_{p,m}^q}$, neighborhoods of which are not simultaneously stabilizable, a good design strategy is to avoid the critical points in system design.

In order to describe Corollary 2.4.9 of Theorem 2.4.4 we need the following definition.

DEFINITION 2.4.8. A compensator is said to stabilize a plant with margin of stability $\varepsilon > 0$ if the closed-loop poles are in the region $\{s \in \mathbb{C} : \operatorname{Re} s < -\varepsilon\}$.

COROLLARY 2.4.9. Let $\xi_1(s), \xi_2(s), \dots$ be a sequence of plants in $S_{p,m}^q$ which approaches a critical point in $\overline{S_{p,m}^q} - \Omega_{p,m}^q$ arbitrarily close. Then the family $\{\xi_i(s), i = 1, 2, \dots\}$ of plants is not simultaneously stabilizable with any margin of stability.

Proof. Suppose the corollary does not hold. Then there exists a compensator which simultaneously stabilizes the given family with a margin of stability $\varepsilon > 0$. This, however, implies that the plants $\xi_i(s)$ in the family are not in the associated base locus. Thus the closed-loop poles are continuous with respect to the plant parameters. It follows from the definition of a critical point that there exists an integer N sufficiently large that $\xi_N(s)$ has a closed-loop pole arbitrarily close to a point s^* in the closed right half of the complex plane. However, this is a contradiction. \square

3. Simultaneous stabilization of plants with structured uncertainty. In this section we restrict our attention to single-input single-output plants and consider the following problem.

Problem 3.1. Let P be a subset of $\Omega_{1,1}^n$ for some integer n . Does there exist a dynamic compensator of degree $\leq q$ which stabilizes each and every plant $\sigma \in P$?

Of course if σ_0 is a plant in P and if the subset P represents structured perturbation of σ_0 in $\Omega_{1,1}^n$, the above Problem 3.1 is a robust stabilization problem with structured uncertainty. Also, for the purpose of estimating the complexity of the compensator which solves a particular design problem, it is of interest to consider.

Question 3.2. Is q a priori bounded?

If P is a single point it is well known that an upper bound on q is given by $q \leq n$. On the other hand it has been shown [13], [12] that if P is a pair of points in $\Omega_{1,1}^n$, an upper bound on q does not exist.

3.1. A general ansatz for simultaneous stabilization. In this section, we consider a family P of single-input single-output proper plants of degree n given by

$$(3.1.1) \quad P = \{x_\lambda(s)/y_\lambda(s) : \lambda \in \Lambda, x_\lambda, y_\lambda \in H; x_\lambda, y_\lambda \text{ are coprime and } \deg x_\lambda/y_\lambda = n \text{ for all } \lambda \in \Lambda\}.$$

Assume that P contains at least two distinct plants. Furthermore let us define

$$(3.1.2) \quad \eta_{ij}(s) = x_i(s)y_j(s) - x_j(s)y_i(s), \quad i, j \in \Lambda$$

where Λ is an arbitrary parameter space. We now state without proof the following result from Theorem 5.1 of [13].

PROPOSITION 3.1.1. Let $x_1(s)/y_1(s), x_2(s)/y_2(s)$ be two distinct plants in P . There exists a proper compensator which simultaneously stabilizes each and every plant in P if and only if there exist $\Delta_1(s), \Delta_2(s) \in J$, the set of multiplicative units of H , such that

(i) $\Delta_1 y_2 - \Delta_2 y_1$ and $\Delta_2 x_1 - \Delta_1 x_2$ vanish at $s_1, \dots, s_t \in \mathbb{C}_u$ with multiplicities at least m_1, \dots, m_t , respectively, where s_1, \dots, s_t are the zeros of $\eta_{12}(s)$ in \mathbb{C}_u with multiplicities m_1, \dots, m_t , respectively. (Here $\mathbb{C}_u = \mathbb{C} - \mathbb{C}_s$.)

(ii) $\Delta_2 x_1 - \Delta_1 x_2$ does not vanish at ∞ with multiplicity m_∞ unless $\eta_{12}(s)$ vanishes at ∞ with multiplicity at least m_∞ .

(iii) There exists $\Delta_\lambda \in J$ for all $\lambda \in \Lambda - \{1, 2\}$ such that

$$(3.1.3) \quad \Delta_1 \eta_{\lambda 2} - \Delta_2 \eta_{\lambda 1} = \Delta_\lambda \eta_{12}.$$

In the proof of the next theorem we will introduce four auxiliary return difference polynomials which we denote by $\alpha_i(s)$, $i = 1, 2, 3, 4$.

We now consider the main result of this section described in the following theorem.

THEOREM 3.1.2. *Assume that P contains at least two proper plants and does not contain three plants g_1, g_2, g_3 such that $g_1(s^*) = g_2(s^*) = g_3(s^*)$ for any $s^* \in \mathbb{C} \cup \{\infty\}$. Furthermore assume that Λ is a compact subset of \mathbb{R}^α for some integer α and that the parameters of $x_\lambda(s)/y_\lambda(s)$ depend continuously on λ . A sufficient condition that every plant in P is simultaneously stabilizable by a dynamic compensator is given by the stability of the four polynomials $\alpha_i(s)$, $i = 1, 2, 3, 4$, defined in (3.1.10).*

Remark 3.1.3. For details on the simultaneous stabilization of a finite number of plants, see [12]–[14], [16], [17], [37], [42].

Remark 3.1.4. The above problem is a version of the robust stabilization problem. It is indeed surprising that such a problem may be tackled via simultaneous stabilization of a finite number of plants, in particular, four. We consider Theorem 3.1.2 to be a theoretically significant result, since simultaneous stabilization of four plants by a stable, minimum phase compensator (currently an open problem) is conceptually simpler than simultaneous stabilization of a family of uncountably many plants.

Proof of Theorem 3.1.2. Let us consider the two proper plants $g_1(s), g_2(s)$ in P defined as follows:

$$g_1(s) = x_1(s)/y_1(s), \quad g_2(s) = x_2(s)/y_2(s)$$

where $x_1, x_2, y_1, y_2 \in H$ and x_i, y_i are coprime $i = 1, 2$. Assume that there exists $\Delta_1, \Delta_2, \Delta_\lambda \in J$, $\lambda \in \Lambda - \{1, 2\}$ such that (3.1.3) is satisfied. We now show that P is simultaneously stabilizable by a possibly improper compensator.

By easy algebraic manipulation of (3.1.3) we first obtain the following two identities:

$$(3.1.4) \quad [\Delta_1 x_2 - \Delta_2 x_1] \eta_{\lambda 1} = [\Delta_\lambda x_1 - \Delta_1 x_\lambda] \eta_{12},$$

$$(3.1.5) \quad [\Delta_1 y_2 - \Delta_2 y_1] \eta_{\lambda 1} = [\Delta_\lambda y_1 - \Delta_1 y_\lambda] \eta_{12}.$$

Noting that $\eta_{\lambda 1}(s)$ does not have a zero at any point where $\eta_{12}(s)$ has a zero for all $\lambda \in \Lambda - \{1, 2\}$, we find from (3.1.4) and (3.1.5) that condition (i) of Proposition 3.1.1 is satisfied. Since condition (ii) of Proposition 3.1.1 is imposed in order to make the compensator proper, it follows that P is simultaneously stabilizable by a possibly improper compensator.

Since Λ is compact we now claim that P is actually stabilizable by a proper compensator. This is because there exists a $\sigma_1, \sigma_2 > 0$ such that the closed-loop poles with respect to the improper compensator is in the region $\{s: -\sigma_2 < \operatorname{Re} s < \sigma_1\}$. It follows that we can perturb the improper compensator and obtain the required proper compensator.

From the argument presented so far we conclude that a necessary and sufficient condition for the simultaneous stabilization of P is the existence of $\Delta_1, \Delta_2, \Delta_\lambda \in J$, $\lambda \in \Lambda - \{1, 2\}$ such that (3.1.3) is satisfied.

We now obtain a sufficiency condition for the existence of $\Delta_1, \Delta_2, \Delta_\lambda \in J$, $\lambda \in \Lambda - \{1, 2\}$ such that (3.1.3) is satisfied.

Clearing the denominator, we may assume without any loss of generality that in (3.1.3), $\Delta_1, \Delta_2, \Delta_\lambda, \eta_{\lambda 2}, \eta_{\lambda 1}$ and η_{12} are polynomials of appropriate degrees. By a suitable algebraic manipulation we conclude the existence of two polynomials $\eta'_{\lambda 2}(s)$ and $\eta'_{\lambda 1}(s)$ such that the following two statements are equivalent:

(1) There exist $\Delta_1, \Delta_2, \Delta_\lambda \in J$, $\lambda \in \Lambda - \{1, 2\}$ such that (3.1.3) is satisfied.

(2) There exist stable polynomials Δ'_1, Δ'_2 such that $\Delta'_1 \eta'_{\lambda 2} - \Delta'_2 \eta'_{\lambda 1}$ is stable for all $\lambda \in \Lambda - \{1, 2\}$.

Let us now write

$$\begin{aligned}\eta'_{\lambda 1}(s) &= a_0(\lambda) + a_1(\lambda)s + \cdots + a_n(\lambda)s^n, \\ \eta'_{\lambda 2}(s) &= b_0(\lambda) + b_1(\lambda)s + \cdots + b_n(\lambda)s^n.\end{aligned}$$

Since the coefficients of $\eta'_{\lambda 1}$ and $\eta'_{\lambda 2}$ are linear combinations of coefficients of $\eta_{\lambda 1}$ and $\eta_{\lambda 2}$, respectively, it follows that $a_i(\lambda)$ and $b_i(\lambda)$ are continuous functions of λ . Since Λ is compact, these functions attain a maximum and a minimum. Assume $a_i(\lambda) \in [\alpha_i, \beta_i]$, $b_i(\lambda) \in [\gamma_i, \delta_i]$ for $\alpha_i \leq \beta_i$, $\gamma_i \leq \delta_i$, $i = 0, \dots, n$. Let us denote the columns of the $(2q+2) \times (n+q)$ matrix

$$(3.1.6) \quad \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_n & & 0 \\ \gamma_0 & \gamma_1 & \cdots & \gamma_n & & \\ & \alpha_0 & \cdots & \alpha_{n-1} & \alpha_n & \\ & \gamma_0 & \cdots & \gamma_{n-1} & \gamma_n & \\ & & & \vdots & \vdots & \\ & & & \alpha_0 & \alpha_1 & \cdots & \alpha_n \\ 0 & & & \gamma_0 & \gamma_1 & \cdots & \gamma_n \end{bmatrix}$$

by $v_0^T, \dots, v_{n+q-1}^T$, respectively, and the columns of the above matrix (3.1.6) by α_i replaced with β_i , γ_i replaced with δ_i , $i = 0, \dots, n$, with $u_0^T, \dots, u_{n+q-1}^T$, respectively. Let us denote Δ'_1/Δ'_2 by

$$(3.1.7) \quad \Delta'_1/\Delta'_2(s) = \left[\sum_{i=0}^q c_i s^i \right] / \left[\sum_{i=0}^{q-1} d_i s^i + s^q \right]$$

where

$$(3.1.8) \quad \underline{\psi} \triangleq [c_0 \quad d_0 \quad \cdots \quad c_{q-1} \quad d_{q-1} \quad c_q \quad 1],$$

$$(3.1.9) \quad \underline{s}^T \triangleq [1, s, s^2, \dots, s^{n+q}]^T.$$

By Theorem 1.3 of [15] it follows that a sufficient condition for the family of polynomials $\Delta'_1 \eta'_{\lambda 2} - \Delta'_2 \eta'_{\lambda 1}$ for all $\lambda \in \Lambda - \{1, 2\}$ to be stable is given by the stability of the following four polynomials:

$$(3.1.10) \quad \begin{aligned}\alpha_1(s) &= \underline{\psi} [v_0^T v_1^T u_2^T u_3^T v_4^T \cdots] \underline{s}^T, \\ \alpha_2(s) &= \underline{\psi} [u_0^T u_1^T v_2^T v_3^T u_4^T \cdots] \underline{s}^T, \\ \alpha_3(s) &= \underline{\psi} [v_0^T u_1^T u_2^T v_3^T v_4^T \cdots] \underline{s}^T, \\ \alpha_4(s) &= \underline{\psi} [u_0^T v_1^T v_2^T u_3^T u_4^T v_4^T \cdots] \underline{s}^T.\end{aligned} \quad \square$$

Remark 3.1.5. The four polynomials in (3.1.10) have the same structure as we would obtain as return difference polynomials from four single-input single-output plants compensated by a stable, minimum phase compensator with the exception that the associated $(2q+2) \times (n+q)$ matrices in (3.1.10) are not the Sylvester matrices corresponding to any single-input single-output plants.

Remark 3.1.6. The proof of the above theorem relies on some recent results due to Kharitonov [28] (see [15] for details).

3.2. Simultaneous stabilization of one-parameter family of plants by interpolation methods. Let $x_1(s)/y_1(s)$ and $x_2(s)/y_2(s)$; $x_1, x_2, y_1, y_2 \in H$ be a pair of proper but not strictly proper plants in $\Omega_{1,1}^n$ written in their coprime representation. For ease of exposition and notational simplicity we assume that the zeros of $x_1 y_2 - x_2 y_1(s)$ in \mathbb{C}_u

are simple and are in the interior of \mathbb{C}_u . We now consider a linear one-parameter family of plants described as follows:

$$(3.2.1) \quad F = \{g_\lambda(s): g_\lambda(s) = [\lambda x_1 + (1-\lambda)x_2]/[\lambda y_1 + (1-\lambda)y_2], \lambda \in [0, 1], g_\lambda \in \Omega_{1,1}^n \forall \lambda\}.$$

In order to state the main result of this section we need the following notation.

Let a_1, \dots, a_t denote the zeros of $x_1 y_2 - x_2 y_1$ in \mathbb{C}_u . Let us define b_i as follows:

$$(3.2.2) \quad b_i = x_2/x_1(a_i) \quad \text{if the multiplicity of } a_i \text{ as a common zero of } x_1(s) \text{ and } x_2(s) \\ \text{is less than or equal to the multiplicity of } a_i \text{ as a common} \\ \text{zero of } y_1(s) \text{ and } y_2(s) \\ = y_2/y_1(a_i) \quad \text{otherwise}$$

for $i = 1, \dots, t$. Let us also define $s_i = (a_i - 1)/(a_i + 1)$ and $z_i = (\sqrt{b_i} - 1)/(\sqrt{b_i} + 1)$, where the branch cut for the square root is taken to be the nonpositive real axis. Furthermore let k be the largest real number such that

$$(3.2.3) \quad [1 - k^2 z_i z_j]/[1 - s_i s_j]_{i,j=1}^t$$

is nonnegative definite. The main result of this section is now described.

THEOREM 3.2.1. *The following three statements are equivalent:*

- (1) *F is simultaneously stabilizable by some dynamic compensator.*
- (2) *F is simultaneously stabilizable by some dynamic compensator of degree $\leq 3n - 2$.*

$$(3.2.4) \quad (3) \quad k > 1.$$

Remark 3.2.2. We find Theorem 3.2.1 quite surprising. In fact, whereas the conjecture that "pairs of simultaneously stabilizable plants of bounded McMillan degree have simultaneously stabilizing compensators of bounded McMillan degree" is false (see [13]), the conjecture that "simultaneously stabilizable linear one-parameter families of plants of bounded McMillan degree have simultaneously stabilizing compensators of bounded McMillan degree" is true. In this sense, stabilizing a discrete r -tuple of plants (in particular a pair of plants) simultaneously appears to be a much more difficult problem.

Remark 3.2.3. Results similar to Theorem 3.2.1 are unknown for the multiparameter family of plants.

The following corollary of Theorem 3.2.1 is also quite interesting.

COROLLARY 3.2.4 (Gain margin in problem). The family of plants

$$(3.2.5) \quad \{\lambda g(s): \deg g(s) = n, \lambda \in [0, 1]\}$$

is simultaneously stabilizable if and only if it is simultaneously stabilizable by a compensator of degree $\leq 3n - 2$.

Thus Theorem 3.2.1 and Corollary 3.2.4 recover a result previously due to Khargonekar and Tannenbaum [27] on the gain margin problem. Furthermore we obtain an explicit bound on the degree of the compensator if one exists.

Proof of Theorem 3.2.1. Let us define

$$(3.2.6) \quad x_\lambda(s) = \lambda x_1(s) + (1-\lambda)x_2(s),$$

$$(3.2.7) \quad y_\lambda(s) = \lambda y_1(s) + (1-\lambda)y_2(s).$$

From Proposition 3.1.1 it follows that a necessary and sufficient condition for the simultaneous stabilization of F is the existence of $\Delta_1, \Delta_2, \Delta_\lambda \in J, \lambda \in (0, 1)$ such that

$$(3.2.8) \quad \Delta_2/\Delta_1(a_i) = b_i, \quad i = 1, \dots, t,$$

$$(3.2.9) \quad \Delta_2/\Delta_1(\infty) \neq x_2/x_1(\infty),$$

$$(3.2.10) \quad \lambda \Delta_1 + (1-\lambda)\Delta_2(s) = \Delta_\lambda(s)$$

for all $\lambda \in (0, 1)$. Condition (3.2.9) is imposed in order to make proper the simultaneously stabilizing compensator. However, since $[0, 1]$ is compact, we may ignore (3.2.9) according to an argument similar to the proof of Theorem 3.1.2.

Thus a necessary and sufficient condition for the simultaneous stabilization of F is the existence of $\Delta_2/\Delta_1(s)$, which interpolates (a_i, b_i) , $i = 1, \dots, t$ and which does not intersect the nonpositive real axis including infinity at any point in \mathbb{C}_u .

Recall that \mathbb{R}^- is the nonpositive real axis including infinity. It is trivial to check that \mathbb{C}_u is conformally equivalent to \mathbb{D} via the following transformation:

$$\begin{aligned}\psi_1: \mathbb{C}_u &\rightarrow \mathbb{D}, \\ s &\mapsto \frac{s-1}{s+1}.\end{aligned}$$

Likewise $\mathbb{C} - \mathbb{R}^-$ is conformally equivalent to \mathbb{D}^- , the open interior of the unit disc, via the following transformation:

$$(3.2.11) \quad \begin{aligned}\psi_2: \mathbb{C} - \mathbb{R}^- &\rightarrow \mathbb{D}^-, \\ s &\mapsto \frac{\sqrt{s}-1}{\sqrt{s}+1}\end{aligned}$$

where the branch-cut for the square root is taken to be the nonpositive real axis.

Thus the above interpolation problem reduces to the following Nevanlinna-Pick interpolation problem [35], [36]: "Does there exist a rational function with real coefficients which defines a map from \mathbb{D} to \mathbb{D}^- in the complex plane, which interpolates the points

$$(3.2.12) \quad \left(\frac{a_i-1}{a_i+1}, \frac{\sqrt{b_i}-1}{\sqrt{b_i}+1} \right)$$

for $i = 1, \dots, t$?"

From standard results in the Nevanlinna-Pick interpolation theory [35], [36] it follows that if (3.2.4) is satisfied there exists a rational function of degree $\leq t-1$ which solves the above interpolation problem. Working backwards, if (3.2.4) is satisfied there exists a rational function $\Delta_2/\Delta_1(s)$ of degree $\leq 2t-2$ that satisfies (3.2.8) and (3.2.10). Since the plants x_1/y_1 and x_2/y_2 are of degree n , it follows that if $k > 1$, the simultaneously stabilizing compensator is of degree $\leq 2t-2-n$. Since $t \leq 2n$ we have $2t-2-n \leq 3n-2$. Thus we have (3) \Rightarrow (2).

On the other hand, if $k \not\geq 1$ it follows that the Nevanlinna-Pick interpolation problem cannot be solved even by an analytic function and therefore the one-parameter family of plants F is simultaneously unstabilizable. Thus we have (1) \Rightarrow (3).

However, this completes the proof since (2) \Rightarrow (1) is clear. \square

3.3. Robust simultaneous stabilization methods. The basic question we pose in this section is the following: Given a subset S of plants in $\Omega_{m,p}^\alpha$ (where α is a given fixed integer) which is simultaneously stabilizable by a dynamic compensator? To what extent can S be perturbed to S' in $\Omega \triangleq \bigcup_{\alpha=0}^\infty \Omega_{m,p}^\alpha$ such that

$$(3.3.1) \quad S \subset S' \subset \Omega$$

and S' is simultaneously stabilizable?

If p_0 is a nominal plant in S , we shall call S a structured perturbation of p_0 . This may result from gross variation of one or more parameters of the plant because of component degradation, a change in the working condition, or failure of one or more components. S' , on the other hand, is an unstructured perturbation of p_0 . This may be introduced at a high frequency of operation as a result of unmodeled dynamics.

In order to describe the main results of this section we need the following assumptions and definitions.

Assume that $p_0(s)$ is a single-input single-output proper plant of degree $\mu_0 \geq 1$ and let $r(s)$ be a 1×1 stable proper transfer function of degree μ . Furthermore assume the following:

(1) All the unstable poles of $p_0(s)$ are simple and their real parts are positive.

(2) The relative degree of $r(s)$ is either 0 or 1 and $|r(j\omega)| > 0$ for all ω . Let us define the following subset S' of Ω :

$$(3.3.2) \quad S' = \{p(s) : |p(j\omega) - p_0(j\omega)| \leq |r(j\omega)| \forall \omega \text{ and } p(s) \text{ has the same number of unstable poles as does } p_0(s)\}.$$

Also we define the subfamily S of S' as follows:

$$(3.3.3) \quad S \triangleq S' \cap \Omega_{1,1}^{\mu_0 + \mu + 1}.$$

The main result of this section is described in the following theorem.

THEOREM 3.3.1. *The following statements are equivalent:*

(3.3.4) (1) S' is simultaneously stabilizable by a proper dynamic compensator.

(3.3.5) (2) S is simultaneously stabilizable by a proper dynamic compensator.

(3.3.6) (3) S or S' is simultaneously stabilizable by a proper dynamic compensator of degree $\leq 2\mu_0 + \mu - 1$.

Remark 3.3.2. Theorem 3.3.1 is an adaptation of a recent result due to Kimura [29] and is therefore not entirely surprising. We remark however that Theorem 3.3.1 is the only known result on the simultaneous stabilization of a structured multiparameter family of plants other than the sufficiency condition in Theorem 3.1.2.

Proof of Theorem 3.3.1. Condition (3.3.4) of course implies (3.3.5) since S is a subfamily of S' . To see that (3.3.5) implies (3.3.4) we proceed as follows. Let $C(s)$ be a stabilizer of $p_0(s)$. Assume that S' is simultaneously unstabilizable. It follows from Lemm 1 of [29] that there exists a $p(s)$ in S' such that for some $\omega = \omega_0$ we have

$$(3.3.7) \quad |[p(j\omega_0) - p_0(j\omega_0)]C(j\omega_0)| \geq |1 + p_0(j\omega_0)C(j\omega_0)|.$$

We now claim that there exists $p^*(s)$ in S such that

$$(3.3.8) \quad C(j\omega_0)[p_0(j\omega_0) - p^*(j\omega_0)] = [1 + p_0(j\omega_0)C(j\omega_0)].$$

Such a plant $p^*(s)$ therefore cannot be stabilized by the controller $C(s)$ since the closed-loop system has a pole at $j\omega_0$. S would therefore be simultaneously unstabilizable by $C(s)$. To prove the claim we proceed as follows.

Using Nevanlinna-Pick interpolation methods [35], [36], we construct a proper rational function $\psi(s)$ with real coefficients of degree 1 with the following property:

$$(3.3.9) \quad |\psi(j\omega)| = 1 \quad \forall \omega,$$

$$(3.3.10) \quad \arg \psi(j\omega_0) = \arg [1 + p_0(j\omega_0)C(j\omega_0)] - \arg [C(j\omega_0)r_1(j\omega)]$$

where $r_1(s)$ is any stable, proper rational function of degree μ and $|r_1(j\omega)| = |r(j\omega)|$ for all ω . Let us define

$$(3.3.11) \quad p_1(s) = p_0(s) - r_1(s)\psi(s).$$

It follows from (3.3.7), (3.3.9), and (3.3.11) that

$$(3.3.12) \quad \begin{aligned} |C(j\omega_0)[p_0(j\omega_0) - p_1(j\omega_0)]| &= |C(j\omega_0)| |r_1(j\omega_0)| \\ &= |C(j\omega_0)| |r(j\omega_0)| \\ &\geq |1 + p_0(j\omega_0)C(j\omega_0)|. \end{aligned}$$

Moreover, using (3.3.10) we have

$$(3.3.13) \quad \arg [(C(j\omega_0))[p_0(j\omega_0) - p_1(j\omega_0)]] = \arg [1 + p_0(j\omega_0)C(j\omega_0)].$$

From (3.3.12), (3.3.13) it follows that there exists an $\varepsilon \in [0, 1]$ such that

$$(3.3.14) \quad p^*(s) = p_0(s) - \varepsilon r_1(s)\psi(s)$$

satisfies condition (3.3.8). Finally $p^*(s)$ clearly belongs to S , implying that S is simultaneously unstabilizable.

The condition (3.3.6) of course implies (3.3.4). To see that (3.3.4) \Rightarrow (3.3.6), note that if S' is simultaneously stabilizable, it follows from [29] that there exists a strictly bounded real function $u(s)$ which satisfies an interpolation condition. The degree of $u(s)$ depends only on the total number of interpolating points (see [41, pp. 139–143]) and is therefore one less than the number of unstable poles of $p_0(s)$. In turn, the degree of the compensator (if it exists) is $\leq 2t + \mu + 1$, where t is the number of unstable poles of $p_0(s)$. Clearly $t \leq \mu_0$ and we have the required inequality. \square

4. Simultaneous pole placement of multimode linear dynamical systems by dynamic compensation.

4.1. Statement and motivation of the problem. The pole placement problem arises from an interest in knowing if we can arbitrarily tune the frequency response of a dynamical system by a proper choice of a feedback compensator. Of course if the plant is multi-model, we need to consider an r -tuple of $p \times m$ rational transfer function matrices $G_1(s), \dots, G_r(s)$ of McMillan degrees n_1, \dots, n_r , respectively, modeling all the various modes of the m -input p -output plant. In fact, the r modes of a strictly proper plant can be topologized as points in the set

$$(4.1.1) \quad \Sigma = \Sigma_{m,p}^{n_1} \times \dots \times \Sigma_{m,p}^{n_r}$$

where

$$(4.1.2) \quad \Sigma_{m,p}^{n_i} = \{p \times m \text{ strictly proper } G_i(s), \deg G_i(s) = n_i\}.$$

(For details on the above topology, see [7], [23], and [24].) A set s_1, \dots, s_{n_i+q} of unordered self-conjugate set of complex numbers can be topologized as a point in \mathbb{R}^{n_i+q} via the coefficients of the polynomial

$$(4.1.3) \quad \prod_{j=1}^{n_i+q} (s - s_j).$$

We now address the following problem, which constitutes an important result of this section.

Problem 4.1.1 (Generic simultaneous pole placement at a generic set of poles). Is it true that for all r -tuple $G_1(s), \dots, G_r(s)$ of $p \times m$ real, strictly proper plants of McMillan degree $n_i, i = 1, \dots, r$, respectively, except perhaps those contained in a proper algebraic set, we can arbitrarily assign a set of r self-conjugate sets of $n_i + q$ poles, $i = 1, \dots, r$, with the possible exception of a proper algebraic subset of poles, of the closed-loop systems $G_i(s)[I + K(s)]^{-1}, i = 1, \dots, r$, respectively, by a real, proper compensator $K(s)$ of McMillan degree q ?

Roughly speaking, Problem 4.1.1 refers to the question of whether or not it is possible to place the closed-loop poles of almost all r -tuples of plants almost everywhere by a dynamic feedback compensator. Clearly, the exact location of the closed-loop poles is unimportant. Otherwise numerical errors in the location of the closed-loop poles would affect the closed-loop response of the system. Consideration of a generic

r -tuple of plants, although a restriction, enables us to compute explicit bounds on the degree of the compensator. We remark, however, that in this section we say nothing about the pole placement of a nongeneric r -tuple of plants. (For an explicit parameterization of the generic set and a discussion on the nongeneric pole placement problem, see [16].)

As noted in [4], a general method of solving pole placement problems is, first, to parameterize, as an open dense subset of a compact space, compensators with certain of their discrete invariants fixed, and second, to parameterize, for fixed $s_0 \in \mathbb{C}$ and fixed plant $G(s)$, the subset of those compensators $K(s)$ which place a pole of $G(s)[I + K(s)G(s)]^{-1}$ at s_0 . Now that we have done this, the pole placement problem becomes a problem of intersection theory on this compact space. For m , p , and q fixed the first phase of this program has been completed in § 2. We now analyze the second phase of this program.

4.2. Analysis of the simultaneous pole placement map. As an immediate extension of the pole placement map in (2.2.5), we define the simultaneous pole placement map

$$(4.2.1) \quad \chi: \overline{S_{p,m}^q} - B \rightarrow \mathbb{RP}^{n_1+q} \times \cdots \times \mathbb{RP}^{n_r+q}$$

defined by

$$(4.2.2) \quad \chi(K(s)) = ([c_{1,0}, \cdots, c_{1,n_1+q}], [c_{2,0}, \cdots, c_{2,n_2+q}], \cdots, [c_{r,0}, \cdots, c_{r,n_r+q}])$$

where

$$(4.2.3) \quad c_{i,0} + c_{i,1}s + \cdots + c_{i,n_i+q}s^{n_i+q}$$

is the characteristic polynomial of the closed-loop system

$$(4.2.4) \quad G_i(s)[I + K(s)G_i(s)]^{-1}$$

for $i = 1, \cdots, r$, respectively. The set B , as before, is the base locus and is described as follows:

$$(4.2.5) \quad B \triangleq \bigcup_{i=1}^r \bigcap_{H_i} \overline{\chi_i^{-1}(H_i)}$$

where H_i is a hyperplane in \mathbb{RP}^{n_i+q} . Moreover if proj_i is the projection map

$$(4.2.6) \quad \text{proj}_i: \mathbb{RP}^{n_1+q} \times \cdots \times \mathbb{RP}^{n_r+q} \rightarrow \mathbb{RP}^{n_i+q}$$

for $i = 1, \cdots, r$ we define

$$(4.2.7) \quad \chi_i(\cdot) \triangleq \text{proj}_i \chi(\cdot).$$

Finally $\overline{\chi_i^{-1}(H_i)}$ is the closure of $\chi_i^{-1}(H_i)$ inside $\overline{S_{p,m}^q}$. The base locus B can be equivalently described as the closure of $\{K(s) \in \overline{S_{p,m}^q} : \det[I + K(s)G_i(s)] = 0 \text{ for some } i \in \{1, \cdots, r\}\}$ where the closure is taken inside $\overline{S_{p,m}^q}$. Problem 4.1.1 can be equivalently described as follows: "For almost all r -tuples of plants in Σ (described in 4.1.1), is it true that the simultaneous pole placement map is almost surjective?"

Remark 4.2.1. If χ is continuous, then almost-surjectivity of χ would imply that χ is surjective. However, since the base locus is not necessarily empty, χ is not, in general, continuous at all points.

Consider a set of r self-conjugate tuples of $n_i + q$ distinct poles given by

$$(4.2.8) \quad \{s_{i1}, s_{i2}, \cdots, s_{i,n_i+q}\}$$

$i = 1, \dots, r$. Of course we may define the associated closed-loop characteristic polynomials as follows:

$$(4.2.9) \quad K_i(s) = \prod_{j=1}^{n_i+q} (s - s_{ij}) = \sum_{j=0}^{n_i+q} c_{ij} s^j$$

for $i = 1, 2, \dots, r$. The point $[c_{i,0}, \dots, c_{i,n_i+q}]$ in \mathbb{RP}^{n_i+q} can be defined as the intersection of the hyperplanes $H_{ij}, j = 1, n_i + q$, where H_{ij} is the hyperplane in \mathbb{RP}^{n_i+q} orthogonal to

$$(4.2.10) \quad [1 \quad s_{ij} \quad s_{ij}^2 \cdots s_{ij}^{n_i+q}].$$

A necessary condition for pole placement at the locations described by (4.2.8) is therefore given by the condition that

$$(4.2.11) \quad \Omega = \bigcap_{i=1}^r \bigcap_{j=1}^{n_i+q} \chi_i^{-1}(H_{ij}) \neq \emptyset.$$

Remark 4.2.2. Even when Ω is nonempty there may not exist a simultaneously pole assigning compensator. This is because Ω may be contained inside the base locus B and therefore for every compensator inside Ω , one of the associated r closed-loop characteristic polynomials would vanish identically. Moreover Ω may contain only points in $\overline{S_{p,m}^q} - S_{p,m}^q$ which are the infinite points of the compactification.

We have therefore established the following theorem.

THEOREM 4.2.3. *Given an r -tuple of proper plants, there exists a finite compensator of degree q which places the poles of the i th plant $G_i(s)$ at conjugate set of distinct $s_{ij}, j = 1, \dots, n_i + q$, for all $i = 1, \dots, r$ if and only if*

$$(4.2.12) \quad (\Omega - B) \cap S_{p,m}^q$$

is nonempty. \square

Example 4.2.4. For the purpose of this example, let us assume that the base field is complex. Consider the problem of placing all eight poles, in the closed loop, of a 2×2 proper plant of degree 7 by a 2×2 proper compensator of degree 1. From Example 2.1.2 we know that $S_{2,2}^1$ is parameterized in $\text{Grass}_{\mathbb{C}}(2, 8)$ by a set of four hypersurfaces described by (2.1.17), (2.1.18) and (2.1.19). Moreover the algebraic curve (2.1.17) describes a Schubert hypersurface (see [31]). Equation (2.1.18) is not itself a Schubert hypersurface but it is a hypersurface homologous to a Schubert hypersurface. To see this, let ϕ be the functional in \mathbb{R}^4 describing the hypersurface (2.1.18). In the dual space of \mathbb{R}^4 , ϕ is deformable to a functional ϕ_s describing a Schubert hypersurface. This follows by considering the deformation

$$(4.2.13) \quad t\phi + (1-t)\phi_s \quad \text{where } 0 \leq t \leq 1$$

which is clearly an element of $(\mathbb{R}^4)^*$ for all $t \in [0, 1]$. Similarly the quadratic curve (2.1.19) is homologous to a pair of Schubert hypersurfaces.

It has been shown by Byrnes [5] that in order to place all eight poles in the closed loop, we must intersect $S_{2,2}^1$ in $\text{Grass}_{\mathbb{C}}(2, 8)$ by eight Schubert hypersurfaces, each corresponding to a given pole. In other words, the problem reduces to computing the intersection of 11 Schubert hypersurfaces together with a union of two Schubert hypersurfaces in the 12-dimensional $\text{Grass}_{\mathbb{C}}(2, 8)$. The problem is to compute the number of points in the intersection.

It follows from Schubert calculus [31] that if the hypersurfaces are in the "general position," then the number of points in the intersection remains invariant with respect to perturbations. The intersection number is computed by considering the homology class σ of the Schubert hypersurfaces as elements of $H_{22}^*(\text{Grass}(m, m+p), \mathbb{Z})$ and

computing $2\sigma^{12}$ in the above homology. By Schubert calculus [31], σ^{12} is the generic intersection of 12 Schubert hypersurfaces in $\text{Grass}_{\mathbb{C}}(2, 8)$ and is given by $d(2, 6)$, where

$$(4.2.14) \quad d(m, p) = \frac{1! \cdots (p-1)! 1! \cdots (m-1)! (mp)!}{1! \cdots (m+p-1)!}.$$

Thus, if we assume general position, there are at most 264 complex compensators which place the poles. Note that not all of the 264 compensators are necessarily finite; some may even be in the base locus.

Remark 4.2.5. In general, ascertaining whether or not certain points in the intersection of several Schubert hypersurfaces of a Grassmannian are indeed in the base locus or are infinite points is difficult. In the remainder of this section we shall study this problem under the special cases $\min(m, p) = 1$, and in § 5 we shall assume $\min(m, p) \geq 1$, $q = 0$.

4.3. Analysis of simultaneous pole placement map under the special case $\min(m, p) = 1$. Let us now reinterpret Theorem 4.2.3 under the special case when $\min(m, p) = 1$. We shall see that for this special case, simple linear algebraic techniques enable us to ascertain whether or not (4.2.12) is nonempty. Without any loss of generality let us assume that $m = 1$. Otherwise we can transpose the plants and the compensator. Consider an r -tuple of $p \times 1$ plants of McMillan degree n_i , $i = 1, \dots, r$, given by

$$(4.3.1) \quad \left[\sum_{j=0}^{n_i} d_{1,j}^i s^j, \dots, \sum_{j=0}^{n_i} d_{p,j}^i s^j \right]^T \left[\sum_{j=0}^{n_i} d_{p+1,j}^i s^j \right]^{-1}$$

for $i = 1, \dots, r$. Consider the $1 \times p$ compensator given by

$$(4.3.2) \quad \left[\left[\sum_{j=0}^q a_{1,j} s^j, \dots, \sum_{j=0}^q a_{p,j} s^j \right]^T \left[\sum_{j=0}^q a_{p+1,j} s^j \right]^{-1} \right]^T.$$

We now ask the following simultaneous pole assignability question.

Question 4.3.1. Do there exist suitable choices of the coefficients $a_{k,j}$, $j = 0, \dots, q$; $k = 1, \dots, p+1$ such that

$$(4.3.3) \quad \sum_{k=1}^{p+1} \left[\sum_{j=0}^{n_i} d_{k,j}^i s^j \right] \left[\sum_{j=0}^q a_{k,j} s^j \right] \equiv \sum_{j=0}^{n_i+q} c_{ij} s^j$$

for all $i = 1, \dots, r$ and where c_{ij} s are defined in (4.2.9)?

We would now like to write (4.3.3) as an equation that involves only the coefficients. Consider the notation

$$(4.3.4) \quad \underline{a} = a_i, \dots, a_{p+1}$$

where

$$(4.3.5) \quad \underline{a}_k = a_{k,0}, \dots, a_{k,q}.$$

Furthermore let us define for $k = 1, \dots, p+1$, the following:

$$(4.3.6) \quad D_k^i = \begin{bmatrix} d_{k,0}^i & \cdots & d_{k,n_i}^i \\ d_{k,0}^i & & d_{k,n_i}^i \end{bmatrix} \begin{matrix} \uparrow \\ q+1 \\ \downarrow \end{matrix},$$

$$(4.3.7) \quad \underline{D}^i = [(D_1^i)^T \cdots (D_{p+1}^i)^T]^T,$$

and

$$(4.3.8) \quad \underline{s}_{ij} = [1 \quad s_{ij} \quad \cdots \quad s_{ij}^{n_i+q}].$$

We may now write (4.3.3) as

$$(4.3.9) \quad \underline{a} \underline{D}^i \underline{s}_{ij}^T = \sum_{l=0}^{n_i+q} c_{il} s_{ij}^l = 0$$

for $j = 1, \dots, n_i + q$; $i = 1, \dots, r$.

It follows from (4.3.9) that a necessary condition for the existence of a compensator which places the poles of each of the plants $G_i(s)$ at distinct conjugate points s_{ij} , $j = 1, \dots, n_i + q$ for all $i = 1, \dots, r$ is given by

$$(4.3.10) \quad \bigcap_{i=1}^r \bigcap_{j=1}^{n_i+q} \text{Ker } \underline{D}^i \underline{s}_{ij} \neq \emptyset.$$

Condition (4.3.10) is a necessary and sufficient condition for solvability of (4.3.9) for a suitable vector \underline{a} . However, the vector \underline{a} may correspond to a compensator (4.3.2) in the base locus B which needs to be avoided.

The base locus is the set of compensators for which $K_i(s) \equiv 0$ (defined in (4.2.9)) for some $i = 1, \dots, r$ and is given by

$$(4.3.11) \quad \bigcup_{j=1}^r \text{Ker } \underline{D}^j.$$

Analogously to Theorem 4.2.3 it follows that we have Theorem 4.3.2.

THEOREM 4.3.2. Assume $\min(m, p) = 1$. An r -tuple of proper plants (4.3.1) of degree n_i , $i = 1, \dots, r$ is simultaneously pole assignable by a proper, or possibly improper, compensator (4.3.2) of degree q at conjugate set of $n_i + q$ complex points for all $i = 1, \dots, r$ if and only if

$$(4.3.12) \quad \bigcap_{i=1}^r \bigcap_{j=1}^{n_i+q} \text{Ker } \underline{D}^i \underline{s}_{ij} - \bigcup_{i=1}^r \text{Ker } \underline{D}^i$$

is nonempty. \square

Remark 4.3.3. The fact that (4.3.12) is nonempty does not imply that it contains a finite compensator. In fact we must ensure separately that $a_{p+1,q} \neq 0$ in order for the associated compensator to be finite. This is done in § 4.4.

Finally we describe the base locus condition purely as a rank computation under the case $\min(m, p) = 1$ and derive a necessary and sufficient condition when the base locus is empty.

THEOREM 4.3.4. Assume $\min(m, p) = 1$. The base locus associated with the pole placement map (4.2.1) for a generic r -tuple of plants is empty if and only if

$$(4.3.13) \quad (q+1) \max(m, p) \leq \min_{i \in \{1, \dots, r\}} n_i.$$

Proof. The matrix \underline{D}^i is of order $(q+1)(m \times p) \times (n_i + q + 1)$ and is of full rank for a generic r -tuple of plants (see [12], [17] for a proof). Clearly $\text{Ker } \underline{D}^i$ is nonempty for some i if

$$(4.3.14) \quad (q+1)(m+p) > n_i + q + 1$$

for some $i = 1, \dots, r$. \square

Remark 4.3.5. Since a necessary condition for simultaneous pole placement of a generic r -tuple of $\min(m, p) = 1$ plants of McMillan degrees n_i , $i = 1, \dots, r$, respectively, is given by

$$(4.3.15) \quad qm + qp + mp \geq \sum n_i + rq$$

we can infer from Theorem 4.3.4 that in most cases of interest, i.e., when (4.3.15) is satisfied, the base locus is nonempty.

To sum up, so far in this section we have studied the simultaneous pole placement problem as an intersection problem in an appropriate compactified space of compensators $\overline{S}_{p,m}^q$. The set of compensators Ω which indeed places the poles simultaneously would in general include infinite compensators and compensators in the base locus. In order to obtain a finite compensator we must exclude the infinite compensators and the base locus from the set Ω . Furthermore we have also analyzed the case $\min(m, p) = 1$ and obtained descriptions of the set Ω , and the base locus B , explicitly.

4.4. Simultaneous pole placement of proper plants by dynamic compensation. In this section we propose to address the simultaneous pole placement Problem 4.1.1. In particular we obtain an upper bound for the degree of the simultaneously pole assignable compensator as a function of n_i , m , p and r . The main result of this section is now stated as follows.

THEOREM 4.4.1. *A generic r -tuple of proper plants $(G_1(s), \dots, G_r(s)$, $\deg G_i(s) = n_i$, $i = 1, \dots, r$, can be pole assigned arbitrarily by a dynamic compensator provided*

$$(4.4.1) \quad r \leq \max(m, p).$$

Indeed if (4.4.1) holds then we have the following:

(a) *A compensator of McMillan degree q would assign an arbitrary r set of $n_i + q$ self-conjugate poles, $i = 1, \dots, r$ provided*

$$(4.4.2) \quad (q+1)[\max(m, p) + 1 - r] > \sum_{i=1}^r n_i.$$

(b) *A compensator of McMillan degree q would assign an arbitrary r set of $n_i + q$ self-conjugate poles, $i = 1, \dots, r$ except perhaps a proper algebraic subset, provided that*

$$(4.4.3) \quad (q+1)[\max(m, p) + 1 - r] \geq \sum_{i=1}^r n_i + (1 - r).$$

It may be noted that even when $r = 1$, (4.4.2) and (4.4.3) are different. This distinction is fundamental and is a reflection of the fact that since the pole placement map is not continuous in general, the exact pole placement and the approximate pole placement problems are different. Thus a generic pair of 1×2 plants of degree one can be pole assignable by a gain feedback almost everywhere. However, for specific pole locations, the associated gain may either be infinite or is in the base locus and therefore cannot be assigned. Indeed, for $r = 1$ we have the following corollary.

COROLLARY 4.4.2. *A generic $p \times m$ proper plant of degree n is pole assignable if*

$$(4.4.4) \quad (q+1) \max(m, p) \geq n + 1$$

and pole assignable almost everywhere if

$$(4.4.5) \quad (q+1) \max(m, p) \geq n.$$

In order to prove Theorem 4.4.1, we need the following lemma which also indicates that (4.4.2), (4.4.3) represent tight bounds.

LEMMA 4.4.3. *Assume $\min(m, p) = 1$. A necessary and sufficient condition for simultaneous pole placement of a generic r -tuple of proper plants is given by (4.4.2). A necessary and sufficient condition for simultaneous pole placement of a generic set of poles for a generic r -tuple of proper plants is given by (4.4.3).*

We now proceed to prove Lemma 4.4.3.

Proof of Lemma 4.4.3. We assume $m = 1$ without any loss of generality and consider the r -tuple of plants given by (4.3.1). There exists a compensator of type (4.3.2) which places the poles of the i th plant $G_i(s)$ at $s_{i,1}, \dots, s_{i,n_i+q}$ provided that, in the notation of § 4.3,

$$(4.4.6) \quad \underline{a}[\underline{D}^1 \underline{D}^2 \dots \underline{D}^r] = [c_{1,0}, \dots, c_{1,n_1+q}, c_{2,0}, \dots, c_{2,n_2+q}, \dots, c_{r,0}, \dots, c_{r,n_r+q}]$$

can be solved for a suitable vector of compensator parameters \underline{a} . By a simple algebraic manipulation, we might describe the vector \underline{a} which satisfies (4.4.6) as the set

$$(4.4.7) \quad S_1 - S_2$$

where

$$(4.4.8) \quad S_1 = \bigcap_{i=1}^r \text{Ker } \tilde{D}^i,$$

$$(4.4.9) \quad S_2 = \bigcup_{i=1}^r \text{Ker } D^i,$$

$$(4.4.10) \quad \tilde{D}^i = \text{col}(\tilde{Q}_1^i, \dots, \tilde{Q}_{p+1}^i),$$

$$(4.4.11) \quad \tilde{Q}_k^i = \text{diag}(c_{i,n_i+q}, \dots, c_{i,n_i+q}) \begin{bmatrix} d_{k0}^i & \dots & d_{kn_i}^i & 0 \\ & d_{k0}^i & \dots & d_{kn_i}^i \\ 0 & d_{k0}^i & \dots & d_{kn_i-1}^i \end{bmatrix} - [0, 0, \dots, 0, d_{kn_i}^i]^T [c_{i0}, \dots, c_{i,n_i+q}].$$

Note that the matrix D^i has already been defined in (4.3.7). Furthermore, note that the set S_2 parameterizes the compensators in the base locus and that $S_1 - S_2$ contains finite or infinite compensators which simultaneously place the poles.

Now that we have described the set $S_1 - S_2$, the following two questions seem to be relevant.

Question 4.4.4. Is the set $S_1 - S_2$ nonempty?

Question 4.4.5. Does the set $S_1 - S_2$ contain vector \underline{a} for which $a_{m+p,q} \neq 0$?

Of course Question 4.4.4 would ascertain if there indeed exists a finite or infinite compensator which simultaneously places the poles. Question 4.4.5 would analyze the possibility of such a compensator being finite. The following two lemmas answer the above two questions.

LEMMA 4.4.6. *The set $S_1 - S_2$ is nonempty if and only if*

$$(4.4.12) \quad \text{column span of } D^i \not\subset \text{column span of } [\tilde{D}^1 \dots \tilde{D}^r]$$

for all $i = 1, \dots, r$.

Proof. (Necessity.) Suppose the contrary. Then there exists $i = i_0$ such that

$$(4.4.13) \quad \text{column span of } D^{i_0} \subset \text{column span of } [\tilde{D}^1 \dots \tilde{D}^r].$$

Hence,

$$(4.4.14) \quad \text{Ker } D^{i_0} \supset \text{Ker } [\tilde{D}^1, \dots, \tilde{D}^r] = S_1.$$

Therefore,

$$(4.4.15) \quad S_1 \subset S_2$$

so that $S_1 - S_2$ is empty.

(Sufficiency.) By assumption,

$$(4.4.16) \quad S_1 \not\subset \text{Ker } D^i$$

for all $i = 1, \dots, r$, so that $S_1 \not\subset S_2$. Hence, $S_1 \cap S_2$ is a proper subspace of S_1 . Thus $S_1 - S_2$ is nonempty. \square

LEMMA 4.4.7. *The set $S_1 - S_2$ contains a vector g for which $a_{m+p,q} \neq 0$ if and only if*

$$S_1 - S_2 \neq \emptyset$$

and

$$(4.4.17) \quad (0, 0, \dots, 0, 1)^T \notin \text{column span of } [\tilde{D}^1, \dots, \tilde{D}^r].$$

Proof. (Necessity.) $S_1 - S_2 \neq \emptyset$ is clearly a necessary condition. Assume on the other hand that the vector $(0, \dots, 0, 1)^T \in \text{column span of } [\tilde{D}^1, \dots, \tilde{D}^r]$. Every element of $S_1 - S_2$ is orthogonal to column span of $[\tilde{D}^1, \dots, \tilde{D}^r]$ and hence orthogonal to $(0, \dots, 0, 1)^T$. Thus if $g \in S_1 - S_2$, then $a_{m+p,q} = 0$.

(Sufficiency.) Suppose on the contrary that every vector g of $S_1 - S_2$ is such that $a_{m+p,q} = 0$. Hence either $(0, \dots, 0, 1)^T \perp S_1 - S_2$ or $S_1 - S_2$ is empty. If $S_1 - S_2$ is non-empty then $S_1 \cap S_2$ is a proper subspace of S_1 and therefore $(0, 0, \dots, 0, 1)^T \perp S_1$ or equivalently $(0, 0, \dots, 0, 1)^T \in \text{column span of } (\tilde{D}^1, \dots, \tilde{D}^r)$. \square

The proof of Lemma 4.4.3 now follows.

LEMMA 4.4.8. *For some $i \in \{1, 2, \dots, r\}$ if there exists $v \in \text{column span of } D^i$, then there exists some choice of $c_{i,j}$ for which $v \in \tilde{D}^i$.*

Proof. Let $f_0^i, \dots, f_{n_i+q}^i$ be the columns of D^i . Let the given vector v be described as

$$(4.4.18) \quad v = \alpha_0 f_0^i + \dots + \alpha_{n_i+q} f_{n_i+q}^i$$

for some choice of $\alpha_0, \dots, \alpha_{n_i+q}$. Let $\tilde{f}_0^i, \dots, \tilde{f}_{n_i+q-1}^i$ be the columns of \tilde{D}^i , where

$$(4.4.19) \quad \tilde{f}_j^i = c_{i,n_i+q} f_j^i - c_{i,j} f_{n_i+q}^i$$

for $j = 0, \dots, n_i + q - 1$. Writing (4.4.19) as

$$(4.4.20) \quad f_j^i = -\frac{1}{c_{i,n_i+q}} \tilde{f}_j^i + \frac{c_{i,j}}{c_{i,n_i+q}} f_{n_i+q}^i$$

for $j = 0, \dots, n_i + q - 1$ and substituting (4.4.20) in (4.4.18), we have

$$(4.4.21) \quad v = \sum_{j=0}^{n_i+q-1} \left[-\frac{\alpha_j \tilde{f}_j^i}{c_{i,n_i+q}} + \frac{\alpha_j c_{i,j}}{c_{i,n_i+q}} f_{n_i+q}^i \right] + \alpha_{n_i+q} f_{n_i+q}^i.$$

It suffices to show therefore the existence of $c_{i,j}$ such that

$$(4.4.22) \quad \sum_{j=0}^{n_i+q} \alpha_j c_{i,j} = 0.$$

Of course (4.4.22) can be satisfied by some choice of $c_{i,j}$. \square

LEMMA 4.4.9. *Assume $\min(m, p) = 1$. A generic r -tuple of proper plants can be simultaneously pole assigned by a proper compensator if and only if (4.4.2) is satisfied.*

Proof. (Sufficiency.) For a generic r -tuple of plants $[D^1 \dots D^r]$ is a matrix of full rank. Thus

$$(4.4.23) \quad \text{column span of } D^i \not\subset \text{column span of } [\tilde{D}^1, \dots, \tilde{D}^r]$$

for all values of $c_{1,0}, \dots, c_{1,n_1+q}, \dots, c_{r,0}, \dots, c_{r,n_r+q}$ and for all $i = 1, \dots, r$. Hence by Lemma 4.4.6, $S_1 - S_2$ is nonempty for the right-hand side of (4.4.6). Moreover, since (4.4.2) is satisfied, we may conclude for a generic r -tuple of plants that

$$(4.4.24) \quad (0, 0, \dots, 0, 1)^T \notin \text{column span of } (D^1, \dots, D^r)$$

so that the hypothesis of Lemma 4.4.7 is also satisfied for the right-hand side of (4.4.6). (Necessity.) If

$$(4.4.25) \quad (q+1)[\max(m, p) + 1 - r] \leq \sum_{i=1}^r n_i,$$

then for a generic r -tuple of plants

$$(4.4.26) \quad (0, \dots, 0, 1)^T \in \text{column span of } [D^1, \dots, D^r].$$

Let v_i be the projection of $(0, \dots, 0, 1)^T$ on the column span of D^i . By Lemma 4.4.8 $v_i \in \text{column span of } \tilde{D}^i$ for some choice of $c_{i,j}$, $i = 1, \dots, r$, $j = 0, \dots, n_i + q$. Hence

$$(4.4.27) \quad [0, \dots, 0, 1]^T \in \text{column span of } [\tilde{D}^1, \dots, \tilde{D}^r]$$

for some choice of $c_{i,j}$, $i = 1, \dots, r$, $j = 0, \dots, n_i + q$. It follows from Lemma 4.4.7 that there does not exist a finite proper compensator which places the poles corresponding to the above choice of $c_{i,j}$. \square

LEMMA 4.4.10. Assume $\min(m, p) = 1$. A generic r -tuple of proper plants can be all simultaneously pole assigned, with the possible exception of a proper algebraic subset of poles, by a proper compensator if and only if (4.4.3) is satisfied.

Proof. (Necessity.) Let us assume that (4.4.3) is not satisfied. For a fixed $i = i_0$, we have

$$(4.4.28) \quad \begin{aligned} & \dim [\text{column span of } D^{i_0} \cap \text{column span of } (\tilde{D}^1, \dots, \tilde{D}^r)] \\ & \geq \dim [\text{column span of } D^{i_0}] + \dim [\text{column span of } (\tilde{D}^1, \dots, \tilde{D}^r)] \\ & \quad - (q+1)(\max(m, p) + 1). \end{aligned}$$

It has been shown in [12] that for a generic set of $c_{i,j} - s$,

$$\dim [\text{column span of } (\tilde{D}^1, \dots, \tilde{D}^r)] = \sum n_i + rq.$$

From the assumption we conclude that

$$(4.4.29) \quad \dim [\text{column span of } (\tilde{D}^1, \dots, \tilde{D}^r)] - (q+1)(m+p) \geq 0.$$

Thus from (4.4.28) and (4.4.29) it may be inferred that generically

$$(4.4.30) \quad [\text{column span of } D^{i_0}] \subset [\text{column span of } (\tilde{D}^1, \dots, \tilde{D}^r)].$$

Hence it follows from Lemma 4.4.6 that for a generic $c_{i,j}$, $S_1 - S_2$ is empty. Hence a pole assigning compensator does not exist.

(Sufficiency.) It is necessary to show that $S_1 - S_2$ is nonempty for a generic $c_{i,j}$. Suppose on the contrary that

$$(4.4.31) \quad \text{column span of } D^{i_0} \subset \text{column span of } [\tilde{D}^1, \dots, \tilde{D}^r]$$

for some $i = i_0$. For a generic $c_{i,j}$, $j = 0, \dots, n_i + q$

$$(4.4.32) \quad \text{column span of } D^{i_0} = \text{column span of } [\tilde{D}^{i_0}, v_0]$$

where

$$(4.4.33) \quad v_0 = (0, \dots, 0, d_{1,n_{i_0}}^{i_0}, \dots, 0, \dots, 0, d_{r,n_{i_0}}^{i_0})^T$$

or equivalently the matrix

$$(4.4.34) \quad [v_0, \tilde{D}^1, \dots, \tilde{D}^r]$$

is not of full rank. However, for a generic $c_{i,j}$, (4.4.34) is of full rank which is a contradiction. Thus $S_1 - S_2$ is nonempty for a generic $c_{i,j}$. Moreover, for an algebraic set of $c_{i,j}$

$$(4.4.35) \quad [0, \dots, 0, 1]^T \in \text{column span of } (\tilde{D}^1, \dots, \tilde{D}^r).$$

This algebraic set misses the point $c_{i,n_i+q} = 1$, $c_{i,j} = 0$ for all $j = 0, \dots, n_i$, $i = 1, \dots, r$, so that it is a proper algebraic set. Thus for a generic set of $c_{i,j}$

$$(4.4.36) \quad (0, 0, \dots, 0, 1)^T \notin \text{column span of } [\tilde{D}^1, \dots, \tilde{D}^r].$$

Thus by Lemma 4.4.7, $S_1 - S_2$ does indeed contain, a vector \underline{a} for which $a_{m+p,q} \neq 0$. Thus there exists a finite (proper) compensator in $S_1 - S_2$. \square

The proof of Lemma 4.4.3 clearly follows from Lemmas 4.4.9 and 4.4.10.

Proof of Theorem 4.4.1. In order to prove Theorem 4.4.1 we need the results of the following two lemmas which are now stated without proof.

LEMMA 4.4.11 [40]. *Given an r -tuple of $p \times m$ plants $G_i(s)$ of degree n_i , each with n_i simple poles, there is an open dense set of $1 \times p$ vectors $v \in \mathbb{R}^p$ such that $vG_i(s)$ has degree n_i for all i .*

LEMMA 4.4.12 [17]. *Given an r -tuple of $p \times m$ plants $G_i(s)$, there exists a constant gain output feedback K such that the closed-loop systems $G_i(s)[I + KG_i(s)]^{-1}$ have distinct simple poles.*

Theorem 4.4.1 now follows by choosing $(v, K) \in \mathbb{R}^p \times \mathbb{R}^{mp}$. We have a mapping from an open dense set

$$(4.4.37) \quad \Phi_{(v,K)}: \sum_{m,p}^{n_1} \times \dots \times \sum_{m,1}^{n_r} \rightarrow \sum_{m,1}^{n_1} \times \dots \times \sum_{m,1}^{n_r},$$

$$(4.4.38) \quad \Phi_{(v,K)}(G_i(s))_{i=1}^r = (vG_i(s)[I + KG_i(s)]^{-1})_{i=1}^r$$

which is rational in the Hankel parameters (H_{ij}) of (G_i) . Applying Lemmas 4.4.11, 4.4.12, and (4.4.3) to the case $\min(m, p) = 1$, i.e.,

$$(4.4.39) \quad \sum_{m,1}^{n_1} \times \dots \times \sum_{m,1}^{n_r}$$

gives, via composition with Φ , an open dense set of

$$(4.4.40) \quad \sum_{m,p}^{n_1} \times \dots \times \sum_{m,p}^{n_r}$$

which can be simultaneously pole assigned. \square

4.5. Examples. In this section we consider three illustrative examples.

Example 4.5.1. Let $r = n = m = p = 1$ and $q = 0$, so that (4.4.5) is satisfied and by Corollary 4.4.2 almost all poles can be placed. Let the plant be written as

$$(4.5.1) \quad \frac{d_{10} + d_{11}s}{d_{20} + d_{21}s}$$

and the compensator be written as a_{10}/a_{20} . The sets S_1 and S_2 defined by (4.4.8) and (4.4.9) may be described as

$$(4.5.2) \quad S_1 = \text{Ker} \begin{bmatrix} c_{11}d_{10} - c_{10}d_{11} \\ c_{11}d_{20} - c_{10}d_{21} \end{bmatrix},$$

$$(4.5.3) \quad S_2 = \text{Ker} \begin{bmatrix} d_{10} & d_{11} \\ d_{20} & d_{21} \end{bmatrix}$$

where

$$(4.5.4) \quad c_{10} + c_{11}s$$

is the associated return difference equation. For a generic plant, S_2 contains the trivial vector so that

$$(4.5.5) \quad S_1 - S_2$$

is nonempty and the base locus is empty. Moreover when

$$(4.5.6) \quad c_{11}d_{10} \neq c_{10}d_{11}$$

the condition of Lemma 4.4.7 is satisfied and there exists a finite compensator which places the poles. Condition (4.5.6) implies that the pole $-d_{10}/d_{11}$ is not assignable and hence, almost all but not all, poles can be placed.

Example 4.5.2. Let us consider the case $r = n = m = p = q = 1$, so that from Corollary 4.4.2 it follows that a generic plant is pole assignable. Let the plant be written as

$$(4.5.7) \quad (d_{10} + d_{11}s)/(d_{20} + d_{21}s)$$

and the compensator be written as

$$(4.5.8) \quad (a_{10} + a_{11}s)/(a_{20} + a_{21}s).$$

The sets S_1, S_2 are given by

$$(4.5.9) \quad S_1 = \text{Ker} \begin{bmatrix} c_{12}d_{10} & c_{12}d_{11} \\ c_{12}d_{20} & c_{12}d_{21} \\ -c_{10}d_{11} & c_{12}d_{10} - c_{11}d_{11} \\ -c_{10}d_{21} & c_{12}d_{20} - c_{11}d_{21} \end{bmatrix},$$

$$(4.5.10) \quad S_2 = \text{Ker} \begin{bmatrix} d_{10} & d_{11} & 0 \\ d_{20} & d_{21} & 0 \\ 0 & d_{10} & d_{11} \\ 0 & d_{20} & d_{21} \end{bmatrix}$$

where the return difference equation is given by

$$(4.5.11) \quad c_{10} + c_{11}s + c_{12}s^2 = 0.$$

For a generic plant, S_2 is nonempty and hence the base locus is nonempty and is of dimension 1. Since S_1 is a subspace of dimension 2 we infer that $S_1 - S_2$ is nonempty. Moreover, for a generic plant the matrix

$$(4.5.12) \quad \begin{bmatrix} 0 & d_{10} & d_{11} & 0 \\ 0 & d_{20} & d_{21} & 0 \\ 0 & 0 & d_{10} & d_{11} \\ 1 & 0 & d_{20} & d_{21} \end{bmatrix}$$

is nonsingular and therefore the condition (4.4.17) is satisfied. Hence there exists a finite compensator which places the poles.

Example 4.5.3. Consider the case $r = m = 2$, $q = 0$, $p = n_1 = n_2 = 1$. Since the condition (4.4.3) is satisfied, it follows from Theorem 4.4.1 that almost all poles can be simultaneously assignable for a generic pair of plants. Let the plants be

$$(4.5.13) \quad [d_{30}^i + d_{31}^i s]^{-1} [d_{10}^i + d_{11}^i s \quad d_{20}^i + d_{21}^i s]$$

for $i = 1, 2$ and let the compensator be

$$(4.5.14) \quad [a_{10}/a_{30} \quad a_{20}/a_{30}].$$

The sets S_1 and S_2 are given by

$$(4.5.15) \quad S_1 = \text{Ker} \begin{bmatrix} c_{11}d_{10}^1 - c_{10}d_{11}^1 & c_{21}d_{10}^2 - c_{20}d_{11}^2 \\ c_{11}d_{20}^1 - c_{10}d_{21}^1 & c_{21}d_{20}^2 - c_{20}d_{21}^2 \end{bmatrix}$$

and

$$(4.5.16) \quad S_2 = \text{Ker} \begin{bmatrix} d_{10}^1 & d_{11}^1 \\ d_{20}^1 & d_{21}^1 \\ d_{30}^1 & d_{31}^1 \end{bmatrix} \cup \text{Ker} \begin{bmatrix} d_{10}^2 & d_{11}^2 \\ d_{20}^2 & d_{21}^2 \\ d_{30}^2 & d_{31}^2 \end{bmatrix}.$$

The base locus is a union of two subspaces of dimension 1 for a generic pair of plants. It may also be checked that generically condition (4.4.17) is satisfied so that almost all poles can indeed be placed.

4.6. The stable stabilization problem. It has been shown that to simultaneously stabilize a pair of plants is equivalent to solving a well-known problem considered by Youla, Bongiorno, and Lu [43]: When can a single plant be stabilized by a stable compensator? Motivated by the solution of Youla, Bongiorno, and Lu [43], we consider the following question.

Question 4.6.1. When is a stabilizing compensator of a multi-input multi-output plant stable?

As pointed out by Vidyasagar and Viswanadham [42], the problem of simultaneously stabilizing $r + 1$ plants by a stable or unstable compensator is equivalent to the problem of simultaneously stabilizing r plants using a stable compensator (see also Saeks and Murray [37] for $r = 1$). It may be pointed out that Question 4.6.1 is interesting in its own right since the unstable compensator might result in poor overall sensitivity of the feedback system to variations in plant parameters.

The following theorem is now stated and proved as an immediate consequence of Theorem 4.4.1 (see [12] for details).

THEOREM 4.6.2. *A generic r -tuple of proper plants $(G_1(s), \dots, G_r(s))$, $\deg G_i(s) = n_i$, $i = 1, \dots, r$ can be stabilized by a stable proper dynamic compensator if*

$$(4.6.1) \quad r < \max(m, p).$$

Indeed if (4.6.1) holds then it is enough to choose a compensator of McMillan degree = q , where q satisfies

$$(4.6.2) \quad q[\max(m, p) - r] + \max(m, p) \geq \sum_{i=1}^r n_i.$$

It may be noted that the bound (4.6.1) is consistent with the following lemma due to Vidyasagar and Viswanadham [42].

LEMMA 4.6.3 (Vidyasagar and Viswanadham [42]). *A generic r -tuple of plants is stabilizable generically if and only if a generic $(r - 1)$ -tuple of plants is stabilized by a stable compensator.*

Due to the above lemma the following corollary can be stated.

COROLLARY 4.6.4. Assume $\min(m, p) = 1$. A necessary and sufficient condition for generic stabilization by a stable compensator of an r -tuple of proper or strictly proper plants is given by (4.6.1).

Proof of Theorem 4.6.2. Consider the compensator in (4.3.2) with the choice of

$$(4.6.3) \quad (a_{m+p,0}, \dots, a_{m+p,q})$$

in such a way that

$$(4.6.4) \quad \sum_{j=0}^q a_{m+p,j} s^j$$

is a stable polynomial. By fixing the vector $(a_{m+p,0}, \dots, a_{m+p,q})$ we can obtain the associated pole placement map χ

$$(4.6.5) \quad \chi: \mathbb{R}^{(q+1)m} \rightarrow \mathbb{R}^{\sum n_i + r(q+1)},$$

$$(4.6.6) \quad \chi(a_{1,0}, \dots, a_{1,q}, \dots, a_{m,0}, \dots, a_{m,q}) = (c'_{1,0}, \dots, c'_{r,n_r+q})$$

with an associated Sylvester matrix of order $(q+1)m$ by $\sum n_i + r(q+1)$. By a treatment analogous to the proof of Theorem 4.4.1 we compute the rank of the affine map χ and the result follows. \square

Proof of Corollary 4.6.4. Sufficiency is clear from Theorem 4.6.2. To prove necessity, let $r = \max(m, p)$. If stabilizability by a stable compensator is generically satisfied, by Lemma 4.6.3 there exists a generic $(m+p)$ -tuple of simultaneously stabilizable $\min(m, p) = 1$ plants. By Theorem 4.4.1 and Lemma 4.4.3, this is a contradiction. \square

5. Simultaneous pole placement of multimode linear dynamical systems by constant gain feedback.

5.1. Statement and motivation of the problem. In § 4 we considered the simultaneous pole placement problem by a dynamic compensator and analyzed the case $\min(m, n) = 1$ in considerable detail. In this section we consider gain feedback, i.e., $q = 0$ without the restriction $\min(m, p) = 1$. Specifically, we consider special cases of the Problem 4.1.1 described as follows.

Simultaneous pole placement problem 5.1.1. Given an r -tuple $G_1(s), \dots, G_r(s)$ of $p \times m$ real, proper transfer functions, $\deg G_i(s) = n_i$, $i = 1, \dots, r$, does there exist an $m \times p$ real gain K such that the closed-loop systems $G_1(s)[I + KG_1(s)]^{-1}, \dots, G_r(s) \times [I + KG_r(s)]^{-1}$ have poles in a prescribed r set of the conjugate set of complex numbers $s_{i,1}, s_{i,2}, \dots, s_{i,n_i+q}$, for $i = 1, \dots, r$?

A parameterization to the solution of Problem 5.1.1 above is obtained in [14]. In particular, the set of simultaneously pole assignable plants have been described qualitatively as a semialgebraic subset in the space of r -tuples of plants. However, in order to get an indication about the relative size of the space of simultaneously stabilizable r -tuples of plants as a subset of the space of r -tuples of plants we consider the following.

Generic simultaneous pole placement at a generic set of poles 5.1.2. Is it true that for all r -tuples $G_1(s), \dots, G_r(s)$ of $p \times m$ real, proper transfer functions, except perhaps those contained in a proper algebraic set, we can arbitrarily assign all but possibly a proper algebraic subset of poles of the closed-loop systems $G_i(s)[I + KG_i(s)]^{-1}$, $i = 1, 2, \dots, r$, by a real gain feedback K ?

First, it is quite evident that the conditions of pole placement involve solving a set of simultaneous polynomial equations. It is also clear that the existence of a solution to these equations is equivalent to the surjectivity of the associated pole placement map χ . In this chapter, the pole placement map is analyzed for the various special cases.

Consider an infinitesimal analysis of χ , viz., a calculation of the Jacobian $d\chi$ on k^{mp} , the mp -dimensional space of feedback gains and on a certain submanifold of k^{mp} . By adapting a technique due to Brockett and Byrnes [3] we can show the following.

THEOREM 5.1.3. *A generic r -tuple of strictly proper $p \times m$ real transfer functions of McMillan degree $= n_i, i = 1, \dots, r$, may be pole assigned arbitrarily if*

$$(5.1.1) \quad m + p - 1 \geq \sum_{i=1}^r n_i.$$

Moreover, a generic r -tuple of proper $p \times m$ transfer functions of McMillan degree $= n_i, i = 1, \dots, r$, may be pole assigned arbitrarily if

$$(5.1.2) \quad m + p - r - 1 \geq \sum_{i=1}^r n_i.$$

For the case of a single proper plant, Theorem 5.1.3 specializes to the following corollary.

COROLLARY 5.1.4 (Brockett and Byrnes [3]). *A generic strictly proper $p \times m$ real transfer function of McMillan degree $= n$ may be pole assigned arbitrarily if*

$$(5.1.3) \quad m + p - 1 \geq n.$$

COROLLARY 5.1.5 (Kimura [30]). *A generic strictly proper $p \times m$ real transfer function of McMillan degree $= n$ is pole assignable almost arbitrarily (i.e., a generic n -tuple of poles may be assigned) if*

$$(5.1.4) \quad m + p - 1 \geq n.$$

Using the infinitesimal computation, in particular by showing that the Jacobian of the pole placement map is nonsingular at a given gain K_0 and by applying the dominant morphism theorem over \mathbb{C} , a technique due to Hermann and Martin [26] has been adapted to show the following theorem.

THEOREM 5.1.6. *A generic r -tuple of $p \times m$ strictly proper complex transfer functions may be assigned a generic $\sum_{i=1}^r n_i$ -tuple of complex poles by a complex gain feedback K if and only if*

$$(5.1.5) \quad mp \geq \sum_{i=1}^r n_i.$$

COROLLARY 5.1.7 (Hermann and Martin [26]). *A generic $p \times m$ strictly proper complex transfer function may be assigned a generic n -tuple of complex poles by a complex gain feedback K if and only if*

$$(5.1.6) \quad mp \geq n.$$

Of course over \mathbb{R} , (5.1.5) and (5.1.6) give rise to necessary conditions for arbitrary pole assignment.

In an earlier paper [3], Brockett and Byrnes have shown that the number of complex gains which places a distinct self-conjugate set of n poles for a nondegenerate strictly proper $p \times m$ transfer function assuming $mp = n$, counted with multiplicity, is given by $d(m, p)$ defined in (4.2.14). In this section we describe nondegenerate and twisted r -tuples of proper plants and generalize the results presented by Brockett and Byrnes [3] as follows. Consider a set of distinct complex numbers $s_{ij}, i = 1, \dots, r, j = 1, \dots, n_i$, where for any $i = i_0$ the set $\{s_{i_0j}\}$ is self-conjugate.

THEOREM 5.1.8. *Assuming $mp = \sum_{i=1}^r n_i$ and that there exists an r -tuple of proper plants that are nondegenerate and twisted at s_{ij} , $i = 1, \dots, r$, $j = 1, \dots, n_i$ (to be defined in Definitions 5.3.4 and 5.3.5), the set of complex gains that places the poles s_{ij} for a generic r -tuple of real proper plants is finite and the number is given by $d(m, p)$ (defined in (4.2.14)).*

In particular, the number (4.2.14) is odd if and only if (5.1.7) is satisfied, and we have the following.

COROLLARY 5.1.9. *Under the hypothesis of Theorem 5.1.8, it is possible to assign any arbitrary r set of the self-conjugate set of n_i distinct poles, $i = 1, \dots, r$, for a generic r -tuple of real proper plants by a real feedback gain if and only if*

$$(5.1.7) \quad \min(m, p) = 1 \quad \text{or} \quad \min(m, p) = 2 \quad \text{and} \quad \max(m, p) = 2^r - 1.$$

Remark 5.1.10. The intersection number $d(m, p)$ counts the number of self-conjugate complex gains (finite or infinite) which places the poles provided that the associated Schubert hypersurfaces are in general position. In [3] it was shown that for a nondegenerate $p \times m$ strictly proper transfer function if $mp = n$ then the associated n Schubert hypersurfaces were indeed in general position. In this section we show that if $mp = \sum_{i=1}^r n_i$ and if the r -tuple of proper plants are nondegenerate and twisted, then the associated mp Schubert hypersurfaces are in general position and intersect in $d(m, p)$ finite gains.

In general it is difficult to ascertain if an r -tuple of proper plants is nondegenerate and twisted. However, we have the following.

LEMMA 5.1.11. *Assume $m = p = 2$, $n = 1$, $r = 4$. There exists an r -tuple of proper plants that are nondegenerate and twisted.*

In the above case $m = p = 2$, $n = 1$, $r = 4$ if the four Schubert hypersurfaces are in general position, there are $d(m, p) = 2$ complex gains which place the notes. Although two is not an odd number, in this case we are able to show the existence of a real gain which places the poles.

THEOREM 5.1.12. *Assume $m = p = 2$, $n = 1$, $r = 4$. There exists a finite gain K such that the generic simultaneous pole placement for an r -tuple of strictly proper plants is possible.*

In fact it is known from Schubert calculus that under the hypothesis of Theorem 5.1.12 there are exactly two conjugate hyperplanes that are either in the base locus or place the poles. The proof of the above theorem follows by showing that the base locus misses one of the two hyperplanes and that both the gains are real.

Consider, on the other hand, the following special pole placement problem with the assumption that

$$(5.1.8) \quad s_{i,j} = 0, \quad i = 1, \dots, r, \quad j = 1, \dots, n_i.$$

The special pole placement is known as the dead beat control problem, which is stated as follows.

Dead beat control problem 5.1.13. Given an r -tuple of proper plants, $G_1(s), \dots, G_r(s)$ each of a given McMillan degree n_i , $i = 1, \dots, r$, with m inputs, p outputs, does there exist a gain K such that $\det[I + KG_i(s)]$ have zeros only at $s = 0$ for every $i = 1, \dots, r$?

Notice that in the above problem the poles to be placed are not distinct. Hence the Schubert calculus as applied to Theorem 5.1.8 would not be directly applicable. A result due to Anderson and Byrnes [2] has been adapted to show the following theorem.

THEOREM 5.1.14. *A necessary condition for placing all the $\sum_{i=1}^r n_i$ poles of an r -tuple of proper $m \times p$ plants of McMillan degree n_i , $i = 1, \dots, r$, respectively, at the origin by a gain feedback K is given by*

$$(5.1.9) \quad mp \geq \sum_{i=1}^r n_i.$$

COROLLARY 5.1.15 (Anderson and Byrnes [2]). *A necessary condition for placing all the n poles of a generic $m \times p$ strictly proper plant $G(s)$ at the origin by a gain feedback K is given by $mp \geq n$.*

It is unknown whether (5.1.9) is a necessary condition for generic simultaneous stabilization as well. Present knowledge is limited to $\min(m, p) = 1$, in which case (5.1.9) is known to be a necessary and sufficient condition. On the other hand, by posing the generic simultaneous dead beat control problem, not as a Schubert intersection problem, but as an intersection problem in the associated homology ring of real or complex Grassmannians, we show the following theorem.

THEOREM 5.1.16. *A sufficient condition for simultaneous dead beat control by a real feedback gain of a generic r -tuple of m -input p -output real proper plants is given by $\min(m, p) = 1$ or $\min(m, p) = 2$, $\max(m, p) = 2^r - 1$ under the condition that $mp = \sum_{i=1}^r n_i$ and that there exists an r -tuple of nondegenerate and twisted proper plants.*

We remark that although we conjecture the existence of an r -tuple of nondegenerate and twisted proper plant whenever $mp = \sum_{i=1}^r n_i$, we have only shown this result when $\min(m, p) = 1$ or when $m = p = 2$, $r = 4$, $n_i = 1$, $i = 1, \dots, 4$.

The main results of this section are now summarized. First, by restricting the simultaneous pole placement map χ onto rank 1 gains and via infinitesimal analysis of χ we obtain a sufficient condition for simultaneous pole assignment of an r -tuple of proper/strictly proper plants by a feedback gain. Furthermore, using infinitesimal analysis of χ over the complex base field and using the dominant morphism theorem, we obtain a necessary and sufficient condition for simultaneous pole assignment by a complex feedback gain. Next we pose the simultaneous pole assignment problem as a Schubert intersection problem and show that if $mp = \sum_{i=1}^r n_i$, then for a nondegenerate and twisted r -tuple of plants, there exists $d(m, p)$ complex conjugate gains which place the poles. If $d(m, p)$ is odd there exists a real gain which places the poles. Even when $d(m, p)$ is not odd we show that for a special case there exists a real gain which places the poles. Finally we analyze the dead beat control problem as a particular case of the pole assignment problem.

5.2. Analysis of the simultaneous pole placement map via rank 1 gains. Let k denote the real or complex field. We now consider a specialization of the simultaneous pole placement map (4.2.1) assuming $q = 0$. To say that the gain K places the poles of $G_1(s), \dots, G_r(s)$ is to say that

$$(5.2.1) \quad \begin{aligned} \chi(K) &= (c'_{1,1}, \dots, c'_{1,n_1}, \dots, c'_{r,1}, \dots, c'_{r,n_r}), \\ c'_{ij} &= c_{ij}/c_{i0}, \quad i = 1, \dots, r, \quad j = 1, \dots, n_i \end{aligned}$$

is surjective. The coefficients $c_{i,j}$ are defined in (4.2.3) and are also the numerator of the rational function

$$(5.2.2) \quad \det[I + KG_i(s)] \quad \text{for all } i = 1, \dots, r.$$

Proof of Theorem 5.1.6. Let us represent generically the k th entry of $G_i(s)$ by $g_{k,j}^{(i)}(s)/g^{(i)}(s)$ where $g^{(i)}(s)$, $g_{k,j}^{(i)}(s)$ are polynomials of degree n_i . The derivative of the

pole placement map (4.2.1) is given by

(5.2.3)

$$d\chi|_{K_0}(K) = (\text{coefficients of the numerator of trace } [(K + K_0)G_i(s)], i = 1, \dots, r).$$

Let $G_{ij}, j = 0, 1, \dots$, be the Laurent series coefficients corresponding to $G_i(s)$. Regarding G_{ij} as vectors in k^{mp} , and expressing the map $d\chi|_{K_0}$ with respect to the standard basis for the space of $m \times p$ matrices, if and only if $G_{i0}, \dots, G_{i,n-1}$ are linearly independent, there exists a gain K such that $d\chi$ is of full rank at $K_0 = 0$. If $mp \geq \sum_{i=1}^r n_i$, it is possible to realize $G_1(s), \dots, G_r(s)$, $\deg G_i(s) = n_i$ so that $G_{i0}, \dots, G_{i,n-1}$ are linearly independent.

By the dominant morphism theorem [38], a necessary and sufficient condition that the image of χ is almost surjective for generic $G_i(s), i = 1, \dots, r$, is given by (5.1.5). \square

LEMMA 5.2.1. *For a $p \times m$ r -tuple of real proper plants $(G_1(s), \dots, G_r(s))$, $\deg G_i(s) = n_i$ chosen generically, there corresponds an r -tuple of generic $1 \times (m + p - 1)$ real proper transfer functions $\tilde{g}_i(s)$ of McMillan degree n_i , where $\tilde{g}_i(s)$ is obtained from*

$$(5.2.4) \quad (g_1^{(i)}, \dots, g_{m+p}^{(i)}) \triangleq [1, 0, \dots, 0 \mid 1, 0, \dots, 0] \begin{bmatrix} G_i(s) & 0 \\ 0 & G_i^T(s) \end{bmatrix}$$

by deleting $g_{m+1}^{(i)}$ for each $i = 1, \dots, r$.

Proof. A generic r -tuple of $p \times m$ proper plants with distinct simple poles may be written as

$$(5.2.5) \quad G_i(s) = \sum_{j=1}^{n_i} R_{i,j}/(s + s_{i,j}) + R_i^{(0)}$$

where $R_{i,1}, \dots, R_{i,n_i}$ is a self-conjugate set of complex rank 1 matrices for any $i = 1, \dots, r$ and $R_i^{(0)}$ is the value of $G_i(s)$ at infinity. From (5.2.4) and (5.2.5) we may write

$$(5.2.6) \quad (g_1^{(i)}, \dots, g_{m+p}^{(i)}) = \sum_{j=1}^{n_i} (k_2^{(0)} R_{ij} \quad k_1^{(0)} R_{ij}^T)/(s + s_{ij}) + (k_2^{(0)} R_i^{(0)} \quad k_1^{(0)} R_i^{(0)T})$$

where

$$(5.2.7) \quad k_1^{(0)T} k_2^{(0)} = \begin{bmatrix} 0 & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 0 & \cdot & \cdot & 0 \end{bmatrix}.$$

For a fixed $i \in \{1, \dots, r\}$ and for all $j = 1, \dots, n_i$ all the coefficients in $(k_2^{(0)} R_{i,j} \mid k_1^{(0)} R_{i,j}^T), (k_2^{(0)} R_i^{(0)} \mid k_1^{(0)} R_i^{(0)T})$, except the $(m+1)$ th coefficients, can be obtained arbitrarily for rank 1 matrices $R_{i,j}$ chosen generically. Indeed, for an r -tuple of plants $G_1(s), \dots, G_r(s)$ chosen generically, the matrices $R_{i,j}, i = 1, \dots, r, j = 1, \dots, n_i$, can be chosen generically so that (5.2.6) takes the form

$$(5.2.8) \quad (g_1^{(i)}, \dots, g_m^{(i)}, g_1^{(i)}, g_{m+2}^{(i)}, \dots, g_{m+p}^{(i)})$$

where

$$(5.2.9) \quad \tilde{g}_i(s) = (g_1^{(i)}, \dots, g_m^{(i)}, g_{m+2}^{(i)}, \dots, g_{m+p}^{(i)})$$

and corresponds to a generic $1 \times (m + p - 1)$ transfer function of McMillan degree n_i . \square

Proof of Theorem 5.1.3. Consider the following notation: If $p(s)/q(s)$ is a proper rational function of degree n , let us denote by $\{p(s)/q(s)\}$ the vector of n coefficients of the numerator of $p'(s)/q'(s) = p(s)/q(s)$ where $p'(s)$ is a monic polynomial of degree n .

Consider the restriction of the simultaneous pole placement map χ to the submanifold M of rank 1 matrices. Let us write a rank 1 matrix K in its dyadic representation

$$(5.2.10) \quad K = k_1^T k_2$$

where

k_1 is a $1 \times m$ nonzero vector,

k_2 is a $1 \times p$ nonzero vector.

Consider the r -tuple of proper plants $G_1(s), \dots, G_r(s)$. The map χ is written as

$$(5.2.11) \quad \chi(K) = (\{1 + k_2 G_i(s) k_1^T\}, i = 1, \dots, r).$$

Let M be a submanifold of rank 1 $m \times p$ matrices of dimension $m + p - 1$. The derivative $d\chi$ acting on $T_k M$ may be obtained as

$$(5.2.12) \quad d\chi|_{K_0}(K) = \left(\left\{ (1, 0, \dots, 0 | 1, 0, \dots, 0) \begin{bmatrix} G_i(s) & 0 \\ 0 & G_i^T(s) \end{bmatrix} \begin{bmatrix} k_1^T \\ k_2^T \end{bmatrix} \right\} \right),$$

$i = 1, \dots, r$

where K_0 is given by (5.2.7). By (5.2.6) and (5.2.12)

$$(5.2.13) \quad d\chi|_{K_0}(K) = [\{(g_1^{(i)}, \dots, g_{m+p}^{(i)})(k_1 \quad k_2)^T\}, i = 1, \dots, r].$$

By scaling the first entry of the vector k_2 as 1 and calling it $k_2' = (1, k_2'')$ it is possible to write, using (5.2.8) and (5.2.9),

$$(5.2.14) \quad d\chi|_{K_0}(K) = (\{\tilde{g}_i(s) k^T + g_1^{(i)}(s)\}, i = 1, \dots, r)$$

where

$$(5.2.15) \quad k = (k_1 \quad k_2'').$$

A generic $g_i(s)$ may be written as

$$(5.2.16) \quad \tilde{g}_i(s) = (g_1^{(i)'}, \dots, g_m^{(i)'}, g_{m+2}^{(i)'}, \dots, g_{m+p}^{(i)'}) / g_0^{(i)}$$

where

$$(5.2.17) \quad g_j^{(i)} = g_j^{(i)'}/g_0^{(i)}, \quad j = 1, \dots, m, m+2, \dots, m+p$$

$g_0^{(i)}, g_j^{(i)'}$ for all i, j are polynomials of degree n .

Using the notation in (5.2.16) it is possible to write

$$(5.2.18) \quad d\chi|_{K_0}(K) = \text{coefficients of} \\ (g_1^{(i)'}, \dots, g_m^{(i)'}, g_{m+2}^{(i)'}, \dots, g_{m+p}^{(i)'}) k^T + (g_1^{(i)}), i = 1, \dots, r).$$

Consider now the matrix representation of the linear map $d\chi|_{K_0}$ in the coefficients of $(g_1^{(i)'}, \dots, g_m^{(i)'}, g_{m+2}^{(i)'}, \dots, g_{m+p}^{(i)'})$. With respect to the standard basis, we obtain a matrix M_1 of order $(m + p - 1)$ by $(\sum n_i + r)$, for generic $g_j^{(i)'}$, and for all i, j , the matrix M_1 is of full rank. Hence for a generic r -tuple of proper plants it follows using Lemma 5.2.1 that $d\chi|_{K_0}$ is surjective if

$$(5.2.19) \quad m + p - 1 \geq \sum n_i + r.$$

If, on the other hand, $G_1(s), \dots, G_r(s)$ are strictly proper, then $g_j^{(i)}$ for all i, j are polynomials of degree $n-1$ so that the corresponding matrix M is of order $(m+p-1)$ by $(\sum n_i)$. Hence for a generic r -tuple of strictly proper plants $d\chi|_{\kappa_0}$ is surjective if

$$(5.2.20) \quad (m+p-1) \geq n_i.$$

Finally, by the implicit function theorem, if $d\chi|_{\kappa_0}$ has full rank, then image χ contains an open neighborhood around $\chi(K_0)$. We now claim that χ is surjective. To see that it is, consider scaling either k_2 or k_1 in (5.2.10) by a scale factor λ . By choosing a rank 1 matrix K_0 such that $\chi(K_0)$ contains a neighborhood around the origin and multiplying by an arbitrary scale factor λ , an arbitrary r -tuple of a closed-loop characteristic polynomial can be achieved.

5.3. Simultaneous pole placement as a Schubert intersection problem. In this section we generalize the pole placement strategy introduced by Brockett and Byrnes [3] to analyze the simultaneous pole placement problem. Though repetitive, the main construction is described as follows.

Consider the basic multivariable feedback equations

$$(5.3.1) \quad G(s)u = y, \quad u = -Ky$$

where u is an m vector and y is a p vector. The closed-loop poles are given by the set of all $s^* \in \mathbb{C}$ where $\det[I + G(s^*)K] = 0$, i.e.,

$$(5.3.2) \quad \det \begin{bmatrix} G(s^*) & -I_p \\ I_m & K \end{bmatrix} = 0$$

or equivalently

$$(5.3.3) \quad \dim[\text{row span of } [G(s^*) \quad -I_p] \cap \text{row span of } [I_m \quad K]] > 0.$$

Row span of $[G(s^*) \quad -I_p]$ is a p plane in $m+p$ space and is therefore a point in $\text{Grass}(p, m+p)$. Every plant $G(s)$ gives rise to a curve $\mathcal{G}: \mathbb{CP}^1 \rightarrow \text{Grass}(p, m+p)$ defined by $\mathcal{G}(s) = \text{row span of } [G(s) \quad -I_p]$. We denote the above curve in the Grassmannian by the graph $[G(s)]$. Likewise, row span of $[I_m \quad K]$ is a point in $\text{Grass}(m, m+p)$. Note that $\text{Grass}(m, m+p)$ is in fact the compactification $\bar{S}_{p,m}^0$ as described in § 2. Not every point in $\text{Grass}(m, m+p)$ comes from a feedback gain. The point

$$(5.3.4) \quad \text{row span of } [K_1 \quad K_2] \in \text{Grass}(m, m+p)$$

corresponds to the gain $K_1^{-1}K_2$ provided $\det K_1 \neq 0$. On the other hand, the set of points

$$(5.3.5) \quad \text{row span of } [K_1 \quad K_2], \quad \det K_1 = 0$$

are the infinite gains in $\text{Grass}(m, m+p)$.

If W is a point in $\text{Grass}(p, m+p)$ we define $\sigma(W)$ as follows:

$$(5.3.6) \quad \sigma(W) \triangleq \{W_1 \in \text{Grass}(m, m+p) \mid \dim(W_1 \cap W) > 0\}.$$

$\sigma(W)$ is classically known as the Schubert hypersurface. It follows from (5.3.3) that $\sigma([G(s^*) \quad -I_p])$ parameterizes the set of gains (finite or infinite) which places the poles of $G(s)$ at $s = s^*$. Thus if the objective is to place the poles at distinct points s_1, \dots, s_n , the pole placement problem is to ensure that

$$(5.3.7) \quad \bigcap_{i=1}^n \sigma([G(s_i) \quad -I_p])$$

contains a finite gain. The contribution of Brockett and Byrnes [3] is that (5.3.7) indeed contains a finite gain for a nondegenerate strictly proper plant.

We now consider Problem 5.1.1 under the assumption $mp = \sum_{i=1}^r n_i$. The simultaneous pole placement problem is interpreted as a geometric intersection problem and we ask the following question.

Question 5.3.1.

$$(5.3.8) \quad \text{When is } \bigcap_{i=1}^r \bigcap_{j=1}^{n_i} \sigma[G_i(s_{i,j}) | -I_p] \text{ nonempty?}$$

DEFINITION 5.3.2. A set of mp, m planes in $m+p$ space $W_i, i=1, \dots, mp$ are in general position if $\dim \bigcap_{i=1}^{mp} \sigma(W_i) = 0$.

Consider an r -tuple of proper plants $G_1(s), \dots, G_r(s)$ of McMillan degree $n_i, i=1, \dots, r$. Let $s_{ij}, i=1, \dots, r, j=1, \dots, n_i$, be the poles to be placed. It is clear that the necessary condition for the simultaneous pole placement problem is that (5.3.8) is nonempty. Conversely, however, if (5.3.8) is nonempty it does not imply that there exists a finite gain K which places the poles. First of all the gain placing the poles might be infinite as defined in (5.3.5) or otherwise the gain K might be in the base locus of the pole placement map, i.e.,

$$(5.3.9) \quad \det \begin{bmatrix} N_i(s) & -D_i(s) \\ K_1 & K_2 \end{bmatrix} \equiv 0$$

for some $i=1, \dots, r$.

Remark 5.3.3. Consideration of the base locus for the analysis of the pole placement problem is a new ingredient of this section and does not appear in [3]. This is because the plants we consider are proper as opposed to being strictly proper.

The following two definitions generalize the notion of nondegeneracy introduced in [3].

DEFINITION 5.3.4. An r -tuple of m -input p -output proper plants is nondegenerate at s_{ij} if no Schubert hypersurface $\sigma(W)$ in Grass $(m, m+p)$, where

$$(5.3.10) \quad W \in \bigcap_{i=1}^r \bigcap_{j=1}^{n_i} \sigma[G_i(s_{i,j}) | -I_p],$$

contains at least one of the curves $\text{graph}[G_i(s)]$ for all $i=1, \dots, r$.

DEFINITION 5.3.5. An r -tuple of m -input p -output proper plants is twisted at $s_{i,j}$ if and only if the set

$$(5.3.11) \quad \bigcap_{i=1}^r \bigcap_{j=1}^{n_i} \sigma[G_i(s_{i,j}) | -I_p] \cap \sigma[0 \quad -I_p]$$

is empty.

Example 5.3.6. $m=p=n=r=1$.

$$(5.3.12) \quad G(s) = (as+b)/(cs+d).$$

Every $G(s)$ satisfying $ad=bc$ is nondegenerate at any s . Every $G(s)$ satisfying $as_0+b=0$ is twisted at $s=s_0$.

Example 5.3.7. $p=1, m=n, r=1$.

$$(5.3.13) \quad G(s) = \left[\sum_{j=0}^m p_j^{(m+1)} s^j \right]^{-1} \left[\sum_{j=0}^m p_j^{(1)} s^j, \dots, \sum_{j=0}^m p_j^{(m)} s^j \right].$$

Every $G(s)$ satisfying

$$(5.3.14) \quad \det \begin{bmatrix} p_0^1 & \dots & p_m^1 \\ p_0^{m+1} & \dots & p_m^{m+1} \end{bmatrix} \neq 0$$

is nondegenerate at any s_1, \dots, s_n . Every $G(s)$ satisfying

$$(5.3.15) \quad \det \begin{bmatrix} \alpha_1(s_1) & \alpha_1(s_m) \\ \alpha_m(s_1) & \alpha_m(s_m) \end{bmatrix} \neq 0$$

is twisted at s_1, \dots, s_m , where $\alpha_i(s) = (p_0^i + p_1^i s + \dots + p_m^i s^m)$, $i = 1, \dots, m$.

Example 5.3.8. $p = 1$, $n = 1$, $m = r$.

$$G_i(s) = (a_{m+1,i}s + b_{m+1,i})^{-1}(a_{1,i}s + b_{1,i}, \dots, a_{m,i}s + b_{m,i}).$$

The nondegenerate r -tuple is given by

$$(5.3.16) \quad \det \begin{bmatrix} a_{11}s_1 + b_{11} & \cdots & a_{m+1,1}s_1 + b_{m+1,1} \\ \vdots & & \vdots \\ a_{1,i-1}s_i + b_{1,i-1} & \cdots & a_{m+1,i-1}s_{i-1} + b_{m+1,i-1} \\ a_{1,i} & \cdots & a_{m+1,i} \\ b_{1,i} & \cdots & b_{m+1,i} \\ a_{1,i+1}s_i + b_{1,i+1} & \cdots & a_{m+1,i+1}s_{i+1} + b_{m+1,i+1} \\ \vdots & & \vdots \\ a_{1m}s_m + b_{1m} & \cdots & a_{m+1,m}s_m + b_{m+1,m} \end{bmatrix} \neq 0, \quad i = 1, \dots, m.$$

The twisted r -tuples are given by

$$(5.3.17) \quad \det \begin{bmatrix} a_{11}s_1 + b_{11} & \cdots & a_{m1}s_1 + b_{m1} \\ \vdots & & \vdots \\ a_{1m}s_m + b_{1m} & \cdots & a_{mm}s_m + b_{mm} \end{bmatrix} \neq 0.$$

Example 5.3.9. This example shows that for $r = 4$, $n = 1$, $m = p = 2$ a generic 4-tuple of strictly proper plants is degenerate. Consider the generic plant

$$G_i: C\mathbb{P}^1 \rightarrow \text{Grass}(2, 4),$$

$$[s \ 1] \mapsto \text{row span of } \begin{bmatrix} a & b & s+e & 0 \\ c & d & 0 & s+e \end{bmatrix}, \quad ad = bc.$$

Clearly the Schubert hypersurface given by

$$\sigma \left[\text{row span of } \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right]$$

contains the graph of $G_i(s)$ for $i = 1, 2, 3, 4$, since

$$\det \begin{bmatrix} a & b & s+e & 0 \\ c & d & 0 & s+e \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = 0.$$

Note. This example provides a counterexample to the folklore conjecture that a generic r -tuple $(mp = \sum_{i=1}^r n_i)$ of strictly proper plants is nondegenerate.

The following two properties are now established.

LEMMA 5.3.10. *An m -input p -output proper plant $G(s)$ of degree $= n$ is nondegenerate at distinct $s_j, j = 1, \dots, n$, if and only if graph $G(s)$ is not contained in any Schubert hypersurface.*

Proof. (\Rightarrow) Suppose not. Then there exists $W \in \text{Grass}(m, m+p)$ such that $\sigma(W)$ contains row span of $[G(s) \ -I_p]$ for all s so that

$$(5.3.18) \quad W \in \bigcap_{j=1}^n \sigma[G(s_j) \ -I_p]$$

which is a contradiction since $G(s)$ is nondegenerate at $s_j, j=1, \dots, n$.

(\Leftarrow) Suppose not. Then there exists $W \in \bigcap_{j=1}^n \sigma[G(s_j) \ -I_p]$ such that $\sigma(W)$ contains graph $[G(s)]$, which is a contradiction. \square

LEMMA 5.3.11. *An r -tuple of m -input p -output strictly proper plant is nondegenerate at distinct s_{ij} if and only if it is twisted at distinct $s_{ij}, i=1, \dots, r, j=1, \dots, n_i$.*

Proof. If the r -tuple of plants is degenerate at s_{ij} , there exists W in $\text{Grass}(m, m+p)$ satisfying (5.3.10) and $\sigma(W)$ contains graph $[G_i(s)]$ for some $i=1, \dots, r$. Since $G_i(s)$ is strictly proper, $[0 \ -I_p] \in \sigma(W)$ or

$$(5.3.19) \quad W \in \sigma[0 \ -I_p].$$

Thus from (5.3.10) and (5.3.11) it may be concluded that the r -tuple of plants is untwisted at s_{ij} .

Conversely, if the r -tuple of plants is untwisted at $s_{ij}, i=1, \dots, r, j=1, \dots, n_i$, by defining $s_{i,n_i+1} = \infty$ for all $i=1, \dots, r$ there exists $W \in \text{Grass}(m, m+p)$ such that

$$(5.3.20) \quad W \in \bigcap_{i=1}^r \bigcap_{j=1}^{n_i+1} \sigma[G_i(s_{ij}) \ -I_p].$$

To say that (5.3.20) is true is to say that the parameters of W satisfy a set of r polynomials in s of degree n_i at n_i+1 different complex points $s_{i1}, \dots, s_{i,n_i+1}$. Hence the polynomials must be zero polynomials or equivalently $W \in \sigma[G_i(s) \ -I_p]$ for all $i=1, \dots, r$. Thus the r -tuple of plants is degenerate. \square

Finally, consider the following lemma, because of which the set of nondegenerate and twisted r -tuples of plants is interesting to study.

LEMMA 5.3.12. *For an r -tuple of nondegenerate and twisted proper plants, the set of gains contained in the set (5.3.11) is finite and places the poles of the i th plant at distinct $s_{i,j}, j=1, \dots, n_i$.*

Proof. Suppose not. Then the set (5.3.11) contains

$$W = [K_1 \ K_2]$$

which either satisfies the condition $W \in \sigma[G_i(s) \ -I_p]$ for all s and for some $i \in \{1, \dots, r\}$ or satisfies

$$W \in \sigma[0 \ -I_p].$$

The former case contradicts the fact that the r -tuple is nondegenerate. The latter case contradicts the fact that the r -tuple is twisted. \square

Proof of Theorem 5.1.8. Consider the set

$$(5.3.21) \quad V = S_{m,p}^{n_1} \times \dots \times S_{m,p}^{n_r} \times \text{Grass}(m, m+p)$$

which topologizes the r -tuple of m -input p -output proper plants together with the finite or infinite compensators. For a fixed $s_{i,j}, i=1, \dots, r, j=1, \dots, n_i$, let us define the algebraic subset of V as follows:

$$(5.3.22)$$

$$S = \left\{ G_1(s), \dots, G_r(s), V \mid V \in \bigcap_{i=1}^r \bigcap_{j=1}^{n_i} \sigma[G_i(s_{i,j}) \ -I_p] \right. \\ \left. \text{and graph } G_i(s) \in \sigma(V) \text{ for some } i=1, \dots, r \text{ or } [0 \ -I_p] \in \sigma(V) \right\}.$$

Consider the projection of S onto

$$U = S_{m,p}^{n_1} \times \cdots \times S_{m,p}^{n_r}$$

which is an algebraic set S' since $\text{Grass}_{\mathbb{C}}(m, m+p)$ is projective [19].

Thus the set of r -tuples of complex plants that are either degenerate or untwisted forms an algebraic subset S' of U .

Let k be \mathbb{R} or \mathbb{C} . It is well known [24] that U over the field k can be viewed as a manifold of dimension $N = \sum_{i=1}^r (n_i(m+p) + mp)$, and hence the algebraic set S' is described locally by a set of algebraic equations. By [14, Lemma 2.2] the same set of algebraic equations describes the complex plants over the complex variables and the real plants over the real variables. Let

$$(5.3.23) \quad \underline{p}(\underline{z}) = 0$$

describe the set S' . By restricting \underline{z} to its real part, $\underline{x} = \text{Re}(\underline{z})$, the equation

$$(5.3.24) \quad \underline{p}(\underline{x}) = 0$$

describes the set of real plants in S' that is either degenerate or untwisted.

By identity theorem, $\underline{p}(\underline{x}) = 0$ for all $\underline{x} \in \mathbb{R}^N$ if and only if $\underline{p}(\underline{x}) \equiv 0$. However, since by hypothesis there exists one real nondegenerate and twisted plant, it is not in the zero set of $\underline{p}(\underline{x})$. Hence $\underline{p}(\underline{x}) \not\equiv 0$, and there is a contradiction. Thus the set

$$(5.3.25) \quad \{\underline{x} \mid \underline{p}(\underline{x}) = 0\}$$

is a proper algebraic subset of the set of real r -tuple of plants. Hence the set of real nondegenerate and twisted plants is a generic subset of the set of real r -tuple of plants.

Thus by Lemma 5.3.12 there exists a generic real r -tuple of plants for which if the set (5.3.8) is nonempty, it contains finite gains which place the poles of $G_i(s)$ at $s_{i,j}$, $i = 1, \dots, r, j = 1, \dots, n_i$.

By a special property of Schubert hypersurfaces [38], we claim that the space X of all gains in the set (5.3.8) has dimension $\geq mp - \sum_{i=1}^r n_i$. It follows that $\dim X \geq 0$ and we conclude that X is nonempty. In fact, $\dim X = 0$. Otherwise by the same property of Schubert hypersurfaces the space $X \cap \sigma[G_i(s^*)| -I_p]$ is nonempty for some $s^* \in \mathbb{C}$ different from $s_{i,j}$. Hence there exists $W \in X$ such that $\sigma(W)$ contains graph $G_i(s)$, which is a contradiction.

Thus $\dim X = 0$ and the $\sum_{i=1}^r n_i$ Schubert hypersurfaces are indeed in a general position so that by Schubert calculus [38] the number of finite gains which place the poles is given by (4.2.14). \square

Proof of Corollary 5.1.9. The claim clearly follows from the fact that the set of gains which places the poles of $G_i(s)$, at $s_{i,j}$, $i = 1, \dots, r, j = 1, \dots, n_i$, is a complex conjugate set. Moreover, the number (4.2.14) is odd if and only if (5.1.7) is satisfied. Finally since the r -tuple of plants is real, and for a fixed i the set $s_{i,j}$ is a complex conjugate set, the set of equations that describes the set (5.3.11) has the property that if row span of $[K_1 \ K_2]$ is in (5.3.11), then the row span of $[\bar{K}_1 \ \bar{K}_2]$ the conjugate of the row span of $[K_1 \ K_2]$ is in (5.3.11). \square

LEMMA 5.3.13. *If M_1 and M_2 are two matrices such that $\text{rank}(M_1 - M_2) = 1$, then there exists a 2×2 proper rational function $G(s)$ of degree 1 such that $G(s_1) = M_1$ and $G(s_2) = M_2$ for a given s_1, s_2 .*

Proof. Writing

$$(5.3.26) \quad G(s) = R/(s + \lambda) + J$$

we obtain the following pair of simultaneous equations:

$$(5.3.27) \quad R/(s_1 + \lambda) + J = M_1, \quad R/(s_2 + \lambda) + J = M_2.$$

Here, R is a rank 1 matrix which may be solved simultaneously to obtain

$$(5.3.28) \quad R = [(s_1 + \lambda)(s_2 + \lambda)/(s_2 - s_1)](M_1 - M_2).$$

Since $\text{rank}(M_1 - M_2) = 1$, there exists a rank 1 matrix R satisfying (5.3.28). Using (5.3.26) we can now solve for J . \square

Proof of Lemma 5.1.11. For $m = p = 2$, $n = 1$, $r = 4$ there exists an r -tuple of proper plants that is nondegenerate and twisted.

Let M_1, M_2, M_3, M_4 be four 2×2 matrices such that

$$(5.3.29) \quad \text{rank}(M_i - M_j) \neq 1$$

and

$$(5.3.30) \quad \bigcap_{i=1}^4 \sigma[M_i \quad -I] \cap \sigma[0 \quad I]$$

is empty. Clearly the above property is satisfied generically.

Consider the row span of

$$(5.3.31) \quad L_i = [M_i \quad I], \quad i = 1, 2, 3, 4$$

as points in Grass $(2, 4)$. By Schubert calculus it is known that

$$(5.3.32) \quad \bigcap_{i=1}^4 \sigma(L_i)$$

contains exactly two points W_1, W_2 in Grass $(2, 4)$, provided L_i is in general position. Moreover, by considering the cell decomposition of a Grassmannian it can be seen that $W_i, i = 1, 2$ are irreducible (see Milnor and Stasheff [44]). Define

$$(5.3.33) \quad H_j = \{\text{row span of } [M \quad I] \mid (M - M_j) \text{ is of rank 1}, j = 1, 2, 3, 4\}$$

which is an irreducible Schubert hypersurface in Grass $(2, 4)$.

If $H_j \subset \sigma(W_1)$ then since $\sigma(W_1)$ is irreducible $H_j = \sigma(W_1)$. Hence, $L_1, L_2, L_3, L_4 \in H_j$ which, however, contradicts (5.3.29). Hence,

$$(5.3.34) \quad H_j - H_j \cap \sigma(W_1)$$

is nonempty. Since H_j is irreducible, $H_j - H_j \cap \sigma(W_1)$ is dense in H_j . Hence

$$(5.3.35) \quad \bigcap_{i=1}^2 [H_j - H_j \cap \sigma(W_i)]$$

is nonempty and it is possible to choose a point M'_j in it.

Consider $G_j(s)$ passing through M_j and M'_j , where $G_j(s_j) = M_j, j = 1, 2, 3, 4$. By Lemma 5.3.13 such a $G_j(s)$ indeed exists. Moreover, $G_j(s), j = 1, 2, 3, 4$, are clearly nondegenerate. Otherwise either of the plants would be contained in W_1 or W_2 , which is a contradiction. Finally the 4-tuple is twisted by (5.3.30). \square

Proof of Theorem 5.1.12. Define $\sigma_0(\infty)$ to be the set of infinite gains distinct from

$$(5.3.36) \quad [U] = \text{row span of } \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Let $G_1(s)$, $G_2(s)$, $G_3(s)$, $G_4(s)$ be the four plants and let s_1, s_2, s_3, s_4 be the set of real poles to be placed. By Schubert calculus [31], the set

$$(5.3.37) \quad \bigcap_{i=1}^4 \sigma[G_i(s_i) - I_2]$$

contains two points for a generic 4-tuple of plants. Consider the set

$$(5.3.38) \quad \Sigma_{2,2}^1 \times \Sigma_{2,2}^1 \times \Sigma_{2,2}^1 \times \Sigma_{2,2}^1 \times \text{Grass}(2, 4)$$

together with the subset

$$(5.3.39) \quad X = \{(G_1, G_2, G_3, G_4, V) \mid G_i(s) \subset \sigma(V), \text{ for some } i\}.$$

Consider the projection

$$(5.3.40) \quad p_2: X \rightarrow \text{Grass}(2, 4).$$

Let V_g be a generic point in image p_2 ; then $p_2^{-1}(V_g)$ is a generic fiber at V_g . Moreover, it is known [5] that if V_{g_1} and V_{g_2} are two generic points, then

$$(5.3.41) \quad p_2^{-1}(V_{g_1}) \cong p_2^{-1}(V_{g_2}).$$

First

$$(5.3.42) \quad p_2^{-1}(V_g) = \prod_{i=1}^4 p_{2i}^{-1}(V_g)$$

where

$$(5.3.43) \quad p_{2i}^{-1}(V_g) = \{G_i(s) \mid G_i(s) \subset \sigma(V_g)\}.$$

Choosing V_g^* to be the

$$(5.3.44) \quad \text{row span of } \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

we may describe $p_2^{-1}(V_g^*)$ as follows:

$$(5.3.45) \quad p_2^{-1}(V_g^*) = \left\{ (G_1, G_2, G_3, G_4) : G_i = \begin{bmatrix} g_{11}^{(i)} & 0 \\ g_{21}^{(i)} & g_{22}^{(i)} \end{bmatrix} \right\}.$$

Since $G_i(s)$ is a 2×2 plant of degree 1 it is clear that

$$(5.3.46) \quad p_{2i}^{-1}(V_g^*) \cong \Sigma_{2,1}^1 \cup \Sigma_{1,2}^1$$

and hence

$$(5.3.47) \quad \dim(p_{2i}^{-1}(V_g^*)) = 3$$

and

$$(5.3.48) \quad \dim(p_2^{-1}(V_g^*)) = 12.$$

Since $\sigma_0(\infty)$ is of dimension 3, it follows that

$$(5.3.49) \quad \dim p_2^{-1}[\sigma_0(\infty)] = 15.$$

Define

$$(5.3.50) \quad \tilde{Y} = \{G_1, G_2, G_3, G_4, V \mid V \in \sigma_0(\infty), \text{ and } G_i \in \sigma(V) \text{ for some } i = 1, 2, 3, 4\}$$

and consider the projection

$$(5.3.51) \quad p: \tilde{Y} \rightarrow \Sigma_{2,2}^1 \times \Sigma_{2,2}^1 \times \Sigma_{2,2}^1 \times \Sigma_{2,2}^1.$$

From (5.3.49) we conclude that $\dim \tilde{Y} = 15$. If $Y = \text{image } p$ it follows that $\prod_{i=1}^4 \Sigma_{2,2}^1 - Y$ is open and dense. Hence for a generic 4-tuple of plants, there does not exist any element in $\sigma_0(\infty)$ which is in the set (5.3.37).

Since $[U]$ as given by (5.3.36) is in (5.3.37), we have two choices. Either $[U]$ is to be counted twice, or there is a finite gain which places the poles. For a generic r -tuple of plants, the former choice is impossible. To show this, it suffices to construct a particular 4-tuple of plants G_1, G_2, G_3, G_4 for which (5.3.37) contains an element other than $[U]$. \square

5.4. Simultaneous dead beat control problem. Consider the following interpretation of the dead beat control problem defined in (5.1.14). Let the state space realization of the plants $G_1(s), \dots, G_r(s)$ be given by

$$(5.4.1) \quad \begin{aligned} \dot{x}(t) &= A_i x(t) + B_i u(t), \\ y(t) &= C_i x(t) + D_i u(t). \end{aligned}$$

The dead beat control problem may be posed as follows.

Problem 5.4.1. Given an r -tuple of matrix triplets $A_i, B_i, C_i, k = 1, \dots, r$, does there exist a matrix K such that

$$(5.4.2) \quad A_i + B_i K C_i$$

for all $i = 1, \dots, r$ are nilpotent?

In this section, Theorem 5.1.12 has been proved by obtaining a necessary condition to Problem 5.4.1. This proof has been adapted from the proof by Anderson and Byrnes [2] for $r = 1$.

Proof of Theorem 5.1.12. Let N_i be the space of the nilpotent $n_i \times n_i$ matrix. It is known that N_i is a subvariety in $\mathbb{R}^{n_i^2}$ having dimension $n_i^2 - n_i$. For a fixed $B_1, C_1, \dots, B_r, C_r$ consider the mapping

$$(5.4.3) \quad \chi: N_1 \times \dots \times N_r \times \mathbb{R}^{mp} \rightarrow \mathbb{R}^{\sum n_i}$$

defined by

$$(5.4.4) \quad \chi(N_1, \dots, N_r, K) = (N_1 - B_1 K C_1, \dots, N_r - B_r K C_r).$$

By assumption, χ is almost surjective so that

$$(5.4.5) \quad \sum (n_i^2 - n_i) + mp \geq \sum n_i^2$$

or, equivalently, (5.1.9) is a necessary condition. \square

Proof of dead beat control problem 5.1.13. Stated in other words, an r -tuple of plants is dead beat controllable only in the case where

$$(5.4.6) \quad \det \begin{bmatrix} G_i(s) & -I_p \\ I_m & K \end{bmatrix} = 0$$

only at $s = 0$, for all $i = 1, \dots, r$.

Let W be the m plane in Grass $(m, m+p)$ which solves Problem 5.4.1. W may be defined by the common zeros of ϕ_1, \dots, ϕ_p on \mathbb{C}^{m+p} . Let $g_{ji}(s)$ denote the j th row of

$$(5.4.7) \quad [G_i(s) \quad -I_p]$$

so that

$$(5.4.8) \quad \begin{bmatrix} \phi_1 \wedge \cdots \wedge \phi_p(g_{1i}(s) \wedge \cdots \wedge g_{pi}(s))|_{s=0} = 0 \\ \vdots \\ \frac{d^{n_i}}{ds^{n_i}} \phi_1 \wedge \cdots \wedge \phi_p(g_{1i}(s) \wedge \cdots \wedge g_{pi}(s))|_{s=0} = 0 \end{bmatrix}$$

for $i = 1, \dots, r$. If we embed Grass $(m, m+p)$ by the Plucker embedding

$$(5.4.9) \quad \text{Grass}(m, m+p) \subset \text{Proj}(\Lambda^m(\mathbb{C}^{m+p})),$$

then each of the above conditions represents a functional

$$(5.4.10) \quad \phi_{ij}: \Lambda^m(\mathbb{C}^{m+p}) \rightarrow \mathbb{C}$$

for $i = 1, \dots, r, j = 1, \dots, n_i$, where $\ker \phi_{ij}$ is a hyperplane intersection of the Grass $(m, m+p)$, i.e., $H_{ij} = \ker \phi_{ij} \cap \text{Grass}(m, m+p)$. The $H_{ij} - s$ are not necessarily Schubert hypersurfaces but have the same homology class. That is, if $\ker \phi$ corresponds to a Schubert hypersurface, then

$$(5.4.11) \quad [H(\phi)] = [H(\phi_{ij})] \in H_{2mp-2}(\text{Grass}(m, m+p), \mathbb{Z}).$$

To show (5.4.11), it is enough to check that the space of linear functionals in \mathbb{C}^{m+p} is path connected and that ϕ can be deformed to ϕ_{ij} by the path

$$(5.4.12) \quad \begin{aligned} \psi: [0, 1] &\rightarrow \mathbb{C}^{m+p*}, \\ \psi(t) &= t\phi + (1-t)\phi_{ij}. \end{aligned}$$

Indeed, $t\phi + (1-t)\phi_{ij} \neq 0$ for every t . Otherwise either $\phi, \phi_{ij} = 0$ or ϕ and ϕ_{ij} are linearly dependent. Finally by Spianer [39] and Greenberg [18], if the hyperplanes $\ker \phi_{ij}$ are in general position for all i, j then the number of points in Grass $(m, m+p)$ that belongs to H_{ij} is finite and is given by (4.2.14).

Finally, it is claimed that for a nondegenerate and twisted r -tuple of plants, H_{ij} are in general position. Otherwise

$$\dim \bigcap_{i,j} H_{ij} > 0.$$

Hence there exists $[K_1 \ K_2]$ which satisfies (5.4.8) together with another condition

$$(5.4.13) \quad \phi_1 \wedge \cdots \wedge \phi_p(g_{1i_0}(s_0) \wedge \cdots \wedge g_{pi_0}(s_0)) = 0$$

for a fixed $i = i_0$ and $s = s_0 \neq 0$. However, (5.4.8) and (5.4.13) together imply that

$$(5.4.14) \quad \det \begin{bmatrix} G_{i_0}(s) & -I_p \\ K_1 & K_2 \end{bmatrix} \equiv 0$$

since a nonidentically vanishing polynomial of degree n_{i_0} cannot have n_{i_0} roots at the origin together with a root at $s = s_0$. The condition (5.4.14) gives a contradiction to the fact that the r -tuple of plants is nondegenerate.

Finally, since the r -tuple of plants is twisted, we have

$$(5.4.15) \quad \det \begin{bmatrix} 0 & -I_p \\ K_1 & K_2 \end{bmatrix} \neq 0$$

or $\det K_1 \neq 0$. Hence the set of gains $[K_1 \ K_2]$ are finite gains. The proof of the theorem now follows analogous to Corollary 5.1.9. \square

6. Conclusion. The main purpose of this paper has been to introduce robust system design techniques by using algebraic geometric methods. First, we addressed the

question of parameterization and compactified the space of dynamical systems of a given McMillan degree. Next, we considered a family of systems and studied how a family of closed-loop systems degenerates as a function of plant and compensator parameters. Finally, we considered two explicit design problems: the simultaneous stabilization problem and the simultaneous pole placement problem. Many other design problems in this context remain open. We hope that the proposed parameterization of the space of linear dynamical systems will find application in parameter identification and indirect adaptive control problems. In fact, with parameterization of the space of pairs of dynamical systems, the adaptive control problem may be viewed as a degenerating plant/compensator family that is stable in the closed loop.

Acknowledgements. It is my great pleasure to acknowledge the comments and encouragement of Professor C. I. Byrnes during my research work at Harvard University. I also thank the anonymous reviewer.

REFERENCES

- [1] J. ACKERMANN, *Uncertainty and Control*, Lecture Notes Control and Information Sciences 70, Springer-Verlag, Berlin, New York, 1985.
- [2] B. D. O. ANDERSON AND C. I. BYRNES, *Output feedback and generic stabilizability*, SIAM J. Control Optim., 22 (1984), pp. 362–380.
- [3] R. W. BROCKETT AND C. I. BYRNES, *Multivariable Nyquist criteria, root loci and pole placement: a geometric viewpoint*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 271–284.
- [4] C. I. BYRNES, *On compactification of spaces of systems and dynamic compensation*, 22nd IEEE Conference on Decision and Control, San Antonio, CA, 1983, pp. 889–894.
- [5] ———, *Stabilizability of multivariable systems and the Ljusternick–Snirel’mann category of real Grassmannians*, Systems Control Lett., 3 (1983), pp. 255–262.
- [6] B. K. GHOSH AND W. P. DAYAWANSA, *A hybrid parameterization of linear single input single output systems*, Systems Control Lett., 8 (1987), pp. 231–239.
- [7] C. I. BYRNES AND N. E. HURT, *On the moduli of linear dynamical systems*, in Adv. in Math., Suppl. Series, 4 (1978), pp. 83–122; Modern Mathematical Systems Theory, MIR, Moscow, 1978. (In Russian.)
- [8] T. DJAFERIS, *Robust observers and regulation for systems with parameters*, Proc. 23rd Conference on Decision and Control, 1984, pp. 1234–1239.
- [9] J. DOYLE, *Structured uncertainty in control system design*, 24th IEEE Conference on Decision and Control, 1985, pp. 260–265.
- [10] J. DOYLE, J. WALL, AND G. STEIN, *Performance and robustness analysis for structured uncertainty*, Proc. 21st IEEE Conference on Decision and Control, 1982, pp. 629–636.
- [11] W. R. EVANS, *Control system synthesis by root locus method*, Trans. AIEE, 69 (1950), pp. 66–69.
- [12] B. K. GHOSH, *Simultaneous stabilization and pole-placement of a multimode linear dynamical system*, Ph.D. dissertation, Harvard Univ., Cambridge, MA, 1983.
- [13] ———, *Simultaneous partial pole placement—a new approach to multi-mode system design*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 440–443.
- [14] ———, *An approach to simultaneous system design, Part I: Semialgebraic Geometric Methods*, SIAM J. Control Optim., 24 (1986), pp. 480–496.
- [15] ———, *Some new results on the simultaneous stabilizability of a family of single input single output systems*, Systems Control Lett., 6 (1985), pp. 39–45.
- [16] ———, *Transcendental and interpolation methods in simultaneous stabilization and simultaneous partial pole placement problems*, SIAM J. Control Optim., 24 (1986), pp. 1091–1109.
- [17] B. K. GHOSH AND C. I. BYRNES, *Simultaneous stabilization and simultaneous pole placement by non-switching dynamic compensation*, IEEE Trans. Automat. Control, 28 (1983), pp. 735–741.
- [18] M. GREENBERG, *Lectures on Algebraic Topology*, W. A. Benjamin, New York, 1967.
- [19] P. GRIFFITHS AND J. HARRIS, *Principles of Algebraic Geometry*, John Wiley, New York, 1978.
- [20] M. HAZEWINKEL, *On families of linear system: degeneration phenomenon*, in Algebraic and Geometric Methods in Linear Systems Theory, C. I. Byrnes and C. F. Martin, eds., Lectures in Applied Mathematics, 18, American Mathematical Society, Providence, RI, 1980, pp. 157–189.

- [21] M. HAZEWINKEL, *Parameterization problems for spaces of linear input-output systems*, Report PM-R8507, Centre for Mathematics and Computer Science, Amsterdam, the Netherlands, September 1985.
- [22] ———, *Parameterization of the space of all linear systems and numerical trouble*, preprint.
- [23] ———, *Moduli and canonical forms for linear dynamical systems II: the topological case*, Math. Systems Theory, 10 (1977), pp. 363–385.
- [24] M. HAZEWINKEL AND R. E. KALMAN, *On Invariants, Canonical Forms and Moduli for Linear Constant Finite-Dimensional Dynamical Systems*, Lecture Notes in Econo-Math. System Theory 131, Springer-Verlag, Berlin, 1976, pp. 48–60.
- [25] U. HELMKE AND W. MANTHEY, *On closures of spaces of linear dynamical systems*, Report 132, Forschungsschwerpunkt Dynamische Systeme, Universität Bremen, Bremen, Federal Republic of Germany.
- [26] R. HERMANN AND C. F. MARTIN, *Application of algebraic geometry to system theory: the McMillan degree and Kronecker indices of transfer functions as topological and holomorphic invariants*, SIAM J. Control Optim., 16 (1978), pp. 743–755.
- [27] P. P. KHARGONEKAR AND A. TANNENBAUM, *Non-Euclidean metrics and the robust stabilization of systems with parameter uncertainty*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1005–1013.
- [28] V. L. KHARITONOV, *Asymptotic stability of an equilibrium position of a family of systems of linear differential equations*, Differential Equations, 14 (1979), pp. 1483–1485.
- [29] H. KIMURA, *Robust stabilizability for a class of transfer functions*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 788–793.
- [30] ———, *Pole assignability by gain output feedback*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 509–516.
- [31] S. L. KLEIMAN AND D. LAKSOV, *Schubert calculus*, Amer. Math. Monthly, 79 (1972), pp. 1061–1082.
- [32] P. S. KRISHNAPRASAD AND C. F. MARTIN, *On families of systems and deformations*, Internat. J. Control, 38 (1983), pp. 1055–1079.
- [33] N. A. LEHTOMAKI, N. R. SANDELL AND M. ATHANS, *Robustness results in linear quadratic Gaussian multivariable control designs*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 75–93.
- [34] D. MUMFORD, *Algebraic Geometry I: Complex Projective Varieties*, Springer-Verlag, Berlin, New York, 1976.
- [35] R. NEVANLINNA, *Über beschränkte Funktionen, die in gegebenen Punkten vorgeschriebene Werte annehmen*, Ann. Acad. Sci. Fenn., 13 (1919).
- [36] G. PICK, *Über die Beschränkungen analytischer Funktionen, welche durch vorgegebenen Funktionswerte bewiesen sind*, Math. Ann., 77 (1916), pp. 7–23.
- [37] R. SAEKS AND J. J. MURRAY, *Fractional representation, algebraic geometry and the simultaneous stabilization problem*, IEEE Trans. Automat. Controls, AC-27 (1982), pp. 895–903.
- [38] I. R. SHAFAREVITCH, *Basic Algebraic Geometry*, Springer-Verlag, New York, 1974.
- [39] E. SPIANER, *Algebraic Topology*, McGraw-Hill, New York, 1966.
- [40] P. K. STEVENS, *Algebra-geometric methods for linear multivariable feedback systems*, Ph.D. dissertation, Harvard Univ., Cambridge, MA, 1982.
- [41] A. TANNENBAUM, *Invariance and System Theory: Algebraic and Geometric Aspects*, Lecture Notes in Math., 845, Springer-Verlag, Berlin, New York, 1981.
- [42] M. VIDYASAGAR AND N. VISWANADHAM, *Algebraic design techniques for reliable stabilization*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 1085–1095.
- [43] D. YOULA, J. BONGIORNO, AND C. LU, *Single loop feedback stabilization of linear multivariable dynamic plant*, Automatica, 10 (1974), pp. 155–173.
- [44] J. W. MILNOR AND J. D. STASHEFF, *Characteristic Classes*, Ann. of Math. Stud. 76, Princeton University Press, Princeton, NJ, 1974.
- [45] C. I. BYRNES, *On root-loci in several variables: Continuity in the high gain limit*, Systems Control Lett., 1 (1981), pp. 69–73.

ON IMPULSE CONTROL WITH PARTIAL OBSERVATION*

G. MAZZIOTTO†, L. STETTNER‡, J. SZPIRGLAS†, AND J. ZABCZYK§

Abstract. This paper presents an existence result for an impulse control problem with partial observation. The unobserved process evolves between any two successive impulse times as a Feller-Markov process on a locally compact separable state space, and the observation process is of a "signal + white noise" type.

Key words. impulse control, filtering, separation principle

AMS(MOS) subject classifications. 60G40, 93E20

1. Introduction. The impulse control problem with partial information which is solved in this paper has been motivated by the following practical situation. The evolution of a stock of goods when not controlled is modeled as a Feller-Markov process $X = (X_t; t \geq 0)$ with value in some locally compact space E . An impulse control v is a sequence $((T_n, \xi_n); n \in \mathbb{N})$, where T_n denotes the n th ordering time and ξ_n is the control parameter which determines the amount of goods ordered at T_n for $n \in \mathbb{N}$. The stock, when submitted to the control policy v , is denoted by X^v . For every n , $X_{T_n}^v$ is the inventory level at time T_n and $X_{T_n}^{v+}$ denotes the new level just after the ordering. Thus

$$X_{T_n}^{v+} = \phi(X_{T_n}^v, \xi_n)$$

where function ϕ indicates how the ordering amount ξ_n depends on the previous level $X_{T_n}^v$ and the control ξ_n . If $E = \mathbb{R}^d$ the ϕ could be defined as $\phi(x, \xi) = x + \xi$.

Between two successive ordering times, say T_n and T_{n+1} , the process X^v has the same evolution as X , but with the new initial value $X_{T_n}^{v+}$ at T_n .

Now, assume that the inventory level X^v is only known through a noisy observation Y . The observation process $Y = (Y_t; t \geq 0)$ is related to the state process X^v as in the classical filtering model:

$$Y_t = \int_0^t h(X_s^v) ds + W_t$$

where $W = (W_t; t \geq 0)$ is a Brownian motion which does not depend on the evolution of X^v between any two successive ordering times. An impulse control is said to be admissible if it is a nonanticipative functional of the observation Y . To each admissible control $v = (T_n, \xi_n; n \in \mathbb{N})$, we associate an average cost $J(v)$ of the following form:

$$J(v) = E^v \left[\int_0^\infty e^{-\alpha s} f(X_s^v) ds + \sum_{m=1}^{m=\infty} e^{-\alpha T_m} c(X_{T_m}^v, \xi_m) \right]$$

where f represents the storage cost per unit time, α is a positive actualization factor, and c the ordering cost. The impulse control problem with partial observation consists in finding an admissible control for which the average cost J achieves its infimum. *This control is said to be optimal.*

A model of this kind has been studied by Menaldi [21] for the linear Gaussian case with quadratic cost. The solution proposed in the present paper is essentially

* Received by the editors October 14, 1985; accepted for publication May 21, 1986.

† Centre National d'Etudes des Télécommunications, Paris A, Issy les Moulineaux, France.

‡ Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland.

§ Institute of Mathematics, Polish Academy of Sciences; Heriot-Watt University, Edinburgh, United Kingdom.

concerned with the nonlinear case, and strictly speaking it is not an extension of the former. The impulse control of partially observed Markov chains and assimilated processes is treated in a series of papers by Anderson and Friedman [1], [9]. A different impulse control problem with incomplete information is studied by Rishel in [25].

The paper extends results of the paper [18] by Mazziotto and Szpirglas to the case of locally compact state spaces. The assumption that the state space is locally compact and not merely compact is motivated by a rich variety of specific examples as described for instance in the monograph "Contrôle impulsif et inéquations quasi-variationnelles," by A. Bensoussan and J. L. Lions [3b]. The majority of the examples treated there and related to such areas such as control of electricity, inventory control, control of production, quality control, and portfolio selections require noncompact state spaces. In some situations we can reduce the general noncompact case to the compact one. This is possible if we can compactify the locally compact state space in such a way that functions defining the cost functional and observation can be extended to continuous functions on the compactification and the Markov semigroup preserves its Feller property on the space of all continuous functions on the compactification. In general, however, such compactification does not exist. Another important feature of the present paper is that here we dispense with the finite life time of the state process, a technical assumption imposed in [18].

The present paper is organized as follows. In § 2, we formulate the impulse control problem with partial observation, the setting being a modification of that from the paper by Mazziotto and Szpirglas [18]. The basic tool to construct the appropriate model is Girsanov's theorem, which permits work with a fixed observation filtration that does not involve the control. In § 3, a filtering process for the noncontrolled model is studied. The emphasis here is placed on the dependence of the filter on the initial data. It turns out that although the filtering process evolves on a nonlocally compact space, it possesses many properties of a Feller-Markov process on a locally compact state space. Several results of this section are proper developments of those obtained in [28] and [29] by Stettner and Zabczyk. The main properties are expressed in Theorems 3.1 and 3.2.

Section 4 is concerned with a stopping time problem for Markov processes on general state spaces. It is shown in particular that if a process is as regular as the filter process of § 3, then the value function (α -reduite) corresponding to a continuous gain function is also continuous. This result formulated as Theorem 4.3 is a basic step in constructing an optimal separated strategy for our impulse control problem. Continuity of the α -reduite for Feller semigroups on compact space was proved by Robin in [26] and in [8] by El Karoui. For nonlocally compact state spaces, such a result is established in [3] by Bensoussan and Lions under various conditions not fulfilled in our setting. The first result concerning the filtering process in the noncompact case is due to Stettner [28].

Section 5 treats the properties of the filtering process for the impulsively controlled partially observed process. Using the Kallianpur-Striebel formula we show (see Theorem 5.1) that the filtering process has, between any two successive impulsions, the same probability law as the filter of the uncontrolled case.

Section 6 starts from a discussion of a weak version of the Hamilton-Jacobi-Bellman equation of the problem. Results of §§ 3-5 imply that the equation has a continuous solution. Moreover, they allow us to define a separate strategy (i.e., a strategy which can be recursively computed in terms of the filter itself) which is a candidate for the optimality. Proof that this strategy is indeed optimal is contained in Theorem 6.1. The methods developed in these last two sections are analogous to those

of [18], but they do not reduce to trivial extensions. The results proved in this paper were announced in [19].

2. Formulation of the control problem. The model describing the evolutions of the unobserved process X^v subject to the impulse control v is constructed as in the works of Robin [26], Lepeltier and Marchal [14], [15], Menaldi [20], and Nagai [23]. In order to have an observation filtration independent of each possible control, we define the filtering model by using the Girsanov Theorem together with the reference probability method. The approach developed in this section is analogous to that of Mazziotto and Szpirglas in [18].

The observation is represented by the coordinate process $Y = (Y_t; t \geq 0)$ on the space $\Omega = C([0, \infty), \mathbf{R}^d)$ of the continuous functions from $[0, \infty)$ into \mathbf{R}^d (for $d \in \mathbf{N}$ fixed). Let ${}^0\mathcal{G} = ({}^0\mathcal{G}_t; t \geq 0)$ filtration generated by Y , and ${}^0\mathcal{G}_\infty = {}^0\mathcal{A}$ is the Borel σ -field on Ω . As will be seen, the impulse control problem leads us to consider various probabilities which are not equivalent on $(\Omega, {}^0\mathcal{A})$. We must extend certain canonical σ -fields to ensure that various applications as first hitting times of closed sets are stopping times, and to allow the construction of nice versions of various processes of conditional distributions. On the other hand, these σ -fields cannot be too large as we can lose the Markov property as well as the physical meaning of the constructed objects. Let \mathbf{P}^0 be the Wiener probability on $(\Omega, {}^0\mathcal{A})$. Let $\mathcal{N}(\mathbf{P}^0)$ be the set of all the \mathbf{P}^0 -negligible sets which are contained in a σ -field ${}^0\mathcal{G}_t$ for at least one finite time t . By this definition, $\mathcal{N}(\mathbf{P}^0)$ may not contain some \mathbf{P}^0 -negligible sets of ${}^0\mathcal{G}_\infty = {}^0\mathcal{A}$. It is worth noticing that for any probability \mathbf{P} on $(\Omega, {}^0\mathcal{A})$ such that, for every finite time t , the restrictions of \mathbf{P}^0 and \mathbf{P} to the σ -field \mathcal{G}_t are equivalent, $\mathcal{N}(\mathbf{P}) = \mathcal{N}(\mathbf{P}^0)$.

Finally, we define the filtration $\mathcal{G} = (\mathcal{G}_t; t \geq 0)$ to be the smallest right-continuous filtration such that, for every finite time t , \mathcal{G}_t contains \mathcal{G}_t^0 and $\mathcal{N}(\mathbf{P}^0)$, and we set $\mathcal{A} = \mathcal{G}_\infty$. As a matter of fact, for any finite t , the filtration $\mathcal{G}' = (\mathcal{G}_s; s \leq t)$ verifies the usual conditions of [6], and this is true for any other probability \mathbf{P} which is equivalent to \mathbf{P}^0 on \mathcal{G}_t . Such an operation will be called, in the sequel, the finite \mathbf{P}^0 completion of the filtration \mathcal{G}^0 .

A strategy or an impulse control is a sequence $v = (T_n, \xi_n; n \in \mathbf{N})$ of random variables where the T_n 's represent the successive instants of impulse, and the ξ_n 's are the corresponding impulsions assumed to take their values in a compact set U . Not all arbitrary sequences v will be allowed as legitimate strategies; their choice will be limited by the amount of available information. The proper strategies, called admissible control, must depend on the observation Y in a nonanticipative way.

DEFINITION 2.1. An admissible control is a sequence $v = (T_n, \xi_n; n \in \mathbf{N})$ of random variables on (Ω, \mathcal{A}) such that $(T_n; n \in \mathbf{N})$ is an increasing sequence of \mathcal{G} -stopping times such that $T_{n+1} > T_n$ on $\{T_n < \infty\}$, $n \in \mathbf{N}$, and $\lim_n T_n = \infty$, and such that ξ_n is a \mathcal{G}_{T_n} -measurable random variable with values in U , for all $n \in \mathbf{N}$. The set of admissible controls is denoted by \mathcal{V} .

From a classical result on the stopping times with respect to the Brownian filtration (see [6, pp. IV-100] and the same application in [18]), we can associate to each admissible control $v = (T_n, \xi_n; n \in \mathbf{N})$ a sequence $(S_n; n \in \mathbf{N})$ of random variables on $(\Omega, \mathcal{A})^{\otimes 2}$ such that

S_n is $\mathcal{G}_{T_n} \otimes \mathcal{A}$ -measurable for every $n \in \mathbf{N}$;

For \mathbf{P}^0 almost surely, every $w \in \Omega$: $S_n(w, \cdot)$ is a \mathcal{G} -stopping time;

For \mathbf{P}^0 almost surely, every $w \in \Omega$: $T_{n+1}(w) = T_n(w) + S_n(w, \theta_{T_n} w)$,

where $(\theta_t; t \geq 0)$ is the translation semigroup of operators on Ω (such that $Y_t \circ \theta_s = Y_{t+s} - Y_s$, $s, t \geq 0$). The admissible control $(T_n, \xi_n; n \in \mathbb{N})$ will also be denoted $(T_n, S_n, \xi_n; n \in \mathbb{N})$.

In order to represent the evolution of the unobserved process between two successive impulsions, we first consider a standard Markov process (see [5]) in its canonical form $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{\mathcal{F}}_t, \tilde{\theta}_t, X_t, (\tilde{\mathbf{P}}_x; x \in E))$, where $\tilde{\Omega}$ is the set of all the right-continuous left-limited functions from $[0, \infty)$ into $E^\delta = E \cup \{\delta\}$, $\tilde{\mathcal{F}} = (\tilde{\mathcal{F}}_t; t \geq 0)$ is the filtration generated by the coordinate process $X = (X_t; t \geq 0)$, with $\tilde{\mathcal{A}} = \tilde{\mathcal{F}}_\infty$, properly completed with respect to the family of probabilities $(\tilde{\mathbf{P}}_x; x \in E)$ and made right continuous; $(\tilde{\theta}_t; t \geq 0)$ is the translation semigroup on $\tilde{\Omega}$. The state space E will be further assumed to be separable locally compact and equipped with its Borel σ -field \mathcal{E} . The point at infinity is denoted by δ . The set of all the probability laws on E is denoted by $M(E)$, and $\mathcal{M}(E)$ is its Borel σ -field. For any $\mu \in M(E)$, $\tilde{\mathbf{P}}_\mu$ is the probability measure defined on $(\tilde{\Omega}, \tilde{\mathcal{A}})$ by $\tilde{\mathbf{P}}_\mu = \int \mu(dx) \tilde{\mathbf{P}}_x$.

Let $P = (P; t \geq 0)$ be the semigroup on E of the process X . The unobserved process X^v subject to the admissible control $v = (T_n, \xi_n; n \in \mathbb{N})$ is assumed to have free Markovian evolution in the state space E , according to the semigroup P , and to have a forced jump at each time T_n , of amount $\phi(X_{T_n}^v, \xi_n)$, where ϕ is a given transformation from $E^\delta \times U$ into E^δ , the operational function, such that $\phi(\delta, \xi) = \delta$ for all $\xi \in U$. In order to construct the model describing the evolution of X^v , together with the observation process Y , we consider the following space:

$$\tilde{\Omega} = \Omega \times \tilde{\Omega}^{\mathbb{N}}, \quad \tilde{\mathcal{A}} = \mathcal{A} \otimes \tilde{\mathcal{A}}^{\otimes \mathbb{N}}.$$

Let $Y = (Y_t; t \geq 0)$ and $(X^n = (X_t^n; t \geq 0); n \in \mathbb{N})$ be the coordinate processes. To the admissible control $v = (T_n, \xi_n; n \in \mathbb{N})$, we associate the process $X^v = (X_t^v; t \geq 0)$ defined on $(\tilde{\Omega}, \tilde{\mathcal{A}})$ as follows:

$$\forall t \geq 0: \quad X_t^v = \sum_{n=0}^{\infty} X_{t-T_n}^{n+1} I_{\{T_n < t \leq T_{n+1}\}},$$

where T_0 is 0 by convention.

The process X^v is right- and left-limited, and is right-continuous between two successive times T_n, T_{n+1} . We also set $X_\infty^v = \delta$ for all v . The observation filtration generated by Y on $(\tilde{\Omega}, \tilde{\mathcal{A}})$ is again denoted by ${}^0\mathcal{G}$. The filtration $\mathcal{F}^v = (\mathcal{F}_t^v; t \geq 0)$ describing the whole evolution of the systems is defined as in [22], [15], or [18]. For all t , \mathcal{F}_t^v is the σ -field of all the sets $B \in \tilde{\mathcal{A}}$ such that

$$\forall n \in \mathbb{N}: \quad B_n \in \mathcal{A} \otimes \tilde{\mathcal{F}}_\infty^{\otimes n}$$

with the following properties:

$$B_1 = B \cap \{0 \leq t \leq T_1\}, \dots, B_{n+1} = B \cap \{T_n < t \leq T_{n+1}\}$$

for fixed (w_1, \dots, w_n) , the application $w \rightarrow I_{B_n}(w, w_1, \dots, w_n)$ is \mathcal{G} -measurable and for fixed w , the application $(w_1, \dots, w_{n+1}) \rightarrow I_{B_n}(w, w_1, \dots, w_{n+1})$ is $\tilde{\mathcal{A}}^{\otimes n} \otimes \bar{F}_{t-T_n(w)}$ -measurable.

The processes Y and X^v are indeed adapted to the filtration \mathcal{F}^v , but this filtration is a priori larger than the natural filtration of (Y, X^v) . Let $\tilde{\mathcal{A}}^v = \mathcal{F}_\infty^v$.

Now, for a given control v and an initial law μ , let the reference probability be defined on $(\tilde{\Omega}, \tilde{\mathcal{A}})$ by

$$P_\mu^v(d\tilde{w}) = P(dw) K_\mu^v(w; dw_1, \dots, dw_n, \dots),$$

where K_μ^v is the Markovian kernel from (Ω, \mathcal{A}) to $(\tilde{\Omega}, \tilde{\mathcal{A}})^{\otimes N}$ obtained, thanks to the Ionescu-Tulcea Theorem (see [24]), from the following projective family of kernels $(K_\mu^{v,n}; n \in \mathbb{N})$. For all $n \in \mathbb{N}$, $K_\mu^{v,n}$ is the kernel from (Ω, \mathcal{A}) to $(\tilde{\Omega}, \tilde{\mathcal{A}})^{\otimes n}$ defined by

$$K_\mu^{v,n}(w; dw_1, \dots, dw_{n+1}) = \bar{P}_\mu(dw_1) \cdots \bar{P}_{\phi(X_{T_n}^v(w_1, \dots, w_n), \xi_n(w))}(dw_{n+1}),$$

where ϕ is the operational function.

We assume that two simultaneous orders ξ_1 and ξ_2 have the same effect on the inventory level described by X^v as does a unique order, i.e.,

$$\forall x \in E, \quad \forall \xi_1, \xi_2 \in U, \quad \exists \xi \in U \quad \text{such that } \phi(\phi(x, \xi_1), \xi_2) = \phi(x, \xi).$$

The definition of \mathbf{P}_μ^v implies that μ is the initial law of X^v , and for every

$$n \in \mathbb{N}: \quad X_{T_n}^v = \phi(X_{T_n}^v, \xi_n) \quad \mathbf{P}_\mu^v\text{-almost surely.}$$

For each law μ and each admissible control v , let $\mathcal{F}^{v,\mu}$ (respectively, $\mathcal{G}^{v,\mu}$) be the right-continuous finitely \mathbf{P}_μ^v -completion of \mathcal{F}^v (respectively, ${}^0\mathcal{G}$), and let $\tilde{\mathcal{A}}^{v,\mu}$ be the \mathbf{P}_μ^v -completion of $\tilde{\mathcal{A}}^v$.

The reference probability space is by now defined as $(\tilde{\Omega}, \tilde{\mathcal{A}}^v, \mathcal{F}^{v,\mu}, \mathcal{G}^{v,\mu}, \mathbf{P}_\mu^v)$. It can be verified that under the probability \mathbf{P}_μ^v , the filtrations $\mathcal{F}^{v,\mu}$ and $\mathcal{G}^{v,\mu}$ satisfy the following conditional independence property (K):

(K) For every t , the σ -fields $\mathcal{F}_t^{v,\mu}$ and $\mathcal{G}_\infty^{v,\mu}$ are \mathbf{P}_μ^v -conditionally independent given the σ -field $\mathcal{G}_t^{v,\mu}$.

As is shown in [33], this property is of basic importance in order to apply the reference probability method.

Now, let us introduce the change of probability which leads to the filtering model. Given an admissible control v , we define the nonnegative local martingale $L^{v,\mu} = (L_t^{v,\mu}; t \geq 0)$, with respect to $\mathcal{F}^{v,\mu}$ and \mathbf{P}_μ^v , by

$$\forall t \geq 0: \quad L_t^{v,\mu} = \exp \left\{ \int_0^t h(X_s^v) \cdot dY_s - \frac{1}{2} \int_0^t |h(X_s^v)|^2 ds \right\}.$$

The process $L_t^{v,\mu}$ depends on μ only through the stochastic integral with respect to Y . By using results from the parametrized stochastic calculus, as it is developed in [7] and [30], we can choose a version of the stochastic integral, and as a consequence, a version L^v of the above local martingale which does not depend on μ , is everywhere right-continuous from R_+ into \bar{R} , and is adapted to all the filtrations $\mathcal{F}^{v,\mu}$, for $\mu \in M(E)$. The function h being bounded, L^v is therefore on arbitrary finite interval $[0, t]$, an almost surely strictly positive p -integrable ($p \geq 1$) martingale, i.e., $\mu \in M(E)$:

$$L_t^v > 0 \quad \mathbf{P}_\mu^v\text{-almost surely}, \quad E_{\mathbf{P}_\mu^v}(|L_t^v|^p) < \infty,$$

and moreover, $E_{\mathbf{P}_\mu^v}(L_t^v) = 1$. For each finite t , we define a probability $Q_\mu^{v,t}$ on \mathcal{F}_t^v , equivalent to \mathbf{P}_μ^v , by setting

$$\forall A \in \mathcal{F}_t^v: \quad Q_\mu^{v,t}(A) = E_{\mathbf{P}_\mu^v}(L_t^v I_A).$$

The family of probability measures $(Q_\mu^{v,t}; t \geq 0)$ is projective. To conclude that it generates a probability Q_μ^v on $(\tilde{\Omega}, \tilde{\mathcal{A}}^v)$, we must recall the functional structure of this space. From its definition $\tilde{\Omega}$ is the product of the space Ω of all the continuous functions from $[0, \infty)$ into R , by the N -Cartesian product of spaces $\bar{\Omega}$ of the right-continuous left-limited functions from $[0, \infty)$ into E . Then $\tilde{\Omega}$ can be endowed with a topological structure of a Polish space (see [6]) such that $\tilde{\mathcal{A}}$ is the Borel σ -field. In other words, $(\tilde{\Omega}, \tilde{\mathcal{A}})$ is a standard measurable space (see [11, Def. I-3.3]), and therefore there exists

(see [11, Thm. I-3.1]) a probability Q_μ^v on $(\tilde{\Omega}, \tilde{\mathcal{A}}^v)$ such that, for any finite t , the restriction of Q_μ^v to \mathcal{F}_t^v coincides with $Q_\mu^{v,t}$.

With these definitions of the completed filtrations, we see that the filtrations $\mathcal{F}^{v,\mu}$ and $\mathcal{G}^{v,\mu}$ are also right-continuous and Q_μ^v -complete.

To complete the definition of the model, we have to check that the process X^v on $(\tilde{\Omega}, \tilde{\mathcal{A}}^v, \mathcal{F}^{v,\mu}, \mathcal{G}^{v,\mu}, Q_\mu^v)$ enjoys the desired regenerative property, i.e., its evolution between any two successive impulse times remains Markovian with the semigroup P , and that the processes (X^v, Y) represent the classical filtering system, i.e., “observation = signal + white noise.”

This is the purpose of the following two results.

PROPOSITION 2.1. *Let $v = (T_n, S_n, \xi_n; n \in \mathbb{N})$ be an admissible control, and let $\mu \in M(E)$. Then, for any measurable bounded functions f and g on E , the following equality holds Q_μ^v almost surely for every n :*

$$\begin{aligned} E_{Q_\mu^v} \left(e^{-\alpha T_{n+1}} g(X_{T_{n+1}}^v) + \int_{T_n}^{T_{n+1}} e^{-\alpha s} f(X_s^v) ds / \mathcal{F}_{T_n}^v(\tilde{w}) \right) \\ = e^{-\alpha T_n(w)} E_{Q_{X_{T_n}^v(w)}} \left(e^{-\alpha S_n(w, \cdot)} g(X_{S_n(w, \cdot)}^v) + \int_0^{S_n(w, \cdot)} e^{-\alpha s} f(X_s) ds \right) \end{aligned}$$

for Q_μ^v , all \tilde{w} of $\tilde{\Omega}$. In this formula, Q_x and X stand for the probability and the process corresponding to the admissible control such that the first (and the others) impulse time is identically infinite, i.e., for the situation where no control is applied.

Proof. Using the definition of \mathbf{P}_μ^v we have as in [18] the following regenerative property:

$$E_{\mathbf{P}_\mu^v}(U(\theta_{T_n} w, w_n) / \mathcal{F}_{T_n}^v) = E_{\mathbf{P}_{X_{T_n}^v(w)}}(U(w'))$$

for all $U(w, w^x)$ \mathcal{F} -measurable bounded functions.

From the relation

$$T_{n+1}(w) = T_n(w) + S_n(w, \theta_{T_n} w) \quad \forall w \in \Omega,$$

we get for all $t < \infty$

$$T_{n+1}(w) \wedge t = T_n(w) \wedge t + S_n(w, \theta_{T_n} w) \wedge (t - T_n(w))^+.$$

Then it is easy to show, as in [18], that

$$L_{T_{n+1}}^v(\tilde{w}) = L_{T_n}^v(\tilde{w}) L_{S_n(w, \theta_{T_n} w)}(\theta_{T_n} w, w_n) \quad \forall \tilde{w} \in \tilde{\Omega}$$

or, for all $t < \infty$

$$L_{t \wedge T_{n+1}}^v(\tilde{w}) = L_{t \wedge T_n}^v(\tilde{w}) L_{((t - T_n w) \vee 0) \wedge S_n(w, \theta_{T_n} w)}(\theta_{T_n} w, w_n).$$

We first prove the formula of the proposition for a fixed time $t < \infty$. For this we can use the fact that the restriction of Q_μ^v to \mathcal{F}_t^v has L_t^v for Radon-Nikodym derivative with respect to \mathbf{P}_μ^v . Then, the proof is similar to that of [18], and is based on the regenerative property under \mathbf{P}_μ^v . We have the following equalities:

$$\begin{aligned} E_{Q_\mu^v} \left(e^{-\alpha(t \wedge T_{n+1})} g(X_{t \wedge T_{n+1}}^v) + \int_{t \wedge T_n}^{t \wedge T_{n+1}} e^{-\alpha s} f(X_s^v) ds / \mathcal{F}_{t \wedge T_n}^v \right) \\ = I_{\{t \leq T_n\}} e^{-\alpha t} g(x_t^v) + I_{\{t > T_n\}} E_{\mathbf{P}_\mu^v} \\ \cdot \left(L_{t \wedge T_{n+1}}^v (L_{T_n}^v)^{-1} \left(e^{-\alpha(t \wedge T_{n+1})} g(X_{t \wedge T_{n+1}}^v) + \int_{T_n}^{T_{n+1}} e^{-\alpha s} f(X_s^v) ds \right) / \mathcal{F}_{T_n}^v \right) \end{aligned}$$

$$= I_{\{t \leq T_n\}} e^{-\alpha t} g(X_t^v) + I_{\{t > T_n\}} e^{-\alpha T_n} \left(E_{Q_{X_{T_n}^v}} \left(e^{-\alpha(t-T_n) \wedge S_n} g(X_{(t-T_n) \wedge S_n}) + \int_0^{(t-T_n) \wedge S_n} e^{-\alpha s} f(X_s) ds \right) \right).$$

Hence, as $t \uparrow +\infty$, the term in $I_{\{t \leq T_n\}} e^{-\alpha t} \dots$ vanishes, and the term in $I_{\{t > T_n\}} e^{-\alpha T_n} \dots$ converges to $e^{-\alpha T_n} \dots$ (with the convention $e^{-\infty} = 0$). The various processes in this term being either continuous or quasileft continuous and bounded, we get the required equality by Lebesgue dominated convergence.

PROPOSITION 2.2. *Let there be given an admissible control v . For any initial law $\mu \in M(E)$, the processes X^v and Y considered on the space $(\bar{\Omega}, \bar{\mathcal{A}}^v)$ endowed with probability Q_μ^v are such that the process W^v , defined by*

$$\forall t < \infty \quad W_t^v = Y_t - \int_0^t h(X_s^v) ds,$$

is a Brownian motion.

Proof. We apply the Girsanov Theorem [16] on any finite interval $[0, t]$. Then $(W_s^v, s \leq t)$ is a Brownian motion for any finite t , and the conclusion follows.

3. Some properties of the filter. In this section we work on the following filtering model. The unobserved process or signal X is a strong Markov process taking its values in a locally compact space E with a countable basis. Its semigroup $P = (P_t; t \geq 0)$ is assumed to be Feller in the sense that

- (i) $P_t C_0 \subset C_0$ for all $t \geq 0$,
- (ii) $P_t f(x) \rightarrow f(x)$ as $t \downarrow 0$ for arbitrary $x \in E$ and $f \in C_0$,

where C_0 denotes the set of all the real continuous functions on E which vanish at infinity.

Let h be a fixed, bounded, continuous function from E into R^d . The observation is an R^d -valued process $Y = (Y_t; t \geq 0)$ such that the process $W = (W_t; t \geq 0)$, defined by

$$\forall t \geq 0: \quad W_t = Y_t - \int_0^t h(X_s) ds$$

is a d -dimensional standard Brownian motion independent of X .

Both processes X and Y are assumed to be defined on a fixed measurable space (Ω, \mathcal{A}) endowed with a family of probabilities $(P_x; x \in E)$ such that X and Y have the prescribed distributions and $P_x(X_0 = x) = 1$ for all $x \in E$. Let $M(E)$ denote the set of all the probability measures on E considered as a topological space with the weak topology and its Borel σ -field $\mathcal{M}(E)$. For any $\mu \in M(E)$, P_μ denotes the probability on (Ω, \mathcal{A}) defined by $P_\mu = \int_E \mu(dx) P_x$.

Let $\mathcal{F} = (\mathcal{F}_t; t \geq 0)$ (respectively, $\mathcal{G} = (\mathcal{G}_t; t \geq 0)$) be the natural filtration generated by the processes X and Y (respectively, by the process Y). For any $\mu \in M(E)$, \mathcal{A}^μ , \mathcal{F}_t^μ (respectively, \mathcal{G}_t^μ) denote the σ -fields generated by all the subsets which are contained in a P_μ -negligible set of some σ -field \mathcal{F}_s (respectively, \mathcal{G}_s) for s finite, and by the σ -field $\mathcal{F}_{t+} = \bigcap_{\varepsilon > 0} \mathcal{F}_{t+\varepsilon}$ (respectively, $\mathcal{G}_{t+} = \bigcap_{\varepsilon > 0} \mathcal{G}_{t+\varepsilon}$). Moreover, we put $\mathcal{F}^\mu = (\mathcal{F}_t^\mu; t \geq 0)$ and $\mathcal{G}^\mu = (\mathcal{G}_t^\mu; t \geq 0)$.

In summary, the collection $(\Omega, \mathcal{A}^\mu, \mathcal{F}^\mu, \mathcal{G}^\mu, P_\mu, X, Y)$ is a mathematical representation of our filtering model with which we deal in this section.

The filter of X given Y associated to any initial law $\mu \in M(E)$ is defined as the unique, up to a P_μ -negligible set, process $\Pi^\mu = (\Pi_t^\mu; t \geq 0)$ adapted to \mathcal{G}^μ with values

in $M(E)$, and such that for an arbitrary bounded Borel function f on E , the process $\Pi^\mu(f) = (\Pi_t^\mu(f); t \geq 0)$ (where $\Pi_t^\mu(f) = \int_E f(x) \Pi_t^\mu(dx)$, for $t \geq 0$), is the optional projection of the process $f(x) = (f(X_t); t \geq 0)$ with respect to the filtration \mathcal{G}^μ and the probability P_μ (see [6]). Thus, in particular, for arbitrary $t \geq 0$,

$$\Pi_t^\mu(f) = E_\mu(f(X_t) | \mathcal{G}_t^\mu) \quad P_\mu\text{-almost surely,}$$

where $E_\mu(\cdot)$ is the expectation operator with respect to the measure P_μ .

It is well known that the filtering process is Markov and can be described by two types of stochastic equations. In the sequel we mainly use the notion of weak solution or solution in distribution of these equations. To formulate them let $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{\mathbf{P}})$ be an arbitrary probability space with a complete right-continuous filtration $\tilde{\mathcal{F}} = (\tilde{\mathcal{F}}_t; t \geq 0)$. Let $B = (B_t; t \geq 0)$ be a d -dimensional Brownian motion with respect to the filtration $\tilde{\mathcal{F}}$ and the probability $\tilde{\mathbf{P}}$, and $u = (u_t; t \geq 0)$ be an $\tilde{\mathcal{F}}$ -adapted, almost surely continuous process with values in $M(E)$. Moreover, let L and $\mathcal{D}(L)$ denote the infinitesimal generator of the semigroup $P = (P_t; t \geq 0)$ acting on C_0 and its domain. The collection $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{\mathcal{F}}, \tilde{\mathbf{P}}, B, u)$ is said to be a solution of the filtering equation of the first, respectively, second type, if and only if $\tilde{\mathbf{P}}$ -almost surely:

- (I) $u_t(f) = u_0(f) + \int_0^t u_s(Lf) ds + \int_0^t (u_s(fh) - u_s(f)u_s(h)) \cdot dB_s$ for arbitrary $t \geq 0$ and $f \in \mathcal{D}(L)$.
- (II) $u_t(f) = u_0(P_t f) + \int_0^t (u_s(hP_{t-s}f) - u_s(h)u_s(P_{t-s}f)) \cdot dB_s$ for all $t \geq 0$ and all bounded, measurable functions f defined on E .

It follows from the works of Fujisaki, Kallianpur, and Kunita [10] and Kunita [13], that the collection $(\Omega, \mathcal{A}^\mu, \mathcal{G}^\mu, P_\mu, I^\mu, \Pi^\mu)$, is a solution of both the equations with the constant initial condition $u_0 = \mu$ and the Brownian motion I^μ being the following innovation process:

$$\forall t \geq 0: \quad I_t^\mu = Y_t - \int_0^t \Pi_s^\mu(h) ds.$$

The pathwise uniqueness of the solution of (II) has been proved by Kunita [13] for the case of E compact, and his proof carries over to the case of locally compact state spaces E as verified by Stettner and Zabczyk in [29]. The equivalence of (I) and (II), as well as the pathwise uniqueness and uniqueness in law of the solutions, follow from a straightforward extension of the results due to Szpirglas [32]. Consequently an arbitrary solution $u = (u_t; t \geq 0)$ of (I) with the initial condition $u_0 = \mu$, denoted in the sequel as $u^\mu = (u_t^\mu; t \geq 0)$, has the same law as $\Pi^\mu = (\Pi_t^\mu; t \geq 0)$.

We are now going to establish some results concerning the dependence of the filtering process on its initial data. Let \hat{C} denote the space of all continuous bounded functions on $M(E)$, and $\hat{P} = (\hat{P}_t; t \geq 0)$ the operators defined for $\hat{f} \in \hat{C}$ as follows:

$$\forall t \geq 0: \quad \hat{P}_t \hat{f}(\mu) = E_\mu(\hat{f}(\Pi_t^\mu)).$$

We start from the following theorem.

THEOREM 3.1. *If the semigroup $P = (P_t; t \geq 0)$ satisfies (i) and (ii) then*

$$\hat{P}_t \hat{C} \subset \hat{C} \quad \text{for all } t \geq 0.$$

This theorem for the case of compact spaces E is contained in [13], but the second part of the proof from [13] does not generalize to the more general case considered in the present paper. The proof of the following lemma can be found in [13].

LEMMA 3.1. *If $(\mu_n; n \in \mathbb{N})$ is a sequence converging weakly to μ in $M(E)$, then for arbitrary $f \in C$,*

$$\lim_n E(|u_t^{\mu_n}(f) - u_t^\mu(f)|^2) = 0$$

for any fixed t .

To prove Theorem 3.1, let $\hat{f} \in \hat{C}$ and $(f_i; i \in \mathbb{N})$ be a countable set of functions from C_0 dense in C_0 , and assume that $(\mu_n; n \in \mathbb{N})$ converges towards μ weakly. Then, by Lemma 3.1,

$$E(|u_t^{\mu_n}(f) - u_t^\mu(f)|^2) \rightarrow 0 \quad \text{as } n \rightarrow +\infty.$$

By a diagonalization procedure we can construct a subsequence $(\mu_{n_k}; n_k \in \mathbb{N})$ such that for arbitrary $i = 1, 2, \dots$

$$u_t^{\mu_{n_k}}(f_i) \rightarrow u_t^\mu(f_i) \quad P\text{-almost surely.}$$

This implies in turn that P -almost surely

$$u_t^{\mu_{n_k}} \rightarrow u_t^\mu \quad \text{weakly as } k \rightarrow +\infty.$$

Since \hat{f} is a bounded and continuous function on $M(E)$, therefore

$$\begin{aligned} \hat{P}_t \hat{f}(\mu_{n_k}) &= E(\hat{f}(u_t^{\mu_{n_k}})) \\ &\rightarrow E(\hat{f}(u_t^\mu)) = \hat{P}_t \hat{f}(\mu), \end{aligned}$$

and this easily implies continuity of the function $\hat{P}_t \hat{f}$.

We will also need the following theorem in formulation of which ρ stands for the following metric on $M(E)$:

$$\rho(\mu, \nu) = \sum_{i=1}^{\infty} 2^{-i} \frac{|\mu(f_i) - \nu(f_i)|}{1 + \|f_i\|}.$$

THEOREM 3.2. (i) *If $(\mu_n; n \in \mathbb{N})$ is a sequence weakly convergent to μ_∞ , then for arbitrary $\varepsilon > 0, T > 0$, there exists a compact set $\Gamma \subset M(E)$ such that for $n = 1, \dots, \infty$*

$$P_{\mu_n}(\Pi_t^{\mu_n} \in \Gamma \text{ for all } t \leq T) \geq 1 - \varepsilon.$$

(ii) *If $u^\mu = (u_t^\mu; t \geq 0)$ denotes a solution to (I), then for arbitrary $\eta > 0, \varepsilon > 0$, there exists $\delta > 0$ such that for all $\mu \in M(E)$*

$$P\left(\sup_{t < \delta} \rho(\mu, u_t^\mu) < \eta\right) \geq 1 - \varepsilon.$$

The proof of (i) will follow from the next two lemmas.

LEMMA 3.2. *For arbitrary compact set $K \subset E$ and arbitrary numbers $T > 0, \varepsilon > 0, \delta > 0$, we can find a compact set $L \subset E$ containing K such that if $\mu(E \setminus K) \leq \delta$, then*

$$P_\mu\left(\sup_{t \leq T} \Pi_t^\mu(E \setminus L) \geq \varepsilon\right) \leq 2\delta/\varepsilon.$$

Proof. Let D denote the set of all nonnegative dyadic numbers. For arbitrary $t \in D, t \leq T$ the event $\{X_t \in E \setminus L\}$ is contained in the event $B = \{X_s \in E \setminus L \text{ for some } s \in D, s \leq T\}$. Therefore

$$\Pi_t^\mu(E \setminus L) = P_\mu(X_t \in E \setminus L / \mathcal{G}_t^\mu) \leq P_\mu(B / \mathcal{G}_t^\mu) \quad P_\mu\text{-almost surely.}$$

Since the process $(P_\mu(B / \mathcal{G}_t^\mu); t \in D)$ is a \mathcal{G}^μ -martingale, from Doob's inequality we

get the following estimate:

$$\begin{aligned} P_\mu \left(\sup_{t \leq T, t \in D} \Pi_t^\mu(E \setminus L) \geq \varepsilon \right) &\leq P_\mu \left(\sup_{t \leq T, t \in D} P_\mu(B/\mathcal{G}_t^\mu) \geq \varepsilon \right) \\ &\leq \frac{1}{\varepsilon} \sup_{t \leq T, t \in D} E_\mu(P_\mu(B/\mathcal{G}_t^\mu)) \\ &= \frac{1}{\varepsilon} \sup_{t \leq T, t \in D} P_\mu(B). \end{aligned}$$

In order to conclude, we need the following result on Feller processes, which is given by Mackevicius in [17] and can also be found in [29]. For an arbitrary compact set $K \subset E$ and numbers $T > 0$, $\delta > 0$, there exists a compact set L containing K such that

$$\sup_{x \in K} P_x(X_s \in E \setminus L \text{ for some } s \leq T) \leq \delta.$$

Consequently,

$$\begin{aligned} P_\mu(B) &= \int_K P_x(B) \mu(dx) + \int_{E \setminus K} P_x(B) \mu(dx) \\ &\leq \delta \mu(K) + \mu(E \setminus K) \leq 2\delta, \end{aligned}$$

as required.

LEMMA 3.3. *Let $(V_n; n \in \mathbb{N})$ be an increasing sequence of compact subsets of E and $(\lambda_n; n \in \mathbb{N})$ a sequence of positive numbers increasing to one. Then the set*

$$\Gamma = \{\mu \in M(E): \mu(V_n) \geq \lambda_n \text{ for all } n = 1, \dots\}$$

is compact in $M(E)$.

Proof. Define

$$\Gamma_n = \{\mu \in M(E): \mu(V_n) \geq \lambda_n\}, \quad n = 1, \dots;$$

then the Γ_n s are closed subsets of $M(E)$. Thus $\bigcap_n \Gamma_n$ is also closed. Since $\Gamma = \bigcap_n \Gamma_n$, therefore the set is also tight. Consequently Γ is compact.

Proof of Theorem 3.2. (i) We will show that we can take Γ from Lemma 3.3, where $(V_n; n \in \mathbb{N})$ and $(\lambda_n; n \in \mathbb{N})$ have to be properly chosen. Since the family $\{\mu_1, \mu_2, \dots, \mu_\infty\}$ is tight, we can find an increasing sequence of compact sets $(\tilde{V}_n; n \in \mathbb{N})$, such that for all $i = 1, 2, \dots, \infty$

$$\mu_i(E \setminus \tilde{V}_n) \leq \varepsilon 2^{-2n-1} \quad \text{for } n \in \mathbb{N}.$$

By Lemma 3.2, we can find compact sets $V_n \supset \tilde{V}_n$ for every n , such that

$$P_{\mu_i} \left(\sup_{t \leq T} \Pi_t^{\mu_i}(E \setminus V_n) \geq 2^{-n} \right) \leq \varepsilon 2^{-n}$$

for $i = 1, 2, \dots, \infty$.

Therefore taking Γ with $\lambda_n = 1 - 2^{-n}$, $n \in \mathbb{N}$ we get, for $i = 1, 2, \dots, \infty$,

$$\begin{aligned} P_{\mu_i}(\Pi_t^{\mu_i} \in \Gamma \text{ for all } t \leq T) &= P_{\mu_i}(\Pi_t^{\mu_i}(E \setminus V_n) < 2^{-n}; \forall t \leq T, \forall n \in \mathbb{N}) \\ &\geq 1 - \sum_{n \in \mathbb{N}} P_{\mu_i} \left(\sup_{t \leq T} \Pi_t^{\mu_i}(E \setminus V_n) \geq 2^{-n} \right) \\ &\geq 1 - \varepsilon \sum_{n \in \mathbb{N}} 2^{-n} = 1 - \varepsilon, \end{aligned}$$

and the proof of (i) is complete.

(ii) Let us notice that for arbitrary $\delta > 0$,

$$\left\{ \sup_{t \leq \delta, i=1, \dots, k} |u_t^\mu(f_i) - \mu(f_i)| < \eta/2 \right\} \subset \left\{ \sup_{t \leq \delta} \rho(u_t^\mu, \mu) < \eta \right\}$$

where k is an integer such that $2^{-k+2} < \eta$. It is sufficient to show that we can find $\delta > 0$ such that

$$P\left(\sup_{t \leq \delta, i \leq k} |u_t^\mu(f_i) - \mu(f_i)| < \eta/2 \right) \geq 1 - \varepsilon.$$

From the filtering equation (I) we get

$$u_t^\mu(f_i) - \mu(f_i) = \int_0^t u_s^\mu(Lf_i) ds + M_t^i$$

where M^i is the following square integrable martingale:

$$\forall t \geq 0: \quad M_t^i = \int_0^t (u_s^\mu(hf_i) - u_s^\mu(h)u_s^\mu(f_i))dB_s.$$

Thus

$$\begin{aligned} P\left(\sup_{t \leq \delta, i \leq k} |u_t^\mu(f_i) - \mu(f_i)| \geq \eta/2 \right) &\leq \sum_{i=1}^k P\left(\sup_{t \leq \delta} |u_t^\mu(f_i) - \mu(f_i)| \geq \eta/2 \right) \\ &\leq \sum_{i=1}^k P\left(\sup_{t \leq \delta} t \|Lf_i\| \geq \eta/4 \right) + P\left(\sup_{t \leq \delta} |M_t^i| \geq \eta/4 \right). \end{aligned}$$

Since for $\delta > 0$ sufficiently small,

$$\|Lf_i\| < \eta/4\delta \quad \text{for } i = 1, \dots, k$$

and

$$P\left(\sup_{t \leq \delta} |M_t^i| \geq \eta/4 \right) \leq 16\eta^{-2}E(|M_\delta^i|^2) \leq 64\delta\eta^{-2}\|f_i\|^2\|h\|^2,$$

the required estimate independent of the initial condition μ holds.

4. A general optimal stopping result and its application. To construct an optimal strategy for the impulse control problem formulated in § 2 we will need an existence result for optimal stopping problems related to a Markov process \hat{X} on a state space \hat{E} . We use here and in the sequel superscript $\hat{\cdot}$ to underline the fact that introduced objects are related to general separable complete spaces \hat{E} rather than to locally compact spaces E ; as a basic example of \hat{E} , we will take $M(E)$ and \hat{X} will be the filter of § 3.

4.1. An optimal stopping result. Let \hat{B} and \hat{C} denote, respectively, the space of all bounded, Borel functions and the space of all bounded continuous functions, equipped with the sup-norm to topology, defined on a complete metric space \hat{E} . Let $\hat{P} = (\hat{P}_t; t \geq 0)$ be a fixed Markov semigroup acting on \hat{B} . On a probability space $(\Omega, \mathcal{F}_\infty, P)$ with an increasing family $\mathcal{F} = (\mathcal{F}_t; t \geq 0)$ of sub- σ -fields of \mathcal{F}_∞ , let $\hat{X} = (\hat{X}_t; t \geq 0)$ be a stochastic process with values in \hat{E} . \hat{X} is said to be Markov with respect to (\hat{P}, \mathcal{F}) if and only if for arbitrary $\hat{f} \in \hat{C}$, $t \geq s \geq 0$,

$$E(\hat{f}(\hat{X}_t) | \mathcal{F}_s) = \hat{P}_t \hat{f}(\hat{X}_s) \quad P\text{-almost surely.}$$

The following conditions will play the same role in our general situation as the Feller property plays for processes on locally compact spaces E . We will require that

$$(1) \quad \hat{P}_t \hat{C} \subseteq \hat{C} \quad \text{for all } t \geq 0.$$

(2) $\hat{P}_t \hat{f} \rightarrow \hat{f}$ for arbitrary $\hat{f} \in \hat{C}$, as $t \downarrow 0$, uniformly on compact sets.

(3) For arbitrary convergent sequence $x_1, x_2, \dots, x_\infty$ arbitrary $\varepsilon > 0, \delta > 0$, there exists a compact set $\Gamma \subset \hat{E}$ and a sequence of right-continuous processes $\hat{X}^1, \hat{X}^2, \dots$ starting from x_1, x_2, \dots , respectively, Markov with respect to (\hat{P}, \mathcal{F}^n) on $(\Omega^n, \mathcal{F}_\infty^n, P^n)$, $n = 1, 2, \dots, \infty$, such that

$$P^n(\hat{X}_t^n \in \Gamma \text{ for all } t \leq \delta) \geq 1 - \varepsilon, \quad n = 1, 2, \dots.$$

The main results of this section are the following theorems. In their formulation, \mathcal{T} stands for the set of all the \mathcal{F} -stopping times.

THEOREM 4.1. (i) Assume that conditions (1)–(3) hold; then for arbitrary $\hat{\phi} \in \hat{C}$ the following set of inequalities:

$$\begin{aligned} \hat{v} &\geq \hat{\phi}, \\ \hat{v} &\geq e^{-\alpha t} \hat{P}_t \hat{v} \quad \forall t \geq 0 \end{aligned}$$

has a minimal Borel solution \hat{v} which is a continuous function on \hat{E} .

(ii) If $\hat{X} = (\hat{X}_t; t \geq 0)$ is a right-continuous process, Markovian with respect to (\hat{P}, \mathcal{F}) on $(\Omega, \mathcal{F}_\infty, P)$, and ξ is an arbitrary bounded and nonnegative \mathcal{F}_0 -measurable random variable, then

$$(4) \quad E(\xi \hat{v}(\hat{X}_0)) = \sup_{T \in \mathcal{T}} E(\xi e^{-\alpha T} \hat{\phi}(\hat{X}_T)).$$

THEOREM 4.2. If in addition to assumptions of Theorem 4.1, the process \hat{X} is also quasi-left-continuous, then there exists an \mathcal{F} -stopping time T^* such that

$$T^* = \inf \{t \geq 0: \hat{v}(\hat{X}_t) = \hat{\phi}(\hat{X}_t)\} \quad \text{almost surely}$$

and

$$E(\xi \hat{v}(\hat{X}_0)) = E(\xi e^{-\alpha T^*} \hat{\phi}(\hat{X}_{T^*})).$$

A result similar to Theorem 4.1(i) for general state spaces \hat{E} was given by Bensoussan in [2, pp. 315–316]). In particular the continuity of \hat{v} was proved in [2] under condition (1) and condition (2ⁱ).

(2ⁱ) $\hat{P}_t \hat{f} \rightarrow \hat{f}$ for arbitrary $\hat{f} \in \hat{C}$ as $t \downarrow 0$, uniformly on \hat{E} .

Condition (2ⁱ) is, however, stronger than (2) and rarely satisfied for specific examples. Theorem 4.2 follows from Theorem 4.1 and classical results on optimal stopping, given in [4] or in [8] for instance, and therefore its proof will be omitted.

We will prove Theorem 4.1 using discrete-time approximation, as in Mackevicius [17].

Proof of Theorem 4.1(i). For arbitrary $r > 0$, the transformation $T_r: \hat{B} \rightarrow \hat{B}$ given by

$$T_r \hat{f} = \max(\hat{f}, e^{-\alpha r} \hat{P}_r \hat{f})$$

is a contraction from \hat{B} into \hat{B} and also from \hat{C} into \hat{C} . It has therefore a unique fixed point $\hat{v}_r \in \hat{C}$. We check easily that \hat{v}_r is the minimal Borel solution of the problem

$$v \geq \hat{\phi}, \quad v \geq e^{-\alpha r} \hat{P}_r v.$$

Denote $\hat{w}_m = \hat{v}_r$, where $r = 2^{-m}$, for $m = 1, 2, \dots$. Since

$$\hat{v}_r \geq \hat{\phi} \quad \text{and} \quad \hat{v}_r \geq e^{-\alpha r} \hat{P}_r \hat{v}_r \geq e^{-\alpha 2r} \hat{P}_{2r} \hat{v}_r,$$

therefore $\hat{v}_r \geq \hat{v}_{2r}$, and we see that $(\hat{w}_m; m \in \mathbb{N})$ is an increasing sequence of continuous functions converging to a lower semi-continuous function \hat{v} . Note that \hat{v} is the minimal solution of

$$\begin{aligned} v &\geq \hat{\phi}, \\ v &\geq e^{-\alpha r} \hat{P}_r v \quad \text{for all dyadic numbers } r > 0. \end{aligned}$$

The function \hat{w}_m can be easily interpreted as the value function of a discrete time optimal stopping problem (see [27]). Let \mathcal{T}_m denote the set of all the \mathcal{F} -stopping times with values in $\{k2^{-m}; k \in \mathbb{N}\} \cup \{\infty\}$, and consider

$$E(\xi \hat{w}_m(\hat{X}_0)) = \sup_{T \in \mathcal{T}_m} E(\xi e^{-\alpha T} \hat{\phi}(\hat{X}_T))$$

for any \mathcal{F}_0 -measurable integrable random variable ξ .

From this we can immediately deduce part (ii) of Theorem 4.1. Since an arbitrary \mathcal{F} -stopping time T can be approximated by a decreasing sequence of stopping times from $\cup_m \mathcal{T}_m$, since $\hat{w}_m \uparrow \hat{v}$, and since \hat{X} is right-continuous

$$\sup_{T \in \mathcal{T}} [\xi e^{-\alpha T} \hat{\phi}(\hat{X}_T)] = \sup_{T \in \cup_m \mathcal{T}_m} E(\xi e^{-\alpha T} \hat{\phi}(\hat{X}_T)) = \sup_{m \in \mathbb{N}} E(\xi \hat{w}_m(\hat{X}_0)) = E(\xi \hat{v}(\hat{X}_0)).$$

Now let us prove (i). We first have to show that the inequality

$$\hat{v} \geq e^{-\alpha t} \hat{P}_t \hat{v}$$

holds for an arbitrary real positive number t .

For a given t , let $(t_k; k \in \mathbb{N})$ be a decreasing sequence of dyadic numbers converging to t . Then for arbitrary k and m ,

$$\hat{v} \geq e^{-\alpha t_k} \hat{P}_{t_k} \hat{v} \geq e^{-\alpha t_k} \hat{P}_{t_k} \hat{w}_m.$$

Letting $k \rightarrow \infty$ and taking into account that $\hat{w}_m \in \hat{C}$, we obtain from (2) that

$$\hat{P}_{t_k} \hat{w}_m \rightarrow \hat{P}_t \hat{w}_m \quad \text{and} \quad \hat{v} \geq e^{-\alpha t} \hat{P}_t \hat{w}_m.$$

Since $\hat{w}_m \uparrow \hat{v}$, the stipulated property holds.

It is not difficult to check that \hat{v} is the minimal solution of the set of inequalities (i). Finally, we have to show that \hat{v} is continuous. This will be a consequence of property (3) and of the following lemma.

LEMMA 4.1. *Let $\hat{X} = (\hat{X}_t; t \geq 0)$ be a Markov process with respect to (\hat{P}, \mathcal{F}) on $(\Omega, \mathcal{F}_\infty, \mathbf{P})$; let T be an \mathcal{F} -stopping time and T_m its dyadic approximation of order m , defined as follows:*

$$\begin{aligned} T_m &= k2^{-m} \quad \text{if } (k-1)2^{-m} < T \leq k2^{-m}, \quad k \in \mathbb{N}, \\ T_m &= T \quad \text{if } T = 0 \quad \text{or } T = \infty. \end{aligned}$$

Then for arbitrary $\hat{\phi} \in \hat{C}$, and for γ_n defined on \hat{E} by

$$\forall x \in \hat{E}: \quad \gamma_m(x) = \sup_{t \leq 2^{-m}} |e^{-\alpha t} \hat{P}_t \hat{\phi}(x) - \hat{\phi}(x)|$$

we have

$$(5) \quad |E(e^{-\alpha T} \hat{\phi}(X_T)) - E(e^{-\alpha T_m} \hat{\phi}(X_{T_m}))| \leq E(e^{-\alpha T} \gamma_m(X_T)).$$

Let us admit for a while the validity of the preceding lemma in order to achieve the proof of the theorem. From (5) and part (ii), we have

$$E(\hat{v}(\hat{X}_0)) = \sup_{T \in \mathcal{T}} E(e^{-\alpha T} \hat{\phi}(\hat{X}_T)) \leq E(\hat{w}_m(\hat{X}_0)) + \sup_T E(e^{-\alpha T} \gamma_m(\hat{X}_T)).$$

Let us notice that for an arbitrary set $\Gamma \subset \hat{E}$ and numbers $S > 0$,

$$E(e^{-\alpha T} \gamma_m(\hat{X}_T)) \leq \left(\sup_{x \in \Gamma} \gamma_m(x) \right) P(X_t \in \Gamma \text{ for all } t \leq S) + e^{-\alpha S} \sup_{x \in \hat{E}} \gamma_m(x).$$

Now let us consider a convergent sequence x_1, x_2, \dots . Property (3) implies that for arbitrary $\varepsilon > 0$, $\delta > 0$, there exists a compact set Γ such that

$$\hat{v}(x_n) \leq \hat{w}_m(x_n) + (1 - \varepsilon) \sup_{x \in \Gamma} \gamma_m(x) + e^{-\alpha \delta} \sup_{x \in \hat{E}} \gamma_m(x).$$

Property (2) implies that $\sup_{x \in \Gamma} \gamma_m(x) \rightarrow 0$ as $m \uparrow \infty$. Taking this into account, we get

$$\lim_n \hat{v}(x_n) \leq \lim_n \hat{w}_m(x_n).$$

The fact that \hat{w}_m is continuous and that $\hat{w}_m \leq \hat{v}$ implies

$$\hat{v}\left(\lim_n x_n\right) = \lim_n \hat{v}(x_n),$$

the required continuity.

Proof of Lemma 4.1. Let $q(dt, dx)$ denote the distribution of (T, X_T) . Then by the Markov property,

$$E(e^{-\alpha T} \hat{\phi}(\hat{X}_T) - e^{-\alpha T_m} \hat{\phi}(\hat{X}_{T_m})) = \int_0^\infty \int_{\hat{E}} (e^{-\alpha T} \hat{\phi}(x) - e^{-\alpha s(t)} \hat{P}_{s(t)-t} \hat{\phi}(x)) q(dt, dx),$$

where $s(0) = 0$ and $s(t) = (k+1)2^{-m}$ if $k2^{-m} < t \leq (k+1)2^{-m}$ for $k = 1, \dots$, and (5) follows.

4.2. Partially observed optimal stopping. We now apply results of §§3 and 4.1 to an optimal stopping problem with partial observation. Let $\hat{E} = M(E)$ and let F and G be two functions from \hat{C} . Our aim is to maximize the expectation

$$(6) \quad J_\mu(T) = E_\mu \left(\int_0^T e^{-\alpha s} F(\Pi_s^\mu) ds + e^{-\alpha T} G(\Pi_T^\mu) \right)$$

with respect to all \mathcal{G}^μ -stopping times T . This problem will reappear later. It is more general than a classical stopping time problem with partial observation of maximizing—with respect to the same set of stopping times—the payoff

$$(7) \quad E_\mu \left(\int_0^T e^{-\alpha s} f(X_s) ds + e^{-\alpha T} g(X_T) \right),$$

where f and g are bounded and continuous functions defined on E .

Taking into account the definition of the process of conditional distribution Π^μ , we can easily write functional (7) in the form (6) with F and G defined as

$$F(\mu) = \mu(f), \quad G(\mu) = \mu(g).$$

Define

$$H(\mu) = E_\mu \left(\int_0^\infty e^{-\alpha s} F(\Pi_s^\mu) ds \right);$$

then

$$J_\mu(T) = H(\mu) + E_\mu(e^{-\alpha T} (G - H)(\Pi_T^\mu)).$$

Therefore the problem of maximizing $J_\mu(\cdot)$ can be reduced to the simpler case with $F = 0$ and $G \in \hat{C}$ provided that H itself belongs to \hat{C} . However, in the situation which interests us $\hat{P}_t F \in \hat{C}$ for all t , and since

$$H = \int_0^\infty e^{-\alpha s} \hat{P}_s F ds,$$

therefore $H \in \hat{C}$ as needed.

The following theorem will be an easy consequence of the previous results.

THEOREM 4.3. *Under the Feller assumptions (1) and (2) of §4.1, the function \hat{v} on $M(E)$ defined by*

$$\forall \mu \in M(E): \hat{v}(\mu) = \sup_{T \in \mathcal{T}} E_{\mu} \left(\int_0^T e^{-\alpha s} F(\Pi_s^{\mu}) ds + e^{-\alpha T} G(\Pi_T^{\mu}) \right),$$

is continuous and bounded, and the random variable

$$\hat{T} = \inf \{t \geq 0: \hat{v}(\Pi_t^{\mu}) = G(\Pi_t^{\mu})\}$$

is an optimal \mathcal{G}^{μ} -stopping time.

Proof. From what was said earlier it follows that we can assume that $F = 0$ and $G \in \hat{C}$. To apply Theorem 4.1 we must check that the properties (1)–(3) of §4.1 hold in the present situation. Property (1) is exactly the same statement as in Theorem 3.1. To show that (2) holds we apply Theorem 3.2(ii). Let K be a fixed compact subset of $\hat{E} = M(E)$ and let F be an arbitrary function from \hat{C} . For $\varepsilon' > 0$ we can find $\eta > 0$ such that $|F(\mu) - F(\nu)| < \varepsilon'$ for all $\mu \in K$ and $\nu \in \hat{E}$ provided that $\rho(\mu, \nu) < \eta$. It follows from Theorem 3.2(ii) that we can choose $\eta > 0$ in such a way that

$$\begin{aligned} |E_{\mu}(F(\Pi_t^{\mu})) - F(\mu)| &\leq E_{\mu}(|F(\Pi_t^{\mu}) - F(\mu)|; \rho(\mu, \Pi_t^{\mu}) < \eta) \\ &\quad + E_{\mu}(|F(\Pi_t^{\mu}) - F(\mu)|; \rho(\mu, \Pi_t^{\mu}) \geq \eta) \\ &\leq \varepsilon' + 2\varepsilon \|F\|, \end{aligned}$$

provided that $t < \delta$.

Since ε' and ε can be chosen arbitrarily small the uniform convergence of $\hat{P}_t F$ to F on K as $t \downarrow 0$ follows.

Property (3) is given by part (i) of Theorem 3.2. This way the proof of the continuity of \hat{v} is complete. Optimality of the stopping time \hat{T} follows from the pathwise continuity of the process Π^{μ} and Theorem 4.2.

5. The controlled and uncontrolled filters. In this section, we study the filtering process associated to a given admissible control v , defined on the system $(\tilde{\Omega}, \tilde{\mathcal{A}}^v, \mathcal{F}^{v,\mu}, \mathcal{G}^{v,\mu}, Q_{\mu}^v)$, say $\Pi^{v,\mu} = (\Pi_t^{v,\mu}; t \geq 0)$. If the control v is $(T_1 = \infty, \dots)$, and so the system is uncontrolled, we denote by $\Pi = (\Pi_t; t \geq 0)$ the corresponding filtering process.

We prove here that, for any admissible control v , the filter $\Pi^{v,\mu}$ can be expressed in terms of the uncontrolled filter Π by means of regenerative formulae somewhat analogous to those which relate X^v to X .

For a fixed initial law μ , the filtering process of X^v given Y is generally defined on $(\tilde{\Omega}, \tilde{\mathcal{A}}^{v,\mu}, \mathcal{F}^{v,\mu}, \mathcal{G}^{v,\mu}, Q_{\mu}^v)$ as an $M(E)$ -valued process: $\Pi^{v,\mu}(f) = (\Pi_t^{v,\mu}(f); t \geq 0)$ is Q_{μ}^v -undistinguishable from the optional projection of the process $f(X^v) = (f(X_t^v); t \geq 0)$ with respect to the filtration $\mathcal{G}^{v,\mu}$ and the probability Q_{μ}^v . As in [33] or [18], it can be seen that the property (K) of §2 allows us to consider a version of the filtering process defined on the space (Ω, \mathcal{A}) endowed with the right-continuous finitely completed filtration \mathcal{G} and with the restriction of the probability Q_{μ}^v (still denoted by Q_{μ}^v). Moreover, we consider the filtering process as being defined on the space $M(E) \times \Omega$ in order to take into account the dependence on the initial law. Let $\Pi^v = (\Pi_t^v; t \geq 0)$ be the $M(E)$ -valued process defined by the Kallianpur–Striebel formula [12] as follows:

$$\forall t < \infty, \quad \forall f \in b(E), \quad \forall \mu \in M(E) \quad \Pi_t^v(\mu, \cdot)(f) = K_{\mu}^v(L_t^v f(X_t^v)) / K_{\mu}^v(L_t^v).$$

The uncontrolled system is the system associated to the admissible control $v = (\infty, \cdot \cdot \cdot)$, where the first impulse time is identically infinite. Let L be the martingale which governs the corresponding change of probability, as defined in § 2, and let Π^μ be the associated filtering process of $X(\infty, \cdot \cdot \cdot)$ with initial law μ given Y .

PROPOSITION 5.1. *The process Π associated to the uncontrolled system, defined in $M(E) \times \Omega$ by the preceding Kallianpur–Striebel formula, is adapted to the filtration $(\mathcal{M}(E) \otimes \mathcal{G}_t; t \geq 0)$, such that for any $(\mu, w) \in M(E) \times \Omega$, the application $t \rightarrow \Pi_t(\mu, w)$ is right-continuous with respect to the weak topology on $M(E)$. Moreover, for all $\mu \in M(E)$, $\Pi(\cdot, \mu)$ is a version of the filtering process with initial law μ , that is to say: for all $f \in b(E)$ and for all $t \geq 0$,*

$$\Pi_t(\mu, \cdot)(f) = E_{Q_\mu^v}(f(X_t^v) / \mathcal{G}_t^{\mu, v}) = \Pi_t^\mu(f) \quad Q_\mu^v\text{-almost surely for } v = (\infty, \cdot \cdot \cdot).$$

Proof. The fact that $\Pi^v(\mu, \cdot)$ is a version of the filtering process of X^v with initial law μ is an immediate consequence of the property (K) recalled in § 2 (see [33], and also [18]). Moreover, as underlined in § 2, the martingale L^v which governs the change of probability on $(\tilde{\Omega}, \mathcal{A})$ is chosen to be independent of the initial law μ , with all its paths right-continuous, and adapted to all the filtrations $\mathcal{F}^{v, \mu}$, for any given admissible control v . In order to show that Π enjoys the regularity properties of the proposition, we only have to study the process $(K_\mu^v(L_t^v f(X_t^v)); t \geq 0)$ for an arbitrary continuous and bounded function f , and $\mu \in M(E)$. If $v = (\infty, \cdot \cdot \cdot)$, then the process X^v reduces to X^1 and is right-continuous. From the fact that L^v is uniformly integrable on any finite interval, we deduce easily that the process $(K_\mu^v(L_t^v f(X_t^v)); t \geq 0)$ has all its trajectories right-continuous. It is clear that the dependence of the kernels K_μ^v s on μ is measurable, and that if we integrate a P_μ -negligible set of $\mathcal{F}_t^{v, \mu}$ by K_μ^v , we obtain a P -negligible set of \mathcal{G}_t , for $t < \infty$. These remarks lead to the desired adaptation property for Π .

When the system is subject to an admissible control, we have a similar result. More important is the fact that the filtering process associated to an arbitrary admissible control can be expressed in terms of the filtering process of the uncontrolled system. We summarize that expression in the following theorem.

THEOREM 5.1. *The process Π^v associated to a given admissible control $v = (T_n, S_n, \xi_n; n \in \mathbb{N})$, and defined on $M(E) \times \Omega$ by the Kallianpur–Striebel formula, is such that for all $\mu \in M(E)$, $\Pi^v(\mu, \cdot)$ is a version of the filtering process of X^v having μ as initial law. Moreover it satisfies the following properties:*

(i) *The process Π^v is adapted to the filtration $(\mathcal{M}(E) \otimes \mathcal{G}_t; t \geq 0)$ on $M(E) \times \Omega$ and all its paths are right-continuous with respect to the weak topology of $M(E)$ everywhere except on the impulse times T_n . At each instant T_n , the process Π^v admits a right-hand limit and*

$$\Pi_{T_n+}^v = \hat{\phi}(\Pi_{T_n}^v, \xi_n) \quad \text{on } \{T_n < \infty\} \quad \forall n \in \mathbb{N},$$

where $\hat{\phi}$ denotes the continuous function from $M(E) \times E$ into $M(E)$ defined by

$$\forall \mu \in M(E), \quad \forall \xi \in E, \quad \forall f \in b(E): \quad \hat{\phi}(\mu, \xi)(f) = \mu(f(\phi(\cdot, \xi))).$$

(ii) *Let T be a \mathcal{G} -stopping time on Ω such that $T_n \leq T \leq T_{n+1}$ for some $n \in \mathbb{N}$, with the following decomposition:*

$$\forall w \in \Omega: \quad T(w) = T_n(w) + S(w, \theta_{T_n} w),$$

where S is a $\mathcal{G}_{T_n} \otimes \mathcal{A}$ -measurable variable on Ω^2 . Then we have on $\{T_n < \infty\}$

$$\Pi_T^v(\mu, w) = \Pi_{S(w, \theta_{T_n} w)}(\Pi_{T_n+}^v(\mu, w), \theta_{T_n} w).$$

(iii) Given any bounded Borel functions \hat{f} and \hat{g} on $M(E)$, the following identity holds for the time T of assertion (ii):

$$\begin{aligned} E_{Q_\mu^v} \left(\int_{T_n}^T e^{-\alpha s} \hat{f}(\Pi_s^v(\mu, \cdot)) ds + e^{-\alpha T} \hat{g}(\Pi_T^v(\mu, \cdot)) / \mathcal{G}_{T_n}^{\mu, v} \right) \\ = e^{-\alpha T_n(w)} E_{Q_{\Pi_{T_n}^{v+}(\mu, w)}} \left[\int_0^{S(w, \cdot)} e^{-\alpha s} \hat{f}(\Pi_s(\Pi_{T_n}^v(\mu, w), \cdot)) ds \right. \\ \left. + e^{-\alpha S(w, \cdot)} \hat{g}(\Pi_{S(w, \cdot)}(\Pi_{T_n}^v(\mu, w), \cdot)) \right]. \end{aligned}$$

Proof. The regularity result contained in the first assertion is proved as Proposition 5.1. We only have to check the formula for the right-hand limit. Let us remark that the Kallianpur–Striebel formula which defines the process Π^v also holds when we replace the constant time t by a bounded stopping time. Therefore for an arbitrary stopping time T and for any $t \in R_+$

$$\Pi_{t \wedge T}^v(\mu, \cdot)(f) = K_\mu^v(L_{t \wedge T}^v f(X_{t \wedge T}^v)) / K_\mu^v(L_{t \wedge T}^v).$$

Using the fact that $\{T \leq t\} \in \mathcal{G}_t \cap \mathcal{G}_T$, we also have on $\{T \leq t\}$

$$\Pi_T^v(\mu, \cdot) = K_\mu^v(L_T^v f(X_T^v)) / K_\mu^v(L_T^v).$$

By letting $t \uparrow \infty$, we obtain the same formula on $\{T < \infty\}$. With this, we prove the assertion (i) as in [18]. Given an impulse time T_n , we can write the above formula for any stopping time $T_n + \varepsilon$, with $\varepsilon > 0$. Taking the limit in ε , we get on $\{T_n < \infty\}$

$$\begin{aligned} \Pi_{T_n}^v(\mu, \cdot)(f) &= K_\mu^v(L_{T_n}^v f(X_{T_n}^v)) / K_\mu^v(L_{T_n}^v) = K_\mu^v(L_{T_n}^v f(\phi(X_{T_n}^v, \xi_n))) / K_\mu^v(L_{T_n}^v) \\ &= \Pi_{T_n}^v(\mu, \cdot)(f(\phi(\cdot, \xi_n))) = \hat{\phi}'(\Pi_{T_n}^v(\mu, \cdot)(f), \xi_n). \end{aligned}$$

The proof (ii) is similar to the one given in [18], with only minor differences due to the fact that $E_{P_\mu^v}(L_{T_n}^v)$ is not necessarily equal to 1, for every $n \in N$. For each finite t , we can prove as in [18] that

$$L_{t \wedge T_{n+1}}^v(\tilde{w}) = \prod_{k=0}^n L_{((t-T_k(w)) \vee 0) \wedge S_k(w, \theta_{T_k} w)}(\tilde{\theta}_{T_k} \tilde{w}),$$

where $T_0 = 0$ by convention.

Then, if we consider a stopping time T as in the theorem, we deduce from this and from the Kallianpur–Striebel formula (see [18]) that

$$\Pi_{T \wedge t}^v(\mu, w) = \Pi_{((t-T_n(w)) \vee 0) \wedge S(w, \theta_{T_n} w)}(\Pi_{(t \wedge T_n(w))}^v + (\mu, w), \theta_{T_n} w).$$

Taking the limit for $t \uparrow \infty$, we obtain on $\{T < \infty\}$

$$\Pi_T^v(\mu, w) = \Pi_{S(w, \theta_{T_n} w)}(\Pi_{T_n}^v(\mu, w), \theta_{T_n} w).$$

The third assertion is only a weaker formulation of the preceding relation.

6. Separated optimal impulse control. In this section, we prove the existence of a solution for the considered impulse control problem with partial observation. Moreover, the proposed optimal solution is a separated control v , that is to say that we can construct both an impulse control v and a process Π^v , which depend nonanticipatively on each other, such that v is an admissible control and Π^v is the filtering process of the system subject to the control v .

Let us recall the impulse control problem in partial observation to be solved. To each admissible control $v = (T_n, S_n, \xi_n; n \in \mathbb{N})$ and each initial law μ , we associate an average cost $J(v, \mu)$ as follows:

$$J(v, \mu) = E_{Q_\mu^v} \left(\int_0^\infty e^{-\alpha s} f(X_s^v) ds + \sum_{n=1}^\infty e^{-\alpha T_n} C(X_{T_n}^v, \xi_n) \right),$$

where f is a given positive bounded continuous function on E , C is a bounded continuous function on $E \times U$ greater than a positive constant k , and α is a positive number.

To avoid simultaneous orders, it is natural to assume that the ordering cost C verifies the following:

For arbitrary $x \in E$, $\xi_1, \xi_2 \in U$, there exists $\xi \in U$ such that

$$C(x, \xi_1) + C(\phi(x, \xi_1), \xi_2) > C(x, \xi).$$

The average cost $J(v, \mu)$ can be easily expressed by means of the filtering process of X^v given Y :

$$J(v, \mu) = E_{Q_\mu^v} \left(\int_0^\infty e^{-\alpha s} \hat{f}(\Pi_s^v(\mu, \cdot)) ds + \sum_{n=1}^\infty e^{-\alpha T_n} \hat{C}(\Pi_{T_n}^v(\mu, \cdot), \xi_n) \right),$$

where \hat{f} and \hat{C} are the positive bounded continuous functions defined, respectively, on $M(E)$ and $M(E) \times U$ by

$$\hat{f}(\mu) = \mu(f), \quad \hat{C}(\mu, \xi) = \mu(C(\cdot, \xi)) \quad \text{for arbitrary } \mu \in M(E).$$

The following result expresses the Bellman principle for our optimization problem. The proof is based on a technique mainly developed by Robin [26], and also by Bensoussan and Lions [3] and Menaldi [20] in a context of complete observation. The same type of result is stated in [18] for a less general problem in partial observation, with a similar proof. It must be stressed that the fact that the reduite of some continuous function of the filter is continuous is of crucial importance in this theorem.

THEOREM 6.1. *There exists a unique continuous bounded function u on $M(E)$ such that for arbitrary $\mu \in M(E)$:*

$$u(\mu) = \inf_{T \in \mathcal{T}(\mathcal{G})} E_{Q_\mu} \left(\int_0^T e^{-\alpha s} \hat{f}(\Pi_s(\mu, \cdot)) ds + e^{-\alpha T} M u(\Pi_T(\mu, \cdot)) \right),$$

where M is the operator on \hat{C} , the space of all bounded continuous functions on $M(E)$, defined by

$$Mh(\mu) = \inf_{\xi \in U} \hat{C}(\mu, \xi) + h(\hat{\phi}(\mu, \xi)) \quad \forall \mu \in M(E), \quad \forall h \in \hat{C}.$$

Moreover, the infimum is achieved for the following \mathcal{G} -stopping time $T^*(\mu)$, for any initial law μ ,

$$T^*(\mu) = \inf \{t \geq 0: u(\Pi_t(\mu, \cdot)) = M u(\Pi_t(\mu, \cdot))\}$$

and there exists a measurable selection $\mu \rightarrow \xi^*(\mu)$ on $M(E)$ such that

$$M u(\mu) = \hat{C}(\mu, \xi^*(\mu)) + u(\hat{\phi}(\mu, \xi^*(\mu))), \quad \mu \in M(E).$$

Sketch of proof. We give just a sketch of the proof, which is classical (see [26], [20], or [3]). Let us define inductively the sequence $(u^n; n = 0, 1, \dots)$ of functions on

$M(E)$ by

$$u^0(\mu) = E_{Q_\mu^v} \left(\int_0^\infty e^{-\alpha s} \hat{f}(\Pi_s(\mu, \cdot)) ds \right),$$

$$u^n(\mu) = \inf_{T \in \mathcal{T}(\mathcal{G})} E_{Q_\mu^v} \left(\int_0^T e^{-\alpha s} \hat{f}(\Pi_s(\mu, \cdot)) ds + e^{-\alpha T} M u^{n-1}(\Pi_T(\mu, \cdot)) \right) \quad \forall n \in N.$$

From the Feller property of the filter, it is clear that u^0 is continuous and bounded on $M(E)$. Let us prove by induction that all the functions u^n s are continuous and bounded. If u^{n-1} is continuous and bounded, so is $M u^{n-1}$ by a classical argument. Then u^n is the optimal cost function for a stopping time problem associated to the functions \hat{f} and $M u^{n-1}$; it is continuous on $M(E)$ by the result of Theorem 4.3. The rest of the proof is classical (see [26]): it can be shown that the sequence $(u^n; n \in N)$ is decreasing and converges uniformly towards a limit u which is the unique solution of the equation given in the theorem. The second part of the theorem follows again from Theorem 4.3.

Now let us define what will be our optimal separated control. We first construct inductively both a sequence of random variables $v = (T_n, \xi_n; n \in N)$ and a process Π^* with values in $M(E)$ and starting from μ . To this end, we mainly use the random variables T^* and ξ^* defined in Theorem 6.1 and the filter Π associated to the uncontrolled system defined by the Kallianpur–Striebel formula of § 5. We define the commutation set I as in [3] or [18] by

$$I = \{\mu \in M(E) \text{ such that } u(\mu) = M u(\mu)\}.$$

Let μ be given. We distinguish two cases.

Case 1. If $\mu \in I$, we set

$$\begin{aligned} T_1 &= 0, \quad \Pi_0^* = \mu, \quad \xi_1 = \xi^*(\mu), \quad \Pi_0^{*+} = \hat{\phi}(\Pi_0^*, \xi_1), \\ T_2(w) &= T^*(\Pi_0^{*+}, w), \quad \Pi_t^* = \Pi_t(\Pi_0^{*+}, w) \quad \text{for } t \leq T_2(w), \\ \xi_2(w) &= \xi^*(\Pi_{T_1}^*(w)), \quad \Pi_{T_2}^{*+} = \hat{\phi}(\Pi_{T_2}^*, \xi_2), \\ T_3(w) &= T^*(\Pi_{T_2}^{*+}(w), \theta_{T_2} w) + T_2(w), \\ \Pi_t^*(w) &= \Pi_t(\Pi_{T_2}^{*+}(w), \theta_{T_2} w) \quad \text{for } T_2(w) < t \leq T_3(w) \end{aligned}$$

and inductively, for $n \geq 3$,

$$\begin{aligned} \xi_n(w) &= \xi^*(\Pi_{T_n}^*(w)), \quad \Pi_{T_n}^{*+} = \hat{\phi}(\Pi_{T_n}^*, \xi_n) \\ T_{n+1}(w) &= T^*(\Pi_{T_n}^{*+}(w), \theta_{T_n} w) + T_n(w), \\ \Pi_t^*(w) &= \Pi_t(\Pi_{T_n}^{*+}(w), \theta_{T_n} w) \quad \text{for } T_n(w) < t \leq T_{n+1}(w). \end{aligned}$$

Case 2. If $\mu \in I$, we set

$$\begin{aligned} T_1(w) &= T^*(\mu, w), \quad \Pi_t^* = \Pi_t(\mu, w) \quad \text{for } t \leq T_1(w), \quad \xi_1(w) = \xi^*(\Pi_{T_1}^*(w)), \\ \Pi_{T_1}^{*+} &= \hat{\phi}(\Pi_{T_1}^*, \xi_1), \quad T_2(w) = T^*(\Pi_{T_1}^{*+}(w), \theta_{T_1} w) + T_1(w), \end{aligned}$$

and inductively, for $n \geq 2$

$$\begin{aligned} \xi_n(w) &= \xi^*(\Pi_{T_n}^*(w)), \quad \Pi_{T_n}^{*+} = \hat{\phi}(\Pi_{T_n}^*, \xi_n), \\ T_{n+1}(w) &= T^*(\Pi_{T_n}^{*+}(w), \theta_{T_n} w) + T_n(w), \\ \Pi_t^*(w) &= \Pi_t(\Pi_{T_n}^{*+}(w), \theta_{T_n} w) \quad \text{for } T_n(w) < t \leq T_{n+1}(w). \end{aligned}$$

PROPOSITION 6.1. (i) *The control $v = (T_n, \xi_n; n \in N)$ defined by the above construction is admissible.*

(ii) The process Π^* is the filtering process of X^v with initial law μ , given the observation Y .

Proof. We first check that the construction given above defines an $M(E)$ -valued process Π^* which is adapted to the filtration \mathcal{G} , right-continuous except at the instants T_n , where

$$\Pi_{T_n}^{*+} = \lim_{\varepsilon \downarrow 0} \Pi_{T_n+\varepsilon}^* \quad \forall n.$$

From this, we deduce easily that the T_n 's are \mathcal{G} -stopping times and each ξ_n is a \mathcal{G}_{T_n} -random variable. We verify that the sequence $(T_n; n \in N)$ is strictly increasing towards ∞ , as in [18]. This proves (i). Then, using Theorem 5.1, we see that the process Π^* and the filtering process $\Pi^{v,\mu}$ associated to the admissible control v and the initial law μ coincide on every stochastic interval $]T_n, T_{n+1}]$. Therefore they are equal on $[0, \infty[$.

More generally, every control v such that Proposition 6.1 holds is called a separated control. Then, the times T_n appear to be the successive entry times of the process Π^v into the set I .

Finally, by using the results of the preceding sections, we get the separation theorem for the partially observed impulse control problem.

THEOREM 6.2. *The separated control $v = (T_n, \xi_n; n \in N)$ defined above is optimal when μ is the initial law, that is to say*

$$J(v, \mu) = \inf_{w \in \mathcal{V}} J(w, \mu).$$

The proof is quite similar to the one given in [18] and therefore is omitted.

REFERENCES

- [1] R. F. ANDERSON AND A. FRIEDMAN, *Multidimensional quality control problems and quasi variational inequalities*, Trans. Amer. Math. Soc., 246 (1978), pp. 31–76.
- [2] A. BENSOUSSAN, *Stochastic Control by Functional Analysis Methods*, North-Holland, Amsterdam, New York, 1982.
- [3a] A. BENSOUSSAN AND J. L. LIONS, *Applications des inéquations variationnelles au contrôle stochastique*, Dunod, Paris, 1978.
- [3b] ———, *Contrôle impulsif et inéquations quasi-variationnelles*, Dunod, Paris, 1983.
- [4] J. M. BISMUT AND B. SKALLI, *Temps d'arrêt optimal, théorie générale des processus, et processus de Markov*, Z. Wahrsch. Verw. Gebiete, 39 (1977), pp. 301–314.
- [5] R. M. BLUMENTHAL AND R. K. GETTOOR, *Markov Processes and Potential Theory*, Academic Press, New York, London, 1968.
- [6a] C. DELLACHERIE AND P. A. MEYER, *Probabilités et potentiel*, Tome 1, Hermann, Paris, 1975.
- [6b] ———, *Probabilités et potentiel*, Tome 2, Hermann, Paris, 1980.
- [7] C. DOLEANS-DADE, *Intégrales stochastiques dépendant d'un paramètre*, Publ. Inst. Statist. Univ. Paris, 16 (1967), pp. 23–34.
- [8] N. EL KAROUI, *Les aspects probabilistes du contrôle stochastique*, Ecole d'été de probabilités de Saint-Flour VIII, 1979, Lecture Notes in Math., 876, Springer-Verlag, Berlin, New York, 1981.
- [9] A. FRIEDMAN, *On the free boundary of a quasi variational inequality arising in a problem of quality control*, Trans. Amer. Math. Soc., 246 (1978), pp. 95–110.
- [10] M. FUJISAKI, G. KALLIANPUR, AND H. KUNITA, *Stochastic differential equations for the non-linear filtering problem*, Osaka J. Math., 9 (1972), pp. 19–40.
- [11] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, New York, 1981.
- [12] G. KALLIANPUR AND C. STRIEBEL, *Estimation of stochastic systems*, Ann. Math. Statist., 39 (1968), pp. 785–801.

- [13] H. KUNITA, *Asymptotic behaviour of the non-linear filtering errors of Markov processes*, J. Multivariate Anal., 1 (1971), pp. 365–393.
- [14] J. P. LEPELTIER AND B. MARCHAL, *Techniques probabilistes dans le contrôle impulsif*, Stochastics, 2 (1979), pp. 243–286.
- [15] ———, *Theorie générale du contrôle impulsif*, SIAM J. Control Optim., 22 (1984), pp. 646–665.
- [16] R. S. LIPTSER AND A. N. SHIRYAYEV, *Statistics of random processes*, Appl. Math., 5 (1977).
- [17] V. MACKEVICIUS, *Convergence of the value of the game in optimal stopping of Markov processes*, Liet. Mat. Rink., 14 (1974), pp. 113–127.
- [18] G. MAZZIOTTO AND J. SZPIRGLAS, *Separation principle for impulse control with partial information*, Stochastics, 10 (1983), pp. 47–73.
- [19] G. MAZZIOTTO, L. STETTNER, J. SZPIRGLAS, AND J. ZABCZYK, *On Impulse Control with Partial Observation*, Lecture Notes Control. Inform. Sci., 69, Springer-Verlag, Berlin, 1985, pp. 296–308.
- [20] J. L. MENALDI, *On the optimal impulse control problem for degenerate diffusions*, this Journal, 18 (1980), pp. 722–739.
- [21] ———, *The separation principle for impulse control problems*, Proc. Amer. Math. Soc. (1980).
- [22] P. A. MEYER, *Renaissance, recollements, mélanges, ralentissements d'un processus de Markov*, Ann. Inst. Fourier (Grenoble), XXV (1975), pp. 465–497.
- [23] H. NAGAI, *On an impulsive control of additive processes*, Z. Wahrsch. Verw. Gebiete, 53 (1980), pp. 1–16.
- [24] J. NEVEU, *Bases mathématiques des probabilités*, Masson, Paris, 1964.
- [25] R. W. RISHEL, *The minimum principle, separation principle and dynamic programming for partially observed jump processes*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 141–143.
- [26] M. ROBIN, *Contrôle impulsif des processus de Markov*, Thèse, Univ. de Paris, 1978.
- [27] A. N. SHIRYAYEV, *Optimal Stopping Rules*, Appl. Math., 8 (1978), monograph.
- [28] L. STETTNER, *On optimal stopping of Feller Markov processes with incomplete information in locally compact state space*, preprint.
- [29] L. STETTNER AND J. ZABCZYK, *Optimal stopping for Feller processes*, preprint.
- [30] C. STRICKER AND M. YOR, *Calcul stochastique dépendant d'un paramètre*, Z. Wahrsch. Verw. Gebiete, 45 (1978), pp. 109–133.
- [31] D. W. STROOK AND S. R. S. VARADHAN, *Multidimensional diffusion processes*, Springer-Verlag, Berlin, New York, 1979.
- [32] J. SZPIRGLAS, *Sur l'équivalence d'équations différentielles stochastiques à valeurs mesure intervenant dans le filtrage markovien non linéaire*, Ann. Inst. H. Poincaré Sect. B (N.S.), XIV (1978), pp. 33–59.
- [33] J. SZPIRGLAS AND G. MAZZIOTTO, *Modèle général de filtrage non linéaire et équations différentielles stochastiques associées*, Ann. Inst. H. Poincaré Sect. B (N.S.), XV (1979), pp. 147–173.

A DUAL APPROACH TO MULTIDIMENSIONAL L_p SPECTRAL ESTIMATION PROBLEMS*

A. BEN-TAL†, J. M. BORWEIN‡, AND M. TEOULLE§

Abstract. A complete duality theory is presented for the multidimensional L_p spectral estimation problem. The authors use a new constraint qualification (BWCQ) for infinite-dimensional convex programs with linear type constraints recently introduced in [Borwein and Wolkowicz, *Math. Programming*, 35 (1986), pp. 83–96]. This allows direct derivation of the explicit optimal solution of the problem as presented in [Goodrich and Steinhardt, *SIAM J. Appl. Math.*, 46 (1986), pp. 417–426], and establishment of the existence of a simple and computationally tractable unconstrained Lagrangian dual problem. Moreover, the results illustrate that (BWCQ) is more appropriate to spectral estimation problems than the traditional Slater condition (which may only be applied after transformation of the problem into an L_p space [Goodrich and Steinhardt, *op. cit.*] and which therefore yields only necessary conditions).

Key words. spectral estimation, infinite-dimensional convex duality, constraint qualifications, moment problems

AMS (MOS) subject classifications 49, 90C, 93

1. Introduction. A basic problem in spectral estimation is the estimation of a *power spectrum*, a measure $\mu \in \mathbb{R}^n$, with a known support, given a finite collection of measured correlations. This problem arises in a wide range of settings such as geophysics, radio astronomy, sonar, and radar, to mention just a few. (See Kay and Marple [8] and the references therein.) In many of these physical problems, the power spectrum is represented by a density. Let $K \subset \mathbb{R}^n$ be a measure space with a σ -finite positive measure dk and let μ be absolutely continuous with respect to dk with density (Radon–Nikodym derivative) $s(k) = d\mu/dk$. Then the problem of estimating a power spectral density function may be stated as the following: Find a nonnegative integrable s on K which vanishes outside of K and exactly matches the measured correlations:

$$(1.1) \quad r(\delta) = \int_K e^{j\delta \cdot k} s(k) dk, \quad \delta \in \Delta \quad (j := \sqrt{-1})$$

where Δ is a finite subset of \mathbb{R}^n with $0 \in \Delta$, $\Delta = -\Delta$ and $\bar{r}(\delta) = r(-\delta)$ for $\delta \in \Delta$. Even when feasible, this problem does not have a unique solution. To overcome this ambiguity, the selection of such an estimate s from the infinite number of possible choices is done by requiring the spectral estimate to optimize some convex functional.

One popular method is Burg's maximum entropy method [2]. In this method the spectral estimate is required to maximize the entropy functional

$$(1.2) \quad H(s) = \int_K \log s(k)$$

* Received by the editors November 20, 1986; accepted for publication (in revised form) October 21, 1987.

† Department of Industrial and Operation Engineering, University of Michigan, Ann Arbor, Michigan 48105-2117.

‡ Department of Mathematics, Statistics and Computing Science, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5. The research of this author was partially supported by an operating grant from the National Sciences and Engineering Research Council of Canada.

§ Department of Mathematics, University of Maryland, Baltimore County, Catonsville, Maryland 21228. The research of this author was supported by a Postdoctoral Fellowship at the Department of Mathematics, Statistics and Computing Science, Dalhousie University.

subject to the correlation-matching constraints (1.1). It can be shown [2] that if there exists a strictly positive trigonometric polynomial P such that the spectral density $1/P(k)$ satisfies the constraints (1.1), then $s_0(k) = 1/P(k)$ is the unique spectral density function solving (1.2) subject to (1.1) (see also the Appendix). In some applications [8], a strictly positive solution of the form $1/P(k)$ fails to exist. This motivated Goodrich and Steinhardt [5] to suggest in a recent paper, an alternative way for selecting a solution to problem (1.1) by formulating the following minimal L^2 norm problem; find $s^* \in L^2$ solving

$$(1.3) \quad \inf \left\{ \int_K s(k)^2 dk : \int_K e^{j\delta \cdot k} s(k) dk = r(\delta), \delta \in \Delta, S \geq 0 \right\}.$$

Under appropriate conditions, it is shown there that the optimal solution $s^*(k) = \max(0, P(k))$, where P is a trigonometric polynomial. The method of proof used in [5] is the Lagrange multiplier technique (11) applied to problem (1.3). The difficulty with this approach is that the nonnegative orthant in L_2 has an empty interior and thus the usual ‘‘Slater type’’ constraint qualification fails. To overcome this difficulty, it is shown in [5] that problem (1.3) can be reduced to a minimization problem in L_∞ by using the transformation $t(k) = (s(k))^{1/2}$. For the transformed problem the constraint qualification is satisfied and thus a Kuhn–Tucker theorem for infinite-dimensional problems (see Luenberger [13, p. 217]) is applicable.

In this paper we develop a duality theory for problem (1.3) which enables us to derive the results obtained in [5], directly from this duality framework. We prove our results using a constraint qualification recently introduced by Borwein and Wolkowicz [1], called accordingly (BWCQ). The dual problem obtained is a *finite-dimensional concave* program and the optimal dual variables are exactly the parameters of the optimal spectral density s^* . Moreover the simple unconstrained nature of the dual problem is suitable for computational purposes and may lead to the construction of reliable algorithms for computing the L^2 optimal spectral estimate s^* . In § 2 we review some basic results on the *extendibility problem* as developed by Lang and McClellan [11] and state the multidimensional spectral estimation problem (P) in an appropriate vector space. In § 3 we present the new constraint qualification (BWCQ) and obtain an explicit version of it for problems of type (P). In particular we show that the (BWCQ) is equivalent to the extendibility condition given in § 2. In § 4 a complete duality theory for the multidimensional problem is derived, including the explicit solution presented in [5]. In § 5 we specialize the duality results to the one-dimensional time series problem. In this case we prove that (BWCQ) simply reduces to the classical test of positive definiteness of a Toeplitz matrix formed from the correlation samples.

Finally, § 6 contains concluding remarks and briefly discusses how to extend our results when the constraint qualification fails.

2. Multi-dimensional spectral estimation and properties. We collect in this section some preliminary results needed in the sequel. We follow the work of Lang and McClellan [11]. Let Δ be a finite set in \mathbb{R}^n of the form $\Delta = \{0, \pm\delta_1, \dots, \pm\delta_m\}$ and assume that

$$\{e^{jk \cdot \delta} : \delta \in \Delta\} \quad (j \text{ denotes } \sqrt{-1})$$

is a set of linearly independent functions on K a compact subset of \mathbb{R}^n . The measure correlations $r(\delta)$ are *conjugate symmetric* functions, i.e., $r(-\delta) = \bar{r}(\delta)$ for all $\delta \in \Delta$. The set Δ has $(2m+1)$ elements and so a conjugate symmetric function on Δ is characterized by $(2m+1)$ real numbers; thus we may think of $(r(\delta))_{\delta \in \Delta}$ as a $(2m+1)$ vector r in

\mathbb{R}^{2m+1} . Similarly we denote by ψ_k the vector in the $(2m+1)$ -dimensional Euclidean space with components $\psi_k(\delta) = e^{jk \cdot \delta}$.

Let E denote the set of *extendible correlation vectors*, that is, $r \in E$ if

$$(2.1) \quad r = \int_K \psi_k d\mu$$

for some finite positive measure μ on K . We define a real-valued Δ -polynomial by

$$P(k) = \sum_{\delta \in \Delta} p(\delta) e^{-jk \cdot \delta} \quad \text{where } p(-\delta) = \bar{p}(\delta).$$

Here we will also use the vector notation $p \in \mathbb{R}^{2m+1}$ for the coefficients $(p(\delta))_{\delta \in \Delta}$ of a Δ -polynomial $P(k)$. Define the set Q of positive Δ -polynomials $Q := \{p \in \mathbb{R}^{2m+1} : P(k) \geq 0, k \in K\}$. The inner product between a vector r of correlations samples and a vector p of polynomial coefficients is defined as

$$(r, p) = \sum_{\delta \in \Delta} \bar{r}(\delta) p(\delta).$$

Note that if $r = \int_K \psi_k d\mu$ then $(r, p) = \int_K P(k) d\mu$. The following theorem provides a characterization of extendibility.

THEOREM 2.1 [11]. *The vector r is extendible if and only if*

$$(2.2) \quad (r, p) \geq 0 \quad \text{for all } p \in Q.$$

Note that E and Q are closed convex cones and thus (2.2) simply states that $E = Q^+$, the polar cone of Q . A simple property of extendibility [11] is that $\text{int } E \neq \emptyset$ and

$$(2.3) \quad \text{int } E = \{r \in E : (r, p) > 0 \text{ for all } p \in Q, p \neq 0\}.$$

As mentioned in the Introduction, in many applications the power spectrum is not represented merely as a measure but as a density. Let K be a measure space with a σ -finite positive measure dk . For $1 < \alpha < \infty$, $L_\alpha \equiv L_\alpha(K, dk)$ will denote the Lebesgue space of α th power integrable real-valued functions on K , i.e.,

$$L_\alpha = \left\{ f : \int_K |f(k)|^\alpha dk < \infty \right\}.$$

Let μ be absolutely continuous with respect to dk with density (Radon-Nikodym derivative) $s(k) = d\mu/dk$. The constraint correlation measurements (2.1) are now

$$r = \int_K \psi_k s(k) dk, \quad s \in L_\alpha.$$

This leads to the following modified extension theorem for spectral densities.

THEOREM 2.2 [11]. *Suppose that K has a finite measure. If every neighborhood of each point in K has a strictly positive dk -measure (as holds with dk equivalent to Lebesgue measure on K), then $r \in \text{int } E$ if and only if there exists a continuous strictly positive function $s(k)$ such that*

$$r = \int_K \psi_k s(k) dk.$$

Remark 2.1. It is interesting to note that Theorems 2.1 and 2.2 remain valid when the linear independence assumption of the particular functions $\{e^{jk \cdot \delta} : \delta \in \Delta\}$ is replaced by any linear independent set of continuous functions on K .

Remark 2.2. It can be shown easily using Theorem 2.1 that in Theorem 2.2 the statement “there exists a *continuous* strictly positive function $s(k)$ ” can be replaced

by the formally weaker statement “there exists a strictly positive function in L_α .” This weaker version will in fact be sufficient for our purposes (see § 3).

We close this section by formulating our problem (1.3) in the appropriate normed vector space. We denote the nonnegative cone in L_α by $L_\alpha^+ = \{f \in L_\alpha : f \geq 0 \text{ a.e.}\}$. Consider the integral functional $I : L_\alpha \rightarrow \mathbb{R}$ defined by

$$I(s) := \|s\|_\alpha^\alpha = \int_K |s(k)|^\alpha dk$$

and the linear operator $A : L_\alpha \rightarrow \mathbb{R}^{2m+1}$ given by

$$(2.4) \quad s \rightarrow \int_K \psi_k s(k) dk.$$

Then our L_α -density estimation problem is to solve

$$(2.5) \quad (P) \quad \inf \{I(s) : As = r, s \in L_\alpha^+\}.$$

Note that for $1 < \alpha < \infty$, $I(s)$ is a convex Fréchet differentiable functional and so (P) is a linearly constrained convex differentiable optimization problem.

3. A constraint qualification for problem (P). In this section we study programs of type (P) and introduce a duality theory for them using a setting recently developed by Borwein and Wolkowicz [1]. We state a version of their results which will be appropriate for problem (P). Consider the following primal problem:

$$(A) \quad v(P) = \inf \{f(x) : Ax = b, x \in S\}$$

where $f : X \rightarrow \mathbb{R}$ is a differentiable convex functional on X , $A : X \rightarrow \mathbb{R}^N$ is a continuous linear operator, X is a normed space, $b \in \mathbb{R}^N$ and $S \subset X$ is a convex cone. We let X^* denote the continuous dual space of X and $\langle \cdot, \cdot \rangle$ the bilinear form in the duality between X and X^* . The Lagrangian dual problem (B) associated with (A) is defined as

$$(B) \quad v(D) = \sup_{y \in \mathbb{R}^N} \{y^T b + \inf_{x \in S} \{f(x) - \langle y, Ax \rangle\}\}.$$

The main result concerning the dual pair (A) and (B) is the existence of a saddle point (x^*, y^*) , for (A) or equivalently the validity of a strong duality result:

$$(3.1) \quad v(P) = v(D) \quad \text{and} \quad v(D) \text{ is attained by some } y^* \in \mathbb{R}^N.$$

Usually, under the familiar “Slater type” constraint qualification,

$$(SCQ) \quad \text{There exists } x \in \text{int } S \text{ such that } Ax = b$$

the strong duality relation is guaranteed. Unfortunately, in problem (P) (see (2.5)) (SCQ) does not hold, since there $S = L_\alpha^+$ the nonnegative cone in L_α , which has empty interior. However, we shall make use of a recent result given in [1] which guarantees the strong duality relation (3.1). We need the following definitions and notation. Given any set C in X , the *polar cone* of C is the set

$$C^+ = \{x' \in X^* : \langle x', x \rangle \geq 0 \text{ if } x \in C\}.$$

Let P be any generating set for S^+ , i.e.,

$$\text{cl } \{\text{cone } P\} = S^+$$

and define

$$(3.2) \quad P^\circ := \{s^+ \in P : Ax = b, x \in S \Rightarrow \langle s^+, x \rangle = 0\}.$$

We denote by $F = \{x : Ax = b, x \in S\}$ the feasible set of (A) and we assume $F \neq \emptyset$.

THEOREM 3.1 [1]. Assume that $v(P)$ is finite. If (BWCQ): $P^\circ \subset \{0\}$ holds. Then, $v(P) = v(D)$ and $v(D)$ is attained by some $y^* \in \mathbb{R}^N$. \square

Remark 3.1. It is interesting to mention that (BWCQ) is equivalent to the condition $\text{cl cone}(F - S) = X$ (see [1]).

We specialize (BWCQ) to a problem of type (P). Let K be a measure space with σ -finite positive measure dk . Let $X = L_\alpha(K, dk) = L_\alpha$, $1 < \alpha < \infty$, and $S := L_\alpha^+$ be the nonnegative orthant in L_α . Further, the bilinear form

$$\langle s, x \rangle = \int_K s(k)x(k) dk$$

defines the duality between L_α and L_β , where L_β is identified as usual with the dual space of L_α , and $\alpha + \beta = \alpha\beta$, ($1 < \beta < \infty$).

THEOREM 3.2. Suppose L_α is separable. For $1 \leq \alpha < \infty$ (BWCQ) holds if and only if the weak "Slater type" condition

$$(3.3) \quad \exists \hat{x} \in L_\alpha : A\hat{x} = b, \quad \hat{x} > 0 \quad (\text{a.e.})$$

is satisfied.

Proof. First, we show that if (3.3) holds then $P^\circ \subset \{0\}$. Let $\hat{x} \in L_\alpha$ such that $A\hat{x} = b$ and $\hat{x} > 0$ (a.e.). If $s^+ \in P^\circ$, then by Definition (3.2), $s^+ \in S^+$ and $\langle s^+, \hat{x} \rangle = \int_K s^+(k)\hat{x}(k) dk = 0$. Let $A_n := \{k : \hat{x}(k) \geq 1/n\}$, $n = 1, 2, \dots$; then since $s^+ \in S^+$ implies that $s^+ \in L_\beta$ with $s^+ \geq 0$ (a.e.), we have

$$0 \leq \frac{1}{n} \int_{A_n} s^+(k) dk \leq \int_K s^+(k)\hat{x}(k) dk = 0$$

so that $\int_{A_n} s^+(k) dk = 0$. But $\{k : x(k) > 0\} = \bigcup_{n=1}^\infty A_n$; hence

$$0 = \int_{\bigcup A_n} s^+(k) dk = \int_K s^+(k) dk \quad \text{implying } s^+(k) = 0 \quad (\text{a.e. } dk).$$

Next, we show the reverse implication. Since L_α is assumed separable, and since P° is trivial, there is a countable set $\{x_n\}_{n=1}^\infty$ in L_α such that

- (i) $Ax_n = b$, $x_n \in S$;
- (ii) If $0 \neq s^+ \in S^+$, $\langle s^+, x_n \rangle > 0$ for some $n \in \mathbb{N}$.

If $A = 0$ there is nothing to prove; then let us assume $A \neq 0$. We distinguish between two cases.

Case 1. $b \neq 0$. Then since $F \neq \emptyset$, $Ax_n = b$ implies that $\|A\| \|x_n\| \geq \|b\|$ and thus $\|x_n\| \geq c > 0$. Hence the series $\sum_{n=1}^\infty 1/2^n \|x_n\|$ converges, say, to γ . Let $\hat{x} := 1/\gamma \sum_{n=1}^\infty x_n/2^n \|x_n\|$. Since L_α is Banach and $\sum_{n=1}^\infty \|x_n/2^n \|x_n\| \| = 1$, then $1/\gamma \sum x_n/2^n \|x_n\|$ converges. Then (i) shows that \hat{x} is feasible and so satisfies (3.3) if we show $\hat{x} > 0$. Now (ii) shows that $\langle s^+, \hat{x} \rangle > 0$ if $s^+ \neq 0$, $s^+ \in S^+$. In particular, for any E in K of finite measure

$$\int_E \hat{x}(k) dk > 0$$

which easily shows that $\hat{x} > 0$ almost everywhere.

Case 2. $b = 0$. From (ii) $x_n > 0$. Thus with $\hat{x} := \sum_{n=1}^\infty x_n/2^n \|x_n\|$. Part (i) of Theorem 3.2 shows that $A\hat{x} = 0$ and with the same proof as above we also have $\hat{x} > 0$ almost everywhere. \square

Note that the separability assumption on L_α in Theorem 3.2 is needed only to prove the second implication.

An immediate consequence of Theorem 3.2 which associates the notion of extendibility with (BWCQ) is given in the next result.

COROLLARY 3.1. *Suppose K has a finite measure. Assume that every neighborhood of each point in K has a strictly positive dk -measure and $\{e^{jk \cdot \delta} : \delta \in \Delta\}$ is linearly independent on K . Then (BWCQ) holds if and only if $r \in \text{int } E$.*

Proof. The result follows from Theorem 3.2. Indeed, under our assumptions, Theorem 2.2 and Remark 2.2 are applicable and thus $r \in \text{int } E$ if and only if

$$(3.4) \quad \exists s \in L_\alpha : \int_K \psi_k s(k) dk = r, \quad s > 0 \quad (\text{a.e.})$$

which is exactly (3.3) with $b := r \in \mathbb{R}^{2m+1}$ and A as defined in (2.4). \square

4. Duality theory for problem (P). Throughout this section we assume that $\{e^{jk \cdot \delta} : \delta \in \Delta\}$ is a linearly independent set of functions on K , which has finite measure and that every neighborhood of each point in K has a strictly positive dk -measure. Recall also that $0 \in \Delta = -\Delta$ so that Δ has $2m+1$ elements, and $r(-\delta) = \bar{r}(\delta)$ so that the problem (P) is potentially feasible. For a real number a we denote $a_+ = \max(0, a)$. Our results are given in L_α with $1 < \alpha < \infty$.

THEOREM 4.1. *If $r \in \text{int } E$, then*

$$(4.1) \quad \min (P) = \max (D)$$

where the dual problem (D) of (P) is explicitly given by

$$(4.2) \quad \sup_{p \in \mathbb{R}^{2m+1}} \left\{ \sum_{\delta \in \Delta} p(\delta) \bar{r}(\delta) - c(\alpha, \beta) \int_K \left(\sum_{\delta \in \Delta} p(\delta) e^{-jk \cdot \delta} \right)_+^\beta dk \right\}.$$

Moreover if p^* solves (D) then the unique optimal solution of (P) is given by

$$(4.3) \quad s_*(k) = d(\alpha, \beta) \left(\sum_{\delta \in \Delta} p^*(\delta) e^{-jk \cdot \delta} \right)_+^{\beta-1}$$

where $c(\alpha, \beta) = \alpha^{1-\beta}/\beta > 0$ and $d(\alpha, \beta) = \alpha^{1-\beta}$.

Proof. Consider problem (P) defined in (2.5):

$$(4.4) \quad \inf I(s) \quad \text{s.t.} \quad \int_K \psi_k(\delta) s(k) dk = r(\delta), \\ \delta \in \Delta, s \in L_\alpha, \quad s(k) \geq 0 \quad \text{a.e.}$$

To construct its Lagrangian dual we associate a complex dual variable $p(\delta)$ (with $p(-\delta) = \bar{p}(\delta)$) to each complex constraint. In order to work in the real field, we split the real and imaginary parts of (4.4). Observing that for two complex numbers z_1, z_2 we have

$$(4.5) \quad \text{Re } z_1 \text{ Re } z_2 + \text{Im } z_1 \text{ Im } z_2 = \frac{1}{2}(z_1 \bar{z}_2 + \bar{z}_1 z_2).$$

Then, using relation (4.5), the Lagrangian function for (P) can be written

$$L(s, p) = I(s) + \frac{1}{2} \sum_{\delta \in \Delta} \{r(\delta) \bar{p}(\delta) + \bar{r}(\delta) p(\delta)\} \\ - \frac{1}{2} \int_K \left\{ \sum_{\delta \in \Delta} p(\delta) \bar{\psi}_k(\delta) + \sum_{\delta \in \Delta} \bar{p}(\delta) \psi_k(\delta) \right\} s(k) dk.$$

But since $r(\delta)$ and $p(\delta)$ are conjugate symmetric on Δ and $\Delta = -\Delta$, thus $L(s, p)$ is simply

$$L(s, p) = I(s) + \sum_{\delta \in \Delta} r(\delta) \bar{p}(\delta) - \int_K P(k) s(k) dk$$

where

$$P(k) := \sum_{\delta \in \Delta} p(\delta) \bar{\psi}_k(\delta) = \sum_{\delta \in \Delta} p(\delta) e^{-jk \cdot \delta}.$$

A Lagrangian dual for (P) is then

$$(4.6) \quad \sup_{p \in \mathbb{R}^{2m+1}} \left\{ \sum_{\delta \in \Delta} p(\delta) \bar{r}(\delta) + \inf_{s \in L_{\alpha}^+} \{I(s) - \langle s, P \rangle\} \right\}.$$

Since here $I(s) = \int_K |s(k)|^{\alpha} dk$ with $1 < \alpha < \infty$, $I(s)$ is a proper convex integral functional [15], and thus applying Theorem 3A of Rockafellar [16] we have

$$(4.7) \quad \inf_{s \in L_{\alpha}^+} \{I(s) - \langle s, P \rangle\} = \int_K \inf_{s \geq 0} \{s(k)^{\alpha} - P(k)s(k)\} dk.$$

Now it is easily verified that

$$(4.8) \quad \begin{aligned} \inf_{s \geq 0} \{s^{\alpha} - Ps\} &= -\frac{\alpha^{1-\beta}}{\beta} \{\max(0, P)\}^{\beta} \\ &= -c(\alpha, \beta)(P(k))_+^{\beta} = -c(\alpha, \beta) \left(\sum_{\delta \in \Delta} p(\delta) e^{-jk \cdot \delta} \right)_+^{\beta}. \end{aligned}$$

Hence, substituting (4.8) in (4.7), from (4.6) we obtain the dual problem (4.2).

From Theorem 2.2, $r \in \text{int } E$ implies that the feasible set

$$F := \{s \in L_{\alpha} : As = r, s \geq 0\}$$

is nonempty. Also note that F is a closed convex subset of the reflexive space L_{α} , so that the existence of an S^* with minimal L_{α} -norm is guaranteed, and since $I(s)$ is a strictly convex functional, S^* is unique. Moreover, by Corollary 3.1, $r \in \text{int } E$ if and only if (BWCQ) holds, thus Theorem 3.1 is applicable and (4.1) is proved.

The optimal solution s^* given in (4.3) is just the optimality condition for $s = s^*$ to solve the minimization problem (4.8). Now, the dual problem (D) in (4.2) is an unconstrained problem; its supremum is attained say at p^* only if p^* is a critical point of the supremand which is differentiable since $\beta > 1$, i.e., p^* is a solution of

$$(4.9) \quad \bar{r}(\delta) = \int_K \beta c(\alpha, \beta) \left(\sum_{\delta \in \Delta} p^*(\delta) e^{-jk \cdot \delta} \right)_+^{\beta-1} e^{-jk \cdot \delta} dk, \quad \delta \in \Delta.$$

Since $r(\delta)$ and $\psi_k(\delta) = e^{jk \cdot \delta}$ are conjugate symmetric on Δ , and $\Delta = -\Delta$, thus (4.9) is nothing else but

$$r(\delta) = \int_K s_*(k) \psi_k(\delta) dk, \quad \delta \in \Delta$$

with $s_*(k) = d(\alpha, \beta) \left(\sum_{\delta \in \Delta} p^*(\delta) e^{-jk \cdot \delta} \right)_+^{\beta-1}$, and the proof of the theorem is completed. \square

We notice that in the simplest case $\alpha = 2$, the dual problem (4.1) is

$$(4.10) \quad \sup_{p \in \mathbb{R}^{2m+1}} \left\{ \sum_{\delta \in \Delta} \bar{r}(\delta) p(\delta) - \frac{1}{4} \int_K \left(\sum_{\delta \in \Delta} p(\delta) e^{-jk \cdot \delta} \right)_+^2 dk \right\}$$

and when p^* solves (4.10) the optimal L_2 -estimate is given by

$$s_*(k) = \frac{1}{2} \left(\sum_{\delta \in \Delta} p^*(\delta) e^{-jk \cdot \delta} \right)_+.$$

Remark 4.1. The corresponding problem (P) in L_1 , is trivially solved by all feasible s . Indeed

$$r_0 = \int_K s(k) dk = \|s\|_1$$

since $s \geq 0$.

We have shown that we can directly derive the explicit solution of the L_α -estimation problem as presented in [5] by using convex programming duality. Moreover, Theorem 4.1 demonstrates that the multidimensional estimation problem subject to correlation matching constraints can be solved via an *unconstrained finite-dimensional* concave programming problem (D). The desired parameters p^* for the optimal spectrum estimate s^* are precisely the optimal dual variables of problem (D), which can be solved efficiently via nonlinear programming techniques [17].

Consider the case when the spectral support may be approximated by a finite number of points:

$$K = \{k_i \in \mathbb{R}^n : i = 1, \dots, N\}.$$

A measure ν on the discrete support K is completely characterized by its value $\nu(k_i)$ at each point. Thus

$$\int_K \left(\sum_{\delta \in \Delta} p(\delta) e^{-jk \cdot \delta} \right)_+^\beta dk = \sum_{i=1}^N \left(\sum_{\delta \in \Delta} p(\delta) e^{-jk_i \cdot \delta} \right)_+^\beta \nu(k_i)$$

and the dual problem (D) gives rise to the *nonsmooth* optimization problem

$$(\hat{D}) \quad \sup_{p \in \mathbb{R}^{2m+1}} \left\{ \sum_{\delta \in \Delta} r(\delta) p(\delta) + \sum_{i=1}^N \gamma_i \left(\sum_{\delta \in \Delta} p(\delta) e^{-jk_i \cdot \delta} \right) \right\}$$

where

$$\gamma_i(t_i) := -c(\alpha, \beta)(t_i)_+^\beta \nu(k_i).$$

Problems of the above type are typical in the nonsmooth optimization literature and may be solved using existing numerical schemes [9], [12]. It is also worth mentioning that the objective function of problem (\hat{D}) has a very special structure which can be exploited using techniques for convex composite optimization problems (see, e.g., [4], [18]).

5. Duality for the time series case. In this section we specialize our duality results derived in § 4 for the important time series case. While for the multidimensional problem we show that (BWCQ) holds if and only if $r \in \text{int } E$, a condition which is not always easy to check; in the time series case we will show that to satisfy (BWCQ) it is necessary and sufficient to test the positive definiteness of a specific Toeplitz matrix (see Theorem-5.1).

Let $K = [-\pi, \pi]$. Given a finite sequence of complex numbers $\{r_n\}_{|n| \leq m}$ with $r_{-n} = \bar{r}_n$ the problem is now to find a nonnegative finite measure μ on $[-\pi, \pi]$ satisfying

$$(5.1) \quad r_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{jkn} d\mu(k), \quad n = 0, \pm 1, \dots, \pm m.$$

This is the so-called *trigonometric moment problem* (see, e.g., Grenander and Szegő [6]). A necessary condition in order to find a nonnegative finite measure μ satisfying (5.1) is that the *Toeplitz matrix*

$$M = [r_{l-h}]_{l,h=0}^m$$

must be positive semidefinite. This condition is also sufficient (see [6, p. 19]).

We now ask the same question for problems involving densities, i.e., finding an integrable function $s \geq 0$ that satisfies

$$(5.2) \quad r_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{jkn} s(k) dk, \quad n = 0, \pm 1, \dots, \pm m.$$

A necessary condition is that M must be positive definite. In fact this condition is also sufficient as has been shown in [6]. (See also the appendix; this also provides a nice illustration of duality theory.)

For the convenience of the reader, we state and prove this result in the following theorem.

THEOREM 5.1. *The following four statements are equivalent:*

- (a) M is positive definite;
- (b) $r \in \text{int } E$;
- (c) (BWCQ) holds;
- (d) Equation (5.2) has a nonzero measurable solution.

Proof. That (b) and (c) are equivalent is direct from Corollary 3.1. It remains to show that (a) is equivalent to (b) and (d). Suppose that (d) holds. Let d_1, \dots, d_m be given complex numbers, not all zero, and consider, for a feasible s , the following Toeplitz form:

$$\begin{aligned} T_m &:= \sum_{l,h=0}^m \bar{d}_l d_h r_{l-h} = \frac{1}{2\pi} \sum_{l,h=0}^m \int_{-\pi}^{\pi} \bar{d}_l d_h e^{jkl} e^{-jkh} s(k) dk \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \lambda(e^{jk}) \bar{\lambda}(e^{+jk}) s(k) dk \end{aligned}$$

where $\lambda[z] := \sum_{l=0}^m d_l z^l$, $z = e^{jk}$.

Any such nonzero polynomial gives rise to a *nonnegative* trigonometric polynomial (see [6, p. 20]):

$$(5.3) \quad p(k) = \lambda(e^{jk}) \bar{\lambda}(e^{jk}) = \lambda(z) \overline{\lambda(z)}$$

which is nonzero (except finitely) for $|z| = 1$. By (d), s is positive on a set of positive measure. Thus

$$T_m = \frac{1}{2\pi} \int_{-\pi}^{\pi} s(k) p(k) dk > 0$$

and M , being a Hermitian matrix, is positive definite. This gives (a).

Now suppose (a) holds. Let \mathcal{M} be the convex cone of positive semidefinite Toeplitz matrices, and let L be the linear operator which sends $M \in \mathcal{M}$ to the transverse diagonal $r := (\text{Re } r, \text{Im } r)$. Then, using Theorem 6.6 in [14]; $M \in \text{relint } \mathcal{M}$ implies $r = L(M) \in \text{relint } L(\mathcal{M})$ and thus if we show that $\text{relint } L(\mathcal{M}) = \text{relint } E$ then we have shown that (a) implies (b). In fact $L(\mathcal{M}) = E$ as follows from the classical moment spaces theory. That is $r \in E$ is extendible if and only if M is in \mathcal{M} (see Karlin and Studden [7, Thm. 4.1, p. 184]). Finally, (b) implies (d), which follows from Theorem 2.2. \square

Theorem 5.1 is exactly what we need in order to derive our duality results for the time series L_α -norm estimation problem:

$$(P_1) \quad \inf_{S \in L_\alpha^+[-\pi, \pi]} \left\{ \int_{-\pi}^{\pi} s^\alpha(k) dk : \frac{1}{2\pi} \int_{-\pi}^{\pi} s(k) e^{jkn} dk = r_n, n = 0, \pm 1, \dots, \pm m \right\}.$$

The one-dimensional analogue of Theorem 4.1 is given in the following result.

THEOREM 5.2. *Suppose M is positive definite. Then*

$$(5.4) \quad \min (P_1) = \max (D_1)$$

where the dual problem (D_1) of (P_1) is explicitly given by

$$(5.5) \quad \sup_{\lambda \in \mathbb{R}^{2m+1}} \left\{ \sum_{n=-m}^m \bar{r}_n \lambda_n - \frac{\alpha^{1-\beta}}{\beta} \int_{-\pi}^{\pi} \left(\sum_{n=-m}^m \lambda_n e^{-jkn} \right)_+^{\beta} dk \right\}.$$

Moreover if λ^* solves (D_1) then the unique optimal solution of (P_1) is given by

$$(5.6) \quad s_*(k) = \alpha^{1-\beta} \left(\sum_{n=-m}^m \lambda_n^* e^{-jkn} \right)_+^{\beta-1}.$$

Proof. Since M is assumed positive definite it follows from Theorem 5.1 that (BWCQ) holds. Thus theorem 3.1 is applicable. The remaining statements in the theorem follow immediately as a special case of Theorem 4.1. \square

6. Conclusions. A complete duality theory for the multidimensional spectral estimation problem has been developed by using the (BWCQ). The simple unconstrained nature of the dual problem seems appropriate for computational purposes and may lead to the construction of reliable algorithms for computing the L_α optimal spectral estimate. Furthermore, the results in the paper demonstrate that the new constraint qualification of Borwein and Wolkowicz [1] is very natural and is equivalent to the extendibility condition.

We also mention that a duality theory without constraint qualification was also developed in [1]. Following [1], particularly example 5.1 therein, it can be shown that our duality framework can also be extended when the constraint qualification fails. Finally we remark that our duality framework can be used to consider problem (P) with other convex objective functionals. In particular the choice of the “entropy type” functional $I(s) := \int_K s(k) \log s(k) dk$ lead to the explicit solution $s^*(k) = e^{P(k)}$. Note that while the computational tractability of this solution seems less attractive than the L_α norm solution, it provides us automatically with a unique strictly positive optimal solution.

Appendix.

THEOREM. *If M is positive definite, then there exists a unique strictly positive trigonometric polynomial*

$$P(k) = \sum_{n=-m}^m a_n e^{-jkn}, \quad \bar{a}_n = a_{-n} \quad \text{for } |n| \leq m$$

such that $s_0(k) = 1/P(k)$ solved the maximum entropy problem

$$\max \left\{ H(s) = \int_{-\pi}^{\pi} \ln s(k) dk \right\}$$

subject to the constraints (5.2).

Proof. Consider the primal entropy maximum problem

$$p = \sup \left\{ H(s) : \int_{-\pi}^{\pi} e^{jkn} s(k) dk = 2\pi r, n = 0, \pm 1, \dots, \pm m, s \geq 0 \right\}.$$

First we note that, since $H(s) := \int_{-\pi}^{\pi} \ln s(k) dk$ is a strictly concave integral functional, then if an optimal solution exists it is unique. The Lagrangian dual of p (following the same arguments as in Theorem 4.1) is

$$d = \inf_{y \in \mathbb{R}^{2m+1}} \left\{ 2\pi \sum_{n=-m}^m \bar{r}_n y_n + \sup \left\{ H(s) - \int_{-\pi}^{\pi} \left(\sum_{n=-m}^m y_n e^{-jkn} \right) s(k) dk \right\} \right\}.$$

It is easy to verify, using again Theorem 3A of [16], that the inner supremum (where necessarily $s > 0$ a.e.) is

$$\int_{-\pi}^{\pi} \operatorname{Ln} \left(\frac{1}{\sum_{n=-m}^m y_n e^{-jkn}} \right) dk - 2\pi$$

with the optimal $s^*(k) = (1/\sum_{n=-m}^m y_n e^{-jkn})$. Hence the dual is

$$d = \inf_{y \in \mathbb{R}^{2m+1}} \left\{ 2\pi \left(\sum_{n=-m}^m \bar{r}_n y_n - 1 \right) + \int_{-\pi}^{\pi} \operatorname{Ln} \left(\frac{1}{\sum_{n=-m}^m y_n e^{-jkn}} \right) dk \right\}.$$

By Theorems 5.1 and 2.2, a suitable form of Slater's condition is met (in L_1); thus $p = d$ and d is attained. By differentiating in the last equation, we can see that this happens if and only if y^* solves

$$(A1) \quad \bar{r}_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{e^{-jkn}}{\sum_{n=-m}^m y_n e^{-jkn}} dk = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-jkn} s^*(k) dk, \quad n = 0, \pm 1, \dots, \pm m.$$

However, $\bar{r}_n = r_{-n}$; thus (7.1) is equivalent to

$$(A2) \quad r_n = \frac{1}{2\pi} \int e^{jkn} s^*(k) dk, \quad n = 0, \pm 1, \dots, m.$$

It remains to show that M positive definite guarantees the existence of a *finite strictly positive* $s^*(k)$. To this end observe that when M is positive definite (A2) implies that

$$(A3) \quad \sum_{l,h=0}^m \bar{d}_l d_h r_{l-h} = \int_{-\pi}^{\pi} \frac{p(k)}{q(k)} dk > 0$$

with

$$q(k) := \sum_{n=-m}^m y_n e^{-jkn} = \frac{1}{s^*(k)}$$

for all nonzero positive trigonometric polynomials $p(k)$ defined in (5.3). Suppose $q(k) = 0$ for some k . Then using the representation (5.3) which is valid for any non-negative trigonometric polynomial, we can write

$$q(k) = \prod_{\mu=1}^m |e^{-jk} - \alpha_{\mu}|^2$$

where $\alpha_1, \dots, \alpha_m$ denote the zeros of $q(k)$. Also note that there exists at least one $h \in [1, m]$ such that $|\alpha_h| = 1$. Now since (A3) holds for any nonzero positive trigonometric polynomial $p(k)$, by choosing $p(k) = \prod_{\nu \neq h} |e^{-jk} - \alpha_{\nu}|^2$ we have

$$\infty > \int_{-\pi}^{\pi} \frac{p(k)}{q(k)} dk = \int_{-\pi}^{\pi} \frac{dk}{|e^{-jk} - \alpha_h|^2} = \int_{-\pi}^{\pi} \frac{dk}{|e^{-jk} - 1|^2} = \frac{1}{2} \int_{-\pi}^{\pi} \frac{dk}{1 - \cos k} = \infty.$$

Hence $q(k) > 0$ and so $0 < s^*(k) < \infty$. \square

We note that it is formally "obvious" that (A1) holds by differentiation. It is, however, technically nontrivial to justify this step. Indeed a fully satisfactory proof is somewhat extended and requires reorganizing the proof we have given.

It is also important to mention that we can solve (7.2) explicitly to find the parameters $\{y_n\}$ (see [3] for details). The maximum entropy spectrum is then given by

$$s^*(k) = \frac{M_{11}^{-1}}{|(M^{-1}\delta)^T \phi(k)|^2}$$

where $\phi(k) = (1, \dots, e^{jmk})$, $\delta = (1, 0, \dots, 0)$ and M^{-1} is the inverse of M (which exists since M is positive definite).

REFERENCES

- [1] J. BORWEIN AND H. WOLKOWICZ, *A simple constraint qualification in infinite dimensional programming*, Math. Programming, 35 (1986), pp. 83–96.
- [2] J. P. BURG, *Maximum entropy spectral analysis*, Ph.D. thesis, Stanford University, Stanford, CA, 1975.
- [3] J. A. EDWARD AND M. M. FITELSON, *Notes on maximum entropy processing*, IEEE Trans. Inform. Theory, IT-19 (1973), pp. 232–234.
- [4] R. FLETCHER, *A model algorithm for composite nondifferential optimization problems*, Math. Programming Stud., 17 (1982), pp. 67–76.
- [5] B. K. GOODRICH AND A. STEINHARDT, *L_2 spectral estimation*, SIAM J. Appl. Math., 46 (1986), pp. 417–426.
- [6] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications*, Univ. of California Press, Berkeley, Los Angeles, 1958.
- [7] S. KARLIN AND W. STUDDEN, *Tchebycheff Systems; with Applications in Analysis and Statistics*, Wiley-Interscience, New York, 1986.
- [8] S. M. KAY AND S. L. MARPLE, *Spectrum analysis—a modern perspective*, Proc. IEEE, 69 (1981), pp. 1380–1419.
- [9] K. C. KIWIEL, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Math. 1133, Springer-Verlag, Berlin, New York, 1985.
- [10] S. W. LANG AND J. H. MCCLELLAN, *Multidimensional MEM spectral estimation*, IEEE Trans. Acoust. Speech Signal Process., ASSP-30 (1982), pp. 880–887.
- [11] ———, *Spectral estimation for sensor arrays*, IEEE Trans. Acoust. Speech Signal Process., ASSP-31 (1983), pp. 349–358.
- [12] C. LEMARECHAL, *Nondifferentiable Optimization*, in Nonlinear Optimization Theory and Algorithms, L. C. W. Dixon, J. Spedicato, and G. F. Szegö, eds., Birkhäuser, Boston, 1980.
- [13] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
- [14] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [15] ———, *Integrals which are convex functions*, II, Pacific J. Math., 39 (1971), pp. 439–469.
- [16] ———, *Integral Functionals, Normal Integrands and Measurable Selections*, Lecture Notes in Math., 543, Springer-Verlag, Berlin, New York, 1976, pp. 157–207.
- [17] K. SCHITTOWSKI, *Nonlinear Programming Codes*, Lecture Notes Econom. Math. Systems, 183, Springer-Verlag, Berlin, 1980.
- [18] R. S. WOMERSLEY, *Local properties of algorithms for minimizing nonsmooth composite functions*, Math. Programming, 32 (1985), pp. 69–89.

ERRATUM: Differentiability of Relations and Differential Stability of Perturbed Optimization Problems*

JEAN-PAUL PENOT¹

S. Dolecki has kindly pointed out to the author that one of the assumptions of Theorem 5.11 of [1] is wrongly used: the proof of this result is correct when X is finite-dimensional but the assumption of B -tangential compactness of A is slightly more general (for instance, A may be a finite-dimensional submanifold of $W \times X$). Let us give, in the following theorem, a slightly reinforced hypothesis for replacing 5.11(a).

THEOREM. *With the notation of Theorem 5.11 of [1], under the following three conditions:*

- (a) *For each $a \in S(0)$ A is B -tangentially compact at $(0, a)$ in the directions zero and w ;*
- (b) *The perturbation (f, A) is well set in the direction w at zero;*
- (c) *For each $a \in S(0)$ and each $v \neq 0$ in $\bar{D}A(0, a)(0)$, $\underline{d}f(0, a; 0, v) > 0$,*
the following inequality holds:

$$\underline{d}m(0, w) \geq \inf \{ \underline{d}f(0, a; w, x) : a \in S(0), x \in \bar{D}A(0, a)(w) \}.$$

Proof. Let (t_n, w_n) be a sequence of $(0, +\infty) \times W$ with limit $(0, w)$ such that $\underline{d}m(0, w) = \lim t_n^{-1}(m(t_n w_n) - m(0))$. We may assume $m(t_n w_n) < +\infty$ for each $n \in \mathbb{N}$; hence $t_n w_n \in \text{dom } A$. Taking $\varepsilon_n = t_n^2$, we can find an infinite subset K of \mathbb{N} and a sequence $(x_k)_{k \in K}$ with limit $a \in S(0)$ such that $x_k \in S_{\varepsilon_k}(t_k w_k)$ for each $k \in K$.

We may assume that the sequence $(s_k)_{k \in K} := (t_k^{-1} r_k)_{k \in K}$ with $r_k := |x_k - a|$ has a limit s in $[0, +\infty]$. When $s \neq +\infty$ the sequence $(t_k^{-1}(x_k - a))_{k \in K}$ is bounded so that, by assumption (a), it may be assumed to have a limit v , with $|v| = s$. We have $(w, v) \in T_{(0,a)}A$ or $v \in \bar{D}A(0, a)(w)$ and

$$\underline{d}m(0, w) \geq \liminf t_n^{-1}[f(t_n w_n, x_n) - f(0, a)] \geq \underline{d}f(v, a; w, v)$$

and the inequality is satisfied in this case.

Let us now suppose that $s = +\infty$. Then we can write $x_k \in A(r_k s_k^{-1} w_k)$ and $(s_k^{-1} w_k)_{k \in K}$ has limit zero in W . As $r_k^{-1}|x_k - a| = 1$ for each $k \in K$, using assumption (a) we may assume that $(u_k)_{k \in K} := (r_k^{-1}(x_k - a))_{k \in K}$ has a limit u in X , with $|u| = 1$. Then we have $u \in \bar{D}A(0, a)(0)$; hence $\underline{d}f(0, a; 0, u) > 0$. Let us first suppose $m(t_k w_k) > -\infty$ for $k \in K$ large enough. Then we have

$$\begin{aligned} \underline{d}m(0, w) &\geq \liminf_{k \in K} t_k^{-1}(f(t_k w_k, x_k) - f(0, a) - t_k^2) \\ &\geq \liminf_{k \in K} s_k r_k^{-1}(f(r_k s_k^{-1} w_k, a + r_k u_k) - f(0, a)) = +\infty \end{aligned}$$

and again the inequality is satisfied in this case. Finally, when $m(t_k w_k) = -\infty$ for k in an infinite subset L of K , we have $f(t_k w_k, x_k) - f(0, a) \leq 0$ for $k \in L$ large enough since

* Received and accepted by the editors May 9, 1986.

¹ Faculté des Sciences, Avenue de l'Université, 64000 Pau, France.

$f(t_k w_k, x_k) < -t_k^{-2}$ for $k \in L$; hence

$$0 < \underline{d}f(0, a; u) \leq \liminf_{k \in L} s_k r_k^{-1} (f(r_k s_k^{-1} w_k, a + r_k u_k)) - f(0, a) \leq 0,$$

a contradiction. Therefore this case does not occur and the inequality holds true. \square

REFERENCE

- [1] J.-P. PENOT, *Differentiability of relations and differential stability of perturbed optimization problems*, SIAM J. Control Optim., 22 (1984), pp. 529–551.

A GENERALIZATION OF A THEOREM OF ARROW, BARANKIN, AND BLACKWELL*

JOHANNES JAHN†

Abstract. This paper presents a generalization of a known density theorem of Arrow, Barankin, and Blackwell ("Admissible points of convex sets," in *Contributions to the Theory of Games*, H. W. Kuhn and A. W. Tucker, eds., Princeton University Press, Princeton, NJ, 1953). This result is shown to hold even in a real normed space partially ordered by a Bishop-Phelps cone.

Key words. vector optimization, support functionals

AMS(MOS) subject classifications. 90C31, 46A55

1. Introduction. In 1953, Arrow, Barankin, and Blackwell [1] proved a famous theorem concerning the density of the set of minimal solutions of strictly positive support functionals in the set of minimal elements of a compact convex subset of \mathbb{R}^n . This result has some important and interesting consequences in multi-objective optimization. But this theorem is restricted to the space \mathbb{R}^n partially ordered with respect to the componentwise ordering. An alternative proof of this result was also given by Peleg [11].

In the past decades the theorem of Arrow, Barankin, and Blackwell was generalized by several authors. Hartley [6] and Bitran and Magnanti [3] extended this result to other ordering cones in \mathbb{R}^n space. Radner [13] and Majumdar [10] generalized the Arrow-Barankin-Blackwell result to the case of l_∞ . Salz [14] extended this theorem to real normed spaces partially ordered by a convex cone with a base norm, and Borwein [4] presented a special version in a normed space with a weakly compact based cone.

The investigations of this paper are based on the idea of the proof given by Salz [14] and some results of Bishop and Phelps [2] (see also Phelps [12]). It is shown that the Arrow-Barankin-Blackwell Theorem remains true in a real normed space partially ordered by a Bishop-Phelps cone.

2. Minimal elements and strictly positive support functionals. Throughout this paper let X be a real normed space partially ordered by a Bishop-Phelps cone C . This cone can be written as

$$C = \{x \in X \mid \alpha \|x\| \leq l(x)\}$$

for some number $\alpha \in (0, 1]$ and some continuous linear functional $l \in X^*$ with $\|l\| = 1$. Notice in the case $\alpha = 1$ that C may be trivial in general although it may be nontrivial for instance in \mathbb{R}^n equipped with the l_1 norm. Such a cone (with $\alpha \neq 1$) was used by Bishop and Phelps [2] (see also [12]) for the investigation of maximal elements of subsets of a normed space whose partial ordering is induced by this cone C . Obviously, the Bishop-Phelps cone C is convex and it is also pointed, i.e., $C \cap (-C) = \{0\}$. Such a Bishop-Phelps cone is also a so-called nuclear cone introduced by Isac [7, p. 310] in a locally convex Hausdorff space.

* Received by the editors February 2, 1987; accepted for publication (in revised form) October 27, 1987.

† Institut für Angewandte Mathematik, Universität Erlangen-Nürnberg, Martensstr. 3, 8520 Erlangen, Federal Republic of Germany.

In this setting we consider a nonempty subset S of X . Recall that an element $\bar{x} \in S$ is called a *minimal* element of S , if $(\{\bar{x}\} - C) \cap S = \{\bar{x}\}$ (where the expression in parentheses denotes the algebraic difference of sets). It is a well-known fact in the case of a convex set S with a nonempty interior (see, e.g., [8, Thm. 5.4]) that for each minimal element \bar{x} of S there exists a continuous linear functional $l \in C^* \setminus \{0\}$ such that

$$l(\bar{x}) \leq l(x) \quad \text{for all } x \in S$$

where the *dual cone* C^* is defined by

$$C^* := \{x^* \in X^* \mid x^*(x) \geq 0 \text{ for all } x \in C\}.$$

So, in this case minimal elements are supported by a positive continuous linear functional. The converse statement is not true in general. But if we consider the *quasi-interior*

$$C^\# := \{x^* \in X^* \mid x^*(x) > 0 \text{ for all } x \in C \setminus \{0\}\}$$

of the dual cone C^* , it is a known result (see, e.g., [8, Thm. 5.18(b)]) that an element $\bar{x} \in S$ is a minimal element of the set S , if there exists some $l \in C^\#$ with the property

$$l(\bar{x}) \leq l(x) \quad \text{for all } x \in S.$$

So, strictly positive support functionals lead to minimal elements by a minimization procedure.

Before going further, notice that the quasi-interior $C^\#$ of the Bishop–Phelps cone C is always nonempty because

$$l(x) \geq \alpha \|x\| > 0 \quad \text{for all } x \in C \setminus \{0\}$$

and therefore the continuous linear functional l belongs to $C^\#$. Not every pointed convex cone has a nonempty quasi-interior of its dual cone. But by a theorem of Krein and Rutman [9] (see also [5, Prop. 2.8] and [8, Thm. 3.38]) in a real separable normed space X with a closed pointed convex cone C the quasi-interior $C^\#$ of the dual cone is nonempty. So, in this setting strictly positive continuous linear functionals exist.

Since the dual cone C^* and its quasi-interior $C^\#$ play an important role in this paper, let us first remark that the closure of the quasi-interior equals the dual cone.

LEMMA 2.1. *Let X be a real normed space partially ordered by a pointed convex cone C with a nonempty quasi-interior $C^\#$ of the dual cone C^* . Then*

$$\text{cl}(C^\#) = C^*$$

(where cl denotes the closure in the norm topology of X^*).

Proof. It is obvious that $C^\# \subset C^*$, and since C^* is closed we get

$$\text{cl}(C^\#) \subset \text{cl}(C^*) = C^*.$$

In order to show the converse inclusion we take any $l \in C^*$. Since $C^\#$ is nonempty we choose some $l^* \in C^\#$. Then we have

$$l_\lambda := l + \lambda l^* \in C^\# \quad \text{for all } \lambda > 0$$

and we get

$$\lim_{\lambda \rightarrow 0_+} \|l - l_\lambda\| = \lim_{\lambda \rightarrow 0_+} \lambda \|l^*\| = 0.$$

So $l \in \text{cl}(C^\#)$. \square

Since we are concerned with a special ordering cone, namely a Bishop–Phelps cone, next we investigate the relations between this and other cones. For that purpose

recall that a nonempty convex subset B of the pointed convex cone C is called a *base* for C , if every $x \in C \setminus \{0\}$ has a unique representation of the form

$$x = \lambda b \quad \text{for some } \lambda > 0 \text{ and some } b \in B.$$

It is a well-known result in this setting (see, e.g., [8, Lemma 3.3]) that a subset B of the ordering cone is a base for C if and only if there exists a strictly positive linear functional l with

$$B = \{x \in C \mid l(x) = 1\}.$$

Notice that l does not need to be continuous.

The following lemma says that every pointed convex cone with a bounded base is contained in a Bishop–Phelps cone (for a related result on nuclear cones compare also [7, Prop. 5]).

LEMMA 2.2. *Let X be a real normed space partially ordered by a pointed convex cone C with a closed bounded base. Then there exists some continuous linear functional $l \in C^*$ with $\|l\| = 1$ and some real number $\alpha \in (0, 1]$ with the property*

$$(1) \quad C \subset \{x \in X \mid \alpha \|x\| \leq l(x)\}.$$

Proof. The proof of this lemma is contained in the proof of Lemma 1.2 in [12]. Let B denote a closed bounded base for C . Since B is closed and convex and $0 \notin B$, by a known separation argument (see, e.g., [8, Cor. 3.19]) there exist some continuous linear functional $l \in C^*$ with $\|l\| = 1$ and some real number β with

$$0 < \beta \leq \inf_{b \in B} l(b).$$

Since the base B is bounded, there exists some positive real number

$$\gamma := \max \left\{ \beta, \sup_{b \in B} \|b\| \right\}.$$

If we set $\alpha := \beta/\gamma$, we have $\alpha \in (0, 1]$ and we obtain for every $x \in B$

$$\alpha \|x\| = \frac{\beta}{\gamma} \|x\| \leq \beta \leq l(x).$$

This implies

$$B \subset \{x \in X \mid \alpha \|x\| \leq l(x)\}.$$

If we notice that C is the cone generated by B , then the inclusion (1) follows immediately. \square

In general the inclusion (1) is strict but there are also cases where equality holds. For instance, take $X = \mathbb{R}^n$, $C = \mathbb{R}_+^n$ and the l_1 norm in \mathbb{R}^n .

3. The main result. In this section we present a generalization of the Arrow–Barankin–Blackwell Theorem. The proof makes use of the idea of proof given by Salz [14].

THEOREM 3.1. *Let X be a real normed space partially ordered by a Bishop–Phelps cone C . Moreover, let S be a nonempty convex subset of X , and let $\bar{x} \in S$ be a minimal element of S for which the set*

$$\hat{S} := \{x \in S \mid \|x - \bar{x}\| \leq 1\}$$

is weakly compact. Then for every $\varepsilon \in (0, 1)$ there exists some $l_\varepsilon \in C^*$ and some $x_\varepsilon \in S$ with

$$l_\varepsilon(x_\varepsilon) \leq l_\varepsilon(x) \quad \text{for all } x \in S$$

such that

$$\|x_\varepsilon - \bar{x}\| \leq \varepsilon.$$

Proof. Notice that the ordering cone C can be written as

$$C = \{x \in X \mid \alpha \|x\| \leq l(x)\}$$

for some $l \in C^*$ with $\|l\| = 1$ and some $\alpha \in (0, 1]$. Then we define for an arbitrary $\beta \in (0, \alpha)$ the functional $f_\beta : X \rightarrow \mathbb{R}$ by

$$f_\beta(x) = \beta \|x - \bar{x}\| + l(x) \quad \text{for all } x \in X.$$

For each $\beta \in (0, \alpha)$ the functional f_β is weakly lower semicontinuous and therefore it attains its minimum on \hat{S} , i.e., there exists some $x_\beta \in \hat{S}$ with

$$f_\beta(x_\beta) \leq f_\beta(x) \quad \text{for all } x \in \hat{S}.$$

Especially, we get

$$\begin{aligned} f_\beta(x_\beta) &\leq f_\beta(\bar{x}), \\ \beta \|x_\beta - \bar{x}\| + l(x_\beta) &\leq l(\bar{x}), \end{aligned}$$

and

$$(2) \quad \|x_\beta - \bar{x}\| \leq \frac{1}{\beta} l(\bar{x} - x_\beta),$$

respectively. As a necessary optimality condition there exists a subgradient $l_\beta \in X^* \setminus \{0\}$ with

$$l_\beta(x - x_\beta) \geq 0 \quad \text{for all } x \in \hat{S}$$

and

$$\begin{aligned} l_\beta(x - x_\beta) &\leq f_\beta(x) - f_\beta(x_\beta) \\ &= \beta \|x - \bar{x}\| + l(x) - \beta \|x_\beta - \bar{x}\| - l(x_\beta) \\ &\leq \beta \|x - x_\beta\| + l(x - x_\beta) \quad \text{for all } x \in X. \end{aligned}$$

For every $x \in C \setminus \{0\}$ we get

$$\begin{aligned} l_\beta(-x) &\leq \beta \|-x\| + l(-x) \\ &< \alpha \|x\| - l(x) \\ &\leq l(x) - l(x) \\ &= 0. \end{aligned}$$

Consequently we have $l_\beta \in C^*$ and

$$(3) \quad l_\beta(x_\beta) \leq l_\beta(x) \quad \text{for all } x \in \hat{S}.$$

Because \bar{x} is assumed to be a minimal element of S we notice that

$$\bar{x} - x \notin C \quad \text{for all } x \in S \setminus \{\bar{x}\}.$$

Since C is a Bishop-Phelps cone, we conclude

$$\alpha \|\bar{x} - x\| > l(\bar{x} - x) \quad \text{for all } x \in S \setminus \{\bar{x}\}$$

and especially

$$(4) \quad \alpha \|\bar{x} - x\| - l(\bar{x} - x) > 0 \quad \text{for all } x \in \hat{S} \setminus \{\bar{x}\}.$$

Finally, take an arbitrary $\varepsilon \in (0, 1)$. Then we consider the case that the set

$$\tilde{S} := \left\{ x \in \hat{S} \mid l(\bar{x} - x) \geq \frac{\varepsilon}{2} \right\}$$

is nonempty. Since \tilde{S} is weakly compact and the functional on the left-hand side of the inequality (4) is weakly lower semicontinuous, there exists some $\delta > 0$ with

$$\alpha \|\bar{x} - x\| - l(\bar{x} - x) > \delta \quad \text{for all } x \in \tilde{S}.$$

Then we obtain for all $\beta \geq \alpha - \delta$

$$\begin{aligned} \beta \|\bar{x} - x\| &\geq (\alpha - \delta) \|\bar{x} - x\| \\ &> l(\bar{x} - x) + \delta - \delta \|\bar{x} - x\| \\ &\geq l(\bar{x} - x) \quad \text{for all } x \in \tilde{S}. \end{aligned}$$

But this implies, with the inequality (2), that $x_\beta \notin \tilde{S}$, i.e., $l(\bar{x} - x_\beta) < \varepsilon/2$. In the case that the set \tilde{S} is empty it is evident that the previous inequality is satisfied. Hence it follows from (2) for $\beta \geq \alpha - \delta$ and $\beta \geq \frac{1}{2}$

$$\|x_\beta - \bar{x}\| \leq \frac{1}{\beta} \frac{\varepsilon}{2} \leq \varepsilon.$$

Since $\varepsilon < 1$ and S is convex, we conclude with (3) that x_β is also a global minimizer, i.e.,

$$l_\beta(x_\beta) \leq l_\beta(x) \quad \text{for all } x \in S.$$

So, the proof is complete. \square

In vector optimization support points which are obtained by strictly positive support functionals are sometimes called properly minimal. But there are also other definitions in use which coincide with the following concept in special cases (see, e.g., [8]).

DEFINITION 3.2. Let X be a real normed space partially ordered by a pointed convex cone C with a nonempty quasi-interior C^* of the dual cone, and let S be a nonempty subset of X . An element $\bar{x} \in S$ is called a *properly minimal* element of the set S , if there exists some $l \in C^*$ with

$$l(\bar{x}) \leq l(x) \quad \text{for all } x \in S.$$

In the previous section we mentioned that properly minimal elements are minimal but the converse statement is not true in general. But with Theorem 3.1 and Lemma 2.2 we immediately get a density result.

COROLLARY 3.3. Let X be a real normed space partially ordered by a Bishop-Phelps cone, and let S be a weakly compact convex subset of X . Then the set of properly minimal elements of S is dense in the set of minimal elements.

If the space X is even reflexive, then it is sufficient to require that the set S is closed and convex.

COROLLARY 3.4. Let X be a real reflexive Banach space partially ordered by a Bishop-Phelps cone, and let S be a closed convex subset of X . Then the set of properly minimal elements of S is dense in the set of minimal elements.

With Lemma 2.2 we know that, in a real normed space X partially ordered by a pointed convex cone C with a closed bounded base, this cone C is contained in an appropriate Bishop-Phelps cone \tilde{C} . These two cones have a nonempty quasi-interior with the property that $\tilde{C}^* \subset C^*$. Therefore, proper minimality with respect to \tilde{C} implies

proper minimality with respect to the ordering cone C . So Theorem 3.1 can also be applied in the case that X is partially ordered by C , if a minimal element \bar{x} of the considered set S is also minimal with respect to the Bishop-Phelps cone \tilde{C} . This result is summarized in the following corollary.

COROLLARY 3.5. *Let X be a real normed space partially ordered by a pointed convex cone C with a closed bounded base, and let S be a weakly compact convex subset of X . If $\bar{x} \in S$ is a minimal element of S which is also minimal with respect to a Bishop-Phelps cone containing C , then for each $\varepsilon \in (0, 1)$ there exists a properly minimal element $x_\varepsilon \in S$ with $\|x_\varepsilon - \bar{x}\| \leq \varepsilon$.*

4. Some examples. Salz [14] considered a pointed convex cone C with a base norm. In this case the base B for C is given as

$$B := \{x \in C \mid \|x\| = 1\},$$

i.e., the strictly positive linear functional l which describes the base equals the norm on C . It is evident that such a base is closed and bounded, and therefore Corollary 3.3 may be applied. It is pointed out by Salz [14] that the natural ordering cones in the spaces l_1 , $L_1(\Omega, \Sigma, \mu)$ and $ca(\Omega, \Sigma)$ satisfy the assumptions on C . Obviously, this result may also be applied to the naturally ordered space \mathbb{R}^n , if we take the l_1 norm.

Next we turn our attention to the real linear space $L_2[a, b]$ with $-\infty < a < b < \infty$. It is well known that the natural ordering cone

$$C := \{x \in L_2[a, b] \mid x(t) \geq 0 \text{ a.e. on } [a, b]\}$$

has a base, e.g.,

$$B \in \left\{ x \in C \mid \int_a^b x(t) dt = 1 \right\}$$

which is unbounded. But if we restrict ourselves to the smaller ordering cone generated by the set

$$B_\alpha := \left\{ x \in B \mid \int_a^b x(t)^2 dt \leq \alpha \right\}$$

for an arbitrary large $\alpha > 0$, then Corollary 3.5 applies to this special ordering.

Finally, we consider the space X of all real symmetric (n, n) matrices (with a fixed $n \in \mathbb{N}$) partially ordered in the natural sense, i.e., the ordering cone C is given as

$$(5) \quad C := \{A \in X \mid x^T A x \geq 0 \text{ for all } x \in \mathbb{R}^n\}.$$

The space X is a real Hilbert space with the scalar product $\langle \cdot, \cdot \rangle$ defined by

$$(6) \quad \langle A, B \rangle := \text{trace}(A \cdot B) \quad \text{for all } A, B \in X.$$

LEMMA 4.1. *Let X be the real Hilbert space of all real symmetric (n, n) matrices with the scalar product (6). Then the pointed convex cone C given in (5) has a closed-bounded base.*

Proof. Let us first remark that the set C is indeed a pointed convex cone. The set

$$B := \{A \in C \mid \langle A, I \rangle = 1\}$$

(where I denotes the identity matrix) is a base for the ordering cone C . This base is certainly closed. In order to see that it is also bounded, take any $A \in B$. Then we get

$$\|A\|^2 = \langle A, A \rangle = \text{trace}(A^2) < (\text{trace}(A))^2 = \langle A, I \rangle^2 = 1,$$

and the proof is complete. \square

So, with the aid of Lemma 4.1, the result presented in Corollary 3.5 can immediately be applied in this setting. This result may be useful for the investigation of minimal covariance matrices which fit into this setting.

5. Conclusion. The presented density result has potential consequences for theoretical as well as numerical investigations. On the basis of this theorem, it seems to make sense in our setting to use the proper minimality notion instead of the minimality notion for converse duality theorems and sufficient conditions of the Lagrange multiplier type. From a numerical point of view it is certainly permissible for the solution of convex problems to work with strictly positive linear functions for scalarization. In this case the scalarized problems are much simpler to solve since uniqueness assumptions can be avoided.

Acknowledgments. The author thanks Professor R. Nehse and T. Staib for helpful discussions on this subject and M. Petschke for pointing out the last example. Moreover, the author gratefully acknowledges the comments of a referee which improved the presentation of this paper.

REFERENCES

- [1] K. J. ARROW, E. W. BARANKIN, AND D. BLACKWELL, *Admissible points of convex sets*, in Contributions to the Theory of Games, H. W. Kuhn and A. W. Tucker, eds., Princeton Univ. Press, Princeton, 1953.
- [2] E. BISHOP AND R. R. PHELPS, *The support functionals of a convex set*, in Proc. Symposium on Pure Math. VII, Convexity, American Mathematical Society, Providence, RI, 1963, pp. 27–35.
- [3] G. R. BITRAN AND T. L. MAGNANTI, *The structure of admissible points with respect to cone dominance*, J. Optim. Theory Appl., 29 (1979), pp. 573–614.
- [4] J. M. BORWEIN, *The geometry of Pareto efficiency over cones*, Math. Operationsforsch. Statist. Ser. Optim., 11 (1980), pp. 235–248.
- [5] ———, *Continuity and differentiability properties of convex operators*, Proc. London Math. Soc., 44 (1982), pp. 420–444.
- [6] R. HARTLEY, *On cone-efficiency, cone-convexity and cone-compactness*, SIAM J. Appl. Math., 34 (1978), pp. 211–222.
- [7] G. ISAC, *Sur l'existence de l'optimum de Pareto*, Riv. Mat. Univ. Parma, 9 (1983), pp. 303–325.
- [8] J. JAHN, *Mathematical vector optimization in partially ordered linear spaces*, Peter Lang, Frankfurt, 1986.
- [9] M. G. KREIN AND M. A. RUTMAN, *Linear operators leaving invariant a cone in a Banach space*, Trans. Amer. Math. Soc., 10, Providence, RI, 1962, pp. 199–325.
- [10] M. MAJUMDAR, *Some approximation theorems on efficiency prices for infinite programs*, J. Econom. Theory, 2 (1970), pp. 399–410.
- [11] B. PELEG, *Topological properties of the efficient point set*, Proc. Amer. Math. Soc., 35 (1972), pp. 531–536.
- [12] R. R. PHELPS, *Support cones in Banach spaces and their applications*, Adv. in Math., 13 (1974), pp. 1–19.
- [13] R. RADNER, *A note on maximal points of convex sets in l_∞* , in Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, L. M. Le Cam and J. Neyman, eds., Univ. of California Press, Berkeley, CA, 1967.
- [14] W. SALZ, *Eine topologische Eigenschaft der effizienten Punkte konvexer Mengen*, Operat. Res. Verfahren, XXIII (1976), pp. 197–202.

MAXIMUM LIKELIHOOD ESTIMATION OF DISCRETE CONTROL PROCESSES*

JOHN RUST†

Abstract. Consider the following “inverse stochastic control” problem. A statistician observes a realization of a controlled stochastic process $\{d_t, x_t\}$ consisting of the sequence of states x_t , and decisions d_t of an agent at times $t = 1, \dots, T$. The null hypothesis is that the agent’s behavior is generated from the solution to a Markovian decision problem. The inverse problem is to use the data $\{d_t, x_t\}$ to go backward and “uncover” the agent’s objective function U , and his beliefs about the law of motion of the state variables p . The problem is complicated by the fact that the statistician generally only observes a subset x_t of the state variables (x_t, η_t) observed by the agent. This paper formulates the inverse problem as a problem of statistical inference, explicitly accounting for *unobserved state variables* η_t in order to produce a nondegenerate and internally consistent statistical model. Specifically, the functions U and p are assumed to depend on a vector of unknown parameters θ known by the agent but not by the statistician. The agent’s preferences and expectations are uncovered by finding a parameter vector $\hat{\theta}$ that maximizes the likelihood function for the observed sample of data. The difficulty is that neither the dynamic programming problem nor the associated likelihood function has an a priori known functional form. In general the solution is only described recursively via Bellman’s “principle of optimality.” This paper derives a *nested fixed-point* maximum likelihood algorithm that computes $\hat{\theta}$ and the associated value function $V_{\hat{\theta}}$ for a class of *discrete control processes* $\{d_t, x_t\}$, where the control variable d_t is restricted to a finite set of alternatives. Given M independent realizations of $\{d_t, x_t\}$ for T time periods, it is shown that $\hat{\theta}$ converges to the true parameter θ^* with probability 1 and has an asymptotic Gaussian distribution as M (or the number of periods T) approaches infinity. Uniform convergence of the algorithm is established by showing that the estimated value function $V_{\hat{\theta}}$ (a random element in a Banach space B) converges with probability 1 to the true value function and has an asymptotic Gaussian distribution in B .

Key words. stochastic control, inverse problem, maximum likelihood estimation, contraction mappings, Newton–Kantorovich algorithm, nested fixed-point algorithm

AMS(MOS) subject classifications. 62M05, 65J15, 90C40

1. Introduction. The theory of statistical inference for stochastic processes, which began with the monograph by Ulf Grenander (1950), is now a well-developed field (cf. Basawa and Prakasa Rao, 1980). The theory of stochastic control, which began with the work of Richard Bellman (1957), is now also well developed (cf. Gihman and Skorohod, 1979). For reasons which are unclear, there has been little interface between these fields, and only recently has work begun on a theory of statistical inference for *controlled stochastic processes*, i.e., stochastic processes that arise as solutions to well-defined optimization problems. In certain fields such as engineering or economics, observed time series data can be interpreted as realizations of controlled stochastic processes of the form $\{d_t, x_t\}$, where d_t is the agent’s decision and x_t is the agent’s state at time t . The goal of statistical inference for these data is not simply to infer the form of the stochastic process governing the historical evolution of $\{d_t, x_t\}$, but to go deeper and attempt to infer the underlying determinant of this stochastic process, namely, the mathematical objective function of the agent.¹ This type of *structural statistical inference* is required in order to test the null hypothesis that the agent’s behavior corresponds to the solution of the stochastic control problem. If

* Received by the editors June 16, 1986; accepted for publication (in revised form) November 4, 1987. This research was funded by National Science Foundation grant SES-8419570.

† Department of Economics, University of Wisconsin, Madison, Wisconsin 53706.

¹ In engineering this inference problem is also known as an “inverse optimal control” problem. In economics it is known as the problem of “revealed preference.”

indeed such a hypothesis is supported by the data, then structural inference is also required in order to perform *policy experiments* that forecast how the stochastic process governing $\{d_t, x_t\}$ changes in response to changes in parameters or constraints affecting the agent's objective function. While the existing literature on estimation of stochastic processes allows us to consistently estimate the *historical* stochastic process governing $\{d_t, x_t\}$, it is of limited use for forecasting the effects of policy changes that alter the agent's objective function. An alteration in the agent's objective function induces a corresponding shift in the solution to the stochastic control problem, implying that the stochastic process governing $\{d_t, x_t\}$ after the policy change is generally different from the historical process governing $\{d_t, x_t\}$ before the policy change. Marschak (1953) and Lucas (1976) have shown that the existing nonstructural or *reduced-form* statistical methods can produce dramatic inference and forecasting errors under commonly analyzed policy experiments.

To make this somewhat abstract discussion more concrete, consider the following example analyzed in Rust (1987a). The data $\{d_t^m, x_t^m\}$, $t = 1, \dots, T_m$, $m = 1, \dots, M$ consist of monthly observations on the mileage x_t^m of each bus m in the fleet of the Madison (Wisconsin) Metropolitan Bus Company. The agent is *Harold Zurcher*, maintenance manager of Madison Metro, who decides each month whether or not to replace the engine on bus m with a rebuilt engine: $d_t^m = 1$ (replace) versus $d_t^m = 0$ (keep). The null hypothesis is that Zurcher follows an engine replacement strategy that minimizes the expected discounted costs of operating each bus over its lifetime. The statistical problem is to use the data $\{d_t^m, x_t^m\}$ to infer the unknown parameter vector $\theta = (\beta, \theta_1, \theta_2, \theta_3)$, where β is Zurcher's intertemporal discount factor, θ_2 is the cost of a replacement engine, θ_3 is a vector of parameters describing the stochastic evolution of the state variable $\{x_t^m\}$ and θ_1 is a vector of parameters specifying the functional form of the operating cost function $c(x, \theta_1)$ (the conditional expectation of monthly operating and maintenance costs as a function of accumulated mileage since last replacement, x). The cost function c cannot be directly estimated from cost data because operating costs include Zurcher's subjective evaluation of the cost of "lost customer good will" associated with the increased frequency of breakdown of older buses. Structural estimation of this model is required because (1) we want to test the null hypothesis that Zurcher's behavior is in fact consistent with this simple model of optimal replacement, and (2) we want to forecast the effect of certain policy changes on the frequency and timing of engine replacement. For example, we might want to forecast how the demand for replacement engines is affected by changes in replacement costs θ_2 , or by bus utilization intensity (represented by an appropriate change in θ_3). Since engine replacement costs and utilization rates have not changed much in the past, existing reduced-form estimation methods (which, for example, attempt to directly measure replacement costs and utilization rates and include them as independent variables in a regression on the total number of engine replacements) will yield imprecise and unreliable forecasts.

This paper defines a class of controlled stochastic processes, *discrete control processes*, which are explicitly derived as solutions to a class of Markovian decision problems where the choices d_t are restricted to a finite set of alternatives. The inverse problem is to use the data $\{d_t, x_t\}$ to go backward and "uncover" the agent's objective function, U , and his beliefs about the law of motion for the state variables, p . The problem is complicated by the fact that the statistician generally only observes a subset x_t of the vector of state variables (x_t, η_t) observed by the agent. We solve the inverse problem by formulating it as a problem of statistical inference, explicitly accounting for *unobserved state variables* η_t in order to produce a nondegenerate and internally

consistent statistical model. Specifically, the functions U and p are assumed to depend on a vector of unknown parameters θ known by the agent but not by the statistician. The agent's preferences and expectations are uncovered by finding a parameter vector $\hat{\theta}$ that maximizes the likelihood function for the observed sample of data. The difficulty is that neither the dynamic programming problem nor the associated likelihood function have an a priori known functional form. In general the solution is only described recursively via Bellman's "principle of optimality." However, under a strong but statistically testable restriction on the joint process $\{x_t, \eta_t\}$, we can develop a tractable maximum likelihood algorithm, the *nested fixed-point* algorithm, that computes $\hat{\theta}$ and the associated value function $V_{\hat{\theta}}$. Given M independent realizations of $\{d_t, x_t\}$ for T time periods, we show that $\hat{\theta}$ converges to the true parameter θ^* with probability 1 and has an asymptotic Gaussian distribution as M (or the number of periods T) approaches infinity. Uniform convergence of the algorithm is established by showing that the estimated value function $V_{\hat{\theta}}$ (a random element of a Banach space B) converges with probability 1 to the true value function and has an asymptotic Gaussian distribution in B . To our knowledge, the nested fixed-point algorithm allows us to formulate and estimate structural parameters for a class of controlled stochastic processes for which there were no previously known estimation methods.

Section 2 summarizes the existing literature on estimation of controlled stochastic processes. Section 3 defines the class of discrete control processes and derives the probability density for the observable components $\{d_t, x_t\}$ of the controlled process $\{d_t, x_t, \eta_t\}$. Section 4 presents the nested fixed point maximum likelihood algorithm and establishes the asymptotic properties of its estimates of θ and the value function V_{θ} .

2. Summary of existing literature. To put the results of this paper in perspective, it is useful to briefly summarize the existing literature on estimation of controlled stochastic processes. The literature dichotomizes into four cases depending on whether or not the time variable t and the control variable d_t are discrete or continuous. The latter distinction is more important, since under very general conditions we can approximate a continuous-time controlled Markov process arbitrarily closely by a discrete-time controlled Markov process (cf. van Dijk, 1984). If d_t can take any value in some convex subset of a Euclidean space, $\{d_t, x_t\}$ is a *continuous control process*; otherwise if d_t is restricted to a countable set, $\{d_t, x_t\}$ is a *discrete control process* (the intermediate case where certain components of d_t are discrete and others are continuous has not been analyzed). Although continuous control processes can be approximated by discrete control processes, in certain circumstances there are analytical advantages to working directly with the continuous formulation instead of the discrete approximation. In an important contribution, Hansen and Singleton (1982) have developed a practical technique for estimating structural parameters of a class of discrete-time, continuous control processes. Their method uses the generalized method of moments technique (Ferguson (1958), Hansen (1982)) to estimate first-order necessary conditions of the agent's stochastic control problem (stochastic Euler equations), avoiding the need for an explicit solution for the optimal decision rule and analytic formulae for the probability distribution of $\{d_t, x_t\}$. The Hansen-Singleton method depends critically on the assumption that the agent's control variable d_t is continuous in order to derive the first-order necessary conditions by the usual variational methods. Unfortunately, many cases such as the bus engine problem have discrete rather than continuous control variables, ruling out the use of the Hansen-Singleton method. A further limitation is their assumption that all variables entering the agent's objective function are observed

by both the agent and the statistician: “this latter qualification does rule out some models in which the implied Euler equations involve unobservable forcing variables” (Hansen and Singleton (1982, p. 1271)).

The difficulties created by allowing discrete control variables d_t and unobserved state variables η_t are twofold. First, discrete stochastic control problems rarely possess closed-form solutions or convenient first-order conditions amenable to estimation. Instead the solution is almost always determined only recursively, from Bellman’s principle of optimality. The second problem is that the optimal control d_t will generally be a function of all the state variables in the model. This implies that, if the agent’s state variables consist of the vector (x_t, η_t) but the statistician only observes x_t , then we obtain a statistical model of the form

$$(2.1) \quad d_t = f(x_t, \eta_t, \theta)$$

which is generally a highly nonlinear, nonseparable function of both the “explanatory variables” x_t and the vector of “error terms” η_t . These considerations lead to a statistical model in which (1) the dependent variable d_t is finite valued, (2) the data $\{d_t, x_t\}$ are serially dependent, (3) the unobservables η_t enter in a nonlinear, nonseparable fashion, and (4) the model f has no a priori known functional form.

In spite of these difficult statistical problems, pioneering efforts by Gotz and McCall (1984), Miller (1984), Pakes (1985), and Wolpin (1984) have successfully estimated several specific formulations of the discrete control process (2.1). Gotz and McCall estimated a “dynamic retention model” of Air Force officers by maximum likelihood, numerically solving the optimal stopping problem of when to leave the Air Force for private industry. Miller estimated a multi-armed bandit model of occupation choice by numerically computing Gitten’s (1979) “dynamic allocation indices.” Pakes estimated an optimal stopping model of patent renewal behavior of European firms by tabulating the distribution of times at which realizations of a simulated stochastic process of patent returns crossed a parametrically determined optimal stopping barrier. Wolpin estimated a model of Malaysian women’s decisions about the number and timing of births over their fertile period by solving a dynamic programming problem by backward induction from the last year of the fertile period. Although these papers have produced ingenious methods for interfacing stochastic control and statistical estimation theory, none of the methods offers a general approach for estimating dynamic programming models of discrete sequential choice. With the exception of Miller’s work (which is restricted to bandit processes), each of the papers handles only specific finite-horizon binary decision problems. Section 3 develops a class of discrete control processes that allows choice sets with arbitrary (finite) numbers of elements, and handles both finite and infinite horizon Markovian decision problems. In particular, all of the above models can be reformulated and estimated as discrete control processes. This framework can also handle problems that fall outside the domain of any of the above methods. Examples are the bus engine replacement problem (Rust, 1987a), retirement behavior of older workers (Rust, 1988), utilization and retirement of cement kilns (Das, 1987), and women’s choice of contraceptives (Montgomery, 1987).

3. Discrete control processes: main results. At each time period t , $t = 1, 2, 3, \dots$, an agent observes his *state* (x_t, η_t) and chooses an *action* d_t from a finite *choice set* $C(x_t)$ resulting in a known reward or *utility* $U(x_t, \eta_t, d_t)$. The null hypothesis is that the agent’s objective is to maximize the expected discounted value of U over a possibly

infinite horizon.² A statistician observes d_t and a subset of the state variables x_t for time periods $t = 1, \dots, T$. The goal is to use the realization of $\{d_t, x_t\}$ to infer the objective function U and the law of motion for the state variables p . The only clear way to do this is to specify parametric functional forms for U and p , calculate a probability density for $\{d_t, x_t\}$ and estimate the unknown parameters θ by maximum likelihood. The conditional probability of d_t given x_t , $P(d_t|x_t, \theta)$, can be calculated by “integrating out” the unobserved state variables η_t in the agent’s decision rule f in (2.1). However, this creates serious computational problems, since f generally does not have an a priori known functional form, but must be calculated by numerical solution of the agent’s dynamic programming problem. Once f is calculated it must be integrated many times (once for each possible value of x), and this whole process must be repeated for each trial value of θ over the course of a numerical search for the maximum likelihood estimate $\hat{\theta}$. In the interest of computational tractability, we impose the following restrictions.

(A1) U has the following additively separable representation:

(3.1) $U(x, \eta, d) = u(x, d, \theta_1) + w(\eta, d)$, where θ_1 is a vector of parameters known to the agent but unknown to the statistician.

(A2) Let $\varepsilon \equiv \{w(\eta, d) | d \in C(x)\}$ be the vector of *unobserved utility components*. The joint stochastic process $\{x_t, \varepsilon_t, d_t\}$ is a *controlled Markov process* with probability density p :

(3.2) $\Pr \{x_{t+1}, \varepsilon_{t+1} | d_t, x_t, \eta_t, d_{t-1}, x_{t-1}, \eta_{t-1}, \dots\} = p(x_{t+1}, \varepsilon_{t+1} | x_t, \varepsilon_t, d_t, \theta_2, \theta_3)$, where (θ_2, θ_3) is a vector of parameters known to the agent but unknown to the statistician.

(A3) For each (x_t, ε_t) and all $d \in C(x_t)$, the Markov transition density p factors as

(3.3) $p(x_{t+1}, \varepsilon_{t+1} | x_t, \varepsilon_t, d, \theta_2, \theta_3) = q(\varepsilon_{t+1} | x_{t+1}, \theta_2) \pi(x_{t+1} | x_t, d, \theta_3), \quad d \in C(x_t)$.

Assumptions (A1) and (A2) imply that the agent’s decision problem is a Markovian decision problem on the transformed state space with elements (x_t, ε_t) . Assumption (A3) is a *conditional independence* restriction that limits the pattern of dependence in $\{x_t, \varepsilon_t\}$ in two ways. First, x_{t+1} is a sufficient statistic for ε_{t+1} , implying that any statistical dependence between ε_t and ε_{t+1} is transmitted entirely through the vector x_{t+1} . Second, the probability density for x_{t+1} depends only on x_t and not on ε_t . Intuitively, the $\{\varepsilon_t\}$ process can be regarded as noise superimposed on the underlying $\{x_t\}$ process, since ε_t is drawn from the density $q(\varepsilon_t | x_t, \theta_2)$ given the realized value of x_t . Certainly (A1)–(A3) are strong restrictions. The payoff, as we will show, is a substantial simplification of the numerical problems of (1) solving the agent’s stochastic control problem to obtain the decision rule f , and (2) integrating f to obtain the associated conditional choice probability, $P(d|x, \theta)$ (see Fig. 1).

² The agent may not literally solve the control problem in the sense of consciously performing the calculations involved, much the way a good pool player exploits the laws of physics (the angle of incidence equaling the angle of reflection, conservation of momentum, etc.) without being consciously aware of these principles. Recent research in learning algorithms (Wheeler and Narendra, 1986) shows that relatively simple adjustment rules lead to optimal behavior in finite state controlled Markov processes even though the agent has no initial knowledge of his objective function or the law of motion of the state variables. Other arguments, based on evolutionary principles, also support the notion that agents can behave “as if” they had solved a complicated control problem even though they may not be consciously aware of having solved it.

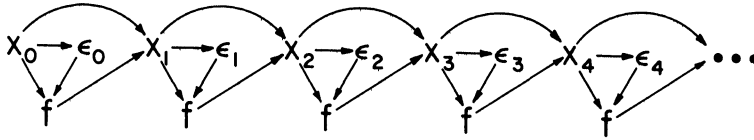


FIG. 1. Pattern of dependence in a controlled Markov process.

Table 1 summarizes the agent's decision problem, formulated as an infinite-horizon Markovian decision problem on the transformed state space (x, ε) .³ The agent chooses a sequence of decision rules or *controls* $f_t(x_t, \varepsilon_t, \theta)$ to maximize expected discounted utility over an infinite horizon. Define the *value function* V_θ by

$$(3.4) \quad V_\theta(x_t, \varepsilon_t) = \sup_{\Pi} E \left\{ \sum_{j=t}^{\infty} \beta^{(j-t)} [u(x_j, f_j, \theta_1) + \varepsilon_j(f_j)] \mid x_t, \varepsilon_t \right\}$$

where $\Pi = \{f_t, f_{t+1}, f_{t+2}, \dots\}$, $f_t(x_t, \varepsilon_t) \in C(x_t)$ for all t , x_t , and ε_t , and where the expectation is taken with respect to the transition density for the controlled stochastic process $\{x_t, \varepsilon_t\}$ determined from Π and the transition density $p(x_{t+1}, \varepsilon_{t+1} \mid x_t, \varepsilon_t, d, \theta_2, \theta_3)$. This treatment implicitly assumes that the controls f_t are nonstochastic and depend only on the current state of the process (x_t, ε_t) . The Markovian structure of the problem, together with Blackwell's theorem (Blackwell, 1968), implies that these assumptions involve no loss in generality.

Problem (3.4) differs from the standard formulation of a Markovian decision problem presented by Blackwell (1968), Denardo (1967), and Bertsekas and Shreve (1978) due to the fact that the utility function $[u(x, d, \theta_1) + \varepsilon(d)]$ will generally not be uniformly bounded in (x, ε) under usual statistical assumptions about the distribution of unobservables (e.g., ε is jointly normally distributed). The unboundedness of the utility function could potentially lead to unbounded values of the value function (3.4)

TABLE 1
Summary of notation for the discrete control problem.

Symbol	Interpretation
$C(x_t)$	Choice set; a finite set of feasible values for the control variable d , when the observed state variable is x_t .
$\varepsilon_t = \{\varepsilon_t(d) \mid d \in C(x_t)\}$	A $\#C(x_t)$ -dimensional vector of "state variables" observed by the agent but not by the statistician. We interpret $\varepsilon_t(d)$ as an unobserved component of utility of alternative d in time period t .
$x_t = \{x_t(1), \dots, x_t(M)\}$	An M -dimensional vector of state variables observed by the agent and the statistician.
$u(x_t, d, \theta_1) + \varepsilon_t(d)$	Realized single period utility obtained in state (x_t, ε_t) when alternative d is chosen.
$p(x_{t+1}, \varepsilon_{t+1} \mid x_t, \varepsilon_t, d, \theta_2, \theta_3)$	Markov transition density for next period state variable when alternative d is chosen and when the current state is (x_t, ε_t) .
$\theta = (\beta, \theta_1, \theta_2, \theta_3)$	The complete $(1 + K_1 + K_2 + K_3)$ vector of parameters to be estimated, where $\beta \in (0, 1)$ is the agent's intertemporal discount factor.

³ Notice that by allowing the observed state variable x to be an element of a complete, separable metric space, the framework also implicitly covers the finite horizon and nonstationary cases as well. In addition, via a well-known transformation of state variables (see, for example, Bertsekas, 1976), the framework also implicitly covers Bayesian control of Markov processes where the agent has imperfect information about transition probabilities or state variables and learns by sequentially updating his or her beliefs according to Bayes rule.

and nonexistence of an optimal policy Π^* . The following additional assumptions are sufficient to insure the existence of an optimal stationary policy $\Pi^* = (f, f, f, \dots)$ with Markovian decision rule f . In order to present the assumptions, we need some additional notation. We require that the number of elements in each choice set $C(x)$ are uniformly bounded in x , so that we have $C(x) \subseteq C = \{1, \dots, N\}$, where N is the least upper bound on the number of elements in each $C(x)$. Define the state space of the controlled process by $S = \{(x, \varepsilon) \mid x \in \Delta, \varepsilon \in R^N\}$, where Δ is a Borel subset of a complete, separable metric space. Typically Δ will be a Borel subset of R^M , as indicated by the notation $x_t = \{x_t(1), \dots, x_t(M)\}$. Strictly speaking, $\varepsilon_t \in R^{\#C(x_t)}$. However, since the dimensionality of ε_t changes with the number of elements in $C(x_t)$ we imbed the state space for ε_t in the common space R^N and fill out the remaining components with values drawn from some arbitrary continuous distribution independent of the other $\#C(x_t)$ components of ε_t . Let λ be a Lebesgue measure on R^N and let μ be a Borel measure on Δ ; the latter need not be nonatomic. Define a measure ν on S in the usual way by $\nu = \mu \times \lambda$. The first assumption guarantees the existence of a probability density that provides the basis for maximum likelihood estimation.

- (A4) For each $d \in C(x_t)$ and $(x_t, \varepsilon_t) \in S$, the conditional probability distribution of $(x_{t+1}, \varepsilon_{t+1})$ given (x_t, ε_t, d) is regular and has a Radon-Nikodym density $p(x_{t+1}, \varepsilon_{t+1} \mid x_t, \varepsilon_t, d, \theta_2, \theta_3)$ with respect to the measure ν on S .

The remaining assumptions guarantee the existence of a stationary optimal policy Π^* to (3.4). Throughout the remainder of the paper we will use the shorthand notation $Ef(x, \varepsilon, d)$ to denote the conditional expectation of a function $f: S \rightarrow R$ with respect to the conditional probability density p , namely,

$$(3.5) \quad Ef(x, \varepsilon, d) \equiv \int_S f(y, \eta) p(y, \eta \mid x, \varepsilon, d, \theta_2, \theta_3) \mu(dy) \times \lambda(d\eta).$$

- (A5) $0 < \beta < 1$;

- (A6) $C(x) \subseteq C = \{1, \dots, N\}$ for all $x \in \Delta$;

- (A7) For each $d \in C(x)$, $x \in \Delta$, $u(x, d, \theta_1)$ is upper semicontinuous at x and has bounded expectation in the sense that $R(x, \varepsilon) \equiv \sum_{j=0}^{\infty} \beta^j R_j(x, \varepsilon) < +\infty$, $(x, \varepsilon) \in S$, where $R_0(x, \varepsilon) = \max_{d \in C(x)} |u(x, d, \theta_1) + \varepsilon(d)|$, $R_{j+1}(x, \varepsilon) = \max_{d \in C(x)} ER_j(x, \varepsilon, d)$, where $ER_j(x, \varepsilon, d)$ is defined in (3.5);

- (A8) For each $x \in C(x)$, $Eh(x, \varepsilon, d)$ is continuous at each point $(x, \varepsilon) \in S$ for all Borel measurable functions $h: S \rightarrow R$ satisfying $|h(x, \varepsilon)| \leq |R(x, \varepsilon)| + 1$, $(x, \varepsilon) \in S$.

THEOREM 3.1. *Under assumptions (A1), (A2), (A4)–(A8) a stationary optimal policy $\Pi^* = (f, f, f, \dots)$ exists for some Borel measurable function $f: S \rightarrow C$. The decision rule f is nonstochastic, Markovian, and is determined from Bellman's equation*

$$(3.6) \quad V_\theta(x, \varepsilon) = \max_{d \in C(x)} [u(x, d, \theta_1) + \varepsilon(d) + \beta EV_\theta(x, \varepsilon, d)]$$

by the identity

$$(3.7) \quad f(x, \varepsilon, \theta) \equiv \operatorname{argmax}_{d \in C(x)} [u(x, d, \theta_1) + \varepsilon(d) + \beta EV_\theta(x, \varepsilon, d)].$$

Theorem 3.1 is a specialization of Theorem 2.1 of Bhattacharya and Majumdar (1985), who extend the basic results of stochastic dynamic programming to allow for unbounded rewards. Note that although $V_\theta(x, \varepsilon)$ is finite for each $(x, \varepsilon) \in S$, it is generally not uniformly bounded in (x, ε) . As a result, we cannot apply the standard

results of Blackwell (1968) and Denardo (1967), who use uniform boundedness to show that V_θ is a fixed point of a contraction mapping on the Banach space B of all uniformly bounded upper semicontinuous functions from S to R . This is unfortunate, since the main numerical method for solving stochastic control problems consists of computing the value function by solving the associated fixed-point problem. Lippman (1975) provided alternative conditions under which Theorem 3.1 holds, where V_θ is the unique fixed point of a contraction mapping on the Banach space B_w of all bounded, upper semicontinuous functions under the weighted supremum norm defined by $\|h\|_w = \sup_{(x, \varepsilon) \in S} |h(x, \varepsilon)| / w(x, \varepsilon)$, where w is a specified weight function satisfying $w \geq 1$. However even if we adopt Lippman's approach, there are two difficulties hampering direct statistical implementation of the model $d_t = f(x_t, \varepsilon_t, \theta)$ given by the solution to (3.2) and (3.3). First, standard distributional assumptions for unobservables imply that ε_t will be continuously distributed on R^N with unbounded support. However, this raises serious dimensionality problems since the optimal stationary policy f will ordinarily be computed by solving for the fixed point V_θ to Bellman's equation (3.2). Since ε is a vector of continuous state variables, it must be *discretized* in order to compute V_θ on a digital computer. The discretization procedure approximates the true value function V_θ , an element of an infinite-dimensional space, by a suitable vector in a high-dimensional Euclidean space. Even if we take a very rough grid approximation to the true continuous distribution of ε_t , the dimension of the resulting finite approximation will be too large to be computationally tractable. Second, since ε_t appears nonlinearly in the unknown function EV_θ , we face the additional problem of integrating out over the ε_t distribution to obtain the conditional choice probabilities $P(d_t | x_t, \theta)$ that enter the likelihood function. The *conditional independence assumption* (A3) enables us to circumvent these problems. The payoff to (A3) is twofold. First, (A3) implies that EV_θ is not a function of ε_t so that the required choice probabilities do not require integration with respect to ε over the unknown function EV_θ . Second, (A3) implies that EV_θ is a fixed point of a separate contraction mapping on the space $\Gamma = \{(x, d) | x \in \Delta, d \in C(x)\}$, eliminating the need to compute the fixed point V_θ on the much larger space S and avoiding the numerical integration required to obtain EV_θ from V_θ . Admittedly, (A3) is a strong assumption adopted in the interest of computational tractability. In § 4 we present a simple Lagrange multiplier test of the validity of (A3).

Before presenting the main results, we need additional notation and a mild regularity condition. Let g be a measurable, bounded, real-valued function from Γ to R . Define the norm of g , $\|g\|_\infty$, in the usual manner by $\|g\|_\infty = \sup_{(x, d) \in \Gamma} |g(x, d)|$. It follows that the set of all measurable, real-valued and $\|\cdot\|_\infty$ -bounded functions on Γ is a Banach space B . Given a vector $r(x) \equiv \{r(x, d) | d \in C(x)\}$, define the function $G(r(x) | x, \theta_2)$ by

$$(3.8) \quad G(r(x) | x, \theta_2) \equiv \int_{\varepsilon} \left[\max_{d \in C(x)} [r(x, d) + \varepsilon(d)] \right] q(\varepsilon | x, \theta_2) \lambda(d\varepsilon).$$

$G(r(x) | x, \theta_2)$ is simply the conditional expectation of the maximum of $[r(x, d) + \varepsilon(d)]$, $d \in C(x)$. McFadden (1981) calls G a *social surplus function*. G has an important property, apparently first noted by Williams (1977) and Daly and Zachary (1979), which is a key to our subsequent results.

THEOREM 3.2. *Suppose the density $q(\varepsilon | x, \theta_2)$ has finite first moments for μ almost all $x \in \Delta$. Then for any vector $r(x) = \{r(x, d) | d \in C(x)\}$, the social surplus function $G(r(x) | x, \theta_2)$ defined in (3.8) has the following properties:*

- (a) For each x , G is a positively linear homogeneous, convex function of $r(x)$.
 (b) G has the additivity property $G(r(x) + \alpha | x, \theta_2) = \alpha + G(r(x) | x, \theta_2)$ for any scalar α , where $r(x) + \alpha \equiv \{r(x, d) + \alpha | d \in C(x)\}$.

(c) Let G_d denote the partial derivative of G with respect to $r(x, d)$, and let $P(d | x, \theta_2)$ denote the conditional probability that $r(x, d) + \varepsilon(d)$ is the largest: $P(d | x, \theta_2) \equiv \int_{\varepsilon} I\{d = \operatorname{argmax}_{j \in C(x)} [r(x, j) + \varepsilon(j)]\} q(d\varepsilon | x, \theta_2)(d\varepsilon | x, \theta_2)$. Then we have

$$(3.9) \quad P(d | x, \theta_2) = G_d(r(x) | x, \theta_2) \text{ as in formula (3.8)}$$

We need one final regularity condition for u and q .

- (A9) $u \in B$ and for each $r \in B$, $EG \in B$ where $G(r(x) | x, \theta_2)$ is given in (3.8) and EG is defined by

$$(3.10) \quad EG(x, d) \equiv \int_y G(r(y) | y, \theta_2) \pi(dy | x, d, \theta_3).$$

Assumption (A9) will be satisfied for most choices of q . A sufficient condition for (A9) to hold is that $q(\varepsilon | x, \theta_2)$ has finite first moments that are uniformly bounded in x . We are now ready to state the main results of this paper.

THEOREM 3.3. Under assumptions (A1)–(A9) the value function V_θ is given by

$$(3.11) \quad V_\theta(x, \varepsilon) = \max_{d \in C(x)} [v_\theta(x, d) + \varepsilon(d)]$$

where $v_\theta: \Gamma \rightarrow R$ is the unique fixed point of the contraction mapping $\Lambda_\theta: B \rightarrow B$ defined by

$$(3.12) \quad \Lambda_\theta(v)(x, d) = u(x, d, \theta_1) + \beta \int_y G(v(y) | y, \theta_2) \pi(dy | x, d, \theta_3), \quad d \in C(x),$$

and where G is the social surplus function defined in (3.8).

THEOREM 3.4. Under assumptions (A1)–(A9) the controlled process $\{d_t, x_t\}$ is Markovian with transition density given by

$$(3.13) \quad \Pr\{d_{t+1}, x_{t+1} | d_t, x_t\} = P(d_{t+1} | x_{t+1}, \theta) \pi(x_{t+1} | x_t, d_t, \theta_3)$$

where the conditional choice probability $P(d | x, \theta)$ is given by

$$(3.14) \quad P(d | x, \theta) = G_d(v_\theta(x) | x, \theta_2)$$

and where v_θ is the fixed point of Λ_θ in (3.12) and G_d is the partial derivative of the social surplus function G in (3.8) with respect to $v_\theta(x, d)$.

The proofs of Theorems 3.3 and 3.4 are given in Appendix 1. Theorem 3.3 shows that under (A3) the dynamic programming problem can be solved by computing the fixed point v_θ to the contraction mapping Λ_θ on the reduced state space Γ rather than the full state space S . Furthermore, the unobserved state variables $\{\varepsilon(d) | d \in C(x)\}$ enter linearly and additively separably inside the max operator in the value function V_θ in (3.11). This implies that the choice probabilities $P(d | x, \theta)$ in (3.14) can be computed using the same formulas used in static discrete choice models with the addition of the expected discounted future utility $\beta EV_\theta(x, d)$ to the usual static utility term $u(x, d, \theta_1)$. To see this, note that EV_θ and v_θ are related by the equations

$$(3.15) \quad v_\theta(x, d) = u(x, d, \theta_1) + \beta EV_\theta(x, d)$$

where EV_θ is the fixed point to the contraction mapping $T_\theta: B \rightarrow B$ defined by

$$(3.16) \quad T_\theta(v)(x, d) = \int_y G([u(y, \theta_1) + \beta v(y)] | y, \theta_2) \pi(dy | x, d, \theta_3).$$

By choosing specific functional forms for q we obtain concrete formulas for the choice probability $P(d|x, \theta)$ and the contraction mapping Λ_θ . For example, if ε has a multivariate extreme value distribution

$$(3.17) \quad q(\varepsilon|x, \theta_2) = \prod_{d \in C(x)} \exp\{-\varepsilon(d) + \theta_2\} \exp\{-\exp\{-\varepsilon(d) + \theta_2\}\},$$

then $P(d|x, \theta)$ is given by the well-known *multinomial logit* formula

$$(3.18) \quad P(d|x, \theta) = \frac{\exp\{v_\theta(x, d)\}}{\sum_{j \in C(x)} \exp\{v_\theta(x, j)\}},$$

and v_θ is given by the fixed point to the contraction mapping Λ_θ :

$$(3.19) \quad \Lambda_\theta(v)(x, d) = u(x, d, \theta_1) + \beta \int_y \log \left[\sum_{j \in C(y)} \exp\{v(y, j)\} \right] \pi(dy|x, d, \theta_3).$$

4. The nested fixed-point maximum likelihood algorithm. Suppose we observe a realization of the process $\{d_t, x_t\}$. We can uncover the agent's preferences u and expectations $p = q\pi$ by finding a parameter $\hat{\theta}$ that maximizes the likelihood function L^f defined by⁴

$$(4.1) \quad L^f(x_1, \dots, x_T, d_1, \dots, d_T|x_0, d_0, \theta) = \prod_{t=1}^T P(d_t|x_t, \theta) \pi(x_t|x_{t-1}, d_{t-1}, \theta_3).$$

The difficulty is that L^f does not have an a priori known functional form: the conditional choice probability $P(d|x, \theta)$ entering (4.1) depends on the value function v_θ which is only implicitly defined as the fixed point of the contraction mapping Λ_θ in (3.12). Theorems 3.3 and 3.4 suggest the following *nested fixed-point* algorithm: an "inner" contraction fixed-point algorithm computes the unknown function v_θ for each value of θ , and an "outer" hill-climbing algorithm searches for the value of θ that maximizes L^f . Theoretically, such an algorithm should be able to estimate a wide range of discrete control processes. However if the approach is to be of any practical use, the algorithm must be able to (1) rapidly compute the fixed point v_θ for any given value of θ , and (2) find the argmax of L^f , $\hat{\theta}$, in as few likelihood function evaluations as possible. In particular L^f should be a smooth (i.e., continuously differentiable) function of θ so that more efficient gradient maximization algorithms can be employed. The smoothness of L^f is also crucial for establishing large sample properties of $\hat{\theta}$ such as asymptotic normality.

Since the underlying "primitive objects" u , q , and π are specified a priori, they can be chosen to be smooth functions of θ . It follows from (4.1) that the issue of smoothness of L^f reduces to the question of the smoothness of $P(d|x, \theta)$ in θ . By Theorems 3.2 and 3.4, $P(d|x, \theta) = G_d(v_\theta(x)|x, \theta_2)$ is a continuously differentiable function of the vector $\{v_\theta(x, d)|d \in C(x)\}$. Therefore the question of smoothness further reduces to the question of finding sufficient conditions under which the mapping $\theta \rightarrow v_\theta$ is a smooth mapping from $R^{1+K_1+K_2+K_3}$ into B . The smoothness of this mapping follows from the implicit function theorem in Banach spaces (cf. Kantorovich and Aiklov (1982, Thm. 3, p. 520)). To see this, note that the pair (v_θ, θ) is a zero of the nonlinear operator $F: B \times R^{1+K_1+K_2+K_3} \rightarrow B$ defined by

$$(4.2) \quad 0 = F(v, \theta) \equiv (I - \Lambda_\theta)(v)$$

⁴ If we have *panel data*, independent realizations of $\{d_t, x_t\}$ for a collection of different agents, then the likelihood function for the panel is simply a product of the individual agent likelihoods in (4.1).

where 0 is the zero element of B and I is the identity operator on B . The implicit function theorem implies that v_θ will be a continuously differentiable function of θ , provided that F has continuous derivatives in θ and the Fréchet derivative of F with respect to v , $\partial F(v, \theta)/\partial v$, is a continuous linear operator on B with continuous inverse at the point $v = v_\theta$. Let $\Lambda'_\theta(v_\theta)$ denote the Fréchet derivative of Λ_θ evaluated at the point v_θ . By Theorem 3.2 and dominated convergence, $\Lambda'_\theta(v_\theta)$ is a linear operator on B with integral representation

$$(4.3) \quad \Lambda'_\theta(v_\theta)(m)(x, d) = \beta \int_y \left[\sum_{j \in C(y)} m(y, j) P(j|y, \theta) \right] \pi(dy|x, d, \theta_3).$$

It is obvious from (4.3) that $\Lambda'_\theta(v_\theta)$ is a continuous linear operator with norm $\|\Lambda'_\theta(v_\theta)\|_\infty \leq \beta < 1$. By the Banach inverse theorem (Kantorovich and Aiklov (1982, Thm. 3, p. 154)), it follows that $\partial F(v, \theta)/\partial v$ is a continuous linear operator with continuous inverse given by

$$(4.4) \quad \left[\frac{\partial F(v, \theta)}{\partial v} \right]^{-1} = [I - \Lambda'_\theta(v)]^{-1} = \sum_{i=0}^{\infty} [\Lambda'_\theta(v)]^i$$

where the latter series is the *Neumann expansion* of $[I - \Lambda'_\theta(v)]^{-1}$.

THEOREM 4.1. *Under regularity conditions (A10)–(A13), $\partial v_\theta/\partial \theta$ exists and is a continuous function of θ given by*

$$(4.5) \quad \frac{\partial v_\theta}{\partial \theta} = [I - \Lambda'_\theta(v)]^{-1} \left[\frac{\partial \Lambda_\theta(v)}{\partial \theta} \right] \Big|_{v=v_\theta}.$$

The regularity conditions (A10)–(A13) guarantee that the operator Λ_θ is a continuously differentiable function of θ , and are listed in Appendix 2.

Equation (4.2) suggests an efficient method for computing the fixed point v_θ , namely, apply Newton's method to find the zero $v_\theta \in B$ of the functional equation $F(v, \theta) = 0$. This leads to an iteration of the form

$$(4.6) \quad v^{j+1} = v^j - [I - \Lambda'_\theta(v^j)]^{-1} [I - \Lambda_\theta](v^j)$$

known as the *Newton–Kantorovich iteration*. Kantorovich established that the sequence $\{v^j\}$ generated by (4.6) converges to the fixed point v_θ at a quadratic rate starting from an initial estimate v^0 in a “domain of attraction” sufficiently close to v_θ .⁵ Since Λ_θ is a contraction mapping, the *contraction iteration*

$$(4.7) \quad v^{j+1} = \Lambda_\theta(v^j)$$

is guaranteed to converge to v_θ at a linear rate starting from any initial estimate v^0 . This suggests the following hybrid method of “polyalgorithm”: we start with contraction iterations (4.7) until we generate an estimate v^j in a domain of attraction of v_θ and then switch to the Newton–Kantorovich iteration (4.6) to rapidly converge to the solution. Convergence and numerical stability of the polyalgorithm are guaranteed by the contraction property for Λ_θ and the fact that $[I - \Lambda'_\theta(v)]$ is necessarily invertible for all $\beta \in (0, 1)$ and $v \in B$. The main work involved in computing the Newton–Kantorovich iteration is solving the linear system involving the operator $[I - \Lambda'_\theta(v^j)]$. However, we obtain as a by-product the derivatives $\partial v_\theta/\partial \theta$ at negligible marginal cost using (4.5).

⁵ The Newton–Kantorovich algorithm is not an “optimal” algorithm: a variant due to Werner (1984) evaluates Λ'_θ at an intermediate point $w = \alpha v^j + (1 - \alpha)\Lambda_\theta(v^j)$, $\alpha \in (0, 1)$, to accelerate convergence. Werner showed that a good choice is $\alpha = \frac{1}{2}$.

Given the derivatives $\partial v_\theta / \partial \theta$, it is a straightforward exercise to compute the derivatives of the likelihood function L^f . This allows us to employ more efficient quasi-Newton gradient maximization algorithms to search for the maximum likelihood estimate $\hat{\theta}$. In particular, the well-known *information equality* implies that the information matrix (the cumulated outer products of the first derivative terms $P(d_{t+1} | x_{t+1}, \theta) \pi(x_{t+1} | x_t, d_t, \theta_3)$) is a good approximation to the negative of the Hessian of L^f in large samples. This idea forms the basis for the BHHH optimization algorithm (Berndt et al., 1974) which only requires first derivatives of the likelihood function.⁶ The nested fixed-point algorithm combines the contraction/Newton-Kantorovich fixed point polyalgorithm and the BHHH/Broyden gradient optimization algorithm in order to obtain an efficient and numerically stable method for computing the maximum likelihood estimate $\hat{\theta}$. The algorithm is summarized in Fig. 2 (where the fixed-point problem is stated in terms of the operator T_θ in (3.16) instead of Λ_θ in (3.12)). Computational experience with the algorithm has verified its favorable theoretical properties. Fixed points as large as several hundred dimensions can be rapidly calculated on an IBM-PC (e.g., in the bus problem a 180-dimensional discretized fixed-point problem can be computed to a tolerance of 10^{-16} within 60 seconds). Fixed points in the hundreds of thousands and even millions of dimensions have been computed in a matter of seconds on supercomputers such as the Cray-2. For further details on the computational performance of the algorithm, see Rust (1987c), (1988).

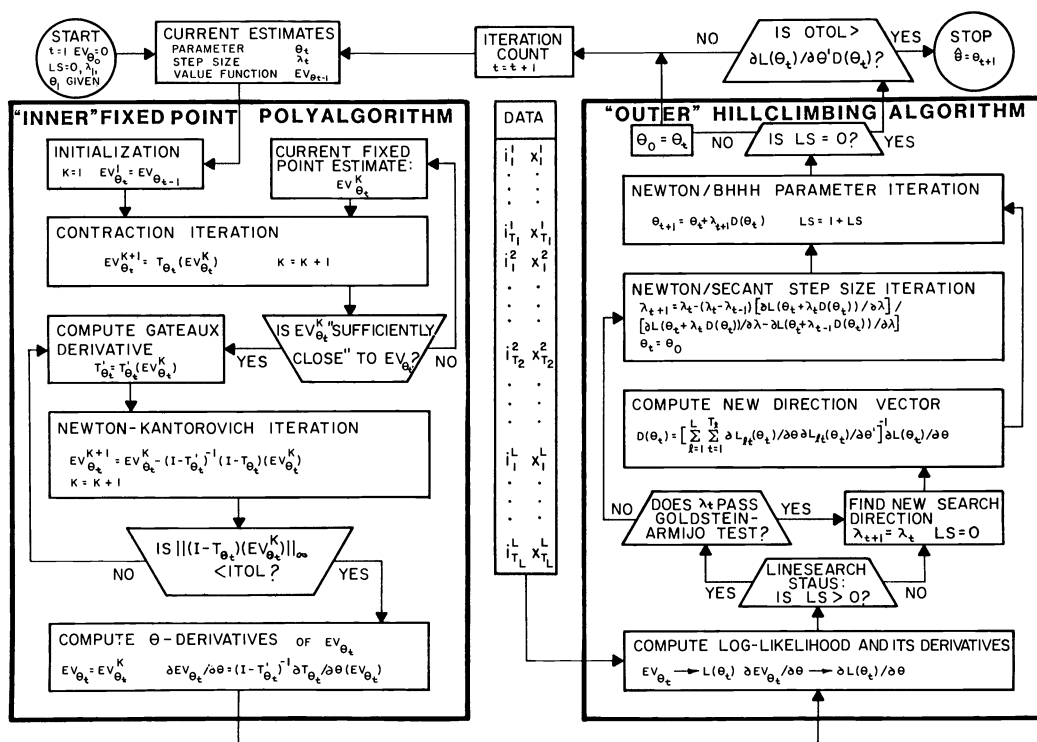


FIG. 2. Nested fixed-point maximum likelihood algorithm.

⁶ Convergence of the BHHH method in small samples can be accelerated by Broyden and Davidson, Fletcher, and Powell updating procedures that adaptively improve the accuracy of the information matrix approximation to the Hessian of L^f .

To conduct inference, we need to know the large sample properties of the maximum likelihood estimates of θ and v_θ . The asymptotic distributions can be derived for three separate cases. In Case One, we have a fixed number of observations T_m for each realization m , $m = 1, \dots, M$, and the number of realizations M tends to infinity. In Case Two, we have a fixed number of realizations M , and the number of observations per realization T_m tends to infinity. In the final case both T_m and M tend to infinity. Currently, most panel data sets have only a limited number of time periods T_m so that approximating the finite sample distribution of the maximum likelihood estimator (MLE) by its asymptotic distribution as $T_m \rightarrow \infty$ is not likely to yield accurate results. Calculation of the asymptotic distribution of the MLE in Cases Two and Three is complicated by the serial dependence in the $\{d_t, x_t\}$ process. Since $\{d_t, x_t\}$ is Markovian, we can show that the log-likelihood function is a zero mean martingale, allowing us to use the law of large numbers and central limit theorems for martingales to establish consistency and asymptotic normality of $\hat{\theta}$. Application of these results appears to require continuous third derivatives of the likelihood (see Billingsley (1961)). To simplify both the assumptions and presentation, we focus on Case One and leave Cases Two and Three to future work. In Case One we further assume the following:

- (A14) The realizations of $\{d_t^m, x_t^m\}$ and $\{d_t^j, x_t^j\}$ are independent if $j \neq m$;
- (A15) The observation period T_m is the same for all realizations $m = 1, \dots, M$;
- (A16) $\{d_t^m, x_t^m\}$ has a unique ergodic distribution $\Omega(d, x, \theta)$ from which the initial observations (d_0^m, x_0^m) are drawn, $m = 1, \dots, M$.

Assumptions (A14)–(A16) imply that the data are independent and *identically* distributed among a cohort of individuals. This enables us to use the simpler IID strong law of large numbers and Lindeberg–Levy central limit theorem to establish the consistency and asymptotic normality of $\hat{\theta}$, requiring only continuous second derivatives of the likelihood.

THEOREM 4.2. *Suppose (A1)–(A16) hold. If additional regularity conditions (A17)–(A35) hold, then we have the following:*

- (1) *The maximum likelihood estimator $\hat{\theta}$ is a well-defined random variable;*
- (2) *$\hat{\theta}$ converges to the true parameter vector θ^* with probability one as the sample size $M \rightarrow \infty$;*
- (3) *The distribution of $\sqrt{M}(\hat{\theta} - \theta^*)$ converges weakly to $N(0, -H(\theta^*)^{-1})$, where $H(\theta^*)$ is given by*

$$\begin{aligned}
 H(\theta^*) = & E\{\partial^2 \log [P(d_t | x_t, \theta^*)] / \partial \theta \partial \theta'\} \\
 & + E\{\partial^2 \log [\pi(x_t | x_{t-1}, d_{t-1}, \theta_3^*)] / \partial \theta \partial \theta'\} \\
 (4.8) \quad & = -E\{[\partial \log P(d_t | x_t, \theta^*) / \partial \theta][\partial \log P(d_t | x_t, \theta^*) / \partial \theta']\} \\
 & - E\{[\partial \pi(x_t | x_{t-1}, d_{t-1}, \theta_3^*) / \partial \theta][\partial \pi(x_t | x_{t-1}, d_{t-1}, \theta_3^*) / \partial \theta']\}.
 \end{aligned}$$

Assumptions (A17)–(A35) are standard regularity conditions guaranteeing compactness of the parameter space, continuity of the second derivatives of the likelihood, and asymptotic uniqueness of the maximum likelihood estimate (a necessary condition for identification). The large number of conditions arise from the need to make separate assumptions on each of the “primitive objects” u , π , and q and are listed in Appendix 2. The proof of Theorem 4.2 closely follows standard treatments of the IID case found, for example, in LeCam (1953), Huber (1967), or White (1982), and is therefore omitted.

Although the parameter estimates $\hat{\theta}$ are the end result of maximum likelihood estimation, in many cases interest focuses not on the parameters themselves but on the form of the estimated value function $v_{\hat{\theta}}$, a B -valued random element. For example, in many cases we would like to compute confidence bands for $v_{\hat{\theta}}$ to judge how precisely we estimate the unknown value function v_{θ^*} .

THEOREM 4.3. *Let $\hat{\theta}$ be a consistent estimator of θ^* with the distribution of $\sqrt{M}(\hat{\theta}_M - \theta^*)$ converging weakly to $N(0, \Sigma)$. Then if (A1)–(A13) hold, we have the following:*

- (1) $v_{\hat{\theta}}$ is a B -valued random element;
- (2) $v_{\hat{\theta}}$ converges to v_{θ^*} with probability 1;
- (3) The distribution of $\sqrt{M}(v_{\hat{\theta}} - v_{\theta^*})$ converges weakly to a Gaussian random element in B with expectation 0 and covariance operator $[\partial v_{\theta^*}/\partial \theta] \Sigma [\partial v_{\theta^*}/\partial \theta']$.

The proof of Theorem 4.3 is given in Rust (1987b).

We conclude the paper by suggesting a simple specification test for the validity of the conditional independence assumption (A3). Recall that (A3) implies that ε_t is independent of ε_{t-1} conditional on x_t , a restriction that seems to be necessary for a computationally tractable estimation algorithm. A natural way to test this assumption is to add some function of last period control variable d_{t-1} to the value function v_{θ} entering the choice probabilities $P(d_t | x_t, \theta)$, say, $v_{\theta}(x_t, d_t) + \theta_4 I\{d_{t-1} = d_t\}$. Under the null hypothesis that (A3) is valid, formula (3.11) of Theorem 3.3 implies that the previous period choice d_{t-1} has no effect on the choice d_t , so the maximum likelihood estimate of θ_4 should converge to 0. However, if (A3) is not true, ε_t will not be independent of ε_{t-1} given x_t . Thus, $d_{t-1} = f(x_{t-1}, \varepsilon_{t-1}, \theta)$ will not be independent of ε_t . It follows that if (A3) does not hold, then the maximum likelihood estimate of θ_4 will converge to a nonzero value. Thus, standard Likelihood Ratio, Lagrange multiplier, or Wald tests of the hypothesis that $\theta_4 = 0$ provide a simple way to test the validity of (A3).

Appendix 1: Proofs of Theorems 3.3 and 3.4.

Proof of Theorem 3.3. First we show that (A3) implies that the conditional expectation $EV_{\theta}(x, \varepsilon, d)$ does not depend on ε , and so can be written as $EV_{\theta}(x, d)$. By (A3) and Fubini's Theorem we have

$$\begin{aligned}
 EV_{\theta}(x, \varepsilon, d) &= \int_S V_{\theta}(y, \xi) p(y, \xi | x, \varepsilon, d, \theta_2, \theta_3) \mu(dy) \times \lambda(d\eta) \\
 (1) \quad &= \int_y \left[\int_{\xi} V_{\theta}(y, \xi) q(\xi | y, \theta_2) \lambda(d\eta) \right] \pi(y | x, d, \theta_3) \mu(dy) \\
 &= EV_{\theta}(x, d).
 \end{aligned}$$

It follows that the ε only enters linearly inside the max operator in Bellman's equation (3.6), so that V_{θ} is represented by (3.11) for some function v_{θ} of the form $v_{\theta} = u + \beta EV_{\theta}$. Substituting the formula for V_{θ} given in (3.11) into Bellman's equation (3.6) we obtain the fixed-point condition for EV_{θ} given in (3.16)

$$\begin{aligned}
 EV_{\theta}(x, d) &= \int_y \int_{\varepsilon} \left[\max_{j \in C(y)} [u(y, j, \theta_1) + \varepsilon(j) + \beta EV_{\theta}(y, j)] \right] q(d\varepsilon | y, \theta_2) \pi(dy | x, d, \theta_3) \\
 (2) \quad &= \int_y G([u(y, \theta_1) + \beta EV_{\theta}(y)] | y, \theta_2) \pi(dy | x, d, \theta_3) \\
 &= T_{\theta}(EV_{\theta})(x, d).
 \end{aligned}$$

Assumption (A9) guarantees that EV_θ is a member of B , since for any $x \in \Delta$ and any measurable function $f: S \rightarrow \Gamma$ we have

$$(3) \quad E\{[u(y, f(y, \varepsilon), \theta_1) + \varepsilon(f(y, \varepsilon))] | x\} \leq G(\|u\|_\infty | x, \theta_2),$$

$$(4) \quad V_\theta(x, \varepsilon) \leq \sup_{\Pi} E \left\{ \sum_{j=t}^{\infty} \beta^{(j-t)} G(\|u\|_\infty | x_j, \theta_2) | x, \varepsilon, \theta_2, \theta_3 \right\},$$

which imply that

$$(5) \quad \|EV_\theta\| < \|EG\|_\infty / (1 - \beta).$$

Since $\|EG\|_\infty < +\infty$ by (A9), it follows that $EV_\theta \in B$. Thus EV_θ is the fixed point of a well-defined mapping T_θ from B to B . To show that T_θ is a contraction mapping, note that for each y and each $g, h \in B$ we have

$$(6) \quad \begin{aligned} & \max_{j \in C(y)} [u(y, j, \theta_1) + \varepsilon(j) + \beta g(y, j)] - \max_{j \in C(y)} [u(y, j, \theta_1) + \varepsilon(j) + \beta h(y, j)] \\ & \leq \max_{j \in C(y)} \beta |g(y, j) - h(y, j)|. \end{aligned}$$

It follows immediately from (6) that

$$(7) \quad \|T_\theta(g) - T_\theta(h)\|_\infty \leq \beta \|g - h\|_\infty$$

so that T_θ is a contraction mapping and EV_θ is the unique fixed point of T_θ in B . Since $v_\theta(x, d) = u(x, d, \theta_1) + \beta EV_\theta(x, d)$, it follows that $v_\theta \in B$. A similar argument shows that v_θ is the fixed point of the contraction mapping Λ_θ defined in (3.12). \square

Proof of Theorem 3.4. Formula (3.14) follows immediately from (3.11) of Theorem 3.3 and (3.9) of Theorem 3.2. The controlled process $\{d_t, x_t\}$ is Markovian since Theorem 3.3 implies that the conditional probability of d_{t+1} given x_{t+1} is given by $P(d_{t+1} | x_{t+1}, \theta)$, and the conditional probability of x_{t+1} given (d_t, x_t) is $\pi(x_{t+1} | x_t, d_t, \theta_3)$ by assumption (A3). \square

Appendix 2: Regularity conditions for Theorems 4.1 and 4.2. Assumptions (A10)–(A13), used in Theorem 4.1, ensure that V_θ is a continuously differentiable function of θ . In what follows, $L(R^k, B)$ denotes the Banach space of all-bounded, linear operators from R^k to B .

$$(A10) \quad \partial u / \partial \theta_1 \in L(R^{K_1}, B) \text{ and is a continuous function of } \theta_1.$$

$$(A11) \quad \text{For any } r \in B, \partial G(r(y) | y, \theta_2) / \partial \theta_2 \text{ exists and is dominated by a } \pi(\cdot | x, d, \theta_3) \text{ integrable function for all } y \text{ except on sets } A(x, d) \text{ of } \pi(\cdot | x, d, \theta_3) \text{ measure zero.}$$

$$(A12) \quad \text{For each } r \in B, \partial EG / \partial \theta_2 \in L(R^{K_2}, B) \text{ and is a continuous function of } \theta \text{ and } r.$$

$$(A13) \quad \text{For each } r \in B, \partial EG / \partial \theta_3 \in L(R^{K_3}, B) \text{ and is a continuous function of } \theta \text{ and } r.$$

Assumptions (A17)–(A35), used in Theorem 4.2, are regularity conditions that guarantee consistency and asymptotic normality of the maximum likelihood estimator $\hat{\theta}$.

$$(A17) \quad \partial^2 u / \partial \theta_1 \partial \theta_1' \in L(R^{K_1^2}, B) \text{ and is a continuous function of } \theta_1.$$

$$(A18) \quad \text{For any } r \in B, \text{ and for any } j, k \in C(y), G_{jk}(r(y) | y, \theta_2) \text{ exists and is dominated by a } \pi(\cdot | x, d, \theta_3) \text{ integrable function for all } y \in \Delta \text{ except on sets } A(x, d) \text{ of } \pi(\cdot | x, d, \theta_3) \text{ measure zero.}$$

(A19) For any $g, r \in B$,

$$\int_y \left\{ \sum_{j,k \in C(y)} g(y, j) g(y, k) G_{jk}(r(y)|y, \theta_2) \right\} \pi(dy|x, d, \theta_3) \in B$$

and is a continuous function of θ and r .

(A20) For any $r \in B$, $\partial^2 G(r(y)|y, \theta_2)/\partial \theta_2 \partial \theta'_2$ exists and is dominated by a $\pi(\cdot|y, d, \theta_3)$ integrable function for all y except on sets $A(x, d)$ of $\pi(\cdot|x, d, \theta_3)$ measure zero.

(A21) For any $r \in B$, $E\{\partial^2 G/\partial \theta_2 \partial \theta'_2\} \in L(R^{K_2}, B)$ and is a continuous function of θ and r .

(A22) For any $r \in B$,

$$\partial^2 \left[\int_y G(r(y)|y, \theta_2) \pi(dy|x, d, \theta_3) \right] / \partial \theta_3 \partial \theta'_3 \in L(R^{K_3}, B)$$

and is a continuous function of θ and r .

(A23) For any $r \in B, j \in C(y)$, $\partial G_j(r(y)|y, \theta_2)/\partial \theta_2$ exists and is dominated by a $\pi(\cdot|x, d, \theta_3)$ integrable function for all y except on sets $A(x, d)$ of $\pi(\cdot|x, d, \theta_3)$ measure zero.

(A24) For any $r, g \in B$,

$$\int_y \left\{ \sum_{j \in C(y)} g(y, j) \partial G_j(r(y)|y, \theta_2)/\partial \theta_2 \right\} \pi(dy|x, d, \theta_3) \in L(R^{K_2}, B)$$

and is a continuous function of θ and r .

(A25) The parameter space Θ is a compact subset of $R^{(1+K_1+K_2+K_3)}$.

(A26) For all $\theta \in \Theta$ and initial distributions Ω , $|\log \pi(y|x, d, \theta_3)|$ has finite expectation with respect to (y, x, d) .

(A27) For any initial distribution Ω , $E\{\log P(d|x, \theta)\}$ and $E\{\log \pi(y|x, d, \theta_3)\}$ have unique maxima at $\theta = \theta^*$ and $\theta_3 = \theta_3^*$, respectively.

(A28) $\partial^2 \log \pi((y|x, d) \theta_3)/\partial \theta_3 \partial \theta'_3$ is a continuous function of θ for $\mu \times \mu \times \kappa$ -almost all $(y, x, d) \in \Delta \times \Gamma$, where κ is a counting measure on R .

(A29) $|\partial^2 \log \pi(y|x, d, \theta_3)/\partial \theta_3 \partial \theta'_3|$ and $|[\partial \log \pi(y|x, d, \theta_3)/\partial \theta_3][\partial \log \pi(y|x, d, \theta_3)/\partial \theta'_3]|$ have finite expectation in (y, x, d) for all $\theta \in \Theta$ and initial distributions Ω .

(A30) For each $r \in B, \mu$ almost all $x \in \Delta$, and all $i, j, k \in C(x)$, the third partial derivatives $G_{ijk}(r(x)|x, \theta_2)$ exist and are continuous in $r(x)$ and θ_2 . Furthermore, for any $g, h \in B$

$$\int_y \left\{ \sum_{i,j,k \in C(y)} g(y, j) h(y, k) G_{ijk}(r(y)|y, \theta_2) \right\} \pi(dy|x, d, \theta_3) \in B.$$

(A31) For each $r \in B, \mu$ almost all $x \in \Delta$, and all $j, k \in C(x)$, the derivative $\partial G_{jk}(r(x)|x, \theta_2)/\partial \theta_2$ exists and is continuous in $r(x)$ and θ_2 . Furthermore, the function

$$\int_y \left\{ \sum_{j,k \in C(y)} g(y, k) \partial G_{jk}(r(y)|y, \theta_2)/\partial \theta_2 \right\} \pi(dy|x, d, \theta_3)$$

is an element of B , for any $g \in B$.

- (A32) For each $r \in B$, μ almost all $x \in \Delta$ and all $j \in C(x)$, the derivative $\partial^2 G_j(r(x)|x, \theta_2)/\partial \theta_2 \partial \theta'_2$ exists and is continuous in $r(x)$ and θ_2 . Furthermore,

$$\int_y \left\{ \sum_{j \in C(y)} \partial^2 G_j(r(y)|y, \theta_2)/\partial \theta_2 \partial \theta'_2 \right\} \pi(dy|x, d, \theta_3) \in L(R^{\kappa_2^2}, B).$$

- (A33) θ^* is interior to Θ , $A(\theta^*)$ is nonsingular, and θ^* is a regular point of $H(\theta^*)$, where $A(\theta^*)$ is defined by

$$A(\theta^*) = E \left\{ \frac{[\partial \log P(d_1|x_1, \theta^*) \pi(x_1|x_0, d_0, \theta_3^*)]}{\partial \theta} \times \frac{[\partial \log P(d_1|x_1, \theta^*) \pi(x_1|x_0, d_0, \theta_3^*)]}{\partial \theta} \right\}.$$

- (A34) For $\mu \times \kappa$ almost all $(x, d) \in \Gamma$, the support of $\pi(\cdot|x, d, \theta_3)$ does not depend on θ .

- (A35) There exists an integrable function $h(y, x, d)$ which satisfies for $\mu \times \mu \times \kappa$ almost all (y, x, d) (relative to the product measure $\mu \times \mu \times \kappa$ on $\Delta \times \Gamma$)

$$h(y, x, d) \geq \pi(y|x, d, \theta_3), \quad \int_y \int_x \int_d h(y, x, d) \mu(dy) \times \mu(dx) \times \kappa(dd) < +\infty.$$

REFERENCES

- A. ARAUJO AND E. GINÉ (1980), *The Central Limit Theorem for Real and Banach Valued Random Variables*, John Wiley, New York.
- I. V. BASAWA AND B. L. S. PRAKASA RAO (1980), *Statistical Inference for Stochastic Processes*, Academic Press, New York.
- R. BELLMAN (1957), *Dynamic Programming*, Princeton Univ. Press, Princeton, NJ.
- E. BERNDT, B. HALL, R. HALL, AND T. HAUSMAN (1974), *Estimation and inference in nonlinear structural models*, Ann. of Economic and Social Measurement, 3/4, pp. 653–665.
- D. BERTSEKAS (1976), *Dynamic Programming and Stochastic Control*, Academic Press, New York.
- D. BERTSEKAS AND S. SHREVE (1978), *Stochastic Optimal Control: the Discrete Time Case*, Academic Press, New York.
- R. N. BHATTACHARYA AND M. MAJUMDAR (1985), *Dynamic programming for discounted and long-run average rewards*, SSRI Working Paper 8416.
- P. BILLINGSLEY (1961), *Statistical Inference for Markov Processes*, Univ. of Chicago Press, Chicago, IL.
- (1979), *Probability and Measure*, John Wiley, New York.
- D. BLACKWELL (1968), *Discounted dynamic programming*, Ann. Math. Statist., pp. 226–235.
- V. BORKAR AND P. VARAIYA (1982), *Identification and adaptive control of Markov chains*, SIAM J. Control Optim., 10, pp. 470–495.
- D. R. COX (1975), *Partial likelihood*, Biometrika, 62, pp. 269–276.
- A. DALY AND S. ZACHARY (1979), *Improved multiple choice models*, in Identifying and Measuring the Determinants of Mode Choice, D. Henscher and Q. Dalvi, eds., Teakfield, London.
- M. DAS (1987), *Utilization and retirement of cement kilns: an econometric analysis*, Ph.D. thesis, Dept. of Economics, Univ. of Wisconsin, Madison, WI.
- E. DENARDO (1967), *Contraction mapping in the theory underlying dynamic programming*, SIAM Rev., pp. 165–177.
- T. A. DOMENCICH AND D. MCFADDEN (1975), *Urban Travel Demand*, North-Holland, Amsterdam, New York.
- N. DUNFORD AND J. SCHWARTZ (1957), *Linear Operators*, John Wiley, New York.
- I. I. GIHMAN AND A. V. SKOROHOD (1979), *Controlled Stochastic Processes*, Springer-Verlag, Berlin, New York.

- P. GILL, W. MURRAY, AND M. WRIGHT (1981), *Practical Optimization*, Academic Press, New York.
- J. C. GITTENS (1979), *A dynamic allocation index for the discounted multi-armed bandit problem*, *Biometrika*, pp. 580-597.
- G. A. GOTZ AND J. J. MCCALL (1984), *A dynamic retention model for air force officers: theory and estimates*, Report R-3028-AF, The Rand Corporation.
- U. GRENANDER (1981), *Abstract Inference*, John Wiley, New York.
- (1950), *Stochastic processes and statistical inference*, *Ark. Mat.*, 1, pp. 177-195.
- V. GUILLEMAN AND A. POLLOCK (1974), *Differential Topology*, Prentice-Hall, Englewood Cliffs, NJ.
- T. S. FERGUSON (1958), *A method for generating best asymptotically normal estimates with application to the estimation of bacterial densities*, *Ann. Math. Statist.*, 29, pp. 1046-1062.
- C. FUTIA (1982), *Invariant distributions and the limiting behavior of Markovian economic models*, *Econometrica*, 50, pp. 1029-1054.
- L. HANSEN (1982), *Large sample properties of generalized method of moment estimators*, *Econometrica*, 50, pp. 1029-1054.
- L. HANSEN AND K. SINGLETON (1982), *Generalized instrumental variables estimation of nonlinear rational expectations models*, *Econometrica*, 50, pp. 1269-1286.
- J. HAUSMAN (1978), *Specification tests in econometrics*, *Econometrica*, 46, pp. 1251-1272.
- J. HECKMAN (1981), *Statistical models for discrete panel data*, in *Structural Analysis of Discrete Data*, C. Manski and D. McFadden, eds., M.I.T. Press, Cambridge, MA.
- (1981), *The incidental parameters problem and the problem of initial conditions in estimating discrete-time, discrete-data stochastic processes*, in *Structural Analysis of Discrete Data*, C. Manski and D. McFadden eds., M.I.T. Press, Cambridge, MA.
- R. HOWARD (1971), *Dynamic Probabilistic Systems Volume I: Markov Models*, John Wiley, New York.
- P. HUBER (1967), *The behavior of maximum likelihood estimates under nonstandard conditions*, in *Proc. Fifth Berkeley Symposium in Mathematical Statistics and Probability*, Univ. of California Press, Berkeley, CA.
- L. KANTOROVICH AND G. AKILOV (1982), *Functional Analysis*, Pergamon Press, Elmsford, NY.
- L. LECAM (1953), *On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates*, *Univ. of California Publications in Statistics*, 1, pp. 277-330.
- S. LIPPMAN (1975), *On dynamic programming with unbounded rewards*, *Management Sci.*, pp. 1225-1233.
- R. E. LUCAS (1976), *Econometric policy evaluation: a critique*, in *The Phillips Curve and Labor Markets*, Carnegie-Rochester Conference on Public Policy 1, K. Brunner and A. K. Meltzer, eds., North-Holland, Amsterdam-New York.
- C. MANSKI AND D. MCFADDEN (1981), *Structural Analysis of Discrete Data with Econometric Applications*, M.I.T. Press, Cambridge, MA.
- J. MARSCHAK (1953), *Economic Measurements for Policy and Prediction*, in *Studies in Econometric Method*, W. C. Hood and T. C. Koopmans, eds., John Wiley, New York.
- D. MCFADDEN (1973), *Conditional logit analysis of qualitative choice behavior*, in *Frontiers of Econometrics*, P. Zarembka, ed., Academic Press, New York.
- (1981), *Econometric models of probabilistic choice*, in *Structural Analysis of Discrete Data*, C. Manski and D. McFadden, eds., M.I.T. Press, Cambridge, MA.
- R. MILLER (1984), *Job matching and occupational choice*, *J. Political Economy*, 92, pp. 1086-1120.
- M. MONTGOMERY (1987), *Lifetime fertility as a controlled stochastic process: an application of Rust's estimation method*, Office of Population Research, Princeton Univ., Princeton, NJ.
- J. ORTEGA AND W. RHEINOLDT (1970), *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York.
- A. PAKES (1985), *Patents as options: some estimates of the value of holding european patent stocks*, *Econometrica*, 54, pp. 755-784.
- J. PRATT (1960), *On interchanging limits and integrals*, *Ann. of Math. Statist.*, pp. 74-77.
- J. RUST (1987a), *Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher*, *Econometrica*, 55, pp. 999-1033.
- (1987b), *The asymptotic distribution of randomly parameterized elements of separable Banach spaces*, SSRI Working Paper #8634, Univ. of Wisconsin, Madison, WI.
- (1987c), *Nested fixed point optimization algorithm documentation manual*, software for IBM-PC, Aptech Systems Inc., Seattle, WA.
- (1988), *A dynamic programming model of retirement behavior*, in *The Economics of Aging*, D. Wise, ed., National Bureau of Economic Research, Univ. of Chicago Press, Chicago, IL.
- T. J. SARGENT (1981), *Interpreting economic time series*, *J. Political Economy*, 89, pp. 213-248.
- R. L. TAYLOR (1978), *Stochastic Convergence of Weighted Sums in Linear Spaces*, Springer-Verlag, Berlin, New York.

- N. N. VAKHANIA (1981), *Probability Distributions on Linear Spaces*, North-Holland, Amsterdam, New York.
- N. M. VAN DIJK (1984), *Controlled Markov Processes: Time Discretization*, CWI Tract 11, Mathematische Centrum, Amsterdam.
- W. WERNER (1984), *Newton-like methods for the computation of fixed points*, Comput. Math. Appl., 10, pp. 77-86.
- R. M. WHEELER AND K. S. NARENDRA (1986), *Decentralized learning in finite Markov chains*, IEEE Trans. Automat. Control., 31, pp. 519-526.
- H. WHITE (1982), *Maximum likelihood estimation of misspecified models*, Econometrica, 50, pp. 1-26.
- P. WHITTLE (1982), *Optimization Over Time: Dynamic Programming and Stochastic Control*, John Wiley, New York.
- H. WILLIAMS (1977), *On the formulation of travel demand models and economic evaluation measures of user benefit*, Environment Planning, A-9, pp. 285-344.
- K. WOLPIN (1984), *An estimable dynamic stochastic model of fertility and child mortality*, J. Political Economy, 92, pp. 1086-1120.

EXISTENCE OF AN OPTIMAL MARKOVIAN FILTER FOR THE CONTROL UNDER PARTIAL OBSERVATIONS*

NICOLE EL KAROUI,[†] DU' HÙ' NGUYEN[‡], AND MONIQUE JEANBLANC-PICQUÉ§

Abstract. This paper concerns the control of diffusions under partial observations. Part I studies the control of the signal process $dX_t = b(t, X_t, U_t) dt + \sigma(t, X_t, U_t) dB_t$ when the observation is $dY_t = h(t, X_t) dt + dW_t$ and when the objective is to maximize a reward function $E\{\int_r^T k(s, X_s, U_s) ds + g(X_T)\}$. The existence of an optimal relaxed control is proved.

Part II studies the separated problem and proves the existence of an optimal Markovian filter. Then, the authors compare the two problems and prove, under mild conditions, that the value functions for the two problems are equal.

Key words. control theory, partially observable diffusions, optimal filter

AMS(MOS) subject classifications. 60G, 60H, 93E

Introduction. In this paper, we are concerned with optimal control problems of the following case. Let X_t denote the signal process which we wish to control and Y_t the observation process. The signal and the observation processes are governed by stochastic differential equations:

$$\begin{aligned} dX_t &= b(t, X_t, U_t) dt + \sigma(t, X_t, U_t) dB_t \quad \text{for } t \geq r, \\ X_t &= X_r \quad \text{for } t \leq r \end{aligned}$$

where X_r has a given distribution μ , and

$$\begin{aligned} dY_t &= h(t, X_t) dt + dW_t \quad \text{for } t \geq r \\ Y_r &= 0 \quad \text{for } t \leq r \end{aligned}$$

where B and W are independent Brownian motions. The problem is to maximize a criterion of the form

$$J(r, \mu) := E \left\{ \int_r^T k(s, X_s, U_s) ds + g(X_T) \right\}$$

where T is a terminal time ($T < \infty$). The control U is A -valued, where A is a compact metric space.

In the first part of this paper, we introduce a wider class of controls. In a customary version of stochastic control under partial observations, the control U_t is a process which is measurable with respect to the σ -field generated by the observation process $(Y_s; s \leq t) = \mathcal{F}_t^Y$.

Here we use controls which are measurable with respect to a σ -field which is larger than $\sigma(Y_s, s \leq t)$ but smaller than the filtration on (Ω, \mathcal{F}, P) . (All the processes X, Y, U are defined on the same space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$.) This idea was used for the first time by Fleming and Pardoux [FP]. Moreover, we work with relaxed controls, i.e., controls (q_t) which are measures. Many authors have used such controls (see Fleming

* Received by the editors June 27, 1986; accepted for publication (in revised form) November 4, 1987.

[†] Ecole Normale Supérieure de Fontenay, 31 avenue Lombart, 92260 Fontenay-aux-Roses, France.

[‡] Khoa Toán, Đại Học Tổng Hợp, Hanoi, Vietnam. This author's research was carried out during a stay at Université Paris-Sud, Orsay.

[§] Département de Mathématiques, Bureau H 43, Ecole Normale Supérieure de Cachan, 61 avenue du président Wilson, 94230, Cachan, France.

and Nisio [FN]). We prove the compactness of the set of laws of controlled processes (X_t, q_t, Y_t) which are called rules. We establish the existence of an optimal control and we study the properties of the value function with respect to the initial value (r, μ) . We end Part I with a comparison of the original problem and the relaxed one: we prove that the two value functions are equal.

In Part II we study the “separated problem.” It is well known that the filter r_t , i.e., the conditional distribution for X_t given \underline{F}_t^Y satisfies the so-called Kushner–Stratanovich equation:

$$(KS) \quad d\pi_t(f) = \pi_t(L^u f) dt + (\pi_t(fh) - \pi_t(f)\pi_t(h))(dY_t - \pi_t(h) dt).$$

We study the control of the solutions to the (KS) equation, with an objective similar to the one in the partially observable case (Part I)

$$J(r, \mu) := E \left\{ \int_r^T \pi_s(K) ds + \pi_T(G) \right\},$$

under a supplementary condition: we retain only the solutions to (KS) which are filters. We transport this problem on a canonical space and we prove the existence of an optimal relaxed control. We establish the stability of the set of controls under conditioning and concatenation and prove the existence of an optimal Markovian filter. Hence we obtain the equations of dynamical programming.

Finally we prove, under a hypothesis of uniqueness of the solutions to the (KS) equation, the equality of the value functions for the initial problem and the separated one. These results generalize those of Fleming and Pardoux, Fleming and Nisio, and Borkar.

Let us now give some details.

Part I. We define exactly the partially observable control problem in § 1 and state the precise assumptions. Zakai’s transformation allows us to consider the observation process as a Brownian motion. We give a martingale formulation to the admissible control problem with a careful study of the different filtrations.

We introduce the relaxed controls in § 2. They are controls which take values in a “generalized action” space, i.e., the space of Radon measures on $\mathbb{R}^+ \times A$ whose projection on \mathbb{R}^+ is the Lebesgue measure. We transport the relaxed control problem on the canonical space (§ 3), and we define controls as probability measures on the canonical space which satisfy a martingale problem and under which the observation process is a Brownian motion. These controls are called rules. We prove (§ 4) the compactness of the set $\mathcal{R}(r, \mu)$ of rules with initial conditions (r, μ) and show that the graph Γ of the set-valued application $(r, \mu) \mapsto \mathcal{R}(r, \mu)$ is closed and contained in a product of compact spaces if (r, μ) belongs to a compact set of $\mathbb{R}^+ \times \mathcal{P}(\mathbb{R}^d)$.

We study the measurability and continuity properties of the value function, under the assumption that the reward functions are continuous. Then we compare the relaxed control problem and the nonrelaxed one, and prove the equality of the two value functions.

Part II. We study the separated problem as a control problem for the solutions to the Kushner–Stratanovich equation (§ 1). We only consider the solutions which are filters for a partially observable problem, and we modelize this restriction as a constraint. We transport the problem on the canonical space and define separated rules. In § 2, we establish the compactness of the separated rules, then we apply the ideas of El Karoui [EK]: we establish the stability of the rules by conditioning and concatenation (§ 3) in order to set the equations of dynamic programming (§ 4) and study the

properties of the “Nisio semigroup.” We apply Krylov’s ideas in § 5 and find an optimal control which is such that the filter is a strong Markovian process. Under a hypothesis on uniqueness of the solutions to the (KS) equation, we are able to prove that the two problems, the original one and the separated one, have the same value function. Under this uniqueness hypothesis, any solution to (KS) is a partially observable (P.O.) filter. The separated problem which we have studied becomes a problem of control of (KS)’s solutions without constraint. If we assume also that h belongs to C_b^2 , we prove that our separated problem is equivalent to the problem of Fleming [F13]. Fleming controls a process π_t such that

$$\pi_t(f) - \pi_r(f) - \int_r^t \pi_s \cdot q_s(Lf)(s) ds$$

is a martingale which admits

$$\int_r^t |\text{cov } \pi_s(f, h)|^2 ds$$

as an increasing process.

PART I

Controls and Rules for the Control Problem Under Partial Observations

1. Partially observable controls.

1.1. Statement of the problem. The control for partially observed diffusions with the following model is usually studied: let X_t be the process which we wish to control (X_t is called the state process or the signal process). We denote by Y_t the observation process and by U_t the control process, where $t \in \mathbb{R}^+$. The state and observation processes are governed by stochastic differential equations:

$$(1.1.1) \quad \begin{aligned} (a) \quad & dX_t = b(t, X_t, U_t) dt + \sigma(t, X_t, U_t) dB_t \\ (b) \quad & dY_t = h(t, X_t) dt + dW_t \\ (c) \quad & \begin{aligned} X_t &= Z \\ Y_t &= 0 \end{aligned} \quad \text{for } t \leq r, \end{aligned}$$

where Z is an \mathbb{R}^d -random variable, independent of (W_t) and of $\sigma(B_{t+r} - B_r, t \geq 0)$, and where B and W are independent standard Wiener processes of dimension m (respectively, k), i.e., (B, W) is a $(m+k)$ -dimensional Wiener process. We denote by \underline{F}_t^Y the σ -algebra $\sigma(Y_s, s \leq t)$. The control process (U_t) is required to be \underline{F}_t^Y -adapted, since the information available to the controller at time t is described by the observations $(Y_s, s \leq t)$. The process U_t is A -valued, where A is a given compact metric space, which is called the “control space.” All these processes are defined on some probability space $(\Omega, \underline{G}_t, Q)$, where \underline{G}_t is an increasing family of σ -algebra. The case where the two Brownian motions are correlated is studied in [EK2].

(1.1.2) The term $\mathcal{U}_s^0 = (\Omega, \underline{G}_t, X_t, Y_t, U_t, B_t, r, Z, Q)$ is called a *strictly admissible P.O. (partially observable) control*.

The problem is to find a strictly admissible P.O. control which *maximizes a criterion* in the form

$$(1.1.3) \quad j(r, \mu; \mathcal{U}_s^0) := E \left\{ \int_r^T k(s, X_s, U_s) ds + g(X_T) \right\}$$

where k and g are the instantaneous and terminal rewards and where μ is the law of Z .

The superscript zero obliges us to remember that we are working with controls U_t which are processes instead of relaxed controls (see § 2).

Under these conditions, the strictly admissible control problem is close to the completely observable strong control problem, where the measurability conditions on control processes are fixed: the modelization space cannot be used by the controller as a parameter. As in the completely observable strong problem, the problem of the existence of an optimal strictly admissible P.O. control has no solution in a general way.

We introduce here a larger class of controls: the relaxed controls and the control rules. We will give a precise formulation of these words in § 2 and 3.

First, we set up some specific notation.

1.2. Conventions. (a) Let $(\Omega, \underline{F}, Q)$ be a probability space and let X, Y be two positive random variables. We denote by $Q(X)$ and $\text{cov } Q(X, Y)$ the expectation and the covariance of X and (X, Y) under Q , i.e.,

$$(1.2.1) \quad \begin{aligned} Q(X) &:= QX := \int X(\omega) dQ(\omega), \\ \text{cov } Q(X, Y) &:= Q(XY) - Q(X)Q(Y). \end{aligned}$$

(b) Let (E, \underline{E}) and (F, \underline{F}) be two measurable spaces and let Z be a positive random variable defined on $(E \times F; \underline{E} \otimes \underline{F})$. If Q is a probability kernel on E into \underline{F} , we set

$$(1.2.2) \quad Q_x(Z) := Q(Z)(x) = \int Z(x, y) Q(x, dy).$$

If P is a probability measure on E , we denote by $P \cdot Q$ the unique probability measure on $E \times F$ such that

$$(1.2.3) \quad \begin{aligned} P \cdot Q(Z) &:= P\{Q(Z)(\cdot)\} := \int Q(Z)(x) dP(x) \\ &:= \int \left\{ \int Z(x, y) Q(x, dy) \right\} dP(x) := P(Q, (Z)). \end{aligned}$$

In particular, if Q does not depend on x , $Q(Z)(x) = Q(Z(x, \cdot))$ and $P \cdot Q$ is the product measure $P \otimes Q$.

1.3. Hypotheses.

(1.3.1) Let A be a compact metric space, called the *action space*.

(1.3.2) The functions $\{b_j, \sigma_{i,j}; j = 1, \dots, d, i = 1, \dots, m\}$ are measurable bounded functions. They are defined on $\mathbb{R}^+ \times \mathbb{R}^d \times A$. Moreover, they are uniformly continuous in the pair (x, a) . We denote by b (respectively, σ) the vector with general term b_j (respectively, the matrix with general term $\sigma_{i,j}$).

(1.3.3) Let h be a bounded measurable function from $\mathbb{R}^+ \times \mathbb{R}^d$ into \mathbb{R}^k which is uniformly continuous in x .

(1.3.4) The elliptic operator L associated with the signal process $(X_t)_{t \leq T}$ is defined on the functions f of $C_b^{1,2}(\mathbb{R}^+ \times \mathbb{R}^d)$:

$$Lf(t, x, a) := \frac{1}{2} \sum_{i,j} a_{i,j} D_{i,j} f(t, x, a) + \sum_j b_j D_j f(t, x, a) + D_t f(t, x).$$

In this formula, $D_{i,j}$ (respectively, D_j, D_t) denotes the operator of derivation with respect to $x_i x_j$ (respectively, x_j, t); $a_{i,j} D_{i,j} f(t, x, a)$ is equal to $a_{i,j}(t, x, a) D_{i,j} f(t, x)$, where $a_{i,j}(t, x, a)$ is the generic term of the $d \times d$ symmetric positive matrix $(\sigma \sigma')(t, x, a)$.

(1.3.5) Let m be a probability measure on A . We denote by $\sigma(t, x, m)$ a square root of the matrix of generic term $m(a_{i,j})(t, x) = \int a_{i,j}(t, x, a) m(da)$ (notation (1.2.2)). In general, $\sigma(t, x, m)$ differs from $m(\sigma)(t, x)$, except if m is a Dirac measure.

(1.3.6) The functions k and g which define the objective are defined on $\mathbb{R}^+ \times \mathbb{R}^d \times A$ (respectively, \mathbb{R}^d). We assume that they are bounded measurable functions.

1.4. Zakai's transformation. The problem of existence of strictly admissible controls associated with nonconstant policy U_t is not easy. We must know the observation in order to determine the adapted control U_t which is used to control the signal X_t , which gives the observation Y_t, \dots , and so on.

All these difficulties disappear when the observation Y_t is only a white noise W_t (i.e., when $h \equiv 0$). In this case, the Brownian B_t related to the signal is independent of the observation; therefore B_t is independent of the control U_t .

DEFINITION 1.4.1. A strictly admissible P.O. control under $h = 0$ is called a W-N (White Noise) strictly admissible control.

It is well known that, with a change of probability, the general case can be reduced to a W-N strictly admissible case; this transformation is due to Zakai.

THEOREM 1.4.2. Let $\mathcal{Q}_{s,h}^0 = (\Omega, \underline{G}_t, X_t, Y_t, U_t, B_t, r, Z, Q)$ be a strictly admissible P.O. control. There exists a probability measure P which is equivalent to Q , defined by

$$P := \exp \left\{ - \int_r^T h'(s, X_s) dY_s + \frac{1}{2} \int_r^T |h(s, X_s)|^2 ds \right\} \cdot Q$$

such that $(\Omega, \underline{G}_t, X_t, Y_t, U_t, B_t, r, Z, P)$ is a W-N strictly admissible control. Conversely, let $\mathcal{Q}_s^0 = (\Omega, \underline{G}_t, X_t, Y_t, U_t, B_t, r, Z, P)$ be a W-N strictly admissible P.O. control, and define

$$(1.4.2.1) \quad L_t := \exp \int_r^t h'(s, X_s) dY_s - \frac{1}{2} \int_r^t |h(s, X_s)|^2 ds.$$

There exists a unique probability measure Q on $(\Omega, \underline{G}_t)$ such that $dQ = L_t dP$ on \underline{G}_t ; moreover, $\mathcal{Q}_{s,h}^0 = (\Omega, \underline{G}_t, X_t, Y_t, U_t, B_t, r, Z, Q)$ is a strictly admissible P.O. control.

Remarks. Here, the prime denotes the transpose of the vector; $(Y_{t+r} - Y_r)_{t \geq 0}$ is a P -Brownian motion.

Proof. This is exactly Girsanov's theorem [SV, Thm. 6.4.1]. The boundedness hypotheses on the coefficients ensure that L_t is an uniformly integrable martingale such that

(1.4.2.2) $\sup_{t \leq T} P(L_t^2) \leq K$ (convention (1.2.1)), where K is a constant which depends only on $\|h\|_\infty$ and T . \square

Let us recall that, if M_t is a bounded Q -martingale, and if $\langle\langle M, Y \rangle\rangle$ denotes the bilinear form associated with the Meyer process [DM2, VII.39] under the probability Q , $M_t - \int_0^t h'(s, X_s) d\langle\langle M, Y \rangle\rangle_s$ is a P -local martingale.

We now establish the properties of the Brownian motion B_t , in order to give another formulation of a W-N strictly admissible control which allows us to set a “good” definition of an admissible control.

PROPOSITION 1.4.3. *Let $(\Omega, \underline{G}_t, P)$ be a probability space and let Y_t, B_t be two Brownian motions. The following conditions are equivalent:*

- (a) *The pair (Y_t, B_t) is a Brownian motion.*
- (b) *B_t is a $\underline{F}_\infty^Y \vee \underline{G}_t$ Brownian motion.*

Proof. In order to prove $a \Rightarrow b$, we must establish that, for $s < t$,

$$P(1_{A_\infty} 1_{G_s} \exp(\theta' B_t - \frac{1}{2} |\theta|^2 t)) = P(1_{A_\infty} 1_{G_s} \exp(\theta' B_s - \frac{1}{2} |\theta|^2 s))$$

where $A_\infty \in \underline{F}_\infty^Y$ and $G_s \in \underline{G}_s$.

The expression $\exp(\theta' B_t - \frac{1}{2} |\theta|^2 t)$ is a martingale which is one plus a stochastic integral with respect to the Brownian motion B_t . We denote this stochastic integral by $\int_0^t H_u dB_u$. We can describe the elements of \underline{F}_∞^Y with stochastic integrals of the form $c + \int_0^v K_u dY_u$. Hence we study the term $P\{(c + \int_0^v K_u dY_u) 1_{G_s} (\int_0^t H_u dB_u)\}$. It is then easy to verify that, if (a) is satisfied, the two terms

$$P\left(1_{G_s} \int_s^t H_u dB_u\right) \quad \text{and} \quad P\left(1_{G_s} \int_s^t H_u dB_u \int_0^v K_u dY_u\right)$$

are equal to zero, and result (b) follows.

Conversely we prove that, if (b) is satisfied, then B_t is independent of \underline{F}_∞^Y . It suffices to show that, for each step function h defined by $h = \sum_i \alpha_i 1_{]t_i, t_{i+1}]}$, where $\alpha_i \in \mathbb{R}^k$,

$$P[\exp(-\sum \alpha_i (B_{t_{i+1}} - B_{t_i})) | \underline{F}_\infty^Y] = \exp[\frac{1}{2} \sum \alpha_i^2 (t_{i+1} - t_i)].$$

Since $\exp(-\int_0^t h'_s dB_s - \frac{1}{2} \int_0^t |h_s|^2 ds)$ is a $\underline{F}_\infty^Y \vee \underline{G}_t$ martingale, it follows that, for each $A_\infty \in \underline{F}_\infty^Y$,

$$P\left(1_{A_\infty} \exp\left(-\int_0^t h'_s dB_s - \frac{1}{2} \int_0^t |h_s|^2 ds\right)\right) = Q(A_\infty) = P(A_\infty).$$

It remains to note that $\int_0^\infty |h_s|^2 ds = \sum \alpha_i^2 (t_{i+1} - t_i)$ and the result (a) follows. \square

1.5. Admissible controls. Let us weaken the measurability conditions on the control U_t . We require the measurability of U_t with respect to a σ -algebra which is larger than \underline{F}_t^Y and smaller than \underline{G}_t . We follow here the ideas of Fleming and Pardoux [FP] which are used again by many authors, such as Bismut [Bi], Borkar [Bo], Fleming and Nisio [FN] and Haussmann [Ha].

DEFINITION 1.5.1. A W-N admissible control is a term $\mathcal{U}^0(r, \mu)$ defined by $\mathcal{U}^0(r, \mu) := (\Omega, \underline{F}_t, \underline{G}_t, X_t, Y_t, U_t, B_t, r, Z, P)$ such that we have the following:

- (a) \underline{F}_t and \underline{G}_t are two right-continuous filtrations such that $\underline{G}_t \supset \underline{F}_t$;
- (b) $(Y_{t+r} - Y_r)$ and $(B_{t+r} - B_r)$ are two Brownian motions with respect to the filtration \underline{G}_{t+r} ;
- (c) U_t and Y_t are \underline{F}_t -adapted, and $Y_t = 0$ for $t \leq r$;
- (d) $(B_{t+r} - B_r)$ is an $\underline{F}_\infty \vee \underline{G}_{t+r}$ Brownian motion;
- (e) Z is a random variable which is independent of \underline{F}_∞ and of $(B_{t+r} - B_r)$; its distribution is μ ;

(f) X_t is a process with continuous paths, which is the solution of the stochastic differential equation (S.D.E.):

$$dX_t = b(t, X_t, U_t) dt + \sigma(t, X_t, U_t) dB_t \quad \text{for } t \geq r,$$

$$X_t = Z \quad \text{for } t \leq r,$$

$$X_t \text{ is } \underline{G}_t\text{-adapted.}$$

We denote by $\mathcal{W}^0(r, \mu)$ the set of W-N admissible controls with initial conditions (r, μ) .

We remark that if (d) is satisfied, then $(B_{t+r} - B_r)$ is a \underline{G}_{t+r} -Brownian motion, since it is \underline{G}_{t+r} adapted.

DEFINITION 1.5.2. A P.O. *admissible control* is a term $\mathcal{U}_h^0(r, \mu)$ which is deduced from a W-N admissible control $\mathcal{U}^0(r, \mu)$ with the change of probability given by $P^h = L_T \cdot P$, where $L_T := \exp \int_r^T h'(s, X_s) dY_s - \frac{1}{2} \int_r^T |h(s, X_s)|^2 ds$. We denote by $j(r, \mu; \mathcal{U}_h^0)$ the reward associated with the control $\mathcal{U}_h^0(r, \mu)$, i.e., $j(r, \mu; \mathcal{U}_h^0) := P^h \{ \int_r^T k(s, X_s, U_s) ds + g(X_T) \}$. We denote by $\mathcal{A}_h^0(r, \mu)$ the set of P.O. admissible controls.

Remark. Under the probability P^h , the process $Y_{t+r} - Y_r - \int_r^{t+r} h(s, X_s) ds$ is a Brownian motion, and $B_{t+r} - B_r$ remains a Brownian motion.

Remark 1.5.3. (a) From the definition (properties d and f) it follows that for each $f \in C_b^{1,2}(\mathbb{R}^+ \times \mathbb{R}^d)$

$$(1.5.3.1) \quad C_t(f) := f(t, X_t) - f(r, X_r) - \int_r^t Lf(s, X_s, U_s) ds \quad (t \geq r)$$

is an $\underline{F}_\infty \vee \underline{G}_t$ martingale and is \underline{G}_t -adapted and $X_t = X_r$ for $t < r$ where X_r is independent of \underline{F}_∞ and has a distribution equal to μ .

(b) Conversely, it is well known that the formulation with a martingale problem is equivalent to an S.D.E., up to a change of probability space [IW].

2. Relaxed partially observable controls. In the preceding section, we have enlarged the set of controls by weakening the measurability conditions.

As in the deterministic case, or as in the case of degenerate diffusions [EHJ] we give a convex formulation of the problem by introducing “randomized” controls which take values in the space of generalized actions.

DEFINITION 2.1. *The space of generalized actions.* The space of generalized actions V is the space of Radon measures on $\mathbb{R}^+ \times A$ whose projection on \mathbb{R}^+ is the Lebesgue measure. We endow V with the topology of vague convergence, checked on the functions which are continuous on (t, a) with a compact support. With this topology, V is a metric compact set. This topology is the same as the stable topology, checked on the functions which are bounded measurable, with a compact support and which are continuous in a . We refer to the paper of Jacod and Memin [JM1] for more details about stable topology.

We will often use the disintegration of an element q of V in the following form:

$$(2.1.1) \quad q(ds, da) = ds q(s, da), \text{ where } q(s, \cdot) \text{ is a measurable kernel with mass equal to 1.}$$

The set of measurable functions v from \mathbb{R}^+ into A is embedded in V in a natural way by the formula $q(ds, da) = ds \delta_{v(s)}(da)$, where δ_z is the Dirac measure at z . We denote by V^0 the set of these atomic measures.

The space V is endowed with its natural σ -algebra \underline{V} , which is the smallest σ -algebra such that the applications $q \rightarrow q(f)$ are measurable for any f bounded, continuous in a and with a compact support. We give to V the natural filtration \underline{V}_t generated by $1_{[0,t]} \cdot q$ for $q \in V$. The following theorem is proved in [GH] or [F12] under the name of the “chattering lemma.” The proof can be found in Appendix B.

THEOREM 2.2. (a) *The space V is a metric compact space.*

(b) *The set V^0 of atomic measures is dense in V . More precisely, there exists a sequence ψ_k of measurable maps from V into V^0 , adapted (i.e., $\psi_k^{-1}(\underline{V}_t) \subset \underline{V}_t$) such that $\psi_k(q)$ converges to q in V for any $q \in V$. Moreover, we can choose $\psi_k(q)$ such that $\psi_k(q)$ are step (with respect to the time) measures.*

By analogy with the preceding section, we define a W-N relaxed control as follows.

DEFINITION 2.3. A W-N relaxed control is a term $\mathcal{U}(r, \mu)$ with the form $\mathcal{U}(r, \mu) := (\Omega, \underline{F}_t, \underline{G}_t, X_t, Y_t, q, r, \mu, P)$ such that we have the following:

- (i) \underline{F}_t and \underline{G}_t are two right-continuous filtrations and $\underline{F}_t \subset \underline{G}_t$;
- (ii) q is a V -valued random variable. We suppose that q is \underline{F}_t -adapted, i.e., $q^{-1}(\underline{V}_t) \subset \underline{F}_t$;
- (iii) $Y_{t+r} - Y_r$ is a \underline{G}_{t+r} -Brownian motion which is \underline{F}_{t+r} -adapted, $Y_t = 0$ for $t \leq r$, X_t is a continuous \underline{G}_t -adapted process;
- (iv) $C_t(f, q) := f(t, X_t) - f(r, X_r) - \int_r^t q_s(Lf)(s, X_s) ds$ is a $\underline{F}_\infty \vee \underline{G}_t$ -martingale if $t \geq r$ and is \underline{G}_t -adapted;
- (v) $X_t = X_r$ for $t \leq r$, where X_r is a random variable which is independent of \underline{F}_∞ .

The distribution of X_r is equal to μ .

The set of W-N relaxed controls is denoted by $\mathcal{W}(r, \mu)$.

Remarks. Remember that $q_s(Lf)(s, x) = \int_A Lf(s, x, a)q(s, da)$.

From Remark 1.5.3 we deduce that each W-N admissible control is a W-N relaxed control.

Conversely, if \mathcal{U} is a W-N relaxed control such that $q \in V^0$ P -almost surely we can associate with \mathcal{U} a W-N admissible control.

DEFINITION 2.4. A P.O. relaxed control is a term $\mathcal{U}_h(r, \mu)$ which is deduced from a W-N relaxed control $\mathcal{U}(r, \mu)$ with the change of probability $P^h = L_T P$, where L_T is defined in Theorem 1.4.2.

The reward associated with $\mathcal{U}_h(r, \mu)$ is equal to

$$j(r, \mu; \mathcal{U}_h) := P^h \left(\int_r^T q_s(k)(s, X_s) ds + g(X_T) \right).$$

The set of P.O. relaxed controls is denoted by $\mathcal{A}_h(r, \mu)$.

In fact, it suffices to work with the canonical space of $((y_t), q, (x_t))$ in order to define the control problem. Thus, we now formalize the statement of the problem with the canonical spaces.

3. Partially observable control rules.

3.1. Notation. Let $(\mathcal{X}, \mathcal{X}_t)$ be the canonical space of continuous functions from \mathbb{R}^+ into \mathbb{R}^d , endowed with its right continuous filtration. A point of \mathcal{X} is denoted by $[x]$ or (x_t) . For each function f in $C_b^{1,2}(\mathbb{R}^+ \times \mathbb{R}^d)$, we define the function C_t on $\mathcal{X} \times V$ by the formula

$$(3.1.1) \quad C_t(f, [x], q) := \begin{cases} f(t, x_t) - f(r, x_r) - \int_r^t q(Lf)(s, x_s) ds & \text{if } t \geq r, \\ 0 & \text{if } t \leq r. \end{cases}$$

$C_t(f, \cdot, \cdot)$ is an $\mathcal{X}_t \otimes \underline{V}_t$ adapted function which is bounded on $[0, T]$ and continuous in $([x], q)$.

3.1.2. The function $\gamma([x], q) := \int_r^T q(k)(s, x_s) ds + g(x_T)$ is a bounded Borelian function.

3.1.3. In the same way, we denote by $(\mathcal{Y}, \underline{\mathcal{Y}}_t)$ the canonical space of continuous functions from \mathbb{R}^+ into \mathbb{R}^k which are equal to zero at 0. $\underline{\mathcal{Y}}_t$ is the canonical right continuous filtration. A point of \mathcal{Y} is denoted by (y_t) or $[y]$. We denote by (\bar{V}, \bar{V}_t) the space $(\mathcal{Y} \times V, \underline{\mathcal{Y}}_t \otimes \underline{V}_t)$. We denote by $(\mathcal{X}, \underline{\mathcal{X}}_t)$ the space $\mathcal{Y} \times V \times \mathcal{X}$, with its filtration $\underline{\mathcal{Y}}_t \otimes \underline{V}_t \otimes \underline{\mathcal{X}}_t$. The σ -algebra $(\bar{V}_t \otimes \{\emptyset, \mathcal{X}\})$ is still denoted by \bar{V}_t . A W-N control rule is a probability on \mathcal{X} which defines a W-N relaxed control.

3.2. Hypothesis. Throughout this work, we will assume the following hypothesis. For each $q \in V$, for each initial condition (r, μ) , there exists one and only one probability measure $S_{(r, \mu)}(q)$ on $(\mathcal{X}, \underline{\mathcal{X}}_t)$ such that we have the following:

$$(3.2.1) \quad \begin{aligned} (i) \quad & C_t(f, [x], q) \text{ is a } \underline{\mathcal{X}}_{t \vee r}\text{-martingale under } S_{(r, \mu)}(q), \\ (ii) \quad & S_{(r, \mu)}(q)(x_r \in A) = \mu(A). \end{aligned}$$

It is well known that the inhomogeneous Markov property holds:

$$(3.2.2) \quad S_{(r, \mu)}(q)(\cdot | \underline{\mathcal{X}}_t) = S_{(t, x_t)}(q)(\cdot) \quad S_{(r, \mu)}(q) \text{ almost surely on the } \sigma\text{-algebra } \sigma(x_s, s \geq t).$$

3.3. Remarks. (a) Hypothesis 3.2 is equivalent to the following: for each constant control q , the S.D.E. $dX_t = q(b)(t, X_t) dt + \sigma(t, X_t, q) dW_t$, where $\sigma(t, x, q)$ is defined as in (1.3.5) has a unique weak solution. It is satisfied if the coefficients b and σ are uniformly Lipschitzian with respect to x , or if the matrix $(a_{i,j}(t, x, q))$ is uniformly elliptic and uniformly continuous with respect to x (see [SV]).

(b) Under the uniqueness hypothesis, the continuity conditions which we assume on the coefficients of the operator L imply that the map $(r, \mu, q) \rightarrow S_{(r, \mu)}(q)$ is continuous. Indeed, it suffices to show that if (r_n, μ_n, q_n) converges to (r, μ, q) , then $S_n = S_{(r_n, \mu_n)}(q_n)$ converges to a probability measure S which is the unique solution to the martingale problem. Since (S_n) is tight, it suffices to show that any cluster point is a solution to the martingale problem. The only difficulty is to verify that if $q_n \rightarrow q$ and $S_n \rightarrow S$, then for each function ψ which is measurable, continuous in (x, a) , and uniformly continuous in x with respect to a , $S_n \int \psi(t, x_t, a) q(t, da) dt$ converges to $S \int \psi(t, x_t, a) q(t, da) dt$. A classical method is to use Skhorochood's representation theorem and the uniform continuity of ψ . Another method is to consider the probability measures $S_n \otimes \delta_{q_n}$ on the space $C(\mathbb{R}^d) \times V$. It is then obvious that $S_n \otimes \delta_{q_n}$ converges to $S \otimes \delta_q$ and the result follows.

(c) In fact, the uniqueness hypothesis is not necessary. If it is not satisfied, we must work with the set $\mathcal{S}(r, \mu, q)$ of the probability measures which verify (3.2.1). Condition (3.2.2) is then modified into: if $S \in \mathcal{S}(r, \mu, q)$, then $S_{(r, \mu)}(\cdot | \underline{\mathcal{X}}_t) \in \mathcal{S}(t, x_t, q)$ if $t \geq r$. It is then possible to show that the set-valued map $(r, \mu, q) \rightrightarrows \mathcal{S}(r, \mu, q)$ is upper semicontinuous (u.s.c.).

The approach we have discussed in the preceding sections allows us to state the following definition.

DEFINITION 3.4. A W-N control rule with initial conditions (r, μ) is a probability measure R on $(\mathcal{X}, \underline{\mathcal{X}}_t)$ such that we have the following:

- (i) $(y_{t+r} - y_r)_t$ is a $\underline{\mathcal{X}}_{t+r}$ -Brownian motion, $y_t = 0$, for $t \leq r$;
- (ii) $C_t(f, [x], q)$ is a $(\bar{V}_\infty \vee \underline{\mathcal{X}}_t, R)$ martingale;
- (iii) $x_t = x_r$ for $t \leq r$; x_r is independent of \bar{V}_∞ , its distribution is equal to μ .

We denote by $\mathcal{R}(r, \mu)$ the set of the W-N control rules with initial conditions (r, μ) .

DEFINITION 3.4.1. A P.O. control rule with initial conditions (r, μ) is a probability measure R^h on $\hat{\mathcal{X}}$, which is deduced from a W-N control rule R with the formula $R^h = L_T \cdot R$, where

$$L_T := \exp \int_r^T h'(s, x_s) dy_s - \frac{1}{2} \int_r^T h^2(s, x_s) ds.$$

The reward associated with this rule is

$$j(r, \mu, R^h) = R^h \left\{ \int_r^T q(k)(s, x_s) ds + g(x_T) \right\}.$$

We denote by $\mathcal{R}^h(r, \mu)$ the set of P.O. control rules with initial conditions (r, μ) . Hypothesis 3.2 allows us to give an easy characterization of the elements of $\mathcal{R}(r, \mu)$.

PROPOSITION 3.5. A probability measure R on $\hat{\mathcal{X}}$ belongs to $\mathcal{R}(r, \mu)$ if and only if R admits a factorization in the form $R = QS_{(r, \mu)}$, where $S_{(r, \mu)}(q)$ is the solution to the martingale problem (3.2.1), and where Q is a probability measure on $(\bar{V}, \bar{\mathcal{V}}_t)$ such that $(y_{t+r} - y_r)_t$ is a $(\bar{\mathcal{V}}_{t+r}, Q)$ Brownian motion and $y_t = 0$ for $t \leq r$.

DEFINITION 3.5. We denote by $\mathcal{Q}(r)$ the set of probability measures on $(\bar{V}, \bar{\mathcal{V}}_t)$ such that $(y_{t+r} - y_r)$ is a $(\bar{\mathcal{V}}_{t+r}, Q)$ Brownian motion and $y_t = 0$ for $t \leq r$.

Proof of Proposition 3.5. Necessity is trivial; therefore we show sufficiency.

From the uniqueness hypothesis 3.2 we know that, conditionally with respect to $\bar{\mathcal{V}}_\infty$, the law of (x_t) is $S_{(r, \mu)}(q)$ and 3.4(ii) and (iii) follow. It is obvious that, for each $A \in \mathcal{F}_t$, $R(A | \bar{\mathcal{V}}_\infty) = S_{(r, \mu)}(q)(A)$ is $\bar{\mathcal{V}}_t$ -adapted since $S_{(r, \mu)}(q)$ on \mathcal{X}_t depends only on $1_{[0, t]}(q)$ for $t \geq r$. Suppose now that M_t is a $\bar{\mathcal{V}}_t$ -bounded martingale. For each $A \in \mathcal{F}_t$ and $B \in \bar{\mathcal{V}}_t$, $R(1_A 1_B (M_{t+s} - M_t)) = R(1_B (M_{t+s} - M_t) S_{(r, \mu)}(\cdot)(A)) = 0$; thus M_t is an \mathcal{F}_{t+r} -martingale. In particular, $y_{t+r} - y_r$ remains a \mathcal{F}_{t+r} -Brownian motion. \square

The set of P.O. rules enables us to describe the optimization problem under partial observations.

PROPOSITION 3.6. Let $\mathcal{U}(r, \mu) = (\Omega, \underline{\mathcal{F}}_t, \underline{\mathcal{G}}_t, X_t, Y_t, B_t, q, r, Z, P)$ be a W-N relaxed control. There exists a W-N rule R such that the two rewards are equal:

$$j(r, \mu; R^h) = R(L_T \gamma) = P(L_T \gamma(X_T, q)).$$

γ is defined in (3.1.2).

Proof. We transport the W-N relaxed problem on a canonical space. Let i be the map from Ω into $\hat{\mathcal{X}}$ defined by $i(\omega) = (Y_t(\omega), q(ds, da)(\omega), X_t(\omega))$. This map is measurable and adapted. The term $(\Omega, i^{-1}(\bar{\mathcal{V}}_t), i^{-1}(\hat{\mathcal{X}}_t), q, Y_t, X_t, P)$ is also a W-N relaxed control with the same reward as \mathcal{U} . The usual rules of probability image show us that Poi^{-1} is a W-N rule. \square

4. Compactness of the set of partially observable rules. We will now use the canonical space and give a “convex compact” frame for the optimization problem under partial observations. We will endow the space of laws of controlled process with the topology of weak convergence.

The semimartingale properties of the process we consider give an easy characterization for the precompactness of the space of laws. We have proved in Remark 3.3 that the uniqueness hypothesis 3.2, the continuity and the boundness of the coefficients imply that the map $(r, \mu, q) \rightarrow S_{(r, \mu)}(q)$ from $[0, T] \times \mathcal{P}(\mathbb{R}^d) \times V$ into $\mathcal{P}(\mathcal{X})$ is continuous.

We have denoted by $\mathcal{P}(E)$ the set of probability measures on E .

PROPOSITION 4.1. The space $\mathcal{R}(r, \mu)$ of W-N rules is a compact subset of $\mathcal{P}(\hat{\mathcal{X}})$. The graph of the set-valued map $(r, \mu) \rightrightarrows \mathcal{R}(r, \mu)$ is closed.

Our reference for set-valued maps is the book by Aubin and Celina [AC]. We have used such maps in a preceding paper [EHJ]. In Appendix A we give the most important definitions and properties of set-valued maps.

Proof of Proposition 4.1. Since the mapping $(r, \mu, q) \rightarrow S_{(r, \mu)}(q)$ is continuous, it suffices to prove that $\mathcal{Q}(r)$ (Def. 3.5) is a compact subset of $\mathcal{P}(\bar{V})$. The restriction to V of the elements of $\mathcal{Q}(r)$ is equal to $\mathcal{P}(V)$. The space V is compact; hence the space $\mathcal{P}(V)$ is compact. The projection of an element of $\mathcal{Q}(r)$ on \mathcal{Y} is a probability measure such that $(y_{t+r} - y_r)$ is a Brownian motion, and $y_t = 0$ for $t \leq r$. The set of these projections is obviously a tight set. Then, $\mathcal{Q}(r)$ is tight in $\mathcal{P}(\bar{V})$.

Notice that, with a similar method, we can show that, if (r, μ) belongs to a compact subset of $\mathbb{R}^+ \times \mathcal{P}(\mathbb{R}^d)$, the set $\{(r, \mu, Q); Q \in \mathcal{Q}(r)\}$ is tight: it suffices to say that the canonical process y_t is a martingale which is equal to 0 at $t = 0$ and whose increasing process is equal to $(t - r)^+$ and is bounded.

It remains to prove that $\mathcal{Q}(r)$ is closed: indeed, the martingale property remains valid when we pass to the limit.

We remark that $y_{t+r} - y_r$ is a $\bar{V}_{t+r} - Q$ Brownian motion equivalent to y_t being a Q -martingale with an increasing process equal to $(t - r)^+$. This proves that if (r_n, Q_n) converges to (r, Q) and if $Q_n \in \mathcal{Q}(r_n)$, then $Q \in \mathcal{Q}(r)$.

It is then easy to verify that the graph of the set-valued map $r, \mu \mapsto \mathcal{R}(r, \mu)$ is closed. \square

REMARKS. The set $\mathcal{R}(r, \mu)$ is convex. It is obvious from Definition 3.4 that if $R_i \in \mathcal{R}(r, \mu)$, and if $R = \sum \alpha_i R_i$ with $\sum \alpha_i = 1$, $\alpha_i \geq 0$, R trivially satisfies (i) and (iii), and $R((C_t - C_s)\Phi_s) = \sum \alpha_i R_i((C_t - C_s)\Phi_s) = 0$; then R satisfies (ii).

COROLLARY 4.2. *The set-valued map $r \mapsto \mathcal{Q}(r)$ is continuous.*

Proof. The continuity is a local property. Then, we can suppose that $r \in [0, T]$ with $T < \infty$. We have seen in the above proof that $\{Q \mid Q \in \mathcal{Q}(r), r \in [0, T]\}$ is tight and closed. Then, this set is compact. The set-valued map $r \mapsto \mathcal{Q}(r)$ from $[0, T]$ into a compact set has a closed graph. Therefore, this map is upper semicontinuous (see Appendix A).

In order to prove the lower semicontinuity, we must approximate each probability $Q \in \mathcal{Q}(r)$ by a sequence $Q_n \in \mathcal{Q}(r_n)$, where r_n is a given sequence which converges to r . (See Appendix A for the definition of l.s.c.).

We consider two cases:

(a) $r_n > r$. Let Q'_n be the image law of Q under the map

$$\begin{aligned} y_t &\rightarrow y_t - y_{r_n} & \text{if } t \geq r_n, \\ &\rightarrow 0 & \text{if } t \leq r_n. \end{aligned}$$

It is easy to see that Q'_n belongs to $\mathcal{Q}(r_n)$.

(b) $r_n < r$. Let Q'_n be the image law of Q under the map $y_t \rightarrow y_t + z_{(t \wedge r - r_n)^+}$, where z is a Brownian motion which is independent of \bar{V}_∞ . Then $Q'_n \in \mathcal{Q}(r_n)$.

With this construction, Q'_n converges to Q . \square

In order to establish the existence of an optimal control we have to prove that the above properties of stability are still valid for the P.O. rules, i.e., that these properties are not removed under the Girsanov transformation.

PROPOSITION 4.3. *Let R_n be a sequence of probability measures on \mathcal{X} such that y_t is an R_n -martingale with a deterministic increasing process $A_n(t)$, where $|A_n(t)| \leq t$.*

Let L_t^n be equal to

$$L_t^n := \exp \int_0^t h'(s, x_s) dy_s - \frac{1}{2} \int_0^t |h(s, x_s)|^2 dA_s^n.$$

If the sequence R_n converges (weakly) to a probability R , and if A_t^n converges to A_t , then the probability measures R_n^h converge to R^h (Definition 2.4).

Proof. Since h is bounded, we have $R_n\{(L_T^n)^2\} \leq K$ (1.4.2.2).

Furthermore, L_t^n is a martingale which satisfies $dL_t^n = L_t^n h'(t, x_t) dy_t$. It is well known [JM2] that the laws of L_t^n on the canonical space $\mathcal{L} = \mathcal{C}([0, T], \mathbb{R}^+)$ are tight, although the coefficients of the stochastic differential equation are not bounded. Therefore, the probability measures $\hat{R}^n = R^n \otimes \delta_{L^n}$ on $\hat{\mathcal{X}} \times \mathcal{L}$ are tight. Let \hat{R} be a cluster point of these measures. y_t remains a martingale on the product space $(\hat{\mathcal{X}} \times \mathcal{L}, \hat{\mathcal{F}}_t \otimes \mathcal{L}_t, \hat{R})$ and its increasing process is equal to A_t . Moreover, if we denote by l_t the canonical coordinate process in \mathcal{L} , the following equation is satisfied: $dl_t = l_t h'(t, x_t) dy_t$. The pathwise uniqueness of the solution of this equation implies that, \hat{R} almost surely,

$$l_t = \exp \int_0^t h'(s, x_s) dy_s - \frac{1}{2} \int_0^t |h|^2(s, x_s) dA_s = L_t.$$

Moreover, for bounded continuous f and for $N < \infty$,

$$\int f(L_T^n \wedge N) dR^n = \int \int f(l_T \wedge N) dR^n \rightarrow \int \int f(l_T \wedge N) dR = \int f(L_T \wedge N) dR.$$

Now the uniform integrability of $R^n((L_T^n)^2) \leq K$ implies that $L_T^n dR^n$ converges weakly to $L_T dR$. \square

Taking Proposition 4.3 into account, we can transfer onto the set of P.O. rules the properties that we have established for W-N rules.

THEOREM 4.4. (a) The set $\mathcal{R}^h(r, \mu)$ of P.O. rules (i.e., $\mathcal{R}^h(r, \mu) = \{L_T \cdot R; R \in \mathcal{R}(r, \mu)\}$ with initial conditions (r, μ)) is a compact convex set.

(b) The set-valued map $(r, \mu) \mapsto \mathcal{R}^h(r, \mu)$ is continuous.

The convexity is obvious: $\sum \alpha_i R_i^h = (\sum \alpha_i R_i) L_T = R L_T$, where $R \in \mathcal{R}(r, \mu)$ since this set is convex, (see § 4.7).

All these results allow us to establish some measurability and regularity properties of the value function: we recall that we assume hypothesis 3.2.

THEOREM 4.5. Let $v(r, \mu)$ be the value function of the P.O. control problem, i.e., $v(r, \mu) := \sup \{R^h(\gamma); R^h \in \mathcal{R}^h(r, \mu)\}$ where

$$\gamma([x], q) := \int_r^T q(k)(s, x_s) ds + g(x_T).$$

Then we have the following:

(a) If k and g are uniformly continuous, the application γ is continuous and v is continuous.

(b) If γ is l.s.c., v is also l.s.c. If γ is measurable, v is universally measurable.

(c) If γ is u.s.c., the set $\mathcal{R}^{h*}(r, \mu)$ of optimal P.O. rules is compact and nonempty. The graph of the set-valued map $(r, \mu) \mapsto \mathcal{R}^{h*}(r, \mu)$ is universally measurable.

Proof. (a) If k and g are uniformly continuous, γ is continuous. The set-valued map $(r, \mu) \mapsto \mathcal{R}^h(r, \mu)$ is continuous and has compact values. Therefore (see Appendix A) $v(r, \mu)$ is continuous since it is a marginal function.

(b) This is also a consequence of the properties of set-valued maps.

(c) If γ is u.s.c., the map $R \rightarrow R(\gamma)$ is also u.s.c. The set $\mathcal{R}^h(r, \mu)$ is compact. Therefore, this map attains its maximum on $\mathcal{R}^h(r, \mu)$: there exists $R^{*h} \in \mathcal{R}^h(r, \mu)$ such that $R^{*h}(\gamma) = \sup \{R^h(\gamma); R^h \in \mathcal{R}^h(r, \mu)\}$. Since $\mathcal{R}^h(r, \mu)$ is the set of probability $\{L_T R, R \in \mathcal{R}(r, \mu)\}$, this last equality implies that there exists $R^* \in \mathcal{R}(r, \mu)$ such that

$$j(r, \mu; R^{*h}) = \sup \{j(r, \mu; R^h); R \in \mathcal{R}(r, \mu)\}.$$

Therefore, the set of optimal rules is not empty. It is easy to verify that this set is closed. \square

Remark. If h and g are u.s.c. uniformly in x , γ is u.s.c.

Some compactness arguments allow us to compare the different control problems.

THEOREM 4.6. *Assume that k and g are uniformly continuous. The relaxed problem and the nonrelaxed problem have the same value function, i.e.,*

$$\begin{aligned} v(r, \mu) &= \sup \{j(r, \mu; R^h); R \in \mathcal{R}(r, \mu)\} \\ &= \sup \{j(r, \mu; \mathcal{U}_h^0); \mathcal{U}_h^0 \in \mathcal{A}_h^0(r, \mu)\} \quad (\text{Definition 1.5.2}). \end{aligned}$$

Proof. We have to prove that any control rule $R^h \in \mathcal{R}^h(r, \mu)$ may be approximated with a sequence of elements of $\mathcal{A}_h^0(r, \mu)$. Let ψ_k be the sequence of maps from V to V^0 that we have defined in Theorem 2.2. Let Q be the probability on $\mathcal{Y} \times V$ defined by $R^h = L_T \cdot R$ and $R = Q \cdot S_{(r, \mu)}$. Let Q_k be the image law of Q under ψ_k . The applications ψ_k are adapted, therefore $(y_{t+r} - y_r)_t$ remains a Q_k -Brownian motion (indeed, the conditional law of $1_{]0, t[} \cdot q$ with respect to \mathcal{Y}_∞ under the probability Q_k is \mathcal{Y}_t -adapted). Q_k converges to Q ; $Q_k S_{(r, \mu)}$ converges to $Q \cdot S_{(r, \mu)}$. Therefore Proposition 4.3 implies that $L_T Q_k S_{(r, \mu)}$ converges to $L_T Q S_{(r, \mu)}$. \square

4.7. An example constructed from the counterexample of Duncan and Varaiya [DV]. Let W be the law of a two-dimensional Brownian motion (B^1, B^2) . We consider the weak solutions of the two-dimensional system, for $0 \leq t \leq T = 1$:

$$S_{(U)}: \begin{cases} dX_t^1 = U(t, \cdot) dt + dB_t^1, \\ dX_t^2 = dB_t^2. \end{cases}$$

Let U_1 and U_2 be two control processes defined as follows:

$$\begin{aligned} U_1(t, \cdot) &= 0, \\ U_2(t, \cdot) &= 0 \quad \text{if } t < \frac{1}{2} \\ &= \text{sgn}(X^{2(\frac{1}{2})}) \quad \text{if } \frac{1}{2} \leq t \leq 1. \end{aligned}$$

The weak solutions of the two problems $S_1 = S_{U_1}$ and $S_2 = S_{U_2}$ are of the form $p^i = \Lambda^{U_i} W$, where Λ^{U_i} is the exponential density. We have $\Lambda_t^1 = 1$ and

$$\begin{aligned} \log \Lambda_t^2 &= 0 \quad \text{if } 0 \leq t \leq \frac{1}{2} \\ &= \text{sgn}(X^{2(\frac{1}{2})})(X_1(t) - X_1(\frac{1}{2})) - \frac{1}{2}(t - \frac{1}{2}) \quad \text{for } t \geq \frac{1}{2}. \end{aligned}$$

We now construct a probability \bar{W} on $\Omega \times \{1, 2\}$ which is the product of W with $\alpha \delta_{\{1\}} + (1 - \alpha) \delta_{\{2\}}$. We extend in a natural way the two Brownians B_i to $\Omega \times \{1, 2\}$ as follows: $\bar{P} = \Lambda_T^U(\omega, i) \cdot \bar{W} = \alpha P^1 \cdot \delta_{\{1\}} + (1 - \alpha) P^2 \cdot \delta_{\{2\}}$, where $\Lambda_T^U(\omega, i) = \Lambda_T^1(\omega) 1_{i=1} + \Lambda_T^2(\omega) 1_{i=2}$ and $T = 1$. Duncan and Varaiya [DV] show that $\alpha \Lambda^1 + (1 - \alpha) \Lambda^2$ is not an "admissible density," i.e., the set of admissible densities is not convex. Here under \bar{P} , B^1 , and B^2 are solutions of the equation

$$dX_t = U(t, i) dt + d\tilde{B}_t(\omega, i)$$

where $U(t, i) = U_i$. Hence, the law of the solution to this equation is the convex combination of the law of X_1 and X_2 .

In the first part we solved a special aspect of the control problem for partially observed diffusions. In order to give more information for the optimal rules and to find a “Markovian” control, it seems natural to introduce the filtering process and to transform the P.O. problem into a problem with complete information for the filter. This problem is well known under the name of “separated problem.” We will study it in Part II.

PART II

Separated Controls.

Existence of An Optimal Markovian Filter

1. Separated controls. Let $\mathcal{U}_h(r, \mu) = (\Omega, \underline{F}_t, \underline{G}_t, X_t, Y_t, q, r, \mu, P^h)$ be a P.O. relaxed control. The filter associated with this control is the conditional law of X_t given the σ -algebra \underline{F}_t under the probability P^h . For the W-N relaxed control \mathcal{U} associated with \mathcal{U}_h , the conditional law of X_t given the σ -algebra \underline{F}_∞ is the probability $S_{(r, \mu)}(q)$. Hence, the conditional law of X_t given \underline{F}_∞ depends only on $1_{[0, t]} \cdot q$. Therefore, this conditional law is equal to the conditional law of X_t given \underline{F}_t . This property allows us to establish that the two filtrations \underline{F}_t and $\underline{F}_t \vee \sigma(X_s, s \leq t)$ satisfy to the K -hypothesis, i.e., for all t $\underline{F}_t \vee \sigma(X_s, s \leq t)$ and \underline{F}_∞ are conditionally independent with respect to \underline{F}_t . It is well known [SM], [FKK], [Yo] that, under the K -hypothesis, the filter (r_t) verifies an S.D.E. (which is measure-valued) called the “Kushner–Stratanovich” equation: for $f \in C_b^{1,2}(\mathbb{R}^+ \times \mathbb{R}^d)$

$$\begin{aligned} (KS) \quad r_t(f) = r_r(f) &+ \int_r^t \int \int_{A \times \mathbb{R}^d} Lf(s, x, a) r_s(dx) q(s, da) ds \\ &+ \int_r^t \{r_s(fh) - r_s(f)r_s(h)\} \{dY_s - r_s(h) ds\}. \end{aligned}$$

The process $Y_t - \int_r^t r_s(h) ds$ is the “innovation process.” It is a $(\underline{F}_t - P)$ Brownian motion.

The separated problem is the problem of control (under complete information) of the measure-valued solution of the equation (KS). As in the case of finite-dimensional S.D.E. we have different notions of control.

Fleming [Fl3] considers a separated problem as a weak control of the weak solutions of (KS), i.e., the solutions associated to any Brownian motion.

We introduce here a less general separated problem. We consider only the solutions of (KS) which are filters for a (P.O.) relaxed model. (We denote these filters by P.O. filters).

As we have already noted in [EHJ], the comparison between the different problems of control (strict P.O. separated ones) can be deduced from weak or strong uniqueness of the filtering equation.

1.1. The filtering equations. We recall the notation in § I.1.1. If m is a probability measure

$$\text{cov } m(f, h) := m(fh) - m(f)m(h) := \text{cov } m(fh).$$

If h is a vectorial function, $\text{cov } m(fh)$ denotes the covariance vector; we denote by $\text{cov}' m(fh)$ the transpose vector.

DEFINITION 1.1.1. Let $(\Omega, \underline{F}_t, P)$ be a probability space endowed with a filtration and $(Y_{t+r} - Y_r)_t$ a Brownian motion.

Let q be an adapted process which takes values in V . An adapted continuous process π_t which is $\mathcal{P}(\mathbb{R}^d)$ -valued is a solution of the (KS) equation if: for all $f \in C_b^{1,2}(\mathbb{R}^d)$

$$(KS) \quad \pi_t(f) = \pi_r(f) + \int_r^t \pi_s \cdot q_s(Lf) ds + \int_r^t \text{cov}' \pi_s(fh) \{dY_s - \pi_s(h) ds\}.$$

It may be possible to construct many solutions of the (KS) equation. Moreover, it is possible that there exist some solutions which are not “filters,” except when we have “weak uniqueness” of the solutions of (KS). In the last section (§ 6) we study the uniqueness problem, this problem has only a partial solution.

Sometimes, it is easier to work with the Zakai's equation. We consider $\sigma_t(f)$, the conditional expectation of $L_t f(X_t)$ given \underline{F}_t , where L_t is defined as in Theorem I.1.4.2. We have

$$(Z) \quad \sigma_t(f) = \sigma_r(f) + \int_r^t \sigma_s q_s(Lf) ds + \int_r^t \sigma'_s(fh) dY_s.$$

It is proved in [SM] and [Ku1] that the two equations (KS) and (Z) are equivalent, i.e.:

If $\pi_r(f) = \sigma_r(f) = \mu(f)$, then $\pi_t(f) = \sigma_t(f)/\sigma_t(1)$ and $d\sigma_t(1) = \langle \pi_t(h)\sigma_t(1), dY_t \rangle$.

1.2. Separated controls. A separated control is a weak control of the solutions of the infinite-dimensional equation (KS), subject to the condition that it be the filter process associated with a P.O. control. (We call this condition the *constraint*.)

DEFINITION 1.2.1. A *separated control* (*S-control*) is a term $\mathcal{V}(r, \mu) := (\Omega, \underline{F}_t, \underline{G}_t, \pi_t, X_t, Y_t, q, r, \mu, P)$ such that we have the following:

- (i) \underline{F}_t and \underline{G}_t are two right-continuous filtration and $\underline{F}_t \subseteq \underline{G}_t$;
- (ii) q is an \underline{F}_t -adapted variable which takes values in V ;
- (iii) $(Y_{t+r} - Y_r)$ is an \underline{F}_{t+r} -adapted process and a \underline{G}_{t+r} -Brownian motion;
- (iv) π_t is an \underline{F}_t -adapted solution of the equation

$$(KS) \quad \begin{aligned} \pi_t(f) = \pi_r(f) + \int_r^t \pi_s q_s(Lf)(s) ds \\ + \int_r^t \text{cov} \pi_s(fh) \{dY_s - \pi_s(h) ds\} \quad P \text{ almost surely,} \end{aligned}$$

$\pi_r(f) = \mu, P$ almost surely.

We also modelize the following constraint:

(v) The term $(\Omega, \underline{F}_t, \underline{G}_t, X_t, Y_t, q, r, \mu, P)$ is a P.O. relaxed control with initial condition equal to (r, μ) .

(vi) For each bounded Borelian function f and for each \underline{F}_t -measurable variable H , $P^h(Hf(X_t)) = P^h(H\pi_t(f))$.

1.2.2. The objective. Let $\mathcal{V}(r, \mu)$ be a separated control and $\lambda_t := \exp \int_r^t \pi'_s(h) dY_s - \frac{1}{2} \int_r^t \pi_s^2(h) ds$. It is a solution of the S.D.E. $d\lambda_s = \pi_s(h)\lambda_s dY_s$.

By analogy with Theorem I.1.4.2, we make the change of probability of Girsanov type $P^\pi = \lambda_T P$.

Let K and G be two bounded measurable functions from $\mathbb{R}^+ \times \mathcal{P}(\mathbb{R}^d) \times A$, (respectively, from $\mathcal{P}(\mathbb{R}^d)$) into \mathbb{R} .

The objective associated with the separated problem is

$$\begin{aligned} J(r, \mu; \mathcal{V}) &:= P^\pi \left\{ \int_r^T q(K)(s, \pi_s) ds + G(\pi_T) \right\} \\ &= P \left\{ \lambda_T \int_r^T q(K)(s, \pi_s) ds + \lambda_T G(\pi_T) \right\}. \end{aligned}$$

1.2.3. Remarks. The constraint (vi) in Definition 1.2.1 that we have imposed on π_t , the solution of the (KS) equation implies that π_t is the filter of the associated P.O. problem. Therefore $\lambda_t \pi_t(f)$ is a version of $P(L_T f(X_t) | \underline{F}_t)$. Hence, the restriction of P^h to \underline{F}_T is equal to P^π .

In particular, if $K(t, m, a) = m(k(t, \cdot, a))$ and $G(m) = m(g)$, where k and g are the reward functions for the P.O. relaxed problem, we have

$$\begin{aligned} J(r, \mu; \mathcal{V}) &= P^\pi \left\{ \int_r^T q_s \pi_s(k)(s) ds + \pi_T(g) \right\} \\ &= P^h \left\{ \int_r^T q_s \pi_s(k)(s) ds + \pi_T(g) \right\} \\ &= P^h \left\{ \int_r^T q_s(k)(s, X_s) ds + g(X_T) \right\} = j(r, \mu; \mathcal{U}). \end{aligned}$$

1.3. Separated rules. We now formalize all the previous definitions on the canonical spaces. We denote by Π the space of continuous applications from \mathbb{R}^+ into $\mathcal{P}(\mathbb{R}^d)$; its canonical right-continuous filtration is $\underline{\Pi}_t$.

Let \mathcal{X} , V , and \mathcal{Y} be the spaces that we have defined in Part I. Here we use many product spaces and we have a lot of notation. We use a bar when we have two spaces: $\bar{V} = \mathcal{Y} \times V$; we use a hat for three spaces: $\hat{\mathcal{X}} = \mathcal{Y} \times V \times \mathcal{X}$; $\hat{\Pi} = \mathcal{Y} \times V \times \Pi$; and we denote by $\tilde{\mathcal{X}}$ the product of the four spaces $\mathcal{Y} \times V \times \Pi \times \mathcal{X}$. We use the same convention for the canonical filtrations and for probability measures. A current point of $\tilde{\mathcal{X}}$ is denoted by (y, q, π, x) or by $([y], q, [\pi], [x])$.

DEFINITION 1.3.1. A *separated rule* is a probability \tilde{R} on $\tilde{\mathcal{X}}$ such that the term $(\tilde{\mathcal{X}}, \hat{\Pi}_t, \hat{\mathcal{X}}_t, \pi_t, x_t, y_t, q, r, \mu, \tilde{R})$ is a separated control. We can write all the conditions we imposed on \tilde{R} :

- (i) $(y_{t+r} - y_r)_t$ is a $\hat{\mathcal{X}}_t - \tilde{R}$ Brownian motion;
- (ii) $C_t(f, [x], q)$ is a $\hat{\Pi}_\infty \vee \hat{\mathcal{X}}_t - \tilde{R}$ martingale;
- (iii) $\tilde{R}(L_t H f(x_t)) = \hat{Q}(\lambda_t H \pi_t(f))$ for each $\hat{\Pi}_t$ -measurable H , where \hat{Q} is the restriction of \tilde{R} to $\hat{\Pi}$;
- (iv) $\pi_r(f) = \mu(f)$; $x_t = x_r$ for $t \leq r$, where the distribution of x_r is equal to μ ; $y_t = 0$ for $t \leq r$.

Recall that $L_t = \exp \int_r^t h'(s, x_s) dy_s - \frac{1}{2} \int_r^t h^2(s, x_s) ds$ and $\lambda_t = \exp \int_r^t \pi'_s(h) dy_s - \frac{1}{2} \int_r^t \pi_s^2(h) ds$.

1.3.2. Remarks. Condition (iii) in Definition 1.3.1 is the constraint $\tilde{R}^h(Hf(x_t)) = \tilde{R}^h(H\pi_t(f))$ (see § 1.2.3).

We have not used the equation (KS); a continuous process π_t which satisfies (iii) is a solution to (KS) (see the Introduction of § 1).

1.3.3. Notation. Let us denote by $\tilde{\mathcal{R}}(r, \mu)$ the space of separated rules with initial conditions (r, μ) and by

$$\hat{\mathcal{Q}}(r, \mu) := \{ \hat{Q} \in \mathcal{P}(\hat{\Pi}) \mid \hat{Q} \text{ is the restriction to } \hat{\Pi} \text{ of an element } \tilde{R} \in \tilde{\mathcal{R}}_{(r, \mu)} \}.$$

We denote by $\Gamma([\pi], q)$ the function $\int_r^T q(K)(s, \pi_s) ds + G(\pi_T)$.

1.3.4. The objective. The objective associated with a separated rule is

$$\begin{aligned} J(r, \mu; \tilde{R}) &:= \tilde{R} \left(\lambda_T \int_r^T q(K)(s, \pi_s) ds + \lambda_T G(\pi_T) \right) \\ &= \hat{Q} \left(\lambda_T \int_r^T q(K)(s, \pi_s) ds + \lambda_T G(\pi_T) \right) \\ &= \hat{Q}(\lambda_T \Gamma). \end{aligned}$$

PROPOSITION 1.3.5. *A probability measure \tilde{R} on $\tilde{\mathcal{X}}$ belongs to $\tilde{\mathcal{R}}(r, \mu)$ if and only if \tilde{R} has a factorization in the form $\tilde{R} = \hat{Q}S_{(r, \mu)}$, where*

- (i) $S_{(r, \mu)}$ is the solution of the martingale problem (1.3.2) of Part I.
- (ii) \hat{Q} is a probability measure on $(\hat{\Pi}, \hat{\Pi}_t)$ such that
 - (a) $(y_{t+r} - y_r)_t$ is a $(\hat{\Pi}_{t+r} - \hat{Q})$ Brownian motion;
 - (b) $\hat{Q}(\lambda_t \pi_t(f)) = \tilde{R}(L_t f(x_t))$.

Proof. The proof is the same as that of Proposition 1.3.5, since property (ii)(b) above is analogous to (ii) of Definition 1.3.4.

When no confusion can arise, we will use the term *separated rules* to describe the probabilities \hat{Q} which satisfy (ii).

2. Compactness of the separated rules. We now prove the compactness of the set of separated rules. We use the same arguments as in § 4 of I, where we have proved the compactness of the set $\mathcal{R}^h = \{R^h = L_T R, R \in \mathcal{R}(r, \mu)\}$. In the separated case, we will study the sets $\tilde{\mathcal{R}}^L(r, \mu)$ and $\hat{\mathcal{Z}}^\lambda(r, \mu)$ which are defined by

$$\begin{aligned} \tilde{\mathcal{R}}^L(r, \mu) &= \{L_T \tilde{R}; \tilde{R} \in \tilde{\mathcal{R}}(r, \mu)\}, \\ \hat{\mathcal{Z}}^\lambda(r, \mu) &= \{\lambda_T \hat{Q}; \hat{Q} \in \hat{\mathcal{Z}}(r, \mu)\}. \end{aligned}$$

We can now prove a theorem which is similar to Theorems 4.5 and 4.6 of Part I.

THEOREM 2.1. (a) *The space $\hat{\mathcal{Z}}^\lambda(r, \mu)$ is convex compact in $\mathcal{P}(\hat{\Pi})$.*
 (b) *The set-valued map $(r, \mu) \mapsto \hat{\mathcal{Z}}^\lambda(r, \mu)$ is u.s.c.*
 (c) *If the function $Q \mapsto Q(\Gamma)$ is u.s.c. the value function $V(r, \mu) := \sup \{\hat{Q}(\lambda_T \int_r^T q_s(K)(s, \pi_s) ds + \lambda_T G(\pi_T)); \hat{Q} \in \hat{\mathcal{Z}}(r, \mu)\}$ is u.s.c.*
 (d) *The set of optimal separated rules is nonempty.*

The main difference between the P.O. case and the separated (S) case is the introduction of the canonical space Π .

Tightness criteria for probability measures on Π are less well known than criteria for probability measures on \mathcal{X} . They have been utilized in [FV], [JM] and [RC]. Let us recall this criterion, which allows us to work with laws of real processes.

THEOREM 2.2. *Let $\mathcal{N}(\Pi)$ be a family of probability measures on Π . Let $(f_k)_{k \in \mathbb{N}}$ be a sequence in $C_0(R^d)$ which is dense. If, for each $k \in \mathbb{N}$, the laws of the processes $(\pi_t(f_k))$ are tight in $C(R^+, R)$ and if, for each fixed t , the image of $\mathcal{N}(\Pi)$ under the map $\pi \mapsto \pi_t$ ($\Pi \rightarrow \mathcal{P}(R^d)$) is tight, then $\mathcal{N}(\Pi)$ is precompact.*

In order to prove Theorem 2.1, we establish three lemmas.

LEMMA 2.2.1. *The set $\tilde{\mathcal{R}}(r, \mu)$ of separated rules with initial conditions (r, μ) is precompact in $\mathcal{P}(\tilde{\mathcal{X}})$.*

Proof of the lemma. It suffices to prove that the marginal sets $\tilde{\mathcal{R}}|_{\mathcal{V}}$, $\tilde{\mathcal{R}}|_{\mathcal{W}}$, $\tilde{\mathcal{R}}|_{\mathcal{X}}$, and $\tilde{\mathcal{R}}|_{\Pi}$ are precompact. As in Part I, we prove the precompactness of the three sets $\tilde{\mathcal{R}}|_{\mathcal{V}}$, $\tilde{\mathcal{R}}|_{\mathcal{W}}$, and $\tilde{\mathcal{R}}|_{\mathcal{X}}$.

In order to establish the same property for $\tilde{\mathcal{R}}|_{\Pi}$, we notice that for $f \in C_b^2$, $\pi_t(f) - \int_r^t \pi_s q_s(Lf) ds$ is a martingale with increasing process equal to

$\int_r^t |\text{cov } \pi_s(fh)|^2 ds$. All the coefficients are bounded and π is a probability; thus the semimartingale $\pi_t(f)$ satisfies the hypotheses of Theorem 1.4.6 in Stroock and Varadhan [SV], therefore the laws of $\pi_t(f_k)$ are tight for f_k as in Theorem 2.2.

We must show that, for each t , the laws of π_t are tight on $\mathcal{P}(R^d)$. Lemma 3.2 of Sznitman [Sz] tells us that it is equivalent to the tightness of the family $\mu_\alpha \in \mathcal{P}(R^d)$ defined by

$$\mu_\alpha(f) = \int \pi_t(f) dR_\alpha \quad \text{where } R_\alpha \in \tilde{\mathcal{H}}(r, \mu).$$

This last integral is equal to $\int f(x_t) dR_\alpha$ and we know that the laws of x_t are a tight set. \square

LEMMA 2.2.2. (a) $\tilde{\mathcal{H}}^L(r, \mu)$ is a compact set.

(b) The graph of the set-valued map $(r, \mu) \Rightarrow \tilde{\mathcal{H}}^L(r, \mu)$ is closed.

Proof. We proceed in the same way as we did in Part I. The only new difficulty is establishing the stability of the constraint in Definition 1.3.1(iii) when we pass to the limit.

(a) Let \tilde{R}_n be a family of $\tilde{\mathcal{H}}(r, \mu)$ which converges to $\tilde{R} \in \mathcal{P}(\tilde{\mathcal{H}})$. It is obvious that (i) $\tilde{R}(\pi_r = \mu) = 1$; (ii) $y_{t+r} - y_r$ is an \tilde{R} -Brownian motion; (iii) $C_t(f, [x], q)$ is an \tilde{R} -martingale (Definition 1.3.1). In order to verify that the constraint is satisfied by \tilde{R} , we show that the probability $L_T \tilde{R}_n$ converges weakly to $L_T \tilde{R}$, this is done as in Proposition I.4.3. The same method shows us that $\lambda_T \tilde{R}_n$ converges weakly to $\lambda_T \tilde{R}$.

(b) Let \tilde{R}_n be a family of $\tilde{\mathcal{H}}(r_n, \mu_n)$ such that $r_n \rightarrow r$, $\mu_n \rightarrow \mu$, and $\tilde{R}_n \rightarrow \tilde{R}$ where $\tilde{R} \in \mathcal{P}(\tilde{\mathcal{H}})$.

As in part (a) we prove that $\tilde{R} \in \tilde{\mathcal{H}}(r, \mu)$. \square

LEMMA 2.2.3. (a) The set $\hat{\mathcal{J}}^\lambda(r, \mu)$ is convex compact in $\mathcal{P}(\hat{\Pi})$.

(b) The graph of the set-valued map $(r, \mu) \Rightarrow \hat{\mathcal{J}}^\lambda(r, \mu)$ is closed.

Proof. We emphasize that we must work with probability measures on $\hat{\Pi}$ which are restrictions of probability measures that belong to $\tilde{\mathcal{H}}(r, \mu)$. The convexity is obvious: $\sum \alpha_i \hat{Q}_i$ is the projection of $\sum \alpha_i \tilde{R}_i$, which belongs to $\tilde{\mathcal{H}}$.

If \hat{Q}_n converges to \hat{Q} , the family \tilde{R}_n associated with \hat{Q}_n belongs to $\tilde{\mathcal{H}}(r, \mu)$; thus we can pick up a subsequence which converges to $\tilde{R} \in \tilde{\mathcal{H}}(r, \mu)$ and the projection of \tilde{R} is equal to \hat{Q} . The same method is used to establish that the graph of the set-valued map $(r, \mu) \Rightarrow \hat{\mathcal{J}}^\lambda(r, \mu)$ is closed. \square

It remains for us to prove Theorem 2.1. Part (a) has been established in Lemma 2.2.3. Let K be a compact set in $R^+ \times \mathcal{P}(R^d)$. With the same method as in Lemma 2.2.2 we establish that $\bigcup_{(r, \mu) \in K} \hat{\mathcal{J}}^\lambda(r, \mu)$ is a compact set. Since the graph of the set-valued map $(r, \mu) \Rightarrow \hat{\mathcal{J}}^\lambda(r, \mu)$ is closed, we obtain the u.s.c. of this set-valued map (see Appendix A). Part (c) follows from the properties of set-valued maps. Part (d) follows from (c) and from the compactness of $\hat{\mathcal{J}}^\lambda(r, \mu)$. \square

3. Conditional rules. The properties of stability under conditioning and concatenation for the laws of the controlled processes are the fundamental statements from which we can deduce the equations of dynamic programming, and the existence of optimal Markovian controls.

For the separated problem, we must prove these properties for the probability measures in $\hat{\mathcal{J}}^\lambda(r, \mu) = \{\lambda_T \cdot \hat{Q}, \text{ where } \hat{Q} \text{ is the restriction to } \hat{\Pi} \text{ of an element } \tilde{R} \in \tilde{\mathcal{H}}(r, \mu)\}$. We will use the characterization of elements in $\hat{\mathcal{J}}^\lambda(r, \mu)$ given in Proposition 1.3.5. We use the techniques developed by Stroock and Varadhan [SV, Chap. 6 § 1]. Let us remark that $V = V_{[0, s]} \times V_{[s, \infty]}$. We recall here the notation of [SV]. Let $s \geq 0$ be given and suppose that Q is a probability measure on $(\hat{\Pi}, \hat{\Pi}^s)$, where $\hat{\Pi}^s = \mathcal{Y}^s \times \mathcal{V}^s \times \mathcal{I}^s$ ($\mathcal{Y}^s = \sigma(y(t), t \geq s) \cdot \cdot \cdot$). If $\hat{\pi} \in C([0, s], R) \times V_{[0, s]} \times C([0, s], \mathcal{P}(R^d))$ and

$Q((x, q, \pi)(s) = \hat{\pi}(s)) = 1$ then there exists a unique probability measure $\delta_{\hat{\pi}} \otimes_s Q$ on $(\hat{\Pi}, \hat{\Pi}_{\leq}^s)$ such that

$$\delta_{\hat{\pi}} \otimes_s Q((x, q, \pi)(t) = \hat{\pi}(t), 0 \leq t \leq s) = 1,$$

$$\delta_{\hat{\pi}} \otimes_s Q(A) = Q(A) \quad \text{for all } A \in \hat{\Pi}_{\leq}^s.$$

We say that $\delta_{\hat{\pi}} \otimes_s Q$ is a *version of Q extended to the σ algebra $\hat{\Pi}_{\leq}^s$* by the probability measure $\delta_{\hat{\pi}}$.

THEOREM 3.1. *Let τ be a $\hat{\Pi}_{\leq}^s$ -stopping time and let $\lambda_{\tau} \cdot \hat{Q} = \hat{Q}^{\lambda}$ be an element of $\hat{\mathcal{Q}}^{\lambda}(r, \mu)$. Let us denote by $\hat{Q}_{\tau}^{\lambda}([\hat{\pi}], \cdot)$ a version of the conditional law of \hat{Q}^{λ} with respect to $\hat{\Pi}_{\leq}^{\tau}$, extended to the σ -algebra $\hat{\Pi}_{\leq}^{\tau}$ by the probability measure $\delta_{[0]}(dy) \otimes \delta_{q[\hat{\pi}]}(dq) \otimes \delta_{[\pi_{\tau}(\hat{\pi})]}(d[\pi])$. Then $\hat{Q}_{\tau}^{\lambda}([\hat{\pi}], \cdot)$ belongs to $\hat{\mathcal{Q}}^{\lambda}(\tau(\hat{\pi}), \pi_{\tau}(\hat{\pi}))$, Q^{λ} almost surely with respect to $\hat{\pi}$.*

Proof. (a) We first note that the probability $\hat{Q}_{\tau}^{\lambda}([\hat{\pi}], \cdot)$ may be written on the form $\lambda_{\tau}^{\tau} \hat{Q}_{\tau}([\hat{\pi}], \cdot)$, where $\hat{Q}_{\tau}([\hat{\pi}], \cdot)$ is a regular conditional probability distribution (r.c.p.d.) of \hat{Q} with respect to $\hat{\Pi}_{\leq}^{\tau}$, with an extension to the σ -algebra $\hat{\Pi}_{\leq}^{\tau}$ given by $\delta_{[0]}(dy) \otimes \delta_{q[\hat{\pi}]}(dq) \otimes \delta_{[\pi_{\tau}(\hat{\pi})]}(d[\pi])$. In this decomposition, we have set

$$\lambda_{\tau}^{\tau} := \exp \int_{\tau}^t h'(s, x_s) dy_s - \frac{1}{2} \int_{\tau}^t |h(s, x_s)|^2 ds.$$

It is easy to check that, under the probability $\hat{Q}_{\tau}([\hat{\pi}], \cdot)$, the process $y_{t+\tau(\hat{\pi})} - y_{\tau(\hat{\pi})}$ is a Brownian motion and that y_{\cdot} is equal to $[0]$ before $\tau(\hat{\pi})$.

(b) In order to prove that $\hat{Q}_{\tau}^{\lambda}([\hat{\pi}], \cdot)$ belongs to $\hat{\mathcal{Q}}^{\lambda}(\tau(\hat{\pi}), \pi_{\tau}(\hat{\pi}))$, it suffices to prove that $\hat{Q}_{\tau}^{\lambda}([\hat{\pi}], \cdot)$ is the restriction of a separated rule, i.e., (Proposition 1.3.5),

$$\hat{Q}_{\tau}([\hat{\pi}], \lambda_{t+\tau(\hat{\pi})}^{\tau(\hat{\pi})} \pi_{t+\tau(\hat{\pi})}(f)) = (\hat{Q}_{\tau}([\hat{\pi}]) S_{\tau(\hat{\pi}), \pi_{\tau}(\hat{\pi})})(L_{t+\tau(\hat{\pi})}^{\tau(\hat{\pi})} f(x_{t+\tau(\hat{\pi})})).$$

Let F be an element of $\hat{\Pi}_{\leq}^{\tau}$. Since \hat{Q}_{τ} is a version of $\hat{Q}|_{\hat{\Pi}_{\leq}^{\tau}}$ for the events after τ , we have

$$\hat{Q}(\lambda_{\tau} F \hat{Q}_{\tau}(\cdot, \lambda_{t+\tau(\cdot)}^{\tau(\cdot)} \pi_{t+\tau(\cdot)}(f))) = \hat{Q}(\lambda_{\tau} F \lambda_{t+\tau}^{\tau} \pi_{t+\tau}(f)).$$

Moreover, \hat{Q} is the law of a separated process; hence, by definition

$$\hat{Q}(\lambda_{t+\tau} F \pi_{t+\tau}(f)) = \hat{Q}_{S_{(r, \mu)}}(L_{t+\tau} F f(x_{t+\tau})).$$

Clearly

$$\hat{Q}_{S_{(r, \mu)}}(L_{t+\tau} F f(x_{t+\tau})) = \hat{Q}(F S_{(r, \mu)}(L_{\tau} L_{t+\tau}^{\tau} f(x_{t+\tau}))).$$

We apply the Markov property for $S_{(r, \mu)}$:

$$\begin{aligned} S_{(r, \mu)}(\cdot) \{L_{t+\tau}^{\tau}(\cdot, [x]) f(x_{t+\tau}(\cdot))\} \\ = S_{(r, \mu)} \{L_{\tau}(\cdot, [x]) S_{\tau(\cdot), x_{\tau}(\cdot)}(\cdot) (L_{t+\tau}^{\tau}(\cdot, [x']) f(x'_{t+\tau}(\cdot)))\}. \end{aligned}$$

The crux of the matter is contained in the observation that, for each $A \in \hat{\mathcal{F}}_{\leq}^{\tau}$ the application $[x] \rightarrow 1_A([\hat{\pi}], [x])$ is $\mathcal{G}_{\tau(\hat{\pi})}^0$ measurable for each $\hat{\pi}$. The point \cdot in the above formula replaces the variable $\hat{\pi}$ and the prime in x' is for the integration variable. Therefore

$$\hat{Q}_{S_{(r, \mu)}}(L_{t+\tau} F f(x_{t+\tau})) = \hat{Q}_{S_{(r, \mu)}}(F L_{\tau} S_{t, x_{\tau}}(\cdot) (f(x'_{t+\tau})) L_{t+\tau}^{\tau}(\cdot, [x'])).$$

The random variable $S_{\tau, x_{\tau}}(\cdot) (f(x_{t+\tau}) L_{t+\tau}^{\tau}(\cdot, [x]))$ depends only on $y_{t+\tau(\cdot)} - y_{\tau(\cdot)}$, q and $(\tau, x_{\tau})(\cdot)$. Hence, its conditional expectation with respect to $\hat{\mathcal{F}}_{\leq}^{\tau}$ is equal to

$$\int \hat{Q}_{\tau}([\hat{\pi}], d[\hat{\pi}_1]) S_{\tau(\hat{\pi}), x_{\tau}(\hat{\pi})}(q(\hat{\pi}_1)) (f(x_{t+\tau(\hat{\pi})}) L_{t+\tau(\hat{\pi})}^{\tau(\hat{\pi})}([\hat{\pi}_1], [x]))$$

where $\hat{Q}_\tau([\hat{\pi}], \cdot)$ is the conditional expectation with respect to $\hat{\pi}_\tau$ which we have defined at the beginning of the proof.

Thus, we see that

$$\hat{Q}S_{(r,\mu)}(L_{t+\tau}Ff(x_{t+\tau})) = \hat{Q}S_{(r,\mu)}(FL_\tau\hat{Q}_\tau(\cdot)S_{\tau,x_\tau}(\cdot)(f(x_{t+\tau})L_{t+\tau}^\tau(\cdot, [x])));$$

indeed, $\hat{Q}S_{(r,\mu)}(L_\tau F(\cdot, x_\tau))$ is equal to $\hat{Q}S_{(r,\mu)}(\lambda_\tau \pi_\tau(F(\cdot, \cdot)))$ if $F(\cdot, x)$ is $\hat{\Pi}_\tau \times \mathcal{X}$ measurable. This property follows from the fact that $\hat{Q}S_{(r,\mu)}$ is a separable rule.

We conclude that

$$\hat{Q}S_{(r,\mu)}(L_{t+\tau}Ff(x_{t+\tau})) = \hat{Q}S_{(r,\mu)}(FL_\tau\hat{Q}_\tau(\cdot)S_{\tau,\pi_\tau}(f(x_{t+\tau})L_{t+\tau}^\tau(\cdot, [x]))),$$

which establishes the result that we wanted to prove. \square

Let τ be a $\hat{\Pi}_\tau$ -stopping time, and \hat{Q} be a probability measure on $\hat{\Pi}$. There exists a regular conditional probability measure (r.c.p.d.) of \hat{Q} with respect to $\hat{\Pi}_\tau$, denoted by $\hat{Q}([\hat{\pi}], \cdot)$, where $[\hat{\pi}]$ belongs to $\hat{\Pi}$. This r.c.p.d. verifies $\hat{Q}(f|\hat{\Pi}_\tau)([\hat{\pi}]) = \int f(y, q, \pi) \hat{Q}([\hat{\pi}], dy dq d\pi)$ for any f that is $\hat{\Pi}_\tau$ -measurable. It is a probability kernel from (Π, Π_τ) into (Π, Π_τ) . Conversely if $\hat{Q}([\hat{\pi}], \cdot)$ is a probability kernel from (Π, Π_τ) into (Π, Π_τ) and if \hat{P} is a probability measure on $(\hat{\Pi}, \hat{\Pi}_\tau)$, there exists a unique probability measure on $(\hat{\Pi}, \hat{\Pi}_\tau)$ denoted by $\hat{P}|\tau|\hat{Q}$ such that we have the following:

- (i) The restriction of $\hat{P}|\tau|\hat{Q}$ to $\hat{\Pi}_\tau$ is equal to \hat{P} ;
- (ii) An r.c.p.d. of $\hat{P}|\tau|\hat{Q}$ with respect to $\hat{\Pi}_\tau$ is equal to $\hat{Q}([\hat{\pi}], \cdot)$.

In particular, if $\hat{Q}_\tau([\hat{\pi}], \cdot)$ is an r.c.p.d. of \hat{Q} with respect to $\hat{\Pi}_\tau$, we have $\hat{Q} = \hat{Q}|\tau|\hat{Q}_\tau$.

We now establish the following theorems.

THEOREM 3.2. $\hat{\mathcal{Q}}^\lambda(r, \mu)$ is stable by concatenation with the following meaning: Let $\hat{Q}^\lambda = \lambda_\tau \hat{Q}$ be an element of $\hat{\mathcal{Q}}^\lambda(r, \mu)$, and τ a $\hat{\Pi}_\tau$ -stopping time, bounded by T . Let $\hat{Q}_\tau([\hat{\pi}]; \cdot)$ be a probability kernel from (Π, Π_τ) into (Π, Π_τ) such that we have the following:

- (i) For all $\hat{\pi} \in \hat{\Pi}$, $\hat{Q}([\hat{\pi}]; \{y(\tau(\hat{\pi}), \cdot) = y(\tau(\hat{\pi}), \hat{\pi})\}, \{\pi(\tau(\hat{\pi}), \cdot) = \pi(\tau(\hat{\pi}), \hat{\pi})\}) = 1$;
- (ii) If $\Delta([\hat{\pi}]; \cdot)$ is a family of probability measures on $\hat{\Pi}$ such that

$$\forall \hat{\pi} \in \hat{\Pi} \quad \Delta([\hat{\pi}]; \{y_s = 0\}, \{\pi_s = \pi(\tau(\hat{\pi}), \hat{\pi})\}, s \leq \tau(\hat{\pi})) = 1,$$

then for all $\hat{\pi} \in \hat{\Pi}$, $\Delta([\hat{\pi}]|\tau|\hat{Q}_\tau)$ belongs to $\hat{\mathcal{Q}}^\lambda(\tau(\hat{\pi}), \pi_\tau(\hat{\pi}))$. Under these hypotheses $\hat{Q}|\tau|\hat{Q}_\tau$ belongs to $\hat{\mathcal{Q}}^\lambda(r, \mu)$.

The proof follows from Theorem 3.1. We check the stability of $\hat{Q}|\tau|\hat{Q}_\tau$ as in Theorem 6.1.2 of [SV]. The only difficulty results from complicated notation. \square

THEOREM 3.3. $\hat{\mathcal{Q}}^\lambda(r, \mu)$ is stable by conditioning with the following meaning. Let \hat{Q}^λ be an element of $\hat{\mathcal{Q}}^\lambda(r, \mu)$ and τ a $\hat{\Pi}_\tau$ -stopping time. If $\hat{Q}_\tau([\hat{\pi}], \cdot)$ is an r.c.p.d. of \hat{Q}^λ with respect to $\hat{\Pi}_\tau$, and Δ is a family of probability measures such that

$$\Delta([\hat{\pi}]; \{y_s = 0\}, \{\pi_s = \pi(\tau(\hat{\pi}), \hat{\pi})\}, s \leq \tau(\hat{\pi})) = 1,$$

then there exists N such that $\hat{Q}^\lambda(N) = 0$ and for all $\hat{\pi} \in \hat{\Pi}$, $\hat{\pi} \notin N$, $\Delta_{[\hat{\pi}]}|\tau|\hat{Q}_\tau^\lambda$ belongs to $\hat{\mathcal{Q}}^\lambda(\tau(\hat{\pi}), \pi_\tau(\hat{\pi}))$.

Remark 3.4. If $\hat{\mathcal{Q}}^\lambda(r, \mu; \tau, \hat{Q}^\lambda)$ is the set of probability measures in $\hat{\mathcal{Q}}^\lambda(r, \mu)$ which coincide with \hat{Q}^λ up to τ , we have a complete description of this set by

$$\hat{\mathcal{Q}}^\lambda(r, \mu; \tau, \hat{Q}^\lambda) = \{\hat{Q}^\lambda|\tau|\hat{Q}_{[\hat{\pi}]}, \hat{Q}_{[\hat{\pi}]} \in \hat{\mathcal{Q}}^\lambda(\tau(\hat{\pi}), \pi_\tau(\hat{\pi}))\}.$$

Indeed, the rules $\hat{Q}^\lambda|\tau|\hat{Q}_{[\hat{\pi}]}$ with $\hat{Q}_{[\hat{\pi}]} \in \hat{\mathcal{Q}}^\lambda(\tau(\hat{\pi}), \pi_\tau(\hat{\pi}))$ belong to $\hat{\mathcal{Q}}^\lambda(r, \mu; \tau, \hat{Q}^\lambda)$ (from Theorems 3.1 and 3.2 and the definition) and if \hat{Q}_1^λ belongs to $\hat{\mathcal{Q}}^\lambda(r, \mu; \tau, \hat{Q}^\lambda)$ we can write $\hat{Q}_1^\lambda = \hat{Q}^\lambda|\tau|\hat{Q}_{1,\tau}^\lambda$ with $\hat{Q}_{1,\tau}^\lambda \in \hat{\mathcal{Q}}^\lambda(\tau(\hat{\pi}), \pi_\tau(\hat{\pi}))$ (Theorem 3.1).

The dynamical programming principle follows, as we will see in the next section.

4. Equations of the dynamic programming principle. The dynamical programming principle follows from the stability of the rules by conditioning and concatenation, as in [EK].

In order to simplify the notation, we set, for each Borelian bounded function $F: \Gamma_u^s(F) := \int_u^s q(K)(t, \pi_t) dt + F(s, \pi_s)$, where K is the instantaneous reward function. With a notation of the same type we have, for a separated rule $\tilde{R}: J(r, \mu; \tilde{R}) = \tilde{R}(L_T \Gamma_r^T(G)) = \hat{Q}(\lambda_T \Gamma_r^T(G))$, where G is the terminal reward function (1.3.4).

THEOREM 4.1. (a) *Let τ be a $\hat{\Pi}_T$ -stopping time, which is bounded by T . Let \hat{Q}^λ be an element of $\hat{\mathcal{Q}}^\lambda(r, \mu)$ and V the value function defined in Theorem 2.1. Then*

$$(4.1.1) \quad \begin{aligned} & \hat{Q}^\lambda \left(\int_r^\tau q(K)(s, \pi_s) ds + V(\tau, \pi_\tau) \right) \\ &= \sup \{ \hat{Q}_1^\lambda(\Gamma_r^T(G)); \hat{Q}_1^\lambda \in \hat{\mathcal{Q}}^\lambda(r, \mu; \tau, \hat{Q}^\lambda) \} \quad (\text{Remark 3.4}); \end{aligned}$$

(b)

$$(4.1.2) \quad V(r, \mu) = \sup \left\{ \hat{Q}^\lambda \left(\int_r^\tau q(K)(s, \pi_s) ds + V(\tau, \pi_\tau) \right); \hat{Q}^\lambda \in \hat{\mathcal{Q}}^\lambda(r, \mu) \right\}.$$

Proof. (a) With the same method as in [EHJ] we prove

$$\begin{aligned} & \hat{Q}^\lambda \left(\int_r^\tau q(K)(s, \pi_s) ds + V(\tau, \pi_\tau) \right) \\ &= \sup \left\{ \hat{Q}^\lambda \left(\int_r^\tau q(K)(s, \pi_s) ds + {}_1\hat{Q}_{\tau(\hat{\pi}), \pi_\tau(\hat{\pi})}^\lambda \left(\int_\tau^T q(K)(s, \pi_s) ds + G(\pi_T) \right) \right) \right\} \end{aligned}$$

where ${}_1\hat{Q}^\lambda$ varies in the set of measurable selectors of $\hat{\mathcal{Q}}^\lambda$

$$= \sup \{ (\hat{Q}^\lambda | \tau | {}_1\hat{Q}^\lambda)(\Gamma_r^T(G)); {}_1\hat{Q}^\lambda \in \hat{\mathcal{Q}}^\lambda(r, \mu) \}.$$

It remains to note that $\{ \hat{Q}^\lambda | \tau | {}_1\hat{Q}^\lambda; {}_1\hat{Q}^\lambda \in \hat{\mathcal{Q}}^\lambda(r, \mu) \} = \hat{\mathcal{Q}}^\lambda(r, \mu; \tau, \hat{Q}^\lambda)$ in order to obtain (4.1.1).

(b) We note that if \hat{Q}^λ belongs to $\hat{\mathcal{Q}}^\lambda(r, \mu)$,

$$\hat{Q}^\lambda(V(\tau, \pi_\tau)) = \sup \{ (\hat{Q}^\lambda | \tau | {}_1\hat{Q}^\lambda)(\Gamma_{\tau(\pi)}^T(G)); {}_1\hat{Q}^\lambda \in \hat{\mathcal{Q}}^\lambda(r, \mu) \}$$

(see the proof of (a)); hence

$$\hat{Q}^\lambda(\Gamma_r^T(V)) = \sup \{ \hat{Q}_1^\lambda(\Gamma_r^T(G)); \hat{Q}_1^\lambda \in \hat{\mathcal{Q}}^\lambda(r, \mu; \tau, \hat{Q}^\lambda) \} \leq V(r, \mu).$$

The reverse inequality follows from (4.1.1). \square

PROPOSITION 4.2. *Assume that K and G are uniformly u.s.c. The set $\mathcal{Q}^{\lambda*}(r, \mu)$ of optimal rules is stable by concatenation and conditioning.*

Proof. The dynamic programming equation enables us to write that, if Q^* is an optimal rule,

$$Q^{\lambda*} \int_r^\tau q(K)(s, \pi_s) ds + V(\tau, \pi_\tau) = Q^{\lambda*} \int_r^\tau q(K)(s, \pi_s) ds + \Gamma_\tau^T(G).$$

Hence $Q^{\lambda*}(V(\tau, \pi_\tau)) = Q^*(\Gamma_\tau^T(G))$. The stability under conditioning follows easily. The stability under concatenation follows from (4.1.2). \square

PROPOSITION 4.3. *Let T'_r be defined by*

$$T'_r F(\mu) = \sup \left\{ \hat{Q}^\lambda \left(\int_r^t q(K)(s, \pi_s) ds + F(t, \pi_t) \right); \hat{Q}^\lambda \in \hat{\mathcal{Q}}^\lambda(r, \mu) \right\}.$$

Then T'_r is a nonhomogeneous semigroup, i.e.,

$$(4.3.1) \quad T'_s(T'_t F)(\mu) = T'_t F(\mu), \quad r \leq s \leq t.$$

Proof and remarks. This result follows immediately from Theorem 4.1. The formulation of the equations of dynamic programming in terms of semigroup is the idea of Nisio [Ni].

In the partially observable case, (4.3.1) is proved under stronger hypotheses on the coefficients by Fleming [F11], when there is linear dependence on the control, and by Fleming and Nisio [FN] when there is no control in the functions $a_{i,j}$. Under these hypotheses, they prove that the semigroup T_t' is a homogeneous semigroup, and that its generator is an extension of the operator \mathcal{G} , defined on

$$\mathcal{D} = \{F: \mathcal{P}(\mathbb{R}^d) \rightarrow \mathbb{R}; F(\mu) = \phi(\langle f_1, \mu \rangle, \dots, \langle f_l, \mu \rangle)\}$$

$$\text{with } \phi \in C_b^\infty(\mathbb{R}^l) \text{ and } f_1, \dots, f_l \in C_b^\infty(\mathbb{R}^l)\}$$

by the formula

$$\begin{aligned} \mathcal{G}F(\mu) = \sup_{a \in A} \left\{ K(\mu, a) + \sum_{i=1}^l \frac{\partial \phi}{\partial z_i}(\langle f_1, \mu \rangle, \dots, \langle f_l, \mu \rangle) \langle Lf_i(\cdot, a), \mu \rangle \right. \\ \left. + \frac{1}{2} \sum_{i,j} \frac{\partial^2 \phi}{\partial z_i \partial z_j}(\langle f_1, \mu \rangle, \dots, \langle f_l, \mu \rangle) \text{cov } \mu(f_i, h) \text{cov } \mu(f_j, h) \right\}. \end{aligned}$$

That is, if $F \in \mathcal{D}$, $\lim_{t \rightarrow 0^+} (T_t' F(\mu) - F(\mu))/t = \mathcal{G}F(\mu)$.

In particular, if we can show (in the homogeneous case) that $V(0, \mu)$ belongs to \mathcal{D} , the equality $V(0, \mu) = (T_0' V)(\mu)$ implies that V satisfies the Hamilton-Jacobi-Bellmann equation

$$(HJB) \quad \mathcal{G}V(\mu) = \frac{\partial V}{\partial t}$$

(see also Bensoussan [Be], Davis and Kohlmann [DK]).

In the case of linear control, Benes and Karatzas [BK1] describe an extension of \mathcal{G} to the functions $H(t, p)$ (where p is a density of probability on \mathbb{R}^d) which are continuous differentiable in t and twice Fréchet differentiable on p . \square

5. Existence of a Markovian optimal filter. Using Krylov's ideas, which are explained in Stroock and Varadhan [SV, Chap. 12], we prove the existence of a family of optimal separated rules such that (π_t) is a strong Markovian process.

THEOREM 5.1. *We suppose that K and G are u.s.c. uniformly. There exists on the space $(\hat{\Pi}, \hat{\Pi}_t)$ an (r, μ) -measurable family of optimal separated rules $\hat{Q}_{(r, \mu)}^\lambda$ such that π_t is a strong Markovian process.*

Proof. The proof is adapted from [SV, p. 293] and is the same as in [EHJ].

Let (λ_n) be a dense subset of $[0, T]$, and (f_n) be a dense subset of $C_0(\mathbb{R}^+ \times \mathcal{P}(\mathbb{R}^d))$. Let $(t_m, \phi_m)_m$ be an enumeration of $\{(\lambda_n, f_k), n \geq 1, k \geq 1\}$. We define by induction the following control problems:

$$\mathcal{Q}_0^*(r, \mu) = \{\hat{Q}^\lambda: \hat{Q}^\lambda \in \hat{\mathcal{Q}}^{*\lambda}(r, \mu)\} \quad (\text{Proposition 4.2}).$$

The set $\mathcal{Q}_0^*(r, \mu)$ is convex compact and the graph of the set-valued map $(r, \mu) \rightrightarrows \mathcal{Q}_0^*(r, \mu)$ is an analytic set:

$$u_m(r, \mu) = \sup \{ \hat{Q}^\lambda(\phi_m(t_m, \pi_{t_m})); \hat{Q}^\lambda \in \mathcal{Q}_{m-1}^*(r, \mu) \},$$

$$\mathcal{Q}_m^*(r, \mu) = \{ \hat{Q}^\lambda \in \mathcal{Q}_{m-1}^*(r, \mu) \text{ such that } \hat{Q}^\lambda(\phi_m(t_m, \pi_{t_m})) = u_m(r, \mu) \}.$$

We have forgotten the subscript λ and the hat for \mathcal{Q}_n^* , since we are a little bit tired.

As we have seen in the preceding sections (§§ 2 and 3), the u.s.c. character of K and G implies that the sets $\mathcal{Q}_n^*(r, \mu)$ are convex compact and not empty, and that u_n

are analytic. Moreover, the sets $\mathcal{Q}_n^*(r, \mu)$ are stable by concatenation and conditioning. Therefore, their intersection $\mathcal{Q}_\infty^*(r, \mu) = \bigcap_{n \geq 0} \mathcal{Q}_n^*(r, \mu)$ is a compact nonempty set of $\mathcal{P}(\hat{\Pi})$ and the graph of the set-valued map $(r, \mu) \mapsto \mathcal{Q}_\infty^*(r, \mu)$ is analytic. We will now show that the restriction of $\mathcal{Q}_\infty^*(r, \mu)$ to $\hat{\Pi}_T$ has at most one member for each pair (r, μ) .

Let Q_1^λ and Q_2^λ be two points in $\mathcal{Q}_\infty^*(r, \mu)$. Then for each pair (p, n) we have $Q_1^\lambda(f_p(\lambda_n, \pi_{\lambda_n})) = Q_2^\lambda(f_p(\lambda_n, \pi_{\lambda_n}))$. Therefore by density of λ_n and by continuity of $Q(f_p(\cdot, \pi))$ we have $Q_1^\lambda(f_p(t, \pi_t)) = Q_2^\lambda(f_p(t, \pi_t))$. But (f_p) is dense in $C_0(\mathbb{R}^+ \times \mathcal{P}(R^d))$ and so $Q_1^\lambda(f(t, \pi_t)) = Q_2^\lambda(f(t, \pi_t))$.

Moreover, if $Q \in \mathcal{Q}_\infty^*(r, \mu)$ we prove the analyticity of the graph of the map $(r, \mu) \mapsto Q(f(t, \pi_t))$.

The stability by conditioning and the above measurability enable us (as in Thm. 6.2.3 of [SV]) to conclude that $Q_1^\lambda(f_1(\pi_{t_1})f_2(\pi_{t_2}) \cdots f_k(\pi_{t_k})) = Q_2^\lambda(f_1(\pi_{t_1}) \cdots f_k(\pi_{t_k}))$; hence the uniqueness of the restriction of $\mathcal{Q}_\infty^*(r, \mu)$ to $\hat{\Pi}_T$ follows. The restriction of the set $\mathcal{Q}_\infty^*(r, \mu)$ to $\hat{\Pi}_T$ is stable by conditioning and concatenation. It has only one element: it is exactly the strong Markov property. \square

6. Comparison between the different control problems. As in the classical case of control of diffusion processes [EHJ], the uniqueness of the solution of controlled equations is an essential property in order to compare the relaxed and nonrelaxed control problem and to show that the value function is l.s.c.

In the (P.O.) case, these problems are solved (Theorems I.4.5 and I.4.6), since we are working under the hypothesis I.3.2 of uniqueness of the solution to the martingale problem on \mathcal{X} .

For the separated problem, we must consider the uniqueness of the solutions to equation (KS).

DEFINITION 6.1. (a) A solution of (KS) is a term $\mathcal{T} := (\Omega, \underline{F}_t, Y_t, \pi_t, q, r, \mu, P)$ such that

$$(KS) \quad d\pi_t(f) = \pi_t \cdot q_t(Lf)(t) dt + \text{cov } \pi_t(f, h)(dY_t - \pi_t(h) dt), \quad \pi_r = \mu$$

for $f \in C_b^{1,2}(\mathbb{R}^+ \times \mathbb{R}^d)$, where Y_t is an \underline{F}_t -Brownian motion.

(b) There is *strong uniqueness* of solution of (KS) if there is at most one process π_t on the space $(\Omega, \underline{F}_t, Y_t, q, r, \mu, P)$ such that (KS) is satisfied.

(c) There is *weak uniqueness* if the following property is satisfied: Let \mathcal{T} and \mathcal{T}' be two solutions associated with processes (Y, q) and (Y', q') such that (Y, q) and (Y', q') have the same law (on $\mathcal{Y} \times V$). Then the laws of the two processes (π_t) and (π'_t) on Π are equal.

The hypothesis of weak uniqueness of the solutions to (KS) implies that each solution is associated with a (P.O.) filter, more precisely, we have the next result.

LEMMA 6.2. Assume that there is weak uniqueness of the solution of (KS). The space $\hat{\mathcal{Q}}(r, \mu)$ (defined in (1.3.3)) is exactly the space of probability measures \hat{Q} on $\mathcal{Y} \times V \times \Pi = \hat{\Pi}$ such that we have the following:

- (i) $(y_{t+r} - y_r)$ is a $\hat{\Pi}_{t+r}$ -Brownian motion;
- (ii) $d\pi_t = \pi_t q_t(Lf)(t) dt + \text{cov } \pi_t(f, h)(dy_t - \pi_t(h) dt), \quad \pi_r = \mu.$

Proof. Let \hat{Q} be an element of $\hat{\mathcal{Q}}(r, \mu)$. The term $(\hat{\Pi}, \hat{\Pi}_t, y_t, q, \pi_t, \mu, \hat{Q})$ is a solution to (KS). Let \hat{R} be the probability measure on $\hat{\Pi} \times \mathcal{X} = \hat{\mathcal{X}}$ described by $\hat{R} = \hat{Q}S_{(r, \mu)}$. Define $\hat{R}^h = L_T \cdot \hat{R}$ (Definition I.1.5.2). The filter $\tilde{r}_t(f) = \hat{R}^h(f(x_t) | \hat{V}_t)$, where $\hat{V}_t = \underline{V}_t \otimes \underline{V}_t$ satisfies (KS), i.e., $(\hat{V}, \hat{V}_t, y_t, q, \tilde{r}_t, \mu, \hat{Q})$ is a solution to (KS), where \hat{Q} is the restriction of \hat{Q} to \hat{V} . The law of the process (y, q, \tilde{r}) on $\hat{\Pi}$ is $\hat{Q} \cdot \delta_{[\tilde{r}]}(d[\pi])$. Thus, under the weak uniqueness hypothesis, we conclude that this law is equal to \hat{Q} . The conditional laws of the two processes π_t and \tilde{r}_t are the same, π is the canonical process,

and the law of \tilde{r} is a Dirac measure. Therefore, the process (π_t) is Q -indistinguishable from the process (\tilde{r}_t) which is \underline{V}_t -adapted. \square

Under the hypothesis of weak uniqueness of the solution to (KS), the separated control problem of Part II can be viewed as a problem of control of the solution to the equation (KS) without any constraint.

We can use the same approximating result as in the P.O. problem and prove the following proposition.

PROPOSITION 6.3. *Assume there is weak uniqueness of the solution to (KS).*

(a) *If the function Γ (1.3.3) is lower semicontinuous, the relaxed and the nonrelaxed separated problem have the same value function.*

(b) *Let V be the value function for the S-problem:*

$$V(r, \mu) := \sup \{J(r, \mu; \tilde{R}); \tilde{R} \in \tilde{\mathcal{R}}(r, \mu)\}.$$

If the functions K and G are continuous, the value function V is continuous.

Proof. Let \hat{Q} be an element of $\hat{\mathcal{Q}}(r, \mu)$. The terms $\tilde{R}^h, \tilde{r}, \tilde{Q}$ have the same meaning as in Lemma 6.2. Let $\psi^k(q)$ be the sequence of mappings which approximate q by step control process, and \tilde{Q}^k be the image law of \tilde{Q} under ψ^k . From the proof of Theorem I.4.6, $L_T(\tilde{Q}^k S_{(r, \mu)})$ converges to $L_T(\tilde{Q} S_{(r, \mu)})$. Let $\tilde{Q}^k = \tilde{Q} \cdot \delta_{[\tilde{r}^k]}(d[\pi])$ and $\tilde{R}^k = \tilde{Q}^k \cdot S_{(r, \mu)}$. From Lemma 2.2.3 the sequence $\tilde{R}^{k,h}$ is relatively compact and all its limit points $\tilde{R}^{\infty, h}$ belong to $\tilde{\mathcal{R}}^L(r, \mu)$, and the restriction of $\tilde{R}^{\infty, L}$ to $\tilde{V} \times \mathcal{X}$ is equal to $L_T(\tilde{Q} \cdot S_{(r, \mu)})$. Since we assume that weak uniqueness holds for (KS), \tilde{R}^∞ almost surely, the canonical process $[\pi_\cdot]$ is indistinguishable from r_t . The proposition follows with classical arguments.

Proof of (b). The continuity of the mapping V is obvious, if we use the properties of set valued functions (see Appendix A). V is the so-called marginal function associated with the set-valued map $(r, \mu) \mapsto \tilde{\mathcal{R}}(r, \mu)$ and with the application $(r, \mu, \hat{Q}) \rightarrow J(r, \mu; \hat{Q})$. The set-valued map is u.s.c. and compact valued, and the application J is continuous. \square

The purpose of the following propositions is to compare the P.O. problem and the S. problem. We follow Krylov [Kr, Chap. 3] and prove an important result: under the hypothesis of weak uniqueness of the solution to (KS), the value function of the S. problem is equal to the value function of the P.O. strictly admissible problem (Definition (1.1.2) in Part I), i.e., the P.O. problem where the controls are processes which are adapted to the filtration of the observation y_t (enlarged with the initial condition).

From now on, we work with the weak uniqueness hypothesis. We introduce some new objects. Let us take $m \in \mathcal{P}(A)$ and a function u defined on $\mathcal{P}(\mathbb{R}^d)$. Let us use the deterministic control $dt m(da)$ and denote $\tilde{R}_{r, \mu}^m$ the rule associated to the constant control m on $[r, T]$, with initial conditions (r, μ) . With notation similar to that of (§ 4), we set

$$\begin{aligned} J_{r, \mu}^m u(\mu) &:= \tilde{R}_{r, \mu}^m(L_t \Gamma_t'(u)) \\ &= \tilde{R}_{r, \mu}^m \left(L_t \left(\int_r^t K(\pi_s, a) m(da) ds + u(\pi_t) \right) \right). \end{aligned}$$

We recall that in these expressions the density L_t depends on r . If there is an ambiguity, we will denote

$$L_t' = \exp \int_r^t h'(s, x_s) dy_s - \frac{1}{2} \int_r^t |h(s, x_s)|^2 ds.$$

Let $G_{r, \mu} u(\mu) = \sup \{J_{r, \mu}^m u(\mu), m \in \mathcal{P}(A)\}$. It is obvious that $G_{r, \mu} G(\mu)$ is the upper bound of $J(r, \mu, \hat{Q})$ with respect to constant measure control. Let $r = t_0 < t_1 < \dots < t_n = T$. Then

$G_{t_0 t_1} \cdots G_{t_{n-1} t_n} G(\mu)$ is the upper bound of $J(r, \mu, \hat{Q})$ with respect to controls q , which are step-constant deterministic controls, i.e., such that $q(t, da) = dt q_i(da)$ for $t \in [t_i, t_{i+1}[$ and $q_i \in \mathcal{P}(A)$. It is easy to verify that if u is a bounded Borelian function, the mapping $(m, \mu) \rightarrow J_{r, \mu}^m(\mu)$ is Borel and $G_{r, t} u(\mu)$ is an analytic application. Moreover, if u is a bounded continuous function, $G_{r, t} u$ is a continuous function. As Krylov, we introduce $\hat{\mathcal{Q}}_s(r, \mu)$, the set of the laws of the processes $(\Omega, \underline{F}_t, Y_t, q, \pi_t, P)$ such that q is a "step feedback" process, i.e., there exists a subdivision $r = t_0 < t_1 \cdots < t_n = T$ of $[r, T]$ such that q is \underline{F}_t^Y -adapted, and we set $V_s(r, \mu) = \sup \{J(r, \mu; \hat{Q}); \hat{Q} \in \hat{\mathcal{Q}}_s(r, \mu)\}$.

(a) First we prove that if q is a step feedback process, we can find a P.O. strictly admissible control with the same gain.

LEMMA 6.4. *Let μ be a probability on \mathbb{R}^d and $(\Omega, \underline{F}_t, Y_t, P)$ be a Brownian motion. Let $q_t([y], [\pi])$ be a $\hat{\Pi}_t$ -adapted control which is constant on the interval $[t_i, t_{i+1}[$ and constant for $t \leq r$. There exists an \underline{F}_t^Y -adapted process, which is the solution to*

$$(6.4.1) \quad \begin{aligned} d\pi_t(f) &= \pi_t \cdot q_t([y], [\pi])(Lf) dt + \text{cov } \pi_t(f, h)(dY_t - \pi_t(h) dt), \\ \pi_t &= \mu \quad \text{if } t \leq r. \end{aligned}$$

Proof. When the control is constant, this result is proved by Kunita [Ku1]. In our case, since y and π are constant before r , we have $q_r([y], [\pi]) = q(\mu)$. Let us denote by π_t^1 the solution of (KS) associated with $q(\mu)$. This solution is $\underline{F}_t^Y \vee \sigma(\mu)$ -measurable. Define $q^1(\cdot, \cdot) = q_{t_1}(y_{t_1}, \pi_{t_1}^1)$. This random variable is $\underline{F}_{t_1}^Y \vee \sigma(\mu)$ -measurable and is independent of $(Y_{t_1+t_1} - Y_{t_1})$. Then we can construct the solution π_t^2 , starting from $\pi_{t_1}^1$ at time t_1 associated with the control $q^1(\cdot)$ and with the Brownian $Y_{t_1+t_1} - Y_{t_1}$ which is independent of the control. This construction is done for $t \in [t_1, t_2]$. Therefore, the process $\pi_t = \pi_t^1$ for $t \in [r, t_1]$, $\pi_t = \pi_t^2$ for $t \in [t_1, t_2]$, is a solution of (6.4.1) on $[r, t_2]$, and is $\underline{F}_t^Y \vee \sigma(\mu)$ -adapted. It remains to define inductively the process π_t . Moreover, since we are working under the weak uniqueness hypothesis, we obtain a P.O. strictly admissible relaxed control associated with the pair (q, π) .

(b) Now we prove the following result.

LEMMA 6.5.

$$V_s(r, \mu) \geq G_{t_0 t_1} \cdots G_{t_{n-1} t_n} G(\mu)$$

for any subdivision $t_0 = R < t_1 \cdots < t_n = T$.

Proof. Let $u_i(\mu) = G_{t_i t_{i+1}} G_{t_{i+1} t_{i+2}} \cdots G_{t_{n-1} t_n} G(\mu)$. Since the set $\{(m, \mu): J_{r, t}^m u(\mu) + \varepsilon \geq G_{r, t} u(\mu)\}$ is analytic, a selection theorem (as in § 5) shows us that there exists an analytic function $\beta_i(\mu)$ which is $\mathcal{P}(A)$ -valued such that for each μ : $u_i(\mu) \leq J_{t_i t_{i+1}}^{\beta_i(\mu)} u_{i+1}(\mu) + \varepsilon$. We define the control q^β with the following formula:

$$q_t^\beta([y], [\pi]) = \beta_i(\pi_{t_i}) \quad \text{for } t \in [t_i, t_{i+1}[.$$

Then we use Lemma 6.4, which enables us to construct a solution of the (KS) equation associated with the control q^β , and such that the solution is \underline{F}_t^Y -adapted. We remark that, under the construction which was done in Lemma 6.4, we have an equality similar to the Lemma 2.14 of Krylov:

$$(6.5.1) \quad \tilde{R} \left(L_{t_2}^{t_1} \left(\int_{t_1}^{t_2} q_s^\beta(K(\pi_s)) ds + u(\pi_{t_2}) \right) \middle| \underline{F}_{t_1}^Y \right) = G_{t_1 t_2}^{q^\beta(t_1)} u(\pi_{t_1}) \quad \text{almost surely.}$$

We add up all the terms with the form (6.5.1) on the intervals $[t_i, t_{i+1}[$ and we find $V_s(r, \mu) \geq J(r, \mu; q^\beta)$. Thus $V_s(r, \mu) \geq G_{t_0 t_1} \cdots G_{t_{n-1} t_n} G(\mu)$.

(c) Let us now show that if q is a step control, i.e., if $q_t = \sum_i 1_{[t_i, t_{i+1}[}(t) q_i(\cdot)$, where q_i is a $\hat{\Pi}_{t_i}$ -adapted $\mathcal{P}(A)$ -valued process, we have

$$(6.5.2) \quad J(r, \mu; q) \leq G_{t_0 t_1} G_{t_1 t_2} \cdots G_{t_{n-1} t_n} G(\mu)$$

with obvious notation. We introduce functions u_j defined by $u_n(\mu) = G(\mu)$ and $u_j(\mu) = G_{t_j t_{j+1}} u_{j+1}(\mu)$. The stability of the rules by concatenation and conditioning and the weak uniqueness hypothesis allow us to set, for $\tilde{R} \in \tilde{\mathcal{R}}(r, \mu)$ associated with the control q ,

$$\tilde{R} \left(L_{t_j t_{j+1}}^{t_j} \left(\int_{t_j}^{t_{j+1}} q_s(K(\pi_s)) ds + u_{j+1}(\pi_{t_{j+1}}) \right) \middle| \tilde{\Pi}_{t_j} \right) = G_{t_j t_{j+1}} u_{j+1}(\pi_{t_j}).$$

Therefore

$$\tilde{R} \left(L_{t_j t_{j+1}}^{t_j} \left(\int_{t_j}^{t_{j+1}} q_s(K(\pi_s)) ds + u_{j+1}(\pi_{t_{j+1}}) \right) \right) \leq \tilde{R}(G_{t_j t_{j+1}} u_{j+1}(\pi_{t_j})),$$

and, adding up all the inequalities, we obtain (6.5.2).

(d) The last step is to prove that any control q may be approximated with step controls, i.e., if q is a relaxed control, there exists a sequence q^n of step controls such that

$$J(r, \mu; q) = \lim J(r, \mu; q^n).$$

The proof is the same as in Krylov.

It remains to gather together all our results. The inequality $V_s(r, \mu) \leq V(r, \mu)$ is obvious. Part (d) proves that $V(r, \mu) = \lim J(r, \mu, q^n)$, where q^n are step controls; thus from (b) and (c), we can state that

$$J(r, \mu, q^n) \leq G_{t_0 t_1} \cdots G_{t_{n-1} t_n} G(\mu) \leq V_s(r, \mu).$$

Then $V_s(r, \mu) = V(r, \mu)$. \square

We have proved the following result.

THEOREM 6.6. *Assume the weak uniqueness of the solution to (KS). Then the strictly admissible S-problem and the relaxed problem have the same value function.*

If we apply these results to the P.O. problem, we obtain the following.

COROLLARY 6.7. *Assume the weak uniqueness of the solution to (KS) and the continuity of k and g . The strictly admissible P.O. problem and the relaxed S-problem are the same.*

Fleming [F13] defines a separated problem (Σ) which seems to be more general than our separated problem (S). A control, in [F13] is a term $(\Omega, \underline{F}_t, P, b, q, \pi)$ such that for each $f \in C_b^{1,2}$

$$d\pi_t(f) = \pi_s \cdot q_s(Lf)(s) + \text{cov}' \pi_s(f, h) db_s$$

where b is a Brownian motion.

Girsanov's theorem implies that, under the change of probability $P^0 = \mu_T \cdot P$, where μ_T satisfies to $d\mu_t = -\mu_t \langle \text{cov } \pi_t(f, h), db_t \rangle$, the term $(\Omega, \underline{F}_t, P, b_t + \int_0^t \pi_s(h) ds, q, \pi_t, P^0)$ is an S-control, under the hypothesis of weak uniqueness of the solution to (KS). Under the assumption that the coordinates of h belong to $C_b^{1,2}$ (we call this condition the *robustness hypothesis*), we can give a weaker form to this problem. We suppose only that π_t is a $\mathcal{P}(\mathbb{R}^d)$ -valued process which satisfies the following: for each $f \in C_b^{1,2}$

$$M_t(\pi, f) := \pi_t(f) - \pi_r(f) - \int_r^t \pi_s \cdot q_s(Lf)(s) ds$$

is a martingale which admits $A_t(\pi, f) := \int_r^t |\text{cov } \pi_s(f, h)|^2 ds$ as an increasing process.

This assertion is a consequence of the following proposition.

PROPOSITION 6.8. *We assume the robustness hypothesis on h . Let $(\Omega, \underline{F}_t, P)$ be a probability space, and let π_t be a $\mathcal{P}(\mathbb{R}^d)$ -valued process defined on Ω which is continuous and which verifies the following: for each $f \in C_b^{1,2}(\mathbb{R}^d)$,*

$$M_t(\pi, f) := \pi_t(f) - \pi_r(f) - \int_r^t \pi_s \cdot q_s(Lf)(s) ds$$

is a local martingale with an increasing process equal to

$$A_t(\pi, f) := \int_r^t |\text{cov } \pi_s(f, h)|^2 ds.$$

Then there exist a larger space $(\hat{\Omega}, \hat{\underline{F}}_t, \hat{P})$ in the form $(\Omega \times \Omega_1, \underline{F}_t \otimes \underline{F}_t^1, P \otimes P_1)$ and a Brownian motion b_t on $\hat{\Omega}$ such that, for each $f \in C_b^{1,2}$,

$$M_t(\pi, f) = \int_r^t \langle \text{cov } \pi_s(f, h), db_s \rangle.$$

Proof. (a) The proof is easy in the one-dimensional case. The robustness hypothesis on h implies that $M_t(\pi, h)$ is a continuous martingale with an increasing process which satisfies $dA_t(\pi, h) = \text{var } \pi_t(h)^2 dt$. The theory of the stochastic integral ensures that the martingale b_t defined by

$$db_t = \frac{1}{\text{var } \pi_t(h)} 1_{\{\text{var } \pi_t(h) \neq 0\}} dM_t(\pi, h) + 1_{\{\text{var } \pi_t(h) = 0\}} d\tilde{b}_t$$

where \tilde{b}_t is a Brownian independent of $(\Omega, \underline{F}_t, P)$ (constructed on $(\Omega_1, \underline{F}_t^1, P)$), is a Brownian motion, and $dM_t(\pi, h) = \text{var } \pi_t(h) db_t$. A simple calculus derived from the formula

$$d(A_t(\pi, f+g) - A_t(\pi, f) - A_t(\pi, g)) = 2 \text{cov } \pi_s(f, h) \text{cov } \pi_s(g, h) ds$$

implies that the martingales $M_t(\pi, f)$ and $M_t(\pi, h)$ have a Meyer process equal to $\text{cov } \pi_s(f, h) \text{var } \pi_s(h) ds$. Hence we can compare $M_t(\pi, f)$ and $Z_t = \int_r^t \text{cov } \pi_s(f, h) db_s$. Let us denote by $\langle\langle \cdot, \cdot \rangle\rangle$ the Meyer process of two martingales:

$$d\langle\langle M_t(\pi, f) - Z_t, M_t(\pi, f) - Z_t \rangle\rangle = 2 \text{cov}^2 \pi_t(f, h) dt - 2d\langle\langle M_t(\pi, f), Z_t \rangle\rangle$$

and

$$\begin{aligned} d\langle\langle M_t(\pi, f), Z_t \rangle\rangle &= \text{cov } \pi_t(f, h) d\langle\langle M_t(\pi, f), b_t \rangle\rangle \\ &= \text{cov } \pi_t(f, h) \frac{1}{\text{var } \pi_t(h)} 1_{\{\text{var } \pi_t(h) \neq 0\}} d\langle\langle M_t(\pi, f), M_t(\pi, h) \rangle\rangle \end{aligned}$$

(we have used \tilde{b}_t independent of $M_t(\pi, f)$). Hence

$$d\langle\langle M_t(\pi, f), Z_t \rangle\rangle = \text{cov}^2 \pi_t(f, h) 1_{\{\text{var } \pi_t(h) \neq 0\}} dt.$$

Since $\text{cov } \pi_t(f, h)$ is equal to zero on $\{\text{var } \pi_t(h) = 0\}$, we have $d\langle\langle M_t(\pi, f), Z_t \rangle\rangle = \text{cov}^2 \pi_t(f, h) dt$. Therefore, the increasing process associated with $M_t(\pi, f) - Z_t$ is equal to zero, and the two martingales are indistinguishable. \square

(b) Let us now solve the general case. The robustness hypothesis on h implies that $M_t(\pi, h)$ is a k -dimensional martingale and that its matricial increasing process

is $D_t(\pi, h)$ with general term equal to $[D_t(\pi, h)]_{i,j} := \int_r^t d_s(\pi, h)_{i,j} ds$, where

$$d_s(\pi, h)_{i,j} = \sum_{\alpha=1}^k \text{cov } \pi_s(h_i, h_\alpha) \text{cov } \pi_s(h_\alpha, h_j).$$

Indeed, it suffices to note that $D_t(\pi, h)$ is the matrix of the quadratic form $\theta \rightarrow A_t(\pi, \langle \theta, h \rangle)$. The matrix $d_s(\pi, h)$ admits a square root with general term $\text{cov } \pi_s(h_i, h_j)$, i.e., the covariance matrix of h under probability π_s . We denote this matrix by $K_{\pi_s}(h)$. Since $K_{\pi_s}(h)$ is a symmetric matrix, it admits a representation of the form $K_{\pi_s}(h) = H_s \Delta_{\pi_s}(h) H_s^*$, where H_s and H_s^* satisfy $H_s H_s^* = I$ and $\Delta_{\pi_s}(h)$ is a diagonal matrix. The theory of stochastic integration proves that $\int_r^t H_s^{-1} dM_s(\pi, h)$ is a martingale associated with an increasing process in the form $\int_r^t H_s^{-1} K_{\pi_s}^2(h) (H_s^{-1})^* ds = \int_r^t \Delta_{\pi_s}^2(h) ds$. Since $\Delta_{\pi_s}^2(h)$ is a diagonal matrix, the coordinates of $\int_r^t H_s^{-1} dM_s(\pi, h)$ are martingales which are pairwise orthogonal. We write $\Delta_{\pi_s}(h)$ in the form

$$\begin{bmatrix} \lambda_s^1(h) & & 0 \\ & \ddots & \\ 0 & & \lambda_s^k(h) \end{bmatrix}.$$

Let us denote by $(\tilde{b}_j)_{j \leq k}$ k -Brownian motions independent of $(\Omega, \underline{F}_t, P)$, which are constructed on a probability space $(\Pi_1, \underline{F}_t^1, P_1)$. We set

$$\hat{b}_j(t) = \int_r^{t+r} \frac{1}{\lambda_s^j(h)} 1_{\{\lambda_s^j(h) \neq 0\}} (H_s^{-1} dM_s(\pi, h))_j + \int_r^{t+r} 1_{\{\lambda_s^j(h) = 0\}} d\tilde{b}_s^j.$$

An easy calculation gives that $\hat{b}_j(t)$ is a Brownian motion which is $\hat{\underline{F}}_{t+r}$ -adapted ($\hat{\underline{F}}_{t+r} = \underline{F}_{t+r} \otimes \underline{F}_{t+r}^1$) and that $H_s^{-1} dM_s(\pi, h) = \Delta_{\pi_s}(h) d\hat{b}_s$, i.e.,

$$dM_s(\pi, h) = K_{\pi_s}(h) db_s \quad \text{if} \quad db_s = H_s d\hat{b}_s.$$

Then b_t is a Brownian motion. \square

We now prove that $dM_s(\pi, f) = \text{cov } \pi_s(f, h) db_s$. First, we remark that, if f and g belong to $C_b^{1,2}$, the Meyer process of $M_t(\pi, f)$ and $M_t(\pi, g)$, which is equal to $C_t(\pi, f, g) = \frac{1}{2}\{A_t(\pi, f+g) - A_t(\pi, f) - A_t(\pi, g)\}$, can easily be calculated:

$$\begin{aligned} dC_t(\pi, f, g) &= \frac{1}{2}\{\langle \text{cov } \pi_t(f+g, h), \text{cov } \pi_t(f+g, h) \rangle \\ &\quad - \langle \text{cov } \pi_t(f, h), \text{cov } \pi_t(f, h) \rangle - \langle \text{cov } \pi_t(g, h), \text{cov } \pi_t(g, h) \rangle\} dt \\ &= \langle \text{cov } \pi_t(f, h), \text{cov } \pi_t(g, h) \rangle dt. \end{aligned}$$

Let Z_t be equal to $\int_r^t \langle \text{cov } \pi_s(f, h), db_s \rangle$. In order to prove that $M_t(\pi, f)$ and Z_t are indistinguishable, it suffices to show that the increasing process of $M_t(\pi, f) - Z_t$ is equal to zero. Notice that $M_t(\pi, f)$ and Z_t are continuous local martingales with the same increasing process $\int_r^t \langle \text{cov } \pi_s(f, h), \text{cov } \pi_s(f, h) \rangle ds$. It remains to study the Meyer process of $M_t(\pi, f)$ and Z_t . We denote this process by $d\langle\langle M_t(\pi, f), Z_t \rangle\rangle$ or $d\langle\langle M_t(\pi, f), dZ_t \rangle\rangle$:

$$\begin{aligned} d\langle\langle M_t(\pi, f), Z_t \rangle\rangle &= \langle\langle dM_t(\pi, f), \langle \text{cov } \pi_t(f, h), db_t \rangle \rangle\rangle \\ &= \langle\langle dM_t(\pi, f), \langle \text{cov } \pi_t(f, h), H_t d\hat{b}_t \rangle \rangle\rangle \\ &= \langle\langle dM_t(\pi, f), \langle H_t^{-1} \text{cov } \pi_t(f, h), d\hat{b}_t \rangle \rangle\rangle \\ &= \langle\langle dM_t(\pi, f), \langle H_t^{-1} \text{cov } \pi_t(f, h), \Delta_{\pi_t}^{-1}(h) H_t^{-1} dM_t(\pi, h) \rangle \rangle\rangle \end{aligned}$$

with the convention that $1/0 = 0$ for the eigenvalues of $\Delta_{\pi_t}^{-1}(h)$, since \tilde{b}_t is independent of $M_t(\pi, f)$. We must now calculate $d\langle\langle M_t(\pi, f), M_t(\pi, h) \rangle\rangle$. This term is equal to

$\sum \text{cov } \pi_t(f, h_i) \text{cov } \pi_t(h_i, h_j) dt$. We can write this equality in a vectorial form: $d\langle M_t(\pi, f), M_t(\pi, h) \rangle = K_{\pi_t}(h) \text{cov } \pi_t(f, h) ds$. We substitute this last equality in the calculation above:

$$\begin{aligned} d\langle M_t(\pi, f), Z_t \rangle &= \langle H_t^{-1} \text{cov } \pi_t(f, h), \Delta_{\pi_t}^{-1}(h) H_t^{-1} K_{\pi_t}(h) \text{cov } \pi_t(f, h) \rangle dt \\ &= \langle \text{cov } \pi_t(f, H_t^{-1}(h)), \Delta_{\pi_t}^{-1}(h) H_t^{-1} H_t \Delta_{\pi_t}(h) \text{cov } \pi_t(f, H_t^{-1}(h)) \rangle dt. \end{aligned}$$

Denote by $I_t(\pi, h)$ the $k \times k$ diagonal matrix, with diagonal terms equal to $1_{\{\lambda_i^t=0\}}$. Then $\Delta_{\pi_t}^{-1}(h) H_t^{-1} H_t \Delta_{\pi_t}(h) = Id - I_t(\pi, h)$. This quite surprising equality results from the convention $1/0=0$ for the eigenvalues of $\Delta_{\pi_t}^{-1}(h)$. Moreover, under the probability π_t , the vector $H_t^{-1}h$ has a covariance matrix equal to $H_t^{-1}K_{\pi_t}(h)H_t = \Delta_{\pi_t}(h)$. Hence the vector $I_t(\pi, h)H_t^{-1}h$ has a covariance matrix equal to zero (under the probability π_t). It follows that $I_t(\pi, h) \text{cov } \pi_t(f, H_t^{-1}h) = \text{cov } \pi_t(f, I_t(\pi, h)H_t^{-1}h) = 0$. We come back to the previous calculation:

$$\begin{aligned} d\langle M_t(\pi, f), Z_t \rangle &= \langle \text{cov } \pi_t(f, H_t^{-1}h), (Id - I_t(\pi, h)) \text{cov } \pi_t(f, H_t^{-1}h) \rangle dt \\ &= \langle \text{cov } \pi_t(f, H_t^{-1}h), \text{cov } \pi_t(f, H_t^{-1}h) \rangle dt \\ &= \langle \text{cov } \pi_t(f, h), \text{cov } \pi_t(f, h) \rangle dt \end{aligned}$$

since H_t is an orthogonal matrix.

We have just established $d\langle M_t(\pi, f), Z_t \rangle = A_t(\pi, f) = \langle Z, Z \rangle_t$. Hence the increasing process of the martingale $M_t(\pi, f) - Z_t$ is equal to zero and the desired theorem follows. \square

THEOREM 6.9. Assume the weak uniqueness of the solution to (KS) and the robustness hypothesis on h , i.e., the coordinates of h belong to $C_b^{1,2}$. Let \mathcal{B} be the set of elements of the form $(\Omega, \underline{F}_t, q, \pi_t, P)$:

$$\begin{aligned} \pi_r &= \mu, \\ \pi_t(f) - \pi_r(f) - \int_r^t \pi_s \cdot q_s(Lf)(s) ds &\text{ is a } P\text{-martingale} \\ \text{with an increasing process equal to } \int_r^t |\text{cov } \pi_s(f, h)|^2 ds. \end{aligned}$$

Then

$$\begin{aligned} \sup \left\{ P \left(\int_r^T q_s(K(s, \pi_s)) ds + G(\pi_T) \right); P \in \mathcal{B} \right\} &= V(r, \mu) \\ &= \sup \{ J(r, \mu, \hat{Q}), \hat{Q} \in \hat{\mathcal{Q}}(r, \mu) \}. \end{aligned}$$

We have defined the same control problem.

Proof. The proof is immediate from Proposition 6.8. \square

Under the hypotheses of Theorem 6.9, the optimal Markovian filter can be associated with a control which is a function, which is Markovian given the filter.

THEOREM 6.10. Assume the same hypotheses as in Theorem 6.9. Let $\hat{Q}_{(r, \mu)}^*$ be a family of optimal rules such that $(\Pi, \underline{\Pi}_t, \pi_t, \hat{Q}_{(r, \mu)}^*)$ is a strong Markovian process. Then, there exists a function $q^*: \mathbb{R}^+ \times \mathcal{P}(\mathbb{R}^d) \rightarrow \mathcal{P}(A)$ which is measurable and such that $(\pi_t, q^*(t, \pi_t), \hat{Q}_{(r, \mu)}^*)$ is a separated filter.

Proof. Under these hypotheses, $\pi_t(f)$ is a semimartingale and an additive functional for each law $\hat{Q}_{(r, \mu)}^*$. Then there exists for each $f \in \mathcal{C}_b^{1,2}$ a function $\beta(f): \mathbb{R}^+ \times \mathcal{P}(\mathbb{R}^d) \rightarrow \mathbb{R}$ such that

$$(6.10.1) \quad \pi_t(f) - \int_r^t \beta(f)(s, \pi_s) ds \text{ is a } \hat{Q}_{(r, \mu)}^* \text{ martingale.}$$

Thus $\hat{Q}_{(r,\mu)}^* \otimes dt$ almost surely $\beta(f)(s, \pi_s) = \pi_s \cdot q_s(Lf)$. Let us denote by B the subset of $(\mathcal{P}(\mathbb{R}^d) \times V)$ defined by $B := \{(m, q) \mid \text{for all } f: \beta(f)(s, m) = m \cdot q_s(Lf)\}$. Obviously it is sufficient to check that 6.10.1 holds for the functions $f_i(x) = e^{(\theta_i, x)}$, where θ_i is a denumerable dense set in \mathbb{R}^d . Since $\{m, q \mid \beta(f_i)(s, m) = m \cdot q_s(Lf_i)\}$ is measurable, the set B is also measurable. From a selection theorem [DY, p. 57], [CV, p. 65] there exists a measurable map $m \rightarrow q_s^*(m)$ such that $\{m, q_s^*(m)\} \in B$. From Theorem 6.9, $\pi_t(f)$ is a solution to (KS) associated with $q^*(s, \pi_s)$ and thus, from the uniqueness hypothesis, it is a separated filter. \square

7. Comments on the uniqueness of the solution to Zakai's equation. The problem of uniqueness of the solution to Zakai's equation has been studied by many authors. Here we only discuss the different results that they have obtained.

In the case where the signal process is an \mathbb{R}^d -valued diffusion with coefficients independent of the control, most recent work has led to the study of the existence of unique solution of partial differential equation (P.D.E.) satisfied by the unnormalized conditional density of the filtering problem. This P.D.E. is not a stochastic equation but an "ordinary" differential equation in which the observed path occurs as a parameter in the coefficients. Davis [Da1] calls this problem "the pathwise filtering problem." The uniqueness result is established under hypotheses on the coefficients: we require ellipticity or (and) coercivity of the matrix $a_{i,j}$ and boundedness assumptions on h and its derivatives (robustness conditions) (Pardoux [Pa2]). When the signal is a Markov process, Kunita [Ku1] also establishes the uniqueness of the solution to Zakai's equation. The boundedness hypothesis on h is too strong (it is not verified in the linear case): if h and its derivatives have a linear (or polynomial) growth, Blankenship, Baras and Hopkins [BBH], Pardoux [Pa3] and Sheu [Sh] establish the uniqueness. Kallianpur and Karandikar [KK] require Hölder conditions on h . Ferreyra [Fe1], [Fe2] has also studied his problem with various hypotheses on the coefficients (e.g., σ may be degenerate). In all these studies, this is the "pathwise" uniqueness which is proved. Kurtz and Ocone [KO] use a martingale problem in order to study Zakai's equation. They assume that the signal process X is a solution to the martingale problem for (A, μ) , where A is a linear operator which satisfies the so-called Echeverria conditions. Under some supplementary hypotheses on h , they establish the uniqueness in law.

The case where there is control in the coefficients of the signal process is studied by Bensoussan [Be] and is generalized by Haussmann [Ha2]. In Bensoussan's work, h is bounded and there is no control in the coefficient σ . We recall here the result obtained by Haussmann under the hypotheses

$$|\sigma(t, x, a) - \sigma(t, \bar{x}, a)| \leq C_R |x - \bar{x}|, \quad |a| \leq R,$$

$$(B_1) \quad |\sigma(t, x, a)| \leq K,$$

$$\sigma(t, x, a)\sigma'(t, x, a) \geq \alpha I$$

$$(B_2) \quad |b(t, x, a)| \leq K(1 + |x| + |a|),$$

$x \rightarrow b(t, x, a)$ is Lipschitzian uniformly with respect to (t, x, a) in a compact set,

$$(B_3) \quad |h(t, x)| + \left| \frac{\partial h}{\partial x_i}(t, x) \right| \leq K(1 + |x|),$$

$$\left| \frac{\partial h}{\partial t}(t, x) \right| \leq K(1 + |x|^2),$$

$$(B_4) \quad |u(t, y)| \leq K \left(1 + \sup_{s \leq t} |y_s| \right), \quad y \in C(0, T; \mathbb{R}^d).$$

Haussmann shows the uniqueness of the solution to the “robust Zakai equation”

$$(7.1) \quad \begin{aligned} \frac{d\mu_t}{dt} - L_{t,\eta}^{\nu*} \mu_t &= 0, \\ \mu_0 &= \rho_0 \end{aligned}$$

where

$$\begin{aligned} L_{t,\eta}^{\nu*} v &= L_{t,\eta}^* v + (\eta_t h_t(x))_{x_i} a_{i,j}(t, x, \eta) v_{x_j}(x) \\ &\quad + \{[(\eta_t h_t(x))_{x_i} a_{i,j}(t, x, \eta)]_{x_j} + \gamma(t, x, \eta)\} v(x) \end{aligned}$$

(η belongs to $C([0, T]; \mathbb{R}^d)$).

$L_{t,\eta}^* v = \frac{1}{2}(a_{i,j}v)_{x_i x_j} - (b_i v)_{x_i}$ is the adjoint to the generator $L_{t,\eta}$,

$$\begin{aligned} \gamma(t, x, \eta) &= \frac{1}{2} (\eta_t h_t(x))_{x_i} a_{i,j}(t, x, \eta) (\eta_t \cdot h_t(x))_{x_j} \\ &\quad - \eta_t \frac{\partial}{\partial t} h_t(x) - L_{t,\eta}(\eta_t \cdot h_t)(s) - \frac{1}{2} |h(t, x)|^2 \end{aligned}$$

where $v_{x_i} = \partial v / \partial x_i$.

If we denote by ψ^η the unique solution to this equation, $\rho_t^\eta(x) := \psi_t^\eta(x) \exp(\eta_t \cdot h_t(x))$ is the unnormalized conditional density of the filtering problem (and is the solution to Zakai's equation).

Haussmann then obtains the uniqueness of the solution to Zakai's equation for controls in the form $u(t, y)$. When the control is in the form $u(t, \pi_t)$, Benes and Karatzas [BK2] show the uniqueness of the solution when there is no control in σ , under robustness conditions on h .

Borkar [Bo] obtains results similar to those of Haussmann when there is no control in the term σ .

The arguments of Haussmann, which utilize the interpretation of the unique solution to the robust system (7.1) can easily be modified in order to study the case when the controls are relaxed and adapted with respect to the filtration $\bar{\mathbb{V}}_t$.

We change hypothesis (B_2) into

$$|b(t, x, a)| \leq K(1 + |x|),$$

$(B'_2) \quad x \rightarrow b(t, x, a)$ is Lipschitzian uniformly with respect to (t, a) and so are the partial derivatives;

and (B_3) becomes

$(B'_3) \quad h$ belongs to C^2 , with a linear growth and its partial derivatives too.

Under these hypotheses, Zakai's equation has a unique solution (uniqueness in law).

Appendix A. Here we recall only the definitions and the properties of set-valued maps. We state a series of definitions concerning the continuity of set-valued maps [AC, p. 41].

DEFINITION [AC, Def. 1, p. 41; Def. 2, p. 43]. Let X and Y be Hausdorff metric spaces, and \mathcal{R} a set-valued map with nonempty values of Y into subsets of X :

(a) \mathcal{R} is upper semicontinuous (u.s.c.) at $y_0 \in Y$ if, for any open N containing $\mathcal{R}(y_0)$, there exists a neighborhood M of y_0 such that $\{\mathcal{R}(y) : y \in M\} \subseteq N$.

(b) \mathcal{R} is lower semicontinuous (l.s.c.) at $y_0 \in Y$ if, for any $x_0 \in \mathcal{R}(y_0)$ and any neighborhood $N(x_0)$ of x_0 , there exists a neighborhood $M(y_0)$ of y_0 such that for all $y \in M(y_0)$, $\mathcal{R}(y) \cap N(x_0) \neq \emptyset$.

(c) \mathcal{R} is continuous at y_0 if it is u.s.c. and l.s.c. at y_0 .

REMARKS. The spaces X and Y are metric. Therefore, condition (b) is equivalent to the following: given any sequence y_n converging to y and any $x \in \mathcal{R}(y)$ there exists a sequence $x_n \in \mathcal{R}(y_n)$ converging to x .

The graph of a u.s.c. set-valued map with closed values from Y to X is closed, i.e., if y_n converges to y and if $x_n \in \mathcal{R}(y_n)$ converges to x , then x belongs to $\mathcal{R}(y)$ [AC, Prop. 2, p. 4]. A set-valued map from Y to a compact set X whose graph is closed is a u.s.c. map [AC, Cor. 1, p. 42]. Moreover, it is a measurable map from Y into $\text{comp}(X)$, where $\text{comp}(X)$ is the space of all compact subsets of X (see Stroock and Varadhan [SV, Chap. 12]).

THEOREM [AC, pp. 51–53]. Let X and Y be separable metric spaces, let \mathcal{R} be a set-valued map from Y to X , and let w be a function from $Y \times X$ to \mathbb{R} . The marginal function $v(y) := \sup \{w(x, y) : x \in \mathcal{R}(y)\}$ and the marginal map $y \mapsto M_y := \{x \in \mathcal{R}(y) : w(x, y) = v(y)\}$ have the following properties:

(a) If w is l.s.c. on $Y \times X$ and \mathcal{R} is l.s.c., then v is l.s.c.

(b) If w is u.s.c. on $Y \times X$ and \mathcal{R} is u.s.c. with compact values, then v is u.s.c.

(c) If w is continuous and \mathcal{R} is continuous with compact values, then the marginal function v is continuous and the marginal set-valued map M is u.s.c.

If we are interested in measurability results, we can weaken the hypotheses. The proof of the following theorems can be found in Castaing and Valadier [CV, p. 86], Dellacherie and Meyer [DM1, Chap. 3], or Stroock and Varadhan [SV, Chap. 12].

PROPOSITION. Let w be a u.s.c. function on $Y \times X$ and let $y \mapsto \mathcal{R}(y)$ be a measurable map of Y into $\text{comp}(X)$. Then we have the following.

(a) The marginal function v is a Borel function and the marginal map M is measurable.

(b) For each probability measure μ on Y

$$\int \mu(dy) v(y) = \sup \left\{ \int \mu(dy) w(y, h(y)), \text{ where } h \text{ varies} \right. \\ \left. \text{in the set of measurable map of } Y \text{ into } X \right. \\ \left. \text{such that } h(y) \in \mathcal{R}(y) \text{ for all } y \in Y \right\}.$$

THEOREM. Let (Y, \mathcal{Y}) be a measurable space and let $\hat{\mathcal{Y}}$ be the universal completion of \mathcal{Y} . Let X be a Polish space endowed with the Borelian σ -algebra $\mathcal{B}(X)$. Let w be a measurable function from $Y \times X$ to \mathbb{R} and let \mathcal{R} be a set-valued map with nonempty values such that $\text{graph } \mathcal{R} \in \hat{\mathcal{Y}} \otimes \mathcal{B}(X)$. Then we have the following:

(a) The marginal function v is $\hat{\mathcal{Y}}$ -measurable;

(b) The graph of the marginal map M is $\hat{\mathcal{Y}} \otimes \mathcal{B}(X)$ -measurable.

Appendix B. We again state Theorem 2.2 and we give its proof.

THEOREM 2.2. There exists a sequence (ψ_k) of measurable functions from V to V^0 , which are adapted (i.e., $(\psi_k)^{-1}(\underline{V}_t) \subset \underline{V}_t$) such that $\psi_k(q)$ converges to q for each $q \in V$ as k converges to ∞ .

Proof. (i) The spaces V_T (sets of measures on $[0, T] \times A$ whose projection on $[0, T]$ is Lebesgue measure) have been studied most thoroughly. The proof of their

compact metrizable properties can be found in [GH] or [Wa, Chap. IV, 2.6]. A projective limit argument allows us to extend these properties to V .

(ii) The density of atomic measures (with support in $[0, T]$) in V_T is proved by an explicit construction in [F11] and [FN]; this construction is described by the “chattering” lemma. We use the same kind of arguments to construct the applications ψ_k of the lemma. The space A being metric compact, we choose a partition $(B_k^{(n)}, k = 1, \dots, k(n))$ into Borelians with diameter less than $1/n$. We denote by $a_k^{(n)}$ a point of $B_k^{(n)}$. We divide $[0, n]$ into intervals $T_j^{(n)}$ of length 2^{-n} , with endpoints $(t_j^{(n)}, t_{j+1}^{(n)})$.

For the sake of simplicity, we omit the superscript n where there is no ambiguity. Let δ be a fixed number $\delta > 0$. In order to construct adapted approximations, we introduce the V_T -adapted processes

$$\begin{aligned} q_k^\delta(t) &= \frac{1}{\delta} \int_{t-\delta}^t q(s, B_k) ds \quad \text{if } t > \delta \\ &= \frac{1}{\delta} \int_0^t q(s, B_k) ds \quad \text{if } t \leq \delta. \end{aligned}$$

We also omit the superscript δ in $q_k^\delta(t)$. The step process $\gamma_{j,k}(t, q) = q_k(t_j)1_{t_j \leq t < t_{j+1}}$ is adapted and satisfies the relation

$$\text{for all } j: \sum_k \gamma_{j,k}(t, q) = 1 \quad \text{if } t \in [t_j, t_{j+1}[\quad \text{and} \quad t_j \geq \delta.$$

We now divide the interval $T_j = [t_j, t_{j+1}[$ into $k(n)$ intervals $T_{j,k}(q)$ of length $(t_{j+1} - t_j)$, where t is any point in T_j . For each n , there exists an index j_n such that $\delta \leq t_{j_n} < \delta + 2^{-n}$. We will denote t_{j_n} by ε . Notice that $\delta \leq \varepsilon \leq \delta + 2^{-n}$. Let $\beta(n, \delta)$ be an arbitrary point in A and set $\alpha(t) = a_k$ (the point in B_k^n that we introduced at the beginning of the proof): if $t \in \bigcup_{j \geq j_n} T_{j,k}(q)$ for $k = 1, 2, \dots, k(n)$, and $\alpha(t) = \beta(n, \delta)$ if $t \notin]\varepsilon, n[$; and let $\psi^{(n,\delta)}(q) = 1_{] \varepsilon, n[}(t) dt \delta_{\alpha(t)} + 1_{[0, \varepsilon]}(t) dt \delta_{\beta(n, \delta)}$, where δ is the Dirac measure on A . All the operations that we have performed are adapted; therefore $\psi^{(n,\delta)}$ is also adapted.

It remains now to show that $\psi^{(n,\delta)}(q)$ converges vaguely to q . We check the convergence for functions f, g , where f is continuous with support in $[0, T]$ and g is continuous on A . Their moduli of continuity are denoted by ω_f and ω_g . If T is bounded by n ,

$$\begin{aligned} & \int_{[0, T] \times A} f(t)g(a) dt q(t, da) - \int_{[0, T] \times A} f(t)g(a) dt \psi^{(n,\delta)}(q)(da) \\ (A1) \quad &= \int_{[0, T] \times A} f(t)g(a) dt q(t, da) - \int_{[0, \varepsilon]} f(t)g(\beta(n, \delta)) dt \\ & \quad - \int_{] \varepsilon, T]} f(t)g(\alpha(t)) dt. \end{aligned}$$

The modulus of this quantity is bounded by

$$\begin{aligned} (A2) \quad & 2\varepsilon \|f\|_\infty \|g\|_\infty + \int_{[0, T]} f(t) \left(\sum_k \int_{B_k} g(a) q(t, da) - g(a_k) q(t, B_k) \right) dt \\ & + \left| \sum_k g(a_k) \left(\int_{] \varepsilon, T]} f(t) q(t, B_k) dt - \int_{\bigcup_j T_{j,k}} f(t) dt \right) \right|. \end{aligned}$$

The second term in this expression is bounded by $\omega_g(1/n)\|f\|_1$. The quantity inside the absolute value in the third term is equal to

$$\begin{aligned}
 & \sum_k g(a_k) \left(\int_{] \varepsilon, T]} f(t) \{q(t, b_k) - q_k(t)\} dt + \sum_j \int_{T_j} f(t) q_k(t) dt - \int_{T_{j,k}} f(t) dt \right) \\
 &= \sum_k g(a_k) \left(\int_{] \varepsilon, T]} f(t) \{q(t, B_k) - q_k(t)\} dt \right) \\
 (A3) \quad & + \sum_k g(a_k) \sum_j f(t_j) \int_{T_j} \{q_k(t) - \gamma_{j,k}(t)\} dt \\
 & - \sum_k g(a_k) \sum_j \int_{T_{j,k}} \{f(t) - f(t_j)\} dt \\
 & + \sum_k g(a_k) \sum_j \int_{T_j} \{f(t) - f(t_j)\} q_k(t) dt.
 \end{aligned}$$

The absolute value of the two last terms is bounded by $2\|g\|_\infty T\omega_f(1/2^n)$.

We now compute the first term of (A3). An easy calculation gives

$$\begin{aligned}
 & \int_\varepsilon^T f(t) \left(q(t, B_k) - \frac{1}{\delta} \int_{t-\delta}^t q(s, B_k) ds \right) dt \\
 (A4) \quad &= \int_\varepsilon^T \left(f(t) q(t, B_k) - \frac{1}{\delta} \int_{t-\delta}^t f(s) q(s, B_k) ds \right) dt \\
 &+ \int_\varepsilon^T \frac{1}{\delta} dt \int_{t-\delta}^t q(s, B_k) (f(s) - f(t)) ds.
 \end{aligned}$$

We denote the first term of the right-hand side of (A4) by C_1^k and the second by C_2^k . By Fubini's theorem, the absolute value of C_2^k is bounded by $\omega_f(\delta) \int_0^T q(s, B_k) ds$ and the absolute value of $\sum_k g(a_k) C_2^k$ is bounded by $T\|g\|_\infty \omega_f(\delta)$. We can write C_1^k in the following form:

$$\int_{T-\delta}^T f(s) q(s, B_k) \left(1 - \frac{T-s}{\delta} \right) ds + \int_{\varepsilon-\delta}^\varepsilon f(s) q(s, B_k) \frac{s+\delta-\varepsilon}{\delta} ds;$$

$\sum_k g(a_k) C_1^k$ is bounded by $\|g\|_\infty \|f\|_\infty 2\delta$. We have proved that the first term of (A3) is bounded by $\|g\|_\infty (2\|f\|_\infty \delta + T\omega_f(\delta))$.

We now compute the second term of (A3):

$$\begin{aligned}
 & \sum_k g(a_k) \sum_j f(t_j) \int_{T_j} (q_k(t) - \gamma_{j,k}(t)) dt \\
 &= \sum_k g(a_k) \sum_j f(t_j) \int_{T_j} (q_k(t) - q_k(t_j)) dt \\
 &= \sum_k g(a_k) \sum_j f(t_j) \frac{1}{\delta} \int_{T_j} dt \int_{t_j}^t (q(s, B_k) - q(s-\delta, B_k)) ds = C_3.
 \end{aligned}$$

Integration by parts yields

$$\begin{aligned}
 C_3 &= \sum_k g(a_k) \sum_j f(t_j) \int_{T_j} \frac{1}{\delta} (t_{j+1} - t) (q(t, B_k) - q(t-\delta, B_k)) dt \\
 &= \sum_j f(t_j) \int_{T_j} \frac{1}{\delta} (t_{j+1} - t) \left(\sum_k g(a_k) (q(t, B_k) - q(t-\delta, B_k)) \right) dt.
 \end{aligned}$$

Hence

$$(A5) \quad |C_3| < 2\|g\|_\infty\|f\|_\infty(2^{-2n}2^n n)/2\delta.$$

We gather together the different results (A1)–(A5):

$$\begin{aligned} |q(fg) - \psi^{(n,\delta)}(q)(fg)| \leq \|g\|_\infty \left\{ 2T\omega_f\left(\frac{1}{2^n}\right) + 2\varepsilon\|f\|_\infty + 2\delta\|f\|_\infty \right. \\ \left. + T\omega_f(\delta) + 2\|f\|_\infty 2^{-n} \frac{n}{2\delta} \right\} + \|f\|_1 \omega_g\left(\frac{1}{n}\right). \end{aligned}$$

It remains to set $\delta = 1/n$ and $\psi_n = \psi^{n,(1/n)}$ to get the expected result.

Acknowledgments. The authors thank the referees for the careful reading of this paper and for the remarks they have made.

REFERENCES

- [AC] J. P. AUBIN AND A. CELINA, *Differential Inclusions*, Springer-Verlag, Berlin, New York, 1984.
- [BBH] J. S. BARAS, G. L. BLANKENSHIP, AND W. E. HOPKINS, *Existence, uniqueness and asymptotic behavior of solutions of Zakai's equation with unbounded coefficients*, IEEE Trans. Automat. Control, 28 (1983), pp. 203–214.
- [BK1] V. E. BENES AND I. KARATZAS, *On the relation of Zakai's and Mortensen's equations*, SIAM J. Control Optim., 21 (1983), pp. 472–489.
- [BK2] ———, *Filtering of diffusions controlled through their conditional measures*, Stochastics, 13 (1984), pp. 1–23.
- [Be] A. BENSOUSSAN, *Maximum principle and dynamic programming: approaches of the optimal control of partially observed diffusions*, Stochastics, 9 (1983), pp. 169–222.
- [Bi] J. M. BISMUT, *Partially observed diffusions and their control*, SIAM J. Control Optim., 20 (1982), pp. 302–309.
- [Bo] V. S. BORKAR, *Existence of optimal control for partially observed diffusions*, Stochastics, 11 (1983), pp. 103–141.
- [BY] P. BREMAUD AND M. YOR, *Changes of filtration and of probability measures*, Z. Wahrsch. Verw. Gebiete, 45 (1978), pp. 269–295.
- [Ch] N. CHRISTOPEIT, *Existence of optimal stochastic controls under partial observations*, Z. Wahrsch. Verw. Gebiete, 51 (1980), pp. 201–213.
- [CH] N. CHRISTOPEIT AND K. HELMES, *Optimal Control for a Class of Partially Observable Systems*, Lecture Notes in Control and Information Sci. 61, Springer-Verlag, Berlin, New York, 1983, pp. 36–60.
- [CK] N. CHRISTOPEIT AND M. KOHLMANN, *Some Recent Results on the Control of Partially Observable Stochastic Systems*, Stochastic Differential Systems, Bonn, 1982; Lecture Notes in Control and Information Sci. 43, Springer-Verlag, Berlin, New York, 1982.
- [CV] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 580, Springer-Verlag, Berlin, New York; 1977.
- [Da1] M. H. DAVIS, *Pathwise Nonlinear Filtering*, in Stochastic Systems, M. Hazewinkel, ed., NATO Advanced Study Institute Series, Reidel, Les Arcs, France, 1980.
- [Da2] ———, *Lectures on Stochastic Control and Nonlinear Filtering*, Tata Institute of Fundamental Research, 1984; Springer-Verlag, Berlin, New York, 1984.
- [DK] M. H. DAVIS AND M. KOHLMANN, *On the nonlinear semigroup of stochastic control under partial observations*, preprint.
- [DM1] C. DELLACHERIE AND P. A. MEYER, *Probabilités et Potentiel*, Chap. I–IV, Hermann, Paris, 1975.
- [DM2] ———, *Probabilités et Potentiel*, Chap. IV–VII, Hermann, Paris, 1980.
- [DV] T. DUNCAN AND P. VARAIYA, *On the solutions of a stochastic control system*, SIAM J. Control, 9 (1971), pp. 354–371.
- [DY] DYNKIN AND YISHKEVICH, *Control Markov Processes*, Springer-Verlag, Berlin, New York, 1979.
- [EK] N. EL KAROUI, *Les aspects probabilistes du contrôle stochastique*, Ecole d'été de Saint-Flour, 1979; Lecture Notes in Math. 876, Springer-Verlag, Berlin, New York, 1981, pp. 74–239.

- [EK] N. EL KAROUI, *Partially observable control of diffusions with correlated noise*, in Proc. EISENHACH, 1986. Lecture Notes and Control and Information Sci., Springer-Verlag, Berlin, New York, to appear.
- [EHJ] N. EL KAROUI, D. HUU NGUYEN, AND M. JEANBLANC-PICQUÉ, *Compactification methods in the control of degenerate diffusions: existence of an optimal control*, Stochastics, 20 (1987), pp. 169–219.
- [ElKo] R. J. ELLIOTT AND M. KOHLMANN, *On the existence of partially observable controls*, Appl. Math. Optim., 9 (1982), pp. 41–66.
- [ELM] N. EL KAROUI, J. P. LEPETIER, AND B. MARCHAL, *Nisio Semigroup Associated to the Control of Markov Processes*, Lecture Notes in Control and Information Sci. 61, Springer-Verlag, Berlin, New York, 1983.
- [Fe1] G. S. FERREYRA, *The Robust Equation Approach to Multidimensional Stochastic Non-linear Filtering*, Probability Theory on Vector Spaces 3, Lublin, 1983; Lecture Notes in Math., Springer-Verlag, Berlin, New York, 1984.
- [Fe2] ———, *On the degenerate parabolic partial differential equations of non-linear filtering*, Comm. Partial Differential Equations, 10 (1985), pp. 555–634.
- [Fl1] W. H. FLEMING, *Nonlinear semigroup for controlled partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 286–301.
- [Fl2] ———, *Generalized Solutions in Optimal Stochastic Control. Differential Games and Control Theory*, Proc. Second Kingston Conference; Lectures Notes in Pure and Applied Math. 30, Marcel Dekker, New York, 1976.
- [Fl3] ———, *Measure-valued processes in the control of partially observable stochastic systems*, Appl. Math. Optim., 6 (1980), pp. 271–285.
- [FM] W. H. FLEMING AND S. K. MITTER, *Optimal control and nonlinear filtering for nondegenerate diffusion processes*, Stochastics, 8 (1982), pp. 63–78.
- [FN] W. H. FLEMING AND N. NISIO, *On stochastic relaxed control for partially observed diffusions*, Nagoya Math. J., 93 (1984), pp. 71–108.
- [FP] W. H. FLEMING AND E. PARDOUX, *Optimal control for partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 261–285.
- [FV] W. H. FLEMING AND M. VIOT, *Some measure valued Markov processes in population genetics theory*, Indiana J. Math., 28 (1979), pp. 817–843.
- [FuNi] Y. FUJITA AND M. NISIO, *Nonlinear semigroups associated with optimal stopping of controlled diffusions under partial observations*, preprint.
- [FKK] M. FUJISAKI, G. KALLIANPUR, AND H. KUNITA, *Stochastic differential equations for the nonlinear filtering problem*, Osaka J. Math., 9 (1972), pp. 19–40.
- [GH] A. GHOULA-HOURI, *Sur la généralisation de la notion de commande d'un système guidable*, RIRO, 4 (1967), pp. 7–32.
- [Ha] U. G. HAUSMANN, *On the existence of optimal controls for partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 385–407.
- [Ha2] ———, *L'équation de Zakai et le problème séparé du contrôle optimal stochastique*, Séminaire de Strasbourg 19, pp. 37–62, Lecture Notes in Math. 1123, Springer-Verlag, Berlin, New York, 1985.
- [Ha3] ———, *Existence of Partially Observable Stochastic Optimal Controls*, Proc. Third Working Conference on Stochastic Differential Systems, Lecture Notes in Control and Information Sci., 36, Springer-Verlag, Berlin, New York, 1981.
- [IW] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, New York, 1981.
- [JM1] J. JACOD AND J. MEMIN, *Sur un type de convergence intermédiaire entre la convergence en loi et la convergence en probabilité*, Séminaire de Strasbourg 15; Lecture Notes in Math. 850, Springer-Verlag, Berlin, New York, 1981.
- [JM2] ———, *Weak and Strong Solutions of Stochastic Differential Equations: Existence and Stability*, Proc. Conf. on Stochastic Integrals, Durham, 1980; Lecture Notes in Math. 851, Springer-Verlag, Berlin, New York, 1981.
- [JM] A. JOFFE AND M. METIVIER, *Weak convergence of sequence of semi-martingales with application to multitype branching processes*, Adv. in Appl. Probab., 18 (1986), pp. 20–65.
- [KK] G. KALLIANPUR AND R. L. KARANDIKAR, *White noise calculus and nonlinear filtering theory*, Ann. Probab., 13 (1985), pp. 1033–1107.
- [Kr] N. V. KRYLOV, *Controlled Diffusion Processes*, Springer-Verlag, Berlin, New York, 1980.
- [Ku1] H. KUNITA, *Asymptotic behaviour of the nonlinear filtering errors of Markov processes*, J. Multivariate Anal., 1 (1971), pp. 365–393.
- [Ku2] ———, *Stochastic differential equations arising from nonlinear filtering*, preprint.

- [KO] T. G. KURTZ AND D. OCONE, *A Martingale Problem for Conditional Distributions and Uniqueness for the Nonlinear Filtering Equations*. Stochastic differential Systems, Lecture Notes in Control 69, Springer-Verlag, Berlin, New York, 1984, pp. 224–234.
- [Ni] M. NISIO, *On nonlinear semigroups for Markov processes associated with optimal stopping*, Appl. Math. Optim., 4 (1978), pp. 1453–1469.
- [Pa1] E. PARDOUX, *Equations of Non-linear Filtering and Applications to Stochastic Control with Partial Observations*, Conf. on Nonlinear Filtering and Stochastic Control, Cortona; Lecture Notes in Math. 972, Springer-Verlag, Berlin, New York, 1983.
- [Pa2] ———, *Stochastic partial differential equations and filtering of diffusion processes*, Stochastics, 3 (1979), pp. 127–168.
- [Pa3] ———, *Equations du filtrage non linéaire de la prédiction et du lissage*, Stochastics 6 (1982), pp. 193–231.
- [RC] S. ROELLY-COPPOLETTA, *A criterion of convergence of measure-valued processes. Application to measure-branching processes*, Stochastics, 7 (1986), pp. 43–66.
- [Sh] S. J. SHEU, *Solution of certain parabolic equations with unbounded coefficients and its applications to nonlinear filtering*, Stochastics, 10 (1983), pp. 31–46.
- [Sz] A. S. SZNITMAN, *Equations de type Boltzmann spatialement homogènes*, Z. Wahrsch. Verw. Gebiete, 66 (1974), pp. 559–592.
- [SM] J. SZPIRGLAS AND G. MAZZIOTTO, *Modèle générale de filtrage non linéaire et équations différentielles stochastiques associées*, Ann. Inst. H. Poincaré Sect. (B) 2 (1979), pp. 147–173.
- [SV] D. W. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin, New York, 1979.
- [Va] P. VARAIYA, *Optimal control of a partially observed stochastic system*, in Stochastic Differential Equations, Vol. 6, H. P. McKean and J. B. Keller, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA; American Mathematical Society, Providence, RI, 1973, pp. 173–187.
- [Wa] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [Yo] M. YOR, *Sur les théories du filtrage et de la prédiction*, Sem. XI, Université de Strasbourg; Lecture Notes in Math. 581, Springer-Verlag, Berlin, New York, 1977.

IDENTIFICATION OF AN INFINITE-DIMENSIONAL PARAMETER FOR STOCHASTIC DIFFUSION EQUATIONS*

S. I. AIHARA† AND Y. SUNAHARA†

Abstract. The identification of an infinite-dimensional unknown parameter for stochastic diffusion equations is considered. After sufficient conditions for the existence of the maximum likelihood estimate are shown, the consistency property of the maximum likelihood estimate is also investigated.

Key words. stochastic diffusion equation, maximum likelihood estimate, consistent estimate, Kalman filter in Hilbert space

AMS(MOS) subject classifications. 93C29, 93E12

1. Introduction. The maximum likelihood (ML) estimate of an unknown parameter for infinite-dimensional stochastic systems was initiated by Balakrishnan [6]. Sufficient conditions for the strong consistency of the unknown constant parameter have been derived for abstract stochastic systems by Bagchi and Borkar [5] and for hyperbolic systems by Aihara and Bagchi [2]. For an infinite-dimensional unknown parameter case, some interesting results are given for *perfect observation data* by using the method of sieves in [4].

In order to clarify our interests, we shall not consider the most general situation. The stochastic heat diffusion equation with an unknown diffusion coefficient is considered as the system model. If the identification problem for the heat diffusion coefficient is explored, we can easily treat the estimation of the other coefficients, for example, convection and/or heat generation coefficients. Furthermore it is possible to consider the general class of stochastic parabolic equations (e.g., population dispersal and biological systems as shown in [7]) by using the same argument presented here.

In this paper, the ML estimation problem for an unknown infinite-dimensional parameter, that is, a coefficient of a stochastic diffusion equation (as shown in (1.3)), is considered under *noisy partial observations*. After showing some conditions for the existence of the optimal ML estimate, we derive the consistency property for the ML estimate.

We set G as the bounded open domain in R^n with the regular boundary Γ and $T =]0, t_f[$. Define

$$(1.1) \quad V = H_0^1(G) \subset H = L^2(G) \subset V' = H^{-1}(G),$$

where the injection of $V \rightarrow H$ is compact (see, e.g., [1, Thm. 6.2, p. 144]) and we will denote by $|\cdot|$ the norm in H .

Consider the following diffusion operator:

$$(1.2) \quad A(a^0) \in \mathcal{L}(V; V')$$

and

$$(1.3) \quad \langle A(a^0)\psi_1, \psi_2 \rangle = \sum_{i=1}^n \int_G a^0(x) \frac{\partial \psi_1}{\partial x_i} \frac{\partial \psi_2}{\partial x_i} dx,$$

* Received by the editors September 22, 1986; accepted for publication (in revised form) November 4, 1987.

† Division of Control Sciences, Faculty of Polytechnic Sciences, Kyoto Institute of Technology, Mastugasaki, Sakyo-ku, Kyoto 606, Japan.

where a^0 is a function with values in R^1 and

$$(C-1) \quad 0 < \alpha \leq a^0(x) \leq \beta \quad \forall x \in \bar{G},$$

$$(C-2) \quad a^0 \in C^2(G).$$

In the sequel, from (C-2), we find that the region of operator $A(a^0)$, which is restricted to H , defines an unbounded operator which is still denoted by $A(a^0)$ with the domain $\mathcal{D}(A) = \{\psi \in V, A(a^0)\psi \in H\} = H_0^1(G) \cap H^2(G)$. $(\Omega, \mathcal{F}, \mathcal{P})$ is a complete probability space and $\{\mathcal{F}_t\}$ is an increasing family of the sub σ -algebra of \mathcal{F} . Let w be an H -valued Brownian motion process with the incremental covariance $Q \in \mathcal{L}_1(H; H)$, where $\mathcal{L}_1(\cdot, \cdot)$ denotes the class of trace operators.

Consider the following stochastic diffusion equation with the m -dimensional observation mechanism:

$$\Sigma_1 \begin{cases} (u(t), \phi) + \int_0^t \langle A(a^0)u(s), \phi \rangle ds = (u_0, \phi) + (w(t), \phi) \quad \forall \phi \in V, \\ y(t) = \int_0^t Bu(s)ds + v(t), \end{cases}$$

where (\cdot, \cdot) and $\langle \cdot, \cdot \rangle$ denote the inner product in H and the duality product on $V \times V'$, respectively, and v is an m -dimensional standard Brownian motion process independent of w and $B \in \mathcal{L}(H; R^m)$.

THEOREM 1.1 [8, Thm. 5.1, p. 189], [20, Thm. 2.1, p. 48], [15]. *Under (C-1)–(C-3) $u_0 \in L^2(\Omega; H)$, there exists a unique solution of Σ_1 such that*

$$(1.4) \quad u \in L^2(\Omega; C(T; H) \cap L^2(T; V)).$$

PROPOSITION 1.1 [10], [22, Thm. 1.4, Chap. 3], [18, Thm. 5.5, Chap. 5]. *Under (C-1) and (C-2), $-A$ generates strongly continuous, contraction semigroup T_t such that*

$$\begin{aligned} \frac{dT_t}{dt} &= -AT_t, \\ T_0 &= I, \\ |T_t \phi|^2 &\leq C e^{-\tilde{\alpha}t} |\phi|^2 \quad \exists C > 0, \quad \tilde{\alpha} > 0 \quad \forall \phi \in H. \end{aligned}$$

LEMMA 1.1. *If we choose the initial condition u_0 as*

$$(1.5) \quad u_0 = \int_{-\infty}^0 T_{-\tau} dw(\tau),$$

the auto-correlation $\Lambda(t-s)$ becomes

$$(1.6) \quad \Lambda(t-s) = E\{u(t)(u(s), \cdot)\} = \int_{-\infty}^{t \wedge s} T_{t-\tau} Q T_{s-\tau} d\tau$$

with

$$(1.7) \quad \int_0^\infty |\Lambda(\tau)|_{\mathcal{L}_2(V'; V)} d\tau < \infty,$$

where $\mathcal{L}_2(X; Y)$ denotes the Hilbert–Schmidt class of operators from X into Y , and $|\cdot|_{\mathcal{L}_2(X; Y)}$ the Hilbert–Schmidt norm.

Proof. From the results of Jerome [13, Thm. 3.2, p. 114], noting that the injection $V \rightarrow H$ is compact, we find that there exist discrete eigenvalues $\{\lambda_i\}$ and their eigenfunctions $\{e_i\}$ which are orthonormal in H with values in V such that

$$(1.8) \quad \begin{aligned} \langle Ae_i, \phi \rangle &= \lambda_i \langle e_i, \phi \rangle \quad \forall \phi \in V, \\ 0 &< \tilde{\alpha} < \lambda_1 \leq \lambda_2 \leq \cdots \rightarrow \infty. \end{aligned}$$

Hence, the contraction semigroup T_t can be rewritten by

$$(1.9) \quad T_t = \sum_{i=1}^{\infty} e^{-\lambda_i t} e_i(e_i, \cdot).$$

Furthermore, from [11, p. 50], the incremental covariance $Q \in \mathcal{L}_1(H; H)$ is also represented by

$$(1.10a) \quad Q = \sum_{i,j=1}^{\infty} q_{ij} e_i(e_j, \cdot)$$

with

$$(1.10b) \quad \sum_{i,j=1}^{\infty} q_{ij} < \infty.$$

Noting that $u(t)$ can be explicitly represented by

$$u(t) = T_t u_0 + \int_0^t T_{t-s} dw(s),$$

from (1.5) and (1.9), we have

$$(1.11) \quad u(t) = \sum_{i=1}^{\infty} \int_{-\infty}^t e^{-\lambda_i(t-s)} e_i(e_i, dw(s)).$$

Then it follows that

$$\begin{aligned} \Lambda(t-s) &= \int_{-\infty}^{t \wedge s} \sum_{i,j=1}^{\infty} q_{ij} e^{-\lambda_i(t-\tau)-\lambda_j(s-\tau)} d\tau e_i(e_j, \cdot) \\ &= \sum_{i,j=1}^{\infty} \frac{q_{ij}}{\lambda_i + \lambda_j} e^{-\lambda_i(t-t \wedge s)-\lambda_j(s-t \wedge s)} e_i(e_j, \cdot). \end{aligned}$$

It is easy to show that

$$|\Lambda(t-s)|_{\mathcal{L}_2(V'; V)}^2 = \sum_{i,j=1}^{\infty} \left(\frac{q_{ij}}{\lambda_i + \lambda_j} \right)^2 e^{-2\lambda_i(t-t \wedge s)-2\lambda_j(s-t \wedge s)} |e_i|_V^2 |e_j|_V^2.$$

From (C-1), noting that $|e_i|_V^2 \leq \text{Const. } \lambda_i$, it follows that

$$|\Lambda(t-s)|_{\mathcal{L}_2(V'; V)}^2 \leq \sum_{i,j=1}^{\infty} \frac{(q_{ij})^2}{(\lambda_i + \lambda_j)^2} e^{-2\lambda_i(t-t \wedge s)-2\lambda_j(s-t \wedge s)} \text{Const. } \lambda_i \lambda_j.$$

Consequently, from (1.10b), using the fact that $\sum (q_{ij})^2 \leq (\sum q_{ij})^2 < \infty$ (see [11, p. 50]), we get

$$\int_0^{\infty} |\Lambda(\tau)|_{\mathcal{L}_2(V'; V)}^2 d\tau \leq \text{Const. } \sum_{i,j=1}^{\infty} \frac{(q_{ij})^2}{(\lambda_i + \lambda_j)^2} < \infty. \quad \square$$

2. Asymptotic behavior of system and filtering equations. The admissible set of an unknown parameter is denoted by Θ , i.e., $a \in \Theta$ implies that a satisfies (C-1) and (C-2) and Θ is a closed, bounded subset of $H^1(G)$. Replacing a^0 by a in Σ_1 , the corresponding solution is denoted by $u(t, a)$.

LEMMA 2.1. *The following estimates can be derived: there exists C (independent of a, b and t_f) > 0 such that*

$$(2.1) \quad \sup_{t \in (-\infty, t_f]} E\{|u(t, a)|^2\} \leq C \quad \forall a \in \Theta,$$

$$(2.2) \quad \sup_{t \in (-\infty, t_f]} E\{|u(t, a)|_V^2\} \leq C \quad \forall a \in \Theta,$$

$$(2.3) \quad E\left\{\sup_{t \in (-\infty, t_f]} |u(t, a) - u(t, b)|^2\right\} \leq C|a - b|^2 \quad \forall a, b \in \Theta.$$

Proof. From (1.11), by using Proposition 1.1, we have

$$(2.4) \quad \begin{aligned} E\{|u(t, a)|^2\} &= \sum_{i,j=1}^{\infty} \int_{-\infty}^t e^{-\lambda_i(t-s) - \lambda_j(t-s)} q_{ij} ds \\ &= \sum_{i,j=1}^{\infty} \frac{q_{ij}}{\lambda_i + \lambda_j} \leq \text{Const.}, \end{aligned}$$

where from (1.8) we use the fact that $0 < \tilde{\alpha} \leq \lambda_i$, for all $i \geq 1$. Furthermore, by using the estimate $|e_i|_V^2 \leq \text{Const. } \lambda_i$ and (1.10b), we also have

$$(2.5) \quad \begin{aligned} E\{|u(t, a)|_V^2\} &= \sum_{i,j=1}^{\infty} \int_{-\infty}^t e^{-(\lambda_i + \lambda_j)(t-s)} q_{ij}(e_i, e_j)_V ds \\ &\leq C \sum_{i,j=1}^{\infty} q_{ij} \frac{(\lambda_i \lambda_j)^{1/2}}{\lambda_i + \lambda_j} \leq \text{Const.} \sum_{i,j=1}^{\infty} q_{ij} \leq \text{Const.} \end{aligned}$$

On the other hand, (1.11) can be represented by

$$(u(t, a), \phi) + \int_{-\infty}^t \langle A(a)u(s, a), \phi \rangle ds = (w(t), \phi) \quad \forall t \in (-\infty, t_f] \quad \forall \phi \in V.$$

Hence, we have

$$(u(t, a) - u(t, b), \phi) + \int_{-\infty}^t \langle A(a)u(s, a) - A(b)u(s, b), \phi \rangle ds = 0 \quad \forall \phi \in V.$$

It follows that

$$\begin{aligned} &|u(t, a) - u(t, b)|^2 + 2 \int_{-\infty}^t \langle A(a)(u(s, a) - u(s, b)), u(s, a) - u(s, b) \rangle ds \\ &= -2 \int_{-\infty}^t \langle A(a - b)u(s, b), u(s, a) - u(s, b) \rangle ds. \end{aligned}$$

From (C-1), noting that $\langle A(a)\phi, \phi \rangle \geq \alpha|\phi|_V^2$ and using Schwartz's inequality, we have

$$(2.6) \quad \begin{aligned} &|u(t, a) - u(t, b)|^2 + 2\alpha \int_{-\infty}^t |u(s, a) - u(s, b)|_V^2 ds \\ &\leq 2 \int_{-\infty}^t |\langle A(a - b)u(s, b), u(s, a) - u(s, b) \rangle| ds \\ &\leq 2 \int_{-\infty}^t |a - b| \left| \sum_{i=1}^n \frac{\partial u(s, b)}{\partial x_i} \right| \left| \sum_{i=1}^n \frac{\partial (u(s, a) - u(s, b))}{\partial x_i} \right| ds \\ &\leq \alpha \int_{-\infty}^t |u(s, a) - u(s, b)|_V^2 ds + \frac{|a - b|^2}{\alpha} \int_{-\infty}^t |u(s, b)|_V^2 ds. \end{aligned}$$

Furthermore, from Poincaré's inequality in [18, Chap. 3, Lemma 3.3], i.e., $\alpha|\psi|_V^2 \cong \tilde{\alpha}|\phi|^2$, by Gronwall's inequality, (2.6) becomes

$$(2.7) \quad E \left\{ \sup_{t \in (-\infty, t_f]} |u(t, a) - u(t, b)|^2 \right\} \leq \frac{|a - b|^2}{\alpha} E \left\{ \int_{-\infty}^{t_f} e^{-\tilde{\alpha}(t-s)} |u(s, b)|_V^2 ds \right\} \quad (\text{by using (2.2)})$$

$$\leq \text{Const. } |a - b|^2. \quad \square$$

From now on, we assume that the initial condition u_0 is given by (1.5). Denoting $E\{u(t, a)|\mathcal{Y}_t\}$ by $\hat{u}(t, a)$ from [19, Cor. 2.13, p. 924], we have

$$(2.8) \quad (\hat{u}(t, a), \phi) + \int_0^t \langle A(a)\hat{u}(s, a), \phi \rangle ds = \int_0^t (P(a)B^* dz(s; a), \phi) \quad \forall \phi \in V,$$

where z is defined by

$$(2.9) \quad z(t; a) = y(t) - \int_0^t B\hat{u}(s, a) ds$$

and $z(t; a)$ becomes a so-called *innovation* process, if $a = a^0$ (true value). Furthermore, from [25, Thm. 1, p. 252] we find that $P(a)$ is a unique solution of the algebraic Riccati equation:

$$(2.10) \quad -\langle A(a)P(a)\psi_1, \psi_2 \rangle - \langle A^*(a)\psi_1, P(a)\psi_2 \rangle + \langle Q\psi_1, \psi_2 \rangle \\ - (P(a)B^*BP(a)\psi_1, \psi_2) = 0 \quad \forall \psi_1, \psi_2 \in V.$$

LEMMA 2.2 [16, Thm. 6.1, p. 171]. *The following estimates can be derived:*

$$(2.11a) \quad P(a) \in \mathcal{L}_1(H; V) \cap \mathcal{L}_1(V'; H),$$

$$(2.11b) \quad P(a) = P^*(a) \geq 0.$$

LEMMA 2.3. *There exists a constant C independent of a and b such that*

$$(2.12) \quad |P(a)|_{\mathcal{L}_2(V'; H)} + |P(a)|_{\mathcal{L}_2(H; V)} \leq C,$$

$$(2.13) \quad |(P(a) - P(b))B^*|_{\mathcal{L}(R^m; H)}^2 \leq C|a - b|^2.$$

Proof. From (2.10), it is easy to derive the following estimate:

$$\alpha(|P(a)|_{\mathcal{L}_2(V'; H)}^2 + |P(a)|_{\mathcal{L}_2(H; V)}^2) + [P(a)B^*BP(a), P(a)] \leq [Q, P(a)],$$

where $[\cdot, \cdot]$ denotes the Hilbert-Schmidt inner product. From the fact that

$$[P(a)B^*BP(a), P(a)] \geq 0,$$

For $\alpha > 0$, $C(\alpha) > 0$, we have

$$\alpha|P(a)|_{\mathcal{L}_2(V'; H)}^2 + \alpha|P(a)|_{\mathcal{L}_2(H; V)}^2 \leq |Q|_{\mathcal{L}_2(H; H)}|P(a)|_{\mathcal{L}_2(H; H)} \\ \leq \frac{\alpha}{2}|P(a)|_{\mathcal{L}_2(V'; H)}^2 + C(\alpha)|Q|_{\mathcal{L}_2(H; H)}^2.$$

Hence, (2.12) can be derived. Denoting P_e by $P(a) - P(b)$, it follows that

$$(2.14) \quad -\langle A(a)P_e\psi_1, \psi_2 \rangle - \langle A^*(a)\psi_1, P_e\psi_2 \rangle - (P_eB^*BP_e\psi_1, \psi_2) \\ - ((P_eB^*BP(b) + P(b)B^*BP_e)\psi_1, \psi_2) \\ = \langle A(a - b)P(b)\psi_1, \psi_2 \rangle + \langle A^*(a - b)\psi_1, P(b)\psi_2 \rangle \quad \forall \psi_1, \psi_2 \in V.$$

By using the same procedure mentioned above, we have

$$(2.15) \quad \begin{aligned} & \alpha |P_e|^2_{\mathcal{L}_2(V;H)} + \alpha |P_e|^2_{\mathcal{L}_2(H;V)} + [P_e B^* B P_e, P_e] + [P(b) B^* B P_e + P_e B^* B P(b), P_e] \\ & \leq |a - b| \{ |P(b)|_{\mathcal{L}_2(V;H)} |P_e|_{\mathcal{L}_2(V;H)} + |P(b)|_{\mathcal{L}_2(H;V)} |P_e|_{\mathcal{L}_2(H;V)} \}. \end{aligned}$$

Note that the third and fourth terms of the left-hand side of (2.15) are positive; then from (2.12), (2.15) becomes

$$(2.16) \quad \begin{aligned} |P(a) - P(b)|^2_{\mathcal{L}_2(H;H)} & \leq C_1 |P(a) - P(b)|^2_{\mathcal{L}_2(V;H)} \\ & \leq C_2 |a - b|^2. \end{aligned} \quad \square$$

LEMMA 2.4. *We need the following condition:*

$$(C-4)^1 \quad B \in \mathcal{L}(V'; R^m).$$

For $a, b \in \Theta$,

$$(2.17) \quad \sup_{t \in T} E \{ |B(\hat{u}(t, a) - \hat{u}(t, b))|^2_{R^m} \} \leq C |a - b|^2,$$

where C is independent of t_f , a , and b .

Proof. Denoting $\hat{e}(t)$ by $\hat{u}(t, a) - \hat{u}(t, b)$, we have

$$(2.18) \quad \begin{aligned} & (\hat{e}(t), \phi) + \int_0^t \{ \langle A(a) \hat{e}(s), \phi \rangle + \langle A(a - b) \hat{u}(s, b), \phi \rangle \} ds \\ & = \int_0^t ((P(a) - P(b)) B^* dy(s), \phi) - \int_0^t ((P(a) - P(b)) B^* B \hat{u}(s, b), \phi) ds \\ & \quad - \int_0^t (P(a) B^* B \hat{e}(s), \phi) ds \quad \forall \phi \in V. \end{aligned}$$

From [25, Thm. 1, p. 252], we find that $-(A(a) + P(a) B^* B)$ generates the exponential stable semigroup such that

$$(2.19) \quad \begin{aligned} \frac{d\tilde{T}_t}{dt} & = -(A(a) + P(a) B^* B) \tilde{T}_t, \quad \tilde{T}_0 = I, \\ |\tilde{T}_t \phi| & \leq M_1 e^{-\alpha_1 t} |\phi| \quad \exists M_1 > 0, \quad \alpha_1 > 0 \quad \forall \phi \in H. \end{aligned}$$

Thus, $\tilde{A}(a) \triangleq A(a) + P(a) B^* B$ has a square root $\tilde{A}^{1/2}(a)$ with its domain V (see Kato [14, p. 282]). Hence, by using the mutual commutability of $\tilde{A}^{1/2}$, $\tilde{A}^{-1/2}$, and \tilde{T}_t , we obtain

$$(2.20) \quad |\tilde{A}^{-1/2}(a) \tilde{T}_t \tilde{A}^{1/2}(a) \phi| = |\tilde{T}_t \phi| \leq M_1 e^{-\alpha_1 t} |\phi|.$$

For convenience of discussions, we use the norm $|\phi|_V^2 = |\tilde{A}^{-1/2} \phi|^2$ in this proof. Hence from (C-4), we can derive the following estimate:

$$(2.21) \quad \begin{aligned} & |\langle A(a - b) \hat{u}(s, b), \tilde{T}_{t-s}^* B^* \rangle|_{R^m} \\ & \leq |a - b| |\hat{u}(s, b)|_V |\tilde{T}_{t-s}^* B^*|_{\mathcal{L}(R^m; V)} \\ & = |a - b| |\hat{u}(s, b)|_V |B \tilde{T}_{t-s}|_{\mathcal{L}(V'; R^m)} \\ & = |a - b| |\hat{u}(s, b)|_V |B \tilde{A}^{1/2}(a) \tilde{A}^{-1/2}(a) \tilde{T}_{t-s} \tilde{A}^{1/2}(a)|_{\mathcal{L}(H; R^m)} \\ & \leq |a - b| |\hat{u}(s, b)|_V |B \tilde{A}^{1/2}(a)|_{\mathcal{L}(H; R^m)} |\tilde{A}^{-1/2}(a) \tilde{T}_{t-s} \tilde{A}^{1/2}(a)|_{\mathcal{L}(H; H)} \end{aligned}$$

¹ A typical example is shown in § 4.

$$\begin{aligned} \text{(from (2.20) and } |B\tilde{A}^{1/2}(a)|_{\mathcal{L}(H;R^m)} &= |B|_{\mathcal{L}(V;R^m)} \leq C, \text{ i.e., (C-4))} \\ &\leq \tilde{C}M_1|a-b||\hat{u}(s, b)|_V e^{-\alpha_1(t-s)}. \end{aligned}$$

On the other hand, from (2.18), it follows that

$$\begin{aligned} \hat{e}(t) &= - \int_0^t T_{t-s}A(a-b)\hat{u}(s, b) ds + \int_0^t \tilde{T}_{t-s}(P(a)-P(b))B^* dy(s) \\ &\quad - \int_0^t \tilde{T}_{t-s}(P(a)-P(b))B^* \hat{u}(s, b) ds. \end{aligned}$$

By using the Schwartz inequality, we have for $C > 0$ (independent of a, b and a^0)

$$\begin{aligned} E\{|B\hat{e}(t)|_{R^m}^2\} &\leq 3E\left\{\left(\int_0^t |B\tilde{T}_{t-s}A(a-b)\hat{u}(s, b)|_{R^m} ds\right)^2\right\} \\ &\quad + 3E\left\{\left(\int_0^t |B\tilde{T}_{t-s}(P(a)-P(b))B^* \hat{u}(s, a^0)|_{R^m} ds\right)^2\right\} \\ &\quad + 3 \int_0^t |B\tilde{T}_{t-s}(P(a)-P(b))B^*|_{\mathcal{L}(R^m;R^m)}^2 ds \\ (2.22) \quad &\quad + 3E\left\{\left(\int_0^t |B\tilde{T}_{t-s}(P(a)-P(b))B^* \hat{u}(s, b)|_{R^m} ds\right)^2\right\} \quad (\text{from (2.19)}) \\ &\leq 3E\left\{\left(\int_0^t |A(a-b)\hat{u}(s, b), \tilde{T}_{t-s}^*B^*|_{R^m} ds\right)^2\right\} \\ &\quad + C|P(a)-P(b)|_{\mathcal{L}(R^m;H)}^2 \\ &\quad \cdot \left(E\left\{\int_0^t e^{-\alpha_1(t-s)}(1+|u(s, a^0)|^2+|\hat{u}(s, b)|^2) ds\right\}\right). \end{aligned}$$

From (2.21) and (2.13) in Lemma 2.3, for $\tilde{C} > 0$, the right-hand side of (2.22) becomes

$$\begin{aligned} &\leq \tilde{C}|a-b|^2\left\{1 + \int_0^t e^{-\alpha_1(t-s)}(E\{|\hat{u}(s, b)|_V^2\} \right. \\ &\quad \left. + E\{|u(s, a^0)|^2\} + E\{|\hat{u}(s, b)|^2\}) ds\right\} \\ &\leq \frac{\tilde{C}}{\alpha_1}|a-b|^2\left\{1 + \sup_{s \in [0, t_f]} (E\{|\hat{u}(s, b)|_V^2\} \right. \\ &\quad \left. + E\{|u(s, a^0)|^2\} + E\{|\hat{u}(s, b)|^2\})\right\}. \end{aligned}$$

Hence, from (2.1) and (2.2) in Lemma 2.1, it follows that

$$E\{|B\hat{e}(t)|_{R^m}^2\} \leq \text{Const. } |a-b|^2,$$

where we use the following relation:

$$E\{|\hat{u}(t, b)|_X^2\} \leq E\{|u(t, b)|_X^2\} \quad \forall b \in \Theta. \quad \square$$

3. The identification problem. The cylindrical measure μ_y induced by y is absolutely continuous with respect to μ_v induced by the Brownian motion process v .

Then, from Liptser and Shiryaev [17, Thm. 7.13, p. 261], the Radon–Nikodym derivative is given by

$$(3.1) \quad \frac{d\mu_y(a^0)}{d\mu_v} = \exp \left\{ \int_0^{t_f} (B\hat{u}(t, a^0))^* dy(t) - \frac{1}{2} \int_0^{t_f} (B\hat{u}(t, a^0))^* B\hat{u}(t, a^0) dt \right\}.$$

The identification problem is to determine the unknown function a^0 so as to maximize the following log-likelihood functional:

$$(3.2) \quad L(t_f, y, a) = \ln \frac{d\mu_y(a)}{d\mu_v},$$

with respect to $a \in \Theta$.

From now on, conditions (C-1)–(C-4) are assumed.

THEOREM 3.1. *There exists at least one optimal ML estimate $\hat{a}_{t_f} \in \Theta$ such that*

$$(3.3) \quad L(t_f, y, \hat{a}_{t_f}) \geq L(t_f, y, a) \quad \forall a \in \Theta.$$

Proof. Let a^n be a minimizing sequence, that is, $-L(t_f, y, a^n) \rightarrow \inf_{a \in \Theta} -L(t_f, y, a)$. Hence, $a^n \in \Theta$, i.e., Θ is a closed bounded subset of $H^1(G)$ and we have

$$(3.4) \quad \|a^n\|_{H^1(G)} \leq \text{Const.}$$

By using the compactness method, from (3.4), we can extract a subsequence still denoted by a^n such that

$$(3.5) \quad a^n \rightarrow \tilde{a} \quad \text{strongly in } H.$$

Then from Lemma 2.4 and (3.5) we get

$$(3.6) \quad B\hat{u}(t, a^n) \rightarrow B\hat{u}(t, \tilde{a}) \quad \text{strongly in } L^2(\Omega; R^m) \quad \forall t.$$

Furthermore, extracting the subsequence of a^n , i.e., $\sum_{n'=1}^{\infty} |a^{n'} - \tilde{a}|^2 < \infty$ and using the Borel–Cantelli lemma, we also have

$$(3.7) \quad B\hat{u}(t, a^n) \rightarrow B\hat{u}(t, \tilde{a}) \quad \text{strongly in } R^n \quad \forall t \quad \text{a.s.}$$

Hence,

$$\lim_{n \rightarrow \infty} \int_0^{t_f} |B\hat{u}(t, a^n)|_{R^m}^2 dt = \int_0^{t_f} |B\hat{u}(t, \tilde{a})|_{R^m}^2 dt \quad \text{a.s.}$$

and subtracting a subsequence of $\hat{u}(t, a^{n'})$, we get

$$\lim_{n' \rightarrow \infty} \int_0^{t_f} (B\hat{u}(t, a^{n'}))^* dy(t) = \int_0^{t_f} (B\hat{u}(t, \tilde{a}))^* dy(t) \quad \text{a.s.}$$

These imply that

$$\begin{aligned} \lim_{n' \rightarrow \infty} -L(t_f, y, a^{n'}) &= -L(t_f, y, \tilde{a}) \\ &\leq \liminf -L(t_f, y, a^{n'}) \\ &= \inf_{a \in \Theta} -L(t_f, y, a). \end{aligned}$$

Then \tilde{a} is an optimal ML estimate. \square

In order to show the consistency property for the ML estimate \hat{a}_{t_f} , we must show that

$$(3.8) \quad g(t, a) = \frac{1}{t} \int_0^t (B\hat{u}(s, a))^* dv(s)$$

is uniformly continuous in a on Θ , uniformly in $t \in [1, \infty)$ almost surely. By using the results of Strait [21, Thm. 3, p. 168], we can extend the techniques used by Bagchi and Borkar [5, Lemma 3.2, p. 212], [9] to the infinite-dimensional unknown parameter case. Then the above statement can be clearly proved in Theorem 3.2.

LEMMA 3.1. *For $N \geq 1$ and $a, b \in \Theta$, we have*

$$(3.9) \quad E \left\{ \left| \frac{N+1}{N} (g(N+1, a) - g(N+1, b)) \right|_{R^1}^4 \right\} \leq \frac{K}{(N+1)^2} |a - b|^4,$$

where K is a constant independent of N , a , and b .

Proof. Noting that $\hat{u}(t)$ is Gaussian, from Lemma 2.4, it follows that

$$\begin{aligned} E\{|B(\hat{u}(s, a) - \hat{u}(s, b))|_{R^m}^4\} &\leq C(E\{|\hat{u}(s, a) - \hat{u}(s, b)|^2\})^2 \\ &\leq \text{Const. } |a - b|^4 \quad \forall s \geq 0. \end{aligned}$$

Hence, from Lemma 4.12 in Liptser and Shiryaev [17, p. 125], we have

$$\begin{aligned} E \left\{ \left| \frac{N+1}{N} (g(N+1, a) - g(N+1, b)) \right|_{R^1}^4 \right\} \\ \leq 36 \left(\frac{N+1}{N} \right)^4 \frac{1}{(N+1)^3} E \left\{ \int_0^{N+1} |B(u(s, a) - u(s, b))|_{R^m}^4 ds \right\}, \end{aligned}$$

and then (3.9) can be derived. \square

THEOREM 3.2.

$$(3.10) \quad \limsup_{t \rightarrow \infty} g(t, a) = 0 \quad a.s.$$

Proof. The Hilbert space $H = L^2(G)$ is separable and then the unknown parameter a can be represented by

$$a(x) = \sum_{i=1}^{\infty} a_i \phi_i(x) \quad \text{for } \phi_i, \text{ orthonormal basis in } H,$$

and furthermore, in order to support $a \in \Theta$, i.e., $a \in C^2(G)$, we need the regularity of ϕ . However, noting that only the H -norm appears in the right-hand side of (2.17) and (3.9), for the convenience of discussion, without loss of generality, it is sufficient to assume from (3.4) that

$$(3.11) \quad \phi_i \in H^1(G) \quad \text{and} \quad \left\| \sum_{i=1}^{\infty} a_i \phi_i(x) \right\|_{H^1(G)} \leq \text{Const.}$$

in order to support that Θ is compact in H . Then, the admissible parameter class Θ is represented as Θ_{l_2} which is compact in l_1 from (3.11), i.e., $\tilde{a} \in \Theta_{l_2}$ implies that

$$(i) \quad a = \sum_{i=1}^{\infty} a_i \phi_i, \quad \tilde{a} = (a_1, a_2, \dots) \in l_2, \quad |\tilde{a}|_{l_2} \leq \text{Const.},$$

$$(ii) \quad a_i \in [0, c_i], \quad \sup |c_i| \leq c, \quad \sum_{i=1}^{\infty} c_i^2 \leq \text{Const.}$$

We denote all binary rational numbers (the form of $l/2^\nu$) in $[0, c]$ by D and then D is dense in $[0, c]$. Hence, in the sequel, a_i is assumed to be a binary rational number in D . Now, we prepare the sequence $m_i(\tilde{m})$ such that

$$m_i(\tilde{m}) = \tilde{m} + i, \quad \tilde{m} \geq 1,$$

and then

$$\sum_{i=1}^{\infty} 2^{-m_i(\tilde{m})} = 2^{-\tilde{m}}.$$

We introduce

$$(3.12) \quad \tilde{a}(j) = \left(\frac{j(1)}{2^{m_1(\tilde{m})}}, \frac{j(2)}{2^{m_2(\tilde{m})}}, \dots \right)$$

and

$$(3.13) \quad \tilde{a}_k(j) = \left(\frac{j(1)}{2^{m_1(\tilde{m})}} + \sum_{l=1}^{k_1} \frac{\alpha_l^1}{2^{m_1(\tilde{m})+l}}, \frac{j(2)}{2^{m_2(\tilde{m})}} + \sum_{l=1}^{k_2} \frac{\alpha_l^2}{2^{m_2(\tilde{m})+l}}, \dots \right),$$

where $\alpha_l^i = 0$ or 1 and $j(i) = 0, 1, 2, \dots, 2^{m_i(\tilde{m})}c - 1$ for each i . It is a routine procedure to show that

$$(3.14) \quad \begin{aligned} & P \left\{ \max_{\substack{0 \leq j(i) \leq 2^{m_i(\tilde{m})}c-1 \\ \text{for every } i}} \sup_{t \in [1, \infty)} |g(t, \tilde{a}_k(j)) - g(t, \tilde{a}(j))|_{R^1} \geq 2^{-\tilde{m}r} \frac{1}{2^r - 1} \right\} \\ & \cong P \left\{ \sum_{i=1}^{\infty} \max_{0 \leq j(i) \leq 2^{m_i(\tilde{m})}c-1} \sup_{t \in [1, \infty)} |g(t, \{a_k(j)\}_i) - g(t, \{a(j)\}_i)|_{R^1} \geq 2^{-\tilde{m}r} \frac{1}{2^r - 1} \right\}, \end{aligned}$$

where

$$\{a_k(j)\}_i = \left(a_1, a_2, \dots, \frac{j(i)}{2^{m_i(\tilde{m})}} + \sum_{l=1}^{k_i} \frac{\alpha_l^i}{2^{m_i(\tilde{m})+l}}, a_{i+1}, \dots \right)$$

and

$$\{a(j)\}_i = \left(a_1, a_2, \dots, \frac{j(i)}{2^{m_i(\tilde{m})}}, a_{i+1}, \dots \right).$$

Noting that $2^{-\tilde{m}r}/(2^r - 1) = \sum_{i=1}^{\infty} 2^{-m_i(\tilde{m})r}$, the right-hand side of (3.14) is evaluated as

$$\begin{aligned} & \cong \sum_{i=1}^{\infty} P \left\{ \max_{0 \leq j(i) \leq 2^{m_i(\tilde{m})}c-1} \sup_{t \in [1, \infty)} |g(t, \{a_k(j)\}_i) - g(t, \{a(j)\}_i)|_{R^1} \geq 2^{-m_i(\tilde{m})r} \right\} \\ & \cong \sum_{i=1}^{\infty} 2^{m_i(\tilde{m})} c P \left\{ \sup_{t \in [1, \infty)} |g(t, \{a_k(j)\}_i) - g(t, \{a(j)\}_i)|_{R^1} \geq 2^{-m_i(\tilde{m})r} \right\} \\ & \cong c \sum_{i=1}^{\infty} 2^{m_i(\tilde{m})} P \left\{ \sum_{q=1}^{k_i} \sup_{t \in [1, \infty)} |g(t, \{a_q(j)\}_i) - g(t, \{a_{q-1}(j)\}_i)|_{R^1} \geq 2^{-m_i(\tilde{m})r} \right\} \\ & \quad \left(\text{from } 2^{-m_i(\tilde{m})r} \geq 2^{-m_i(\tilde{m})r} \sum_{q=1}^{k_i} 2^{-qr} \cdot ((2^r - 1) \vee 1) \right) \\ & \cong c \sum_{i=1}^{\infty} 2^{m_i(\tilde{m})} \sum_{q=1}^{k_i} P \left\{ \sup_{t \in [1, \infty)} |g(t, \{a_q(j)\}_i) - g(t, \{a_{q-1}(j)\}_i)|_{R^1} \right. \\ & \quad \left. \geq 2^{-m_i(\tilde{m})r} \cdot 2^{-qr} \cdot ((2^r - 1) \vee 1) \right\} \\ & \cong c \sum_{i=1}^{\infty} 2^{m_i(\tilde{m})} \sum_{q=1}^{k_i} \sum_{N=1}^{\infty} P \left\{ \sup_{t \in [N, N+1]} |g(t, \{a_q(j)\}_i) - g(t, \{a_{q-1}(j)\}_i)|_{R^1} \right. \\ & \quad \left. \geq 2^{-m_i(\tilde{m})r} \cdot 2^{-qr} \cdot ((2^r - 1) \vee 1) \right\}. \end{aligned}$$

From Lemma 3.1 we also have,

$$E\left\{\frac{(N+1)^4}{N^4}|g(N+1, \{a_q(j)\}_i) - g(N+1, \{a_{q-1}(j)\}_i)|_{R^1}^4\right\} \leq \frac{K}{2^{4(m_i(\tilde{m})+q)}} \frac{1}{(N+1)^2}.$$

Hence, by using the Kolmogorov-Doob inequality in [12, p. 28], it follows that

$$\begin{aligned} & P\left\{\sup_{t \in [N, N+1]} |g(t, \{a_q(j)\}_i) - g(t, \{a_{q-1}(j)\}_i)|_{R^1} \geq 2^{-m_i(\tilde{m})(1/2)-q(1/2)}\right\} \\ (3.15) \quad & \leq E\{|g(N+1, \{a_q(j)\}_i) - g(N+1, \{a_{q-1}(j)\}_i)|_{R^1}^4\} \frac{(N+1)^4}{N^4} 2^{2m_i(\tilde{m})+2q} \\ & \leq \frac{K}{2^{4(m_i(\tilde{m})+q)}} \frac{1}{(N+1)^2} 2^{2m_i(\tilde{m})+2q}. \end{aligned}$$

From (3.15), choosing $r = \frac{1}{2}$ in (3.14), we have

$$\begin{aligned} & P\left\{\max_{\substack{0 \leq j(i) \leq 2^{m_i(\tilde{m})} c-1 \\ \text{for every } i}} \sup_{t \in [1, \infty)} |g(t, \tilde{a}_k(j)) - g(t, \tilde{a}(j))|_{R^1} \geq 2^{-\tilde{m}/2}\right\} \\ & \leq c \sum_{i=1}^{\infty} 2^{3m_i(\tilde{m})} \left\{ \sum_{q=1}^{\infty} \sum_{N=1}^{\infty} 2^{2q} \cdot \frac{K}{2^{4m_i(\tilde{m})+4q}} \cdot \frac{1}{(N+1)^2} \right\} \\ & \leq cK \sum_{i=1}^{\infty} 2^{-m_i(\tilde{m})} \sum_{q=1}^{\infty} 2^{-2q} \cdot \sum_{N=1}^{\infty} \frac{1}{(N+1)^2} \\ & \leq \tilde{C} 2^{-\tilde{m}}. \end{aligned}$$

Hence,

$$\begin{aligned} & P\left\{\bigcup_{\tilde{m} \geq \nu} \left\{ \max_{\substack{0 \leq j(i) \leq 2^{m_i(\tilde{m})} c-1 \\ \text{for every } i}} \sup_{t \in [1, \infty)} |g(t, \tilde{a}_k(j)) - g(t, \tilde{a}(j))|_{R^1} \geq 2^{-\tilde{m}/2} \right\}\right\} \\ & \leq \tilde{C} \sum_{\tilde{m} = \nu}^{\infty} 2^{-\tilde{m}}. \end{aligned}$$

Furthermore, noting that

$$|\tilde{a}_k(j) - \tilde{a}(j)|_{l_2} \leq \sum_{i=1}^{\infty} \sum_{l=1}^{k_i} \frac{\alpha_l^i}{2^{m_i(\tilde{m})+l}} \leq \frac{1}{2^{\tilde{m}}},$$

and that D is dense in $[0, c]$, we find that for any $\varepsilon > 0$ and any $\delta > 0$, there exists $\nu(\delta, \varepsilon)$ such that

$$P\left\{\sup_{\substack{\tilde{a}, \tilde{b} \in \Theta_{l_2} \\ |\tilde{a} - \tilde{b}|_{l_2} \leq 1/2^\nu}} \sup_{t \in [1, \infty)} |g(t, \tilde{a}) - g(t, \tilde{b})|_{R^1} > \varepsilon\right\} < \delta.$$

Hence, we have

$$(3.16) \quad P\left\{\lim_{\nu \rightarrow \infty} \sup_{\substack{\tilde{a}, \tilde{b} \in \Theta_{l_2} \\ |\tilde{a} - \tilde{b}|_{l_2} \leq 1/2^\nu}} \sup_{t \in [1, \infty)} |g(t, \tilde{a}) - g(t, \tilde{b})|_{R^1} = 0\right\} = 1.$$

From the well-known fact that

$$\lim_{t \rightarrow \infty} g(t, \tilde{a}) = 0 \quad \text{a.s.} \quad \forall \tilde{a} \in \Theta_{l_2}$$

and by using the uniform continuity of $g(t, \tilde{a})$ with respect to \tilde{a} —i.e., (3.16) permits us to commute $\sup_{\tilde{a}}$ and \lim_t operations—and noting that Θ_{l_2} is compact in l_2 , (3.10) can be derived. \square

Now, we shall state the consistency property for the maximum likelihood estimate \hat{a}_{t_f} . By using a technique similar to the one proposed by Tugnait [23, Thm. 1, p. 655] and Bagchi and Borkar [5, Thm. 3.1, p. 212], [9], we can derive the following main theorem.

THEOREM 3.3. *Let \mathcal{M} be the set of measure zero outside of which Theorem 3.2 holds. For each $\omega \notin \mathcal{M}$, letting \hat{a}_{t_f} be the maximum likelihood estimate of a^0 , we have the following strong consistency property:*

$$(3.17) \quad \lim_{t_f \rightarrow \infty} \frac{1}{t_f} \int_0^{t_f} |B(\hat{u}(t, \hat{a}_{t_f}) - u(t, a^0))|_{\mathbb{R}^m}^2 dt = \text{tr} [B^* P(a^0) B].$$

Proof. From (3.1), we have

$$\begin{aligned} \frac{d\mu_y(a)}{d\mu_v} = \exp \left\{ -\frac{1}{2} \int_0^{t_f} (B(u(t, a^0) - \hat{u}(t, a)))^* B(u(t, a^0) - \hat{u}(t, a)) dt \right. \\ \left. + \frac{1}{2} \int_0^{t_f} (Bu(t, a^0))^* Bu(t, a^0) dt + \int_0^{t_f} (B\hat{u}(t, a))^* dv(t) \right\}. \end{aligned}$$

By using the measure transformation technique, we obtain

$$\frac{d\mu_y(a)}{d\mu_y(a^0)} = \frac{d\mu_y(a)}{d\mu_v} \bigg/ \frac{d\mu_y(a^0)}{d\mu_v},$$

and then we find that for each $\omega \notin \mathcal{M}$, the ML estimate \hat{a}_{t_f} satisfies

$$(3.18) \quad \frac{d\mu_y(a_{t_f})}{d\mu_y(a^0)} \geq 1.$$

Hence, we need to evaluate $\sup_{a \in \Theta} \{d\mu_y(a)/d\mu_y(a^0)\}$, i.e.,

$$\begin{aligned} \sup_{a \in \Theta} \left\{ \frac{d\mu_y(a)}{d\mu_v} \bigg/ \frac{d\mu_y(a^0)}{d\mu_v} \right\} &\leq \sup_{a \in \Theta} \exp \left\{ -\frac{1}{2} I(t_f, a) + \frac{1}{2} I(t_f, a^0) \right\} \\ &\cdot \sup_{a \in \Theta} \exp \left\{ \int_0^{t_f} (B(\hat{u}(t, a) - \hat{u}(t, a^0)))^* dv(t) \right\}, \end{aligned}$$

where

$$I(t_f, a) = \int_0^{t_f} (B(u(t, a^0) - \hat{u}(t, a)))^* B(u(t, a^0) - \hat{u}(t, a)) dt$$

and

$$I(t_f, a^0) = \int_0^{t_f} (B(u(t, a^0) - \hat{u}(t, a^0)))^* B(u(t, a^0) - \hat{u}(t, a^0)) dt.$$

From Theorem 3.2, we have

$$(3.19) \quad \lim_{t_f \rightarrow \infty} \sup_{a \in \Theta} \frac{1}{t_f} \int_0^{t_f} (B(\hat{u}(t, a) - \hat{u}(t, a^0)))^* dv(t) = 0 \quad \text{a.s.}$$

Furthermore, from (1.7) in Lemma 1.1, and the Gaussian property of u and \hat{u} , processes u and \hat{u} become stationary ergodic as stated in [24, Prop. 6.1, p. 68] and then

$$(3.20) \quad \lim_{t_f \rightarrow \infty} \frac{1}{t_f} I(t_f, a^0) = \lim_{t_f \rightarrow \infty} \frac{1}{t_f} \int_0^{t_f} (B(u(t, a^0) - \hat{u}(t, a^0)))^* B(u(t, a^0) - \hat{u}(t, a^0)) dt \\ = \text{tr} [B^* P(a^0) B].$$

It is well known that the Kalman filter is the minimum mean square error state estimate in the case where the unknown parameter is matched to the true one. Hence

$$(3.21) \quad -E \left\{ \sup_{a \in \Theta} -I(t_f, a) \right\} \geq \text{tr} [B^* P(t_f, a^0) B]$$

where

$$P(t_f, a^0) = E \{ (u(t_f, a^0) - \hat{u}(t_f, a^0))(u(t_f, a^0) - \hat{u}(t_f, a^0), \cdot) \}.$$

The stationary and ergodic property implies that

$$-\lim_{t_f \rightarrow \infty} \sup_{a \in \Theta} \frac{1}{t_f} I(t_f, a) \geq \text{tr} [B^* P(a^0) B].$$

Consequently we have

$$(3.22) \quad \lim_{t_f \rightarrow \infty} \frac{1}{t_f} \ln \left\{ \sup_{a \in \Theta} \frac{d\mu_y(a)}{d\mu_y(a^0)} \right\} \leq 0.$$

Now, let \hat{a}_{t_f} be the maximum likelihood estimate, and then from (3.18) we have

$$-\frac{1}{2} I(t_f, \hat{a}_{t_f}) + \frac{1}{2} I(t_f, a^0) + \int_0^{t_f} (B(\hat{u}(t, \hat{a}_{t_f}) - \hat{u}(t, a^0)))^* dv(t) \geq 0$$

and from (3.22)

$$\lim_{t_f \rightarrow \infty} \frac{1}{t_f} \{I(t_f, \hat{a}_{t_f}) - I(t_f, a^0)\} \geq 0.$$

Hence,

$$\lim_{t_f \rightarrow \infty} \frac{1}{t_f} \int_0^{t_f} (B(\hat{u}(t, \hat{a}_{t_f}) - \hat{u}(t, a^0)))^* dv(t) \geq \lim_{t_f \rightarrow \infty} \frac{1}{2t_f} (I(t_f, \hat{a}_{t_f}) - I(t_f, a^0)) \geq 0.$$

From (3.19), we have

$$\lim_{t_f \rightarrow \infty} \frac{1}{2t_f} (I(t_f, \hat{a}_{t_f}) - I(t_f, a^0)) = 0.$$

This implies (3.17). \square

4. Conclusions. It is very easy to obtain the same result of consistency property for the more general elliptic operator than for $A(a)$ in (1.3), hyperbolic systems, and systems with an unknown boundary region. In such general cases, it should be noted that we need the smoothness property for the unknown coefficients, such as (C-2), and/or for the boundary region, in order to support the existence of the optimal ML estimate. An important example of the operator B which satisfies (C-4) is given by

$$B(\cdot) = \int_G b(x)(\cdot) dx \in R^1,$$

where we assume that

$$b \in H_0^1(G)$$

in order to support (C-4).

The derivation of the necessary conditions for the optimal ML estimate is also important for deriving the numerical algorithm. Regarding this problem, some results have been derived for the boundary parameter identification by Aihara, Sunahara, and Ishikawa [3], using the stochastic backward integral.

Acknowledgments. The authors express their thanks to Professor Arunabha Bagchi at Twente University of Technology, Enschede, the Netherlands for many fruitful discussions on § 3 of this paper.

REFERENCES

- [1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] S. I. AIHARA AND A. BAGCHI, *Parameter identification for hyperbolic stochastic systems*, Memorandum 545, Technische Hogeschool Twente, Enschede, the Netherlands, 1985.
- [3] S. I. AIHARA, Y. SUNAHARA, AND M. ISHIKAWA, *Identification of boundary parameter for stochastic boundary systems*, Stochastic Distributed Parameter System Project Report 8606, Kyoto Institute of Technology, Kyoto, Japan, 1986 Proc. 10th IFAC World Congress, Munich, 1987, to appear.
- [4] A. BAGCHI, *Identification for a hereditary system with distributed delay*, Systems Control Lett., 5 (1985), pp. 339–345.
- [5] A. BAGCHI AND V. BORKAR, *Parameter identification in infinite dimensional linear systems*, Stochastics, 12 (1984), pp. 201–213.
- [6] A. V. BALAKRISHNAN, *Identification and Stochastic Control of a Class of Distributed Systems with Boundary Noise*, Lecture Notes in Economics and Mathematical Systems 107, Springer-Verlag, Berlin, New York, 1975.
- [7] H. T. BANKS, *On a variational approach to some parameter estimation problems*, Tech. Report 85-14, Lefschetz Center for Dynamical Systems, Brown Univ., Providence, RI, May, 1985.
- [8] A. BENSOUSSAN, *Filtrage Optimal des Systèmes Linéaires*, Dunod, Paris, 1971.
- [9] V. BORKAR AND A. BAGCHI, *Parameter estimation in continuous-time stochastic systems*, Stochastics, 8 (1982), pp. 193–212.
- [10] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sciences, Springer-Verlag, New York, 1978.
- [11] I. M. GELFAND AND N. YA. VILENKIN, *Generalized Functions*, Vol. 4, Academic Press, New York, 1964.
- [12] N. IKEDA AND SH. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, New York, 1981.
- [13] J. W. JEROME, *On the L_2 n -width of certain classes of functions of several variables*, J. Math. Anal. Appl., 20 (1967), pp. 110–123.
- [14] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
- [15] P. KOTELenez AND R. F. CURTAIN, *Local behavior of Hilbert space valued stochastic integrals and the continuity of mild solutions of stochastic evolution equations*, Stochastics, 6 (1982), pp. 239–257.
- [16] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1971.
- [17] R. S. LIPTSER AND A. N. SHIRYAEV, *Statistics of Random Processes*, I, Springer-Verlag, Berlin, New York, 1977.
- [18] S. MIZOHATA, *The Theory of Partial Differential Equations*, Cambridge Univ. Press, Cambridge, U.K., 1973.
- [19] J. Y. OUVARD, *Martingale projection and linear filtering in Hilbert spaces*, I: the theory, SIAM J. Control Optim., 16 (1978), pp. 912–937.
- [20] E. PARDOUX, *Equations aux dérivées partielles stochastiques non linéaires monotones*, Thèse, Université de Paris XI, Paris, 1975.
- [21] P. T. STRAIT, *Sample function regularity for Gaussian processes with the parameter in a Hilbert space*, Pacific J. Math., 19 (1966), pp. 159–173.
- [22] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.
- [23] J. K. TUGNAIT, *Continuous-time system identification in compact parameter sets*, IEEE Inform. Theory, IT-31 (1985), pp. 652–659.
- [24] E. WONG, *Stochastic Processes in Information and Dynamical Systems*, McGraw-Hill, New York, 1971.
- [25] J. ZABCZYK, *Remarks on the algebraic Riccati equation in Hilbert spaces*, Appl. Math. Optim., 3 (1976), pp. 251–258.

REPLACEMENT WITH NONCONSTANT OPERATING COST*

R. F. ANDERSON†

Abstract. The long-run average cost problem is considered in the case of a nondecreasing Markov wear process with failure determined by a random threshold. The method of analysis is first, to consider the discounted problem, and then, to let the discount factor go to zero. Simple conditions are given to specify when the do-nothing policy is optimal and when it is not.

Key words. optimal replacement, long-run average cost, optimal stopping, reliability

AMS(MOS) subject classifications. primary 90B25, 62L15; secondary 93E20, 60G40

1. Introduction. Assume that we have a machine that is subjected to deterioration and failure with use. Possible states of the machine will be taken to be x , $0 \leq x < \infty$, where the state $x = 0$ will be considered a new machine and deteriorated states will be taken as $x > 0$. Use of the machine is described by a wear process X_t ; $t \geq 0$, a nondecreasing right-continuous Markov process with left limits. Here X_t is considered the state of the machine at time t subject to the Markovian wear process that is causing the deterioration. Failure of the machine occurs at time $\sigma = \inf \{t: X_t \geq Y\}$, where Y is a positive random variable independent of X_t ; $t \geq 0$. Models of this type have been considered by many authors under the name of shock or wear processes with a random threshold failure. (See Abdel-Hameed [1], [2], Drosen [5], Esary, Marshall, and Proschan [7], Taylor [12], Zuckerman [13].)

We are interested in studying the long-run average optimal replacement problem. A new machine is placed into operation and is subjected to deterioration. At each moment in time before failure we have the option of replacing the current machine by a new one at cost $c(x)$. If the machine fails before a decision to replace it has been made, we must immediately replace it by a new one at cost c_0 . Let $f(x)$ be the operating cost that is taken to be wear-state dependent. Let τ be an X_t ; $t \geq 0$ stopping time and σ be as above. Under the replacement policy τ , the long-run average cost is given by

$$J(\tau) = \frac{E_0 \left[\int_0^{\tau \wedge \sigma} f(X_t) dt + c(X_\tau) I_{(\tau < \sigma)} + c_0 I_{(\tau \geq \sigma)} \right]}{E_0[\tau \wedge \sigma]}.$$

We seek a policy τ that minimizes the cost $J(\tau)$. Let

$$\lambda = \inf_{\tau} J(\tau).$$

The usual method of analysis consists of first solving the parameterized family of stopping problems

$$\mu_\lambda = \inf_{\tau} E_0 \left[\int_0^{\tau \wedge \sigma} (f(X_t) - \lambda) dt + c(X_\tau) I_{(\tau < \sigma)} + c_0 I_{(\tau \geq \sigma)} \right]$$

and then determining a λ so that $\mu_\lambda = 0$. Work in this direction has been done under the assumption $f(x) \equiv k$ and various distributional assumptions on Y . (See Abdel-Hameed [1], [2], Drosen [5], Taylor [12], and Zuckerman [13].)

* Received by the editors April 14, 1986; accepted for publication (in revised form) November 9, 1987.

† Department of Mathematics, University of North Carolina, Charlotte, North Carolina 28223. This work was partially supported by Air Force Office of Scientific Research contract AFOSR-80-0245.

Our method consists of first studying the discounted problem and then letting the discount factor go to zero. (See Robin [9], [10] for previous work using this technique.) What allows this method to succeed is that the uncontrolled process, that is, $\tau \equiv \infty$, consists of letting the machine operate until failure, replacing it by a new one, and letting it operate until failure, etc., can be viewed as a Markov process, which we call the replaced process. This Markov process has an invariant measure depending on the distribution of Y . What is of some interest is that the usual assumption of an exponential rate of convergence to the invariant measure is not required.

In § 2 we establish notation and assumptions and deal with the discounted version of the replacement problem. Section 3 is devoted to the study of the replaced process. Section 4 contains the main technical result for letting the discount go to zero. In § 5 we show that the value function of the discounted replacement problem suitably normalized converges to the value function of the long-run average cost problem. Also simple conditions are given that separate when the do-nothing policy, that is, $\tau \equiv \infty$, is optimal from when it is not.

2. Notation, assumption, and study of the discounted replacement problem. Let $\Omega = D(R^+, R^+)$ be the space of right continuous functions with left limits. Here $R^+ = [0, \infty)$. Let $X_t(\omega) = \omega(t)$ for $\omega \in \Omega$, $\mathcal{F}_t = \sigma(X_s; 0 \leq s \leq t)$, $\mathcal{F}^0 = \mathcal{F}_\infty^0$, and $\mathcal{F}_t, \mathcal{F}_\infty$ be the respective universally completed σ -fields. Let $(\Omega; \mathcal{F}_t, X_t; t \geq 0; P_x; x \in R^+)$ be a homogeneous, nondecreasing, nonnegative Markov process with associated semi-group $T_t; t \geq 0$ defined on $C_b(R^+)$, the space of bounded real-valued continuous functions defined on R^+ with norm taken to be supremum norm. We assume

$$(2.1) \quad T_t; t \geq 0 \text{ is Feller, that is, for } f \in C_b(R^+), T_t f \in C_b(R^+) \text{ and } T_t f \rightarrow f \text{ in supremum norm as } t \rightarrow 0.$$

Let A denote the infinitesimal generator of $T_t; t \geq 0$ and D_A its domain. Assume also that

$$(2.2) \quad X_t; t \geq 0 \text{ is quasi-left continuous, that is, if } \tau_n; n \geq 1 \text{ is a sequence of stopping times with } \tau_n \uparrow \tau, \text{ then } X_{\tau_n} \rightarrow X_\tau \text{ almost surely } P_x \text{ on the set } (\tau < \infty) \text{ (see Dynkin [6a, p. 103]).}$$

Let Y be a positive random variable independent of $X_t; t \geq 0$ with a continuous distribution function $G(y)$. Let $\bar{G}(y) = 1 - G(y)$. Define

$$\sigma = \inf \{t: X_t \geq Y\}$$

and let $H(t) = P_0(\sigma \leq t) = P_0(X_t \geq Y)$. Assume

$$(2.3) \quad G(0) = 0 \quad \text{and} \quad E_0[\sigma] < \infty.$$

Let

$$(2.4) \quad f, c \in C_b(R^+), \quad f, c \geq 0, \quad c_0 > 0 \text{ a constant.}$$

Let $X_t^k; t \geq 0$ and $Y_k, k \geq 1$ be independent copies of $X_t; t \geq 0$ and Y . Define $\sigma_k = \inf \{t: X_t^k \geq Y_k\}$. Suppose $\tau_k; k \geq 1$ are stopping times with respect to $\mathcal{F}_t^k; t \geq 0$, the universal completion of $\sigma(X_s^k; 0 \leq s \leq t)$. Assume for all $n, \sum_{k \geq n} \sigma_k \wedge \tau_k = \infty$ almost surely P_0 . We use the notation $\tilde{\tau} = (\tau_1, \tau_2, \dots)$ and refer to $\tilde{\tau}$ as replacement times. Define for $\alpha > 0$

$$\begin{aligned} \tilde{I}_0^\alpha(\tilde{\tau}) = E_0 & \left[\int_0^{\sigma_1 \wedge \tau} \bar{e}^{\alpha t} f(X_t^1) dt + \bar{e}^{\alpha \sigma_1 \wedge \tau_1} (c(X_{\tau_1}^1) I_{(\tau_1 < \sigma_1)} + c_0 I_{(\tau_1 \geq \sigma_1)}) \right. \\ & \left. + \sum_{n=2}^{\infty} \prod_{j=1}^{n-1} \bar{e}^{\alpha \sigma_j \wedge \tau_j} \left(\int_0^{\sigma_n \wedge \tau_n} \bar{e}^{\alpha t} f(X_t^n) dt + \bar{e}^{\alpha \sigma_n \wedge \tau_n} (c(X_{\tau_n}^n) I_{(\tau_n < \sigma_n)} + c_0 I_{(\tau_n \geq \sigma_n)}) \right) \right]. \end{aligned}$$

Here it is assumed that $P_0(X_0^k = 0) = 1$ for all $k \geq 1$. Let

$$V_0^\alpha = \inf_{\tilde{\tau}} \tilde{I}_0^\alpha(\tilde{\tau}).$$

We seek to find the optimal policy $\hat{\tau}$ so that

$$V_0^\alpha = \tilde{I}_0^\alpha(\hat{\tau}),$$

and we refer to this as the optimal replacement problem.

Let $\tilde{\tau} = (\tau_1, \tau_2, \dots)$ be any replacement sequence and define $\tilde{\tau}^1 = (\tau_2, \tau_3, \dots)$. $\tilde{\tau}^1$ is a replacement sequence and

$$\tilde{I}_0^\alpha(\tilde{\tau}) = E_0 \left[\int_0^{\sigma_1 \wedge \tau_1} \bar{e}^{\alpha t} f(X_t^1) dt + \bar{e}^{\alpha \tau_1 \wedge \sigma} (\tilde{I}_0^\alpha(\tilde{\tau}^1) + c(X_{\tau_1}^1) I_{(\tau_1 < \sigma)} + c_0 I_{(\tau_1 \geq \sigma)}) \right].$$

Define

$$I_0^\alpha(\tau) = E_0 \left[\int_0^{\sigma \wedge \tau} \bar{e}^{\alpha t} f(X_t) dt + \bar{e}^{\alpha \tau \wedge \sigma} (V_0^\alpha + c(X_\tau) I_{(\tau < \sigma)} + c_0 I_{(\tau \geq \sigma)}) \right].$$

It is an easy argument to show that

$$V_0^\alpha = \inf_{\tau} I_0^\alpha(\tau).$$

Remark 2.1. The usual discretization argument will show that for any stopping time τ with respect to \mathcal{F}_t : $t \geq 0$ on the set $(\tau < \infty)$

$$P_0(\sigma \leq \tau | \mathcal{F}_\tau) = P_0(X_\tau \geq Y | \mathcal{F}_\tau) = G(X_\tau)$$

and

$$P_0(\sigma > \tau | \mathcal{F}_\tau) = P_0(X_\tau < Y | \mathcal{F}_\tau) = \bar{G}(X_\tau).$$

On the set $(\tau = \infty)$

$$P_0(\sigma \leq \tau | \mathcal{F}_\tau) = 1 \quad \text{and} \quad P_0(\sigma > \tau | \mathcal{F}_\tau) = 0.$$

The following identities can then be easily established:

$$\begin{aligned} E_0 \left[\int_0^{\sigma \wedge \tau} \bar{e}^{\alpha t} f(X_t) dt \right] &= E_0 \left[\int_0^\tau \bar{e}^{\alpha t} \bar{G}(X_t) f(X_t) dt \right], \\ E_0[\bar{e}^{\alpha \sigma \wedge \tau} c(X_\tau) I_{(\tau < \sigma)}] &= E_0[\bar{e}^{\alpha \tau} c(X_\tau) \bar{G}(X_\tau)], \\ E_0[\bar{e}^{\alpha \sigma \wedge \tau} I_{(\tau \geq \sigma)}] &= E_0 \left[\bar{e}^{\alpha \tau} G(X_\tau) + \int_0^\tau \alpha \bar{e}^{\alpha t} G(X_t) dt \right]. \end{aligned}$$

Define for $x \geq 0$

$$\begin{aligned} J_x^\alpha(\tau) = E_x \left[\int_0^\tau \bar{e}^{\alpha t} (\bar{G}(X_t) f(X_t) + \alpha (V_0^\alpha + c_0) G(X_t)) dt \right. \\ \left. + \bar{e}^{\alpha \tau} (V_0^\alpha + \bar{G}(X_\tau) c(X_\tau) + c_0 G(X_\tau)) \right]. \end{aligned}$$

We see from Remark 2.1 that $J_0^\alpha(\tau) = I_0^\alpha(\tau)$ and so $V_0^\alpha = \inf_{\tau} J_0^\alpha(\tau)$.

THEOREM 2.1. Suppose $c(0) > 0$ and (2.1)–(2.4) are satisfied. Let $V^\alpha(x)$ be the maximal solution of the problem

$$(2.5) \quad \begin{aligned} h &\in C_b(\mathbb{R}^+), \quad h(x) \leq h(0) + \bar{G}(x)c(x) + c_0 G(x), \\ h(x) &\leq \bar{e}^{\alpha t} T_t h(x) + \int_0^t \bar{e}^{\alpha s} T_s (\bar{G}f + \alpha(h(0) + c_0)G)(x) ds; \end{aligned}$$

then

$$(2.6) \quad V^\alpha(x) = \inf_{\tau} J_x^\alpha(\tau)$$

and if

$$(2.7) \quad \hat{\tau} = \inf \{t \geq 0: V^\alpha(X_t) = V^\alpha(0) + \bar{G}(X_t)c(X_t) + c_0 G(X_t)\}$$

then $V^\alpha(x) = J_x^\alpha(\hat{\tau})$. Moreover $V^\alpha(0) = V_0^\alpha = \inf_{\hat{\tau}} \tilde{I}_0^\alpha(\hat{\tau})$ and if $\hat{\tau} = (\hat{\tau}_1, \hat{\tau}_2, \dots)$, where $\hat{\tau}_n$ is defined by (2.7) with X_t^k replacing X_t , then

$$(2.8) \quad V_0^\alpha = \tilde{I}_0^\alpha(\hat{\tau}).$$

Proof. For $b > 0$ and τ a stopping time define

$$(2.9) \quad \begin{aligned} J_x^b(\tau) &= E_x \left[\int_0^\tau \bar{e}^{\alpha t} (\bar{G}(X_t)f(X_t) + \alpha(b + c_0)G(X_t)) dt \right. \\ &\quad \left. + \bar{e}^{\alpha \tau} (b + \bar{G}(X_\tau)c(X_\tau) + c_0 G(X_\tau)) \right], \\ V_b(x) &= \inf_{\tau} J_x^b(\tau). \end{aligned}$$

Under our assumption it follows from Robin [10] or Bensoussan [3] that $V_b(x) \in C_b(\mathbb{R}^+)$ and is the maximal solution of

$$(2.10) \quad \begin{aligned} h &\in C_b(\mathbb{R}^+), \quad h(x) \leq b + \bar{G}(x)c(x) + c_0 G(x), \\ h(x) &\leq \bar{e}^{\alpha t} T_t h(x) + \int_0^t \bar{e}^{\alpha s} T_s (\bar{G}f + \alpha(b + c_0)G)(x) ds. \end{aligned}$$

Moreover if

$$\hat{\tau}_b = \inf \{t: V_b(X_t) = b + \bar{G}(X_t)c(X_t) + c_0 G(X_t)\},$$

then

$$(2.11) \quad V_b(x) = J_x^b(\hat{\tau}_b).$$

Since for $0 < b_0 < b_1$, $J_x^{b_0}(\tau) \leq J_x^{b_1}(\tau)$ for any stopping time τ ,

$$(2.12) \quad V_{b_0}(x) \leq V_{b_1}(x), \quad 0 < b_0 < b_1.$$

Moreover for any $b_0, b_1 \geq 0$

$$J_x^{b_0}(\tau) - J_x^{b_1}(\tau) = E_x \left[\int_0^\tau \alpha \bar{e}^{\alpha t} (b_0 - b_1) G(X_t) dt + \bar{e}^{\alpha \tau} (b_0 - b_1) \right]$$

and therefore

$$(2.13) \quad |V_{b_0}(x) - V_{b_1}(x)| \leq 2(b_0 - b_1).$$

We show the existence of a \bar{b} so that $V_{\bar{b}}(0) = \bar{b}$. To this end we first show there is a $b_0 > 0$ so that $V_{b_0}(0) < b_0$. Recall that $\sigma = \inf \{t: X_t \geq Y\}$. Now

$$H(t) \equiv P_0(\sigma \leq t) = P_0(X_t \geq Y) = E_0[G(X_t)]$$

and so

$$\hat{H}(\alpha) \equiv E_0[\bar{e}^{\alpha\sigma}] = E_0\left[\int_0^\infty \alpha \bar{e}^{\alpha t} G(X_t) dt\right].$$

Since H is not a point mass at 0, $\hat{H}(\alpha) < 1$. Define

$$z_0 = E_0\left[\int_0^\infty \bar{e}^{\alpha t} (f(X_t)\bar{G}(X_t) + \alpha c_0 G(X_t)) dt\right].$$

For $b > 0$, $V_b(0) \leq z_0 + b\hat{H}(\alpha)$. Select $b_0 > 0$ large enough so that $z_0 + b_0\hat{H}(\alpha) < b_0$. By construct then $0 \leq b_1 = V_{b_0}(0) < b_0$, and by (2.12) we have inductively $0 \leq b_k = V_{b_{k-1}}(0) \leq V_{b_{k-2}}(0) = b_{k-1}$, $k \geq 1$. Define $\lim_{k \rightarrow \infty} b_k = \bar{b}$. From (2.13)

$$|V_{b_k}(x) - V_{\bar{b}}(x)| \leq 2|b_k - \bar{b}|$$

and so

$$\bar{b} = \lim_{k \rightarrow \infty} V_{b_k}(0) = V_{\bar{b}}(0).$$

Now from (2.9) and (2.10)

$$(2.14) \quad V_{\bar{b}}(x) = \inf_{\tau} E_x \left[\int_0^{\tau} \bar{e}^{\alpha t} (\bar{G}(X_t)f(X_t) + \alpha (V_{\bar{b}}(0) + c_0)G(X_t)) dt \right. \\ \left. + \bar{e}^{\alpha\tau} (V_{\bar{b}}(0) + \bar{G}(X_{\tau})c(X_{\tau}) + c_0G(X_{\tau})) \right]$$

and $V_{\bar{b}}(x)$ is the maximal solution of

$$(2.15) \quad h \in C_b(R^+), \quad h(x) \leq V_{\bar{b}}(0) + \bar{G}(x)c(x) + c_0G(x), \\ h(x) \leq \bar{e}^{\alpha t} T_t h(x) + \int_0^t \bar{e}^{\alpha s} T_s (\bar{G}f + \alpha (V_{\bar{b}}(0) + c_0)G)(x) ds.$$

Note that (2.15) is different from (2.5); this is the main point remaining.

Let τ be any stopping time. By (2.14) and Remark 2.1

$$V_{\bar{b}}(0) \leq E_0 \left[\int_0^{\sigma \wedge \tau} \bar{e}^{\alpha t} f(X_t) dt + \bar{e}^{\alpha\sigma \wedge \tau} (V_{\bar{b}}(0) + I_{(\tau < \sigma)} c(X_{\tau}) + c_0 I_{(\tau \geq \sigma)}) \right].$$

Therefore for any inspection sequence $\tilde{\tau} = (\tau_1, \tau_2, \dots)$

$$(2.16) \quad V_{\bar{b}}(0) \leq E_0 \left[\int_0^{\sigma_1 \wedge \tau_1} \bar{e}^{\alpha t} f(X_t^1) dt + \bar{e}^{\alpha\sigma_1 \wedge \tau_1} (I_{(\tau_1 < \sigma_1)} c(X_{\tau_1}^1) + c_0 I_{(\tau_1 \geq \sigma_1)}) \right. \\ \left. + \sum_{m=2}^n \prod_{j=1}^{m-1} \bar{e}^{\alpha\sigma_j \wedge \tau_j} \left(\int_0^{\sigma_n \wedge \tau_n} \bar{e}^{\alpha t} f(X_t^n) dt \right. \right. \\ \left. \left. + \bar{e}^{\alpha\sigma_n \wedge \tau_n} (V_{\bar{b}}(0) + I_{(\tau_n < \sigma_n)} c(X_{\tau_n}^n) + c_0 I_{(\tau_n \geq \sigma_n)}) \right) \right].$$

Since $\sum_{k \geq n} \tau_n \wedge \tau_k = \infty$ almost surely P_0 , $V_{\bar{b}}(0) \leq \tilde{I}_0^{\alpha}(\tilde{\tau})$. Note that if $h(\cdot)$ is any solution of (2.5) the same argument shows $h(0) \leq I_0^{\alpha}(\tau)$.

Let $\hat{\tau}_k = \inf \{t: V_{\bar{b}}(X_t^k) = V_{\bar{b}}(0) + \bar{G}(X_t^k)c(X_t^k) + c_0 G(X_t^k)\}$. Since $\hat{\tau}_k$ is optimal for $V_{\bar{b}}$ and by Remark 2.1,

$$(2.17) \quad V_{\bar{b}}(0) = E_0 \left[\int_0^{\hat{\tau}_k \wedge \sigma_k} \bar{e}^{\alpha t} f(X_t^k) dt + \bar{e}^{\alpha \hat{\tau}_k \wedge \sigma_k} (V_{\bar{b}}(0) + c(X_{\hat{\tau}_k}^k) I_{(\hat{\tau}_k < \sigma_k)} + c_0 I_{(\hat{\tau}_k \geq \sigma_k)}) \right].$$

By assumption $c(0) > 0$, and so $V_{\bar{b}}(0) < V_{\bar{b}}(0) + c(0)$. Hence for $\varepsilon > 0$ there is a $\delta > 0$ so that if $0 \leq x \leq \delta$

$$V_{\bar{b}}(x) + \delta < V_{\bar{b}}(0) + \bar{G}(x)c(x) + c_0 G(x).$$

This implies $\hat{\tau}_k > \rho_{k,\delta/2} = \inf \{t: X_t^k \geq \delta/2\}$. Since $X_t: t \geq 0$ is right-continuous, $\tau_{k,\delta/2} > 0$ almost surely P_0 and therefore $\hat{\tau}_k > 0$ almost surely P_0 . Since σ_k is not a point mass at 0, $E_0[\hat{\tau}_k \wedge \sigma_k] \neq 0$ and so $\sum_{k \geq n} \hat{\tau}_k \wedge \sigma_k = \infty$ almost surely P_0 and $\hat{\tau} = (\hat{\tau}_1, \tau_2, \dots)$ is an inspection sequence.

Because at each stage (2.17) is an equality when the $\hat{\tau}_k$'s are used and since $\hat{\tau} = (\hat{\tau}_1, \hat{\tau}_2, \dots)$ is an inspection sequence, we see that $V_{\bar{b}}(0) = \tilde{I}_0^\alpha(\hat{\tau})$, and therefore $V_0^\alpha = V_{\bar{b}}(0)$. Hence for any solution $h(x)$ of (2.5) $h(0) \leq V_{\bar{b}}(0)$ and if $\hat{\tau}$ is optimal for $V_{\bar{b}}$, standard arguments show

$$h(x) \leq E_x \left[\int_0^{\hat{\tau}} \bar{e}^{\alpha t} \bar{G}(X_t) f(X_t) dt + \bar{e}^{\alpha \hat{\tau}} (h(0) + \bar{G}(X_{\hat{\tau}})c(X_{\hat{\tau}}) + c_0 G(x_{\hat{\tau}})) \right] \leq V_{\bar{b}}(x).$$

Hence $V_{\bar{b}}(x)$ is the maximal solution of (2.5).

3. The replaced process. Let $X_t^k: t \geq 0$ and Y^k be as in § 2. As before $\sigma_k = \inf \{t: X_t^k \geq Y^k\}$. Assume $P_0(X_0^k = 0 \text{ for all } k) = 1$ and define $\sigma_0 = 0$. Let $T_k = \sum_{j=0}^k \sigma_j$ and define

$$Z_t = X_{t-T_{k-1}}^k \quad \text{on the set } T_{k-1} \leq t < T_k.$$

We refer to $Z_t: t \geq 0$ as the replaced process. Let $H(t) = P(\sigma \leq t)$, $H^{(k)}(t)$ the k -fold convolution of $H(t)$ and $R(t) = \sum_0^\infty H^{(k)}(t)$. It is standard from renewal theory (see Feller [8]) that (2.3) implies $R(t) < \infty$. From Remark 2.1 for $B \subset R^+$ Borel,

$$\begin{aligned} P_0(Z_t \in B) &= \sum_{k=1}^\infty P_0(T_{k-1} \leq t, X_{t-T_{k-1}}^k \in B, X_{t-T_{k-1}}^k < Y^k) \\ &= \sum_{k=1}^\infty \int_0^t E_0[\bar{G}(X_{t-s}) I_B(X_{t-s})] H^{(k-1)}(ds) \\ &= \int_0^t \int_B p_{t-s}(0, dz) \bar{G}(z) R(ds) \end{aligned}$$

where $p_t(0, dz) = P_0(X_t \in dz)$.

Define Y_x by defining for $y > x$

$$P(Y_x \leq y) = P(Y \leq y | Y > x) = \frac{G(y) - G(x)}{\bar{G}(x)} = G_x(y).$$

Let Y_x^1 have distribution $G_x(y)$ and be independent of the other Y^k 's and the X^k 's. We use the notation P_x to stand for the convention $P_x(X_0^1 = x, X_0^k = 0, k \geq 2) = 1$. Define

$$\sigma_x = \inf \{t: X_t^1 \geq Y_x^1\} \quad \text{and} \quad H_x(t) = P_x(\sigma_x \leq t).$$

By a slight modification of Remark 2.1,

$$(3.1) \quad \begin{aligned} P_x(Z_t \in B) &= E_x[I_B(X_t)\bar{G}_x(X_t)] + \sum_{k=2}^{\infty} \int_0^t E_0[I_B(X_{t-s})\bar{G}(X_{t-s})]H_x * H^{(k-2)}(ds) \\ &= \int_B p_t(x, dy) \frac{\bar{G}(y)}{\bar{G}(x)} + \int_0^t \int_B p_{t-s}(0, dy) \bar{G}(y)H_x * R(ds) \end{aligned}$$

where $p_t(x, dy) = P_x(X_t \in dy)$. A direct calculation will show that the transition probabilities for $Z_t: t \geq 0$ satisfy the Chapman-Kolmogorov equations and therefore $Z_t: t \geq 0$ is Markov.

Remark 3.1. From (3.1) we see

$$1 = \frac{E_x[\bar{G}(X_t)]}{\bar{G}(x)} + \int_0^t E_0[\bar{G}(X_{t-s})]H_x * R(ds)$$

and therefore

$$\int_0^t E_0[\bar{G}(X_{t-s})]H_x * R(ds) = \frac{E_x[G(X_t) - G(x)]}{\bar{G}(x)} = E_x[G_x(X_t)].$$

For $f \in C_b(R^+)$ define $S_t f(z) = E_z[f(Z_t)]$ and so (3.1) gives the identity

$$(3.2) \quad S_t f(x) = \frac{1}{\bar{G}(x)} T_t(\bar{G}f)(x) + \int_0^t T_{t-s}\bar{G}f(0)H_x * R(ds).$$

Suppose

$$(3.3) \quad G \in D_A \quad \text{and} \quad AG(x)/\bar{G}(x) \text{ is bounded and continuous.}$$

Since $T_t: t \geq 0$ is Feller, it is easily seen, from (3.2) and the fact that $H_x(t)$ is weakly continuous in x , that if $f \in C_b(R^+)$ then $S_t f \in C_b(R^+)$. To show that $S_t: t \geq 0$ is strongly continuous, note that by Remark 3.1 and (3.2),

$$|S_t f(x) - f(x)| \leq \left| \frac{T_t(\bar{G}f)(x)}{\bar{G}(x)} - f(x) \right| + \|f\| T_t G_x(x).$$

Now

$$\frac{T_t(\bar{G}f)(x)}{\bar{G}(x)} - f(x) = \frac{T_t(\bar{G}f)(x) - \bar{G}(x)T_t f(x)}{\bar{G}(x)} + T_t f(x) - f(x)$$

and

$$\left| \frac{T_t(\bar{G}f)(x) - \bar{G}(x)T_t f(x)}{\bar{G}(x)} \right| \leq \|f\| T_t G_x(x).$$

Moreover by Dynkin's formula

$$T_t G_x(x) = \frac{1}{\bar{G}(x)} \int_0^t T_s(AG)(x) ds \leq \int_0^t T_s \left(\frac{AG}{\bar{G}} \right)(x) ds \leq t \left\| \frac{AG}{\bar{G}} \right\|.$$

Hence

$$\|S_t f - f\| \leq \|T_t f - f\| + 2t \left\| \frac{AG}{\bar{G}} \right\| \rightarrow 0.$$

Remark 3.2. Let \tilde{A} be the infinitesimal generator of $S_t: t \geq 0$. From (3.1),

$$(3.4) \quad \begin{aligned} \tilde{A}f(x) = & \lim_{t \rightarrow 0} \frac{1}{\bar{G}(x)} \frac{T_t(\bar{G}f)(x) - \bar{G}(x)f(x)}{t} \\ & + \lim_{t \rightarrow 0} \frac{1}{t} \int_0^t T_{t-s}(\bar{G}f)(0)H_x * R(ds). \end{aligned}$$

Now by Remark 3.1,

$$\begin{aligned} \int_0^t T_{t-s}(\bar{G}f)(0)H_x * R(ds) &= \int_0^t E_0[\bar{G}(X_{t-s})(f(X_{t-s}) - f(0))]H_x * R(ds) \\ &\quad + f(0)E_x[G_x(X_t)]. \end{aligned}$$

If $\bar{G}f \in D_A$ and $A(\bar{G}f)(x)/\bar{G}(x)$ is bounded and continuous then

$$\begin{aligned} & \left| \frac{1}{\bar{G}(x)} \frac{T_t \bar{G}f(x) - \bar{G}f(x)}{t} - \frac{A\bar{G}f(x)}{\bar{G}(x)} \right| \\ & \leq \left| \frac{1}{t} \int_0^t \left(T_s \left[\frac{A\bar{G}f}{\bar{G}} \right](x) - \frac{A\bar{G}f(x)}{\bar{G}(x)} \right) ds \right| \\ & \quad + \left| \frac{1}{t} \int_0^t E_x \left[\frac{A\bar{G}f(X_s)}{\bar{G}(X_s)} \left(\frac{\bar{G}(X_s)}{\bar{G}(x)} - 1 \right) \right] ds \right| \\ & \leq \sup_{0 \leq s \leq t} \left\| T_s \left(\frac{A\bar{G}f}{\bar{G}} \right) - \frac{A\bar{G}f}{\bar{G}} \right\| + \left\| \frac{A\bar{G}f}{\bar{G}} \right\| \frac{1}{t} \int_0^t E_x[G_x(X_s)] ds \\ & \leq \sup_{0 \leq s \leq t} \left\| T_s \left(\frac{A\bar{G}f}{\bar{G}} \right) - \frac{A\bar{G}f}{\bar{G}} \right\| + \left\| \frac{A\bar{G}f}{\bar{G}} \right\| \left\| \frac{AG}{\bar{G}} \right\| t / 2. \end{aligned}$$

Also

$$\begin{aligned} & \left| \frac{1}{t} \int_0^t E_0[\bar{G}(X_{t-s})(f(X_{t-s}) - f(0))]H_x * R(ds) \right| \\ & \leq \sup_{0 \leq s \leq t} |T_s f(0) - f(0)| \frac{1}{t} E_x[G_x(X_t)] \\ & \leq \sup_{0 \leq s \leq t} \|T_s f - f\| \left\| \frac{AG}{\bar{G}} \right\| \end{aligned}$$

and

$$\begin{aligned} \left| \frac{1}{t} E_x[G_x(X_t)] - \frac{AG(x)}{\bar{G}(x)} \right| &\leq \left| \frac{1}{t} \int_0^t \left(T_s \left(\frac{AG}{\bar{G}} \right)(x) - \frac{AG(x)}{\bar{G}(x)} \right) ds \right| \\ &\quad + \left| \frac{1}{t} \int_0^t E_x \left[\frac{AG(X_s)}{\bar{G}(X_s)} \left(\frac{\bar{G}(X_s)}{\bar{G}(x)} - 1 \right) \right] ds \right| \\ &\leq \sup_{0 \leq s \leq t} \left\| T_s \left(\frac{AG}{\bar{G}} \right) - \frac{AG}{\bar{G}} \right\| + \left\| \frac{AG}{\bar{G}} \right\|^2 t / 2. \end{aligned}$$

Therefore if $\bar{G}f \in D_A$ and $A\bar{G}f(x)/\bar{G}(x)$ is bounded and continuous then

$$\tilde{A}f(x) = \frac{A\bar{G}f(x)}{\bar{G}(x)} + f(0) \frac{AG(x)}{\bar{G}(x)}.$$

Moreover, a simple check of the above will show that if $f \in D_{\bar{A}}$, then $\bar{G}f \in D_A$ and $A\bar{G}f/\bar{G}$ is continuous and bounded.

THEOREM 3.1. Suppose (2.1) and (2.2) hold and

$$(3.5) \quad E_x[\sigma_x] = \int_0^\infty \frac{E_x[\bar{G}(x_t)]}{\bar{G}(x)} dt < \infty \quad \text{for all } x.$$

Then $Z_t: t \geq 0$ has a unique invariant probability measure Π given by

$$\Pi(A) = \frac{E_0[\int_0^\sigma I_A(X_t) dt]}{E_0[\sigma]} = \frac{E_0[\int_0^\infty I_A(X_t) \bar{G}(X_t) dt]}{E_0[\int_0^\infty \bar{G}(X_t) dt]}.$$

Proof. Remark 2.1 shows the two versions of Π are identical. Recall the convention $P_x(X_0^1 = x, X_0^k = 0, k \geq 2) = 1$. Let $f \in C_b(R^+)$. Since $\sigma_k: k \geq 1$ and $\int_0^{\sigma_k} f(X_t^k) dt: k \geq 1$ are independent and, respectively, identically distributional for $k \geq 2$ with means

$$E_x[\sigma_k] = E_0\left[\int_0^\infty \bar{G}(X_t) dt\right], \quad k \geq 2$$

and

$$E_x\left[\int_0^{\sigma_k} f(X_t) dt\right] = E_0\left[\int_0^\infty f(X_t) \bar{G}(X_t) dt\right], \quad k \geq 2$$

we have by the Strong Law of Large Numbers

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t f(Z_s) ds &= \lim_{n \rightarrow \infty} \frac{1/n \sum_{k=1}^n \int_0^{\sigma_k} f(X_s^k) ds}{1/n \sum_{k=1}^n \sigma_k} \\ &= \frac{E_0[\int_0^\infty f(X_t) \bar{G}(X_t) dt]}{E_0[\int_0^\infty \bar{G}(X_t) dt]} \quad \text{a.s. } P_x \end{aligned}$$

for all x . Hence Π is an invariant measure.

If $\tilde{\Pi}$ is any other invariant measure, for $f \in C_b(R^+)$

$$\begin{aligned} \int_0^\infty f(z) \tilde{\Pi}(dz) &= \int_0^\infty E_z[f(Z_t)] \tilde{\Pi}(dz) = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \int_0^\infty E_z[f(Z_s)] \tilde{\Pi}(dz) ds \\ &= \int_0^\infty \int_0^\infty f(z) \Pi(dz) \tilde{\Pi}(dz) = \int_0^\infty f(z) \Pi(dz). \end{aligned}$$

Remark 3.3. In the article by Robin [9] it is required that

$$|P_x(Z_t \in \Gamma) - \Pi(\Gamma)| \leq C_1 \bar{e}^{\delta_1 t}, \quad \delta_1 > 0$$

and as a consequence

$$\left\| S_t f - \int f(z) \Pi(dz) \right\| \leq C_2 \bar{e}^{\delta_2 t}, \quad \delta_2 > 0.$$

For our case, it can be proved using renewal theory that $\lim_{t \rightarrow \infty} S_t f(x) = \int_0^\infty f(z) \Pi(dZ)$ pointwise, but we will not have in general an exponential rate of convergence.

For what follows we impose the conditions

$$(3.6) \quad P_x(X_t < \infty) = 1 \quad \forall t \text{ and } x, \quad \lim_{t \rightarrow \infty} X_t = \infty \quad \text{almost surely } P_x \quad \forall x;$$

$$(3.7) \quad E_x[\sigma_x] = \frac{E_x[\int_0^\infty \bar{G}(X_s) dt]}{\bar{G}(x)} \leq C \quad \text{independent of } x.$$

Remark 3.4. Let $f \in C_b(R^+)$. The main point of (3.7) is to insure that $E_x[\int_0^\infty \bar{G}(X_t)f(X_t) dt]/\bar{G}(x)$ is continuous and bounded. Clearly

$$\left| \frac{E_x[\int_0^\infty \bar{G}(X_t)f(X_t) dt]}{\bar{G}(x)} \right| \leq \|f\| E_x[\sigma_x] \leq C \|f\|,$$

and continuity follows from the Feller property and the fact that

$$E_x \left[\int_T^\infty \bar{G}(X_t) dt \right] \leq C E_x[\bar{G}(X_T)].$$

Let $f \in C_b(R^+)$ and for $\alpha > 0$ let $u_\alpha(x)$ be the unique solution of $\tilde{A}u_\alpha - \alpha u_\alpha = -f$. It is standard that

$$u_\alpha(z) = \int_0^\infty \bar{e}^{\alpha s} E_z[f(Z_t)] dt.$$

Define

$$\bar{f} = \int_0^\infty f(z) \Pi(dz).$$

THEOREM 3.2. Under the assumption (2.1), (2.2), (3.3), (3.6), and (3.7), we have the following:

- (i) $\lim_{\alpha \rightarrow 0} \alpha u_\alpha(x) = \bar{f}$ uniformly in x ;
- (ii) Let $v_\alpha(x) = u_\alpha(x) - u_\alpha(0)$; then $\lim_{\alpha \rightarrow 0} v_\alpha(x) = v(x)$ uniformly on compact sets where

$$v(x) = \frac{E_x[\int_0^\infty (f(X_t) - \bar{f}) \bar{G}(X_t) dt]}{\bar{G}(x)},$$

and moreover $v(x)$ is the unique solution of

$$-\tilde{A}v = f - \bar{f} \quad \text{with } v(0) = 0.$$

Proof. From (3.1)

$$u_\alpha(x) = \frac{1}{\bar{G}(x)} \int_0^\infty \bar{e}^{\alpha t} T_t(\bar{G}f)(x) dt + \int_0^\infty \bar{e}^{\alpha t} \int_0^t T_{t-s}(\bar{G}f)(0) H_x * R(ds) dt.$$

By (3.7)

$$\left| \frac{1}{\bar{G}(x)} \int_0^\infty \bar{e}^{\alpha t} T_t(\bar{G}f)(x) dt \right| \leq \|f\| E_x[\sigma_x] \leq C \|f\|$$

and so

$$\lim_{\alpha \rightarrow 0} \alpha \frac{1}{\bar{G}(x)} \int_0^\infty \bar{e}^{\alpha t} T_t(\bar{G}f)(x) dt = 0 \quad \text{uniformly in } x.$$

Next

$$\int_0^\infty \bar{e}^{\alpha t} \int_0^t T_{t-s}(\bar{G}f)(0) H_x * R(ds) dt = \int_0^\infty \bar{e}^{\alpha t} T_t(\bar{G}f)(0) dt \hat{H}_x(\alpha) \hat{R}(\alpha)$$

where \hat{H}_x and \hat{R} are the Laplace transforms of H_x and R , respectively. Since

$$\hat{R}(\alpha) = \frac{1}{1 - \hat{H}(\alpha)} \quad \text{where } \hat{H}(\alpha) = E_0[\bar{e}^{\alpha \sigma}]$$

and

$$\lim_{\alpha \rightarrow 0} \frac{1 - \hat{H}(\alpha)}{\alpha} = E_0[\sigma] = E_0 \left[\int_0^\infty \bar{G}(X_t) dt \right],$$

to prove (i) we must still show $\lim_{\alpha \rightarrow 0} \hat{H}_x(\alpha) = 1$ uniformly in x . Now

$$\begin{aligned} 1 - \hat{H}_x(\alpha) &= \frac{\bar{G}(x) - E_x \left[\int_0^\infty \alpha \bar{e}^{\alpha t} (G(X_t) - G(x)) dt \right]}{\bar{G}(x)} \\ &= \frac{E_x \left[\int_0^\infty \alpha \bar{e}^{\alpha t} \bar{G}(X_t) dt \right]}{\bar{G}(x)} \leq \alpha E_x[\sigma_x] \leq \alpha C \end{aligned}$$

and we have uniform convergence in x . To show (ii), first note

$$u_\alpha(0) = E_0 \left[\int_0^\infty \bar{e}^{\alpha t} \bar{G}(X_t) f(X_t) dt \right] \frac{1}{1 - \hat{H}(\alpha)}.$$

Therefore

$$\begin{aligned} u_\alpha(x) - u_\alpha(0) &= \frac{1}{\bar{G}(x)} E_x \left[\int_0^\infty \bar{e}^{\alpha t} \bar{G}(X_t) f(X_t) dt \right] \\ &\quad + E_0 \left[\int_0^\infty \bar{e}^{\alpha t} \bar{G}(X_t) f(X_t) dt \right] \frac{\hat{H}_x(\alpha) - 1}{1 - \hat{H}(\alpha)}. \end{aligned}$$

It is enough to consider $f \geq 0$. Now as $\alpha \rightarrow 0$

$$\frac{1}{\bar{G}(x)} E_x \left[\int_0^\infty \bar{e}^{\alpha t} \bar{G}(X_t) f(X_t) dt \right] \uparrow \frac{1}{\bar{G}(x)} E_x \left[\int_0^\infty \bar{G}(X_t) f(X_t) dt \right]$$

with the limit being bounded and continuous by Remark 3.4. Also

$$\lim_{\alpha \rightarrow 0} E_0 \left[\int_0^\infty \bar{e}^{\alpha t} \bar{G}(X_t) f(X_t) dt \right] \frac{\alpha}{1 - \hat{H}(\alpha)} = \bar{f}$$

and

$$\frac{1 - \hat{H}_x(\alpha)}{\alpha} = \frac{E_x \left[\int_0^\infty \bar{e}^{\alpha t} \bar{G}(X_t) dt \right]}{\bar{G}(x)} \uparrow \frac{E_x \left[\int_0^\infty \bar{G}(X_t) dt \right]}{\bar{G}(x)}.$$

Hence uniformly on compact sets $\lim_{\alpha \rightarrow 0} u_\alpha(x) - u_\alpha(0) = v(x)$.

Since $v(0) = 0$, to show $-\tilde{A}v = f - \bar{f}$, by Remark 3.2 it is enough to prove that $\tilde{G}v \in D_A$ and $-A\tilde{G}v = \tilde{G}(f - \bar{f})$. To this end we have, by the Markov property,

$$\tilde{G}(x)v(x) = E_x \left[\int_0^t \bar{G}(X_s)(f(X_s) - \bar{f}) ds + \bar{G}(X_t)v(X_t) \right].$$

Therefore

$$\begin{aligned} &\left| \frac{T_t \tilde{G}v(x) - \tilde{G}v(x)}{t} + \bar{G}(x)(f(x) - \bar{f}) \right| \\ &\leq \frac{1}{t} \int_0^t \|T_s(\tilde{G}(f - \bar{f}) - \bar{G}(f - \bar{f}))\| ds \rightarrow 0 \quad \text{as } t \rightarrow 0. \end{aligned}$$

Last, for uniqueness, suppose v_1 and v_2 are two solutions of $-\tilde{A}u = f - \bar{f}$, $u(0) = 0$. Let $w = v_1 - v_2$. Then $\tilde{A}w = 0$ and since $w(0) = 0$ it follows that $A\tilde{G}w = 0$. By Dynkin's formula then $\tilde{G}(x)w(x) = E_x[\tilde{G}(X_t)w(X_t)]$ for all t and by (3.6), $\tilde{G}(x)w(x) = \lim_{t \rightarrow \infty} E_x[\tilde{G}(X_t)w(X_t)] = 0$ for all x and so $w(x) = 0$ for all x .

Remark 3.5. It is well known that a necessary condition that $-\tilde{A}v=f$ have a solution is for $\int_0^\infty f(z)\Pi(dz)=0$. By the argument just used in the last part of the proof of Theorem 3.2, it can be proved that $\int_0^\infty f(z)\Pi(dz)=0$ is also a sufficient condition. Moreover any two solutions of $\tilde{A}v=-f$ must differ by a constant.

4. Asymptotics of a stopping problem. Assume that

$$(4.1) \quad f, \Psi \in C_b(R^+) \quad \text{with } f \geq \beta > 0 \text{ and } \Psi \geq 0.$$

Let $G(x)$ be as before. For $\alpha > 0$ and τ a stopping time define

$$(4.2) \quad J_x^\alpha(\tau) = E_x \left[\int_0^\tau \bar{e}^{\alpha t} \bar{G}(X_t) f(X_t) dt + \bar{e}^{\alpha \tau} \bar{G}(X_\tau) \Psi(X_\tau) \right]$$

and let

$$(4.3) \quad J_x(\tau) = E_x \left[\int_0^\tau \bar{G}(X_t) f(X_t) dt + \bar{G}(X_\tau) \Psi(X_\tau) \right].$$

Under the assumption (2.2), (3.6), and (3.7) it is consistent to define $\bar{G}(X_\tau)\Psi(X_\tau)=0$ on the set $(\tau=\infty)$ and note that under (4.1) $\lim_{t \rightarrow \infty} J_x(\tau \wedge t) = J_x(\tau) < \infty$ for any stopping time τ . Define

$$(4.4) \quad u(x) = \inf_\tau J_x(\tau) \quad \text{and} \quad u_\alpha(x) = \inf_\tau J_x^\alpha(\tau).$$

We wish to establish the following generalization of Theorem 3.1 of Robin [9] (see also [9, Remark 3.3].

THEOREM 4.1. *Under (2.1)–(2.3), (3.6), (3.7), and (4.1) we have the following:*

(i) *u is the maximal element of the set of functions*

$$(4.5) \quad \begin{aligned} h &\in C_b(R^+), \quad h(x) \leq \bar{G}(x)\Psi(x), \\ h(x) &\leq T_t h(x) + \int_0^t T_s(\bar{G}f)(x) ds. \end{aligned}$$

(ii) $\hat{\tau} = \inf \{t: u(X_t) = \bar{G}(X_t)\Psi(X_t)\}$ *is optimal, that is*

$$u(x) = J_x(\hat{\tau}).$$

(iii) $u_\alpha \uparrow u$ *uniformly on compact sets.*

The proof will follow after a series of lemmas.

Remark 4.1. The difference between our Theorem 4.1 and Robin's Theorem 3.1 of [9] is the presence of $\bar{G}(x)$ in our case. In Robin's result, which also assumes (4.1), for $\hat{\tau}$ optimal, it must be that $E_x[\hat{\tau}] < \infty$. For our case, since $\lim_{t \rightarrow \infty} J_x(\tau \wedge t) = J_x(\tau)$ we have

$$(4.6) \quad u(x) = \inf \{J_x(\tau): E_x[\tau] > \infty\},$$

but it may happen that for $\hat{\tau}$ optimal, $E_x[\hat{\tau}] = \infty$.

At this stage in the proof it is necessary to introduce the penalized problem and establish properties of it. Let

$\gamma_t: t \geq 0$ be nonanticipating with respect to $\mathcal{F}_t: t \geq 0$ with $0 \leq \gamma_t \leq 1$.

For $\varepsilon > 0$ define

$$J_x^\varepsilon(\gamma) = E_x \left[\int_0^\infty \bar{e}^{1/\varepsilon} \int_0^t \gamma_s ds \left(\bar{G}(X_t) f(X_t) + \frac{1}{\varepsilon} \gamma_t \bar{G}(X_t) \Psi(X_t) \right) dt \right]$$

and let

$$u_\varepsilon(x) = \inf_{\gamma} J_x^\varepsilon(\gamma).$$

Remark 4.2. We claim that

$$u_\varepsilon(x) = \inf \left\{ J_x^\varepsilon(\gamma) : \int_0^\infty \gamma_t dt = \infty \text{ a.s. } P_x \right\}.$$

Suppose $P_x(\int_0^\infty \gamma_t dt < \infty) > 0$ and let $\delta > 0$. It is enough to show there is a $\tilde{\gamma}_t : t \geq 0$ so that $\int_0^\infty \tilde{\gamma}_t dt = \infty$ and

$$J_x^\varepsilon(\tilde{\gamma}) \leq J_x^\varepsilon(\gamma) + \delta.$$

From (3.7) we have

$$J_x^\varepsilon(\gamma) \leq E_x \left[\int_0^\infty \bar{G}(X_t) dt \right] \left(\|f\| + \frac{1}{\varepsilon} \|\Psi\| \right) < \infty,$$

and hence there is a T_0 so that if $T \geq T_0$

$$E_x \left[\int_0^\infty \bar{e}^{1/\varepsilon \int_0^t \gamma_s ds} \left(\bar{G}(X_t) f(X_t) + \frac{1}{\varepsilon} \gamma_t \bar{G}(X_t) \Psi(X_t) \right) dt \right] < \delta/2.$$

Since (3.6) implies that $\lim_{t \rightarrow \infty} E_x[\bar{G}(X_t)] = 0$, we can find a T_1 so that if $T \geq T_1$, $E_x[\bar{G}(X_T) \Psi(X_T)] < \delta/2 \|\Psi\|$. Select $T \geq T_0 \vee T_1$ and define

$$\tilde{\gamma}_t = \begin{cases} \gamma_t, & t \leq T, \\ 1, & t > T. \end{cases}$$

It then is a simple matter of checking to see that

$$J_x^\varepsilon(\tilde{\gamma}) \leq J_x^\varepsilon(\gamma) + \delta.$$

Define

$$J_x^{\varepsilon,n}(\gamma) = E_x \left[\int_0^\infty \bar{e}^{1/\varepsilon \int_0^t \gamma_s ds} \left((\bar{G}(X_t) V \frac{1}{n} f(X_t) + \frac{1}{\varepsilon} \gamma_t \bar{G}(X_t) \Psi(X_t)) \right) dt \right]$$

and

$$J_x^{\varepsilon,\alpha}(\gamma) = E_x \left[\int_0^\infty \bar{e}^{\alpha t} \bar{e}^{1/\varepsilon \int_0^t \gamma_s ds} \left(\bar{G}(X_t) f(X_t) + \frac{1}{\varepsilon} \gamma_t \bar{G}(X_t) \Psi(X_t) \right) dt \right].$$

Let

$$u_{\varepsilon,n}(x) = \inf_{\gamma} J_x^{\varepsilon,n}(\gamma) \quad \text{and} \quad u_{\varepsilon,\alpha}(x) = \inf_{\gamma} J_x^{\varepsilon,\alpha}(\gamma).$$

Section 3.3 of [9] establishes that $u_{\varepsilon,n} \in C_b(R^+)$. See either [3] or [10] to find that $u_{\varepsilon,\alpha}(x) \in C_b(R^+)$.

LEMMA 4.2. (i) $u_{\varepsilon,n}(x) \downarrow u_\varepsilon(x)$.

(ii) $u_{\varepsilon,\alpha}(x) \uparrow u_\varepsilon(x)$.

(iii) $u_\varepsilon(x)$ is the unique nonnegative solution in $C_b(R^+)$ of

$$(4.7) \quad u_\varepsilon(x) = T_1 u_\varepsilon(x) + \int_0^T T_s \left(\bar{G}f - \frac{1}{\varepsilon} (u_\varepsilon - \bar{G}\Psi)^+(x) \right) ds$$

and if

$$\hat{\gamma}_t = \begin{cases} 1, & u_\varepsilon(X_t) \geq \bar{G}(X_t)\Psi(X_t), \\ 0, & u_\varepsilon(X_t) < \bar{G}(X_t)\Psi(X_t), \end{cases}$$

then

$$(4.8) \quad u_\varepsilon(x) = J_x^\varepsilon(\hat{\gamma}).$$

Proof. For (i), note that if $\gamma_t \equiv 1$, $J_x^{\varepsilon,n}(\gamma) \leq \varepsilon/n\|f\| + \|\Psi\| < \infty$.

If, on the other hand, for $P_x(\int_0^\infty \gamma_t dt < \infty) > 0$, then (4.1) implies

$$J_x^{\varepsilon,n}(\gamma) \geq \frac{\beta}{n} E_x \left[\int_0^\infty \bar{e}^{1/\varepsilon \int_0^t \gamma_s ds} dt \right] = \infty,$$

and therefore

$$(4.9) \quad u_{\varepsilon,n}(x) = \inf \left\{ J_x^{\varepsilon,n}(\gamma) : \int_0^\infty \gamma_t dt = \infty \text{ a.s. } P_x \right\}.$$

Since, if $P_x(\int_0^\infty \gamma_t dt = \infty) = 1$ then $J_x^{\varepsilon,n}(\gamma) \downarrow J_x^\varepsilon(\gamma)$ as $n \rightarrow \infty$, it follows from Remark 4.2 and (4.9) that $u_{\varepsilon,n}(x) \downarrow u_\varepsilon(x)$.

For (ii) observe that for all $\gamma_t: t \geq 0$, $J_x^{\varepsilon,\alpha}(\gamma) \uparrow J_x^\varepsilon(\gamma)$ as $\alpha \rightarrow 0$, and so $u_{\varepsilon,\alpha}(x)$ increases to a function $\tilde{u}_\varepsilon(x)$ as $\alpha \rightarrow 0$, where $\tilde{u}_\varepsilon(x) \leq u_\varepsilon(x)$. By taking $\gamma_t \equiv 1$ we obtain the inequality

$$\begin{aligned} \tilde{u}_\varepsilon(x) &\leq u_\varepsilon(x) \leq \bar{G}(x) E_x \left[\int_0^\infty \bar{e}^{1/\varepsilon t} \left(f(X_t) + \frac{1}{\varepsilon} \Psi(X_t) \right) dt \right] \\ &\leq \bar{G}(x) (\varepsilon \|f\| + \|\Psi\|) \end{aligned}$$

and so

$$(4.10) \quad \lim_{x \rightarrow \infty} \tilde{u}_\varepsilon(x) = 0.$$

From [3] or [10] the standard results yield the identity

$$u_{\varepsilon,\alpha}(x) = \bar{e}^{\alpha t} T_t u_{\varepsilon,\alpha}(x) + \int_0^t \bar{e}^{\alpha s} T_s \left(\bar{G}f - \frac{1}{\varepsilon} (u_{\varepsilon,\alpha} - \bar{G}\Psi)^+ \right)(x) ds$$

and so

$$\tilde{u}_\varepsilon(x) = T_t \tilde{u}_\varepsilon(x) + \int_0^t T_s \left(\bar{G}f - \frac{1}{\varepsilon} (\tilde{u}_\varepsilon - \bar{G}\Psi)^+ \right)(x) ds.$$

For any $\gamma_t: t \geq 0$ integration by parts yields that

$$(4.11) \quad \bar{e}^{1/\varepsilon \int_0^t \gamma_s ds} \tilde{u}_\varepsilon(X_t) + \int_0^t \bar{e}^{1/\varepsilon \int_0^s \gamma_u du} \left(\bar{G}(X_s) f(X_s) + \frac{1}{\varepsilon} \gamma_s \tilde{u}_\varepsilon(X_s) - \frac{1}{\varepsilon} (\tilde{u}_\varepsilon - \bar{G}\Psi)^+(X_s) \right) ds$$

is a martingale. Letting

$$\tilde{\gamma}_t = \begin{cases} 0, & \tilde{u}_\varepsilon(X_t) < \bar{G}(X_t)\Psi(X_t), \\ 1, & \tilde{u}_\varepsilon(X_t) \geq \bar{G}(X_t)\Psi(X_t) \end{cases}$$

in (4.11) and after taking expectations, letting $t \rightarrow \infty$, and using (3.6) and (4.10) we obtain $\tilde{u}_\varepsilon(x) = J_x^\varepsilon(\tilde{\gamma})$ and so $\tilde{u}_\varepsilon(x) = u_\varepsilon(x)$.

Parts (i) and (ii) show that $u_\varepsilon(x) \in C_b(R^+)$ and (4.8) have been established, so what is left of (iii) is the uniqueness of nonnegative solutions of (4.7). If $w_\varepsilon(x) \in C_b(R^+)$ and is a nonnegative solution of (4.7), then (4.11) is valid for $w_\varepsilon(x)$ and so $w_\varepsilon(x) \leq J_x^\varepsilon(\gamma)$ for and $\gamma_t: t \geq 0$. Also since (4.10) is valid for $w_\varepsilon(x)$ and if

$$\hat{\gamma}_t = \begin{cases} 0, & w_\varepsilon(X_t) < \bar{G}(X_t)\Psi(X_t), \\ 1, & w_\varepsilon(X_t) \geq \bar{G}(X_t)\Psi(X_t) \end{cases}$$

we obtain $w_\varepsilon(x) = J_x^\varepsilon(\hat{\gamma})$ and so $w_\varepsilon(x) = u_\varepsilon(x)$.

Remark 4.3. Let

$$J_x^n(\tau) = E_x \left[\int_0^\tau \left(\bar{G}(X_t) V \frac{1}{n} \right) f(X_t) dt + \bar{G}(X_\tau) \Psi(X_\tau) \right]$$

and $u_n(x) = \inf_\tau J_x^n(\tau)$; then $u_n(x) \in C_b(R^+)$ and $u_n(x) \downarrow u(x)$. To see this note that for the case $u_n(x)$, Theorem 3.1 of [9] applies and so $u_n \in C_b(R^+)$. Note that for any stopping time τ with $E_x[\tau] < \infty$, $J_x^n(\tau) \downarrow J_x(\tau)$ and therefore $u_n(x) \downarrow$ and $u_n(x) \geq u(x)$. For $\varepsilon > 0$ choose τ so that $0 \leq J_x(\tau) - u(x) < \varepsilon$. Since $\lim_{t \rightarrow \infty} J_x(\tau \wedge t) = J_x(\tau)$ and $\lim_{n \rightarrow \infty} J_x^n(\tau \wedge t) = J_x(\tau \wedge t)$, we can select t_0 and n_0 large enough so that $|J_x^{n_0}(\tau \wedge t) - J_x(\tau)| < \varepsilon$. Hence $u_{n_0}(x) - u(x) < 2\varepsilon$ and $u_n(x) \downarrow u(x)$.

Remark 4.4. Observe that Lemma 4.2 is valid when $\bar{G}(x)\Psi(x)$ is replaced by

$$(4.12) \quad \Psi_\beta(x) = E_x \left[\int_0^\infty \bar{e}^{\beta t} \bar{G}(X_t) \Psi(X_t) dt \right]$$

since

$$\Psi_\beta(x) \leq \frac{1}{\beta} \bar{G}(x) \|\Psi\|.$$

LEMMA 4.3. $u_\varepsilon(x) \downarrow u(x)$ as $\varepsilon \rightarrow 0$.

Proof. An argument due to Menaldi (see [3] or [4]) shows that for $\alpha > 0$, $u_{\varepsilon, \alpha} \downarrow$ as $\alpha \rightarrow 0$. Lemma 4.2 then yields the fact that $u_\varepsilon \downarrow$ as $\varepsilon \rightarrow 0$.

Define $\tau_\varepsilon = \inf \{t: u_\varepsilon(X_t) \geq \bar{G}(X_t)\Psi(X_t)\}$. Lemma 4.2(iii) implies

$$u_\varepsilon(x) = E_x \left[u_\varepsilon(X_{\tau_\varepsilon \wedge t}) + \int_0^{\tau_\varepsilon \wedge t} \bar{G}(X_s) f(X_s) ds \right].$$

On $(\tau_\varepsilon = \infty)$, $u_\varepsilon(X_t) < \bar{G}(X_t)\Psi(X_t)$ for all t and so $\lim_{t \rightarrow \infty} u_\varepsilon(X_{\tau_\varepsilon \wedge t}) = 0$. On $(\tau_\varepsilon < \infty)$ because of (2.2)

$$\lim_{t \rightarrow \infty} u_\varepsilon(X_{\tau_\varepsilon \wedge t}) = u_\varepsilon(X_{\tau_\varepsilon}) \geq \bar{G}(X_{\tau_\varepsilon})\Psi(X_{\tau_\varepsilon}).$$

Hence we can conclude that

$$(4.13) \quad u_\varepsilon(x) \geq J_x(\tau_\varepsilon) \geq u(x).$$

Let $\Psi_\beta(x)$ be as in (4.12). Define $w_{\varepsilon, \beta}(x)$ and $w_\beta(x)$ by substituting Ψ_β in place of $\bar{G}\Psi$ in the definitions of u_ε and u . The same argument that was used to establish (4.13) shows $w_{\varepsilon, \beta}(x) \geq w_\beta(x)$.

Suppose $E_x[\tau] < \infty$ and define

$$\gamma_t^\tau = \begin{cases} 0, & t < \tau, \\ 1, & t \geq \tau. \end{cases}$$

Now for the Ψ_β problem

$$J_x^{\varepsilon, \beta}(\gamma^\tau) - J_x^\beta(\tau) = E_x \left[\int_\tau^\infty \bar{e}^{1/\varepsilon(t-\tau)} \left(\bar{G}(X_t) f(X_t) + \frac{1}{\varepsilon} \Psi_\beta(X_t) \right) dt - \Psi_\beta(X_\tau) \right].$$

Since $\Psi_\beta \in D_A$, it follows that

$$\Psi_\beta(X_{\tau+t}) \bar{e}^{1/\varepsilon t} - \int_0^t \bar{e}^{1/\varepsilon s} \left(A \Psi_\beta(X_{\tau+s}) - \frac{1}{\varepsilon} \Psi_\beta(X_{\tau+s}) \right) ds$$

is a martingale with respect to $\mathcal{G}_s = \mathcal{F}_{\tau+s}$. Hence by taking expectations and letting $t \rightarrow \infty$, we obtain

$$\begin{aligned} -E_x[\Psi_\beta(X_\tau)] &= E_x \left[\int_0^\infty \bar{e}^{1/\varepsilon t} \left(A \Psi_\beta(X_{\tau+t}) - \frac{1}{\varepsilon} \Psi_\beta(X_{\tau+t}) \right) dt \right] \\ &= E_x \left[\int_\tau^\infty \bar{e}^{1/\varepsilon(t-\tau)} \left(A \Psi_\beta(X_t) - \frac{1}{\varepsilon} \Psi_\beta(X_t) \right) dt \right]. \end{aligned}$$

Therefore

$$\begin{aligned} J_x^{\varepsilon, \beta}(\gamma^\tau) - J_x^\beta(\tau) &= E_x \left[\int_0^\infty \bar{e}^{1/\varepsilon(t-\tau)} (\bar{G}(X_t) f(X_t) - A \Psi_\beta(X_t)) dt \right] \\ &\leq \varepsilon \|f - A \Psi_\beta\|. \end{aligned}$$

For the same reason as in (4.6), $w_\beta(x) = \inf \{J_x^\beta(\tau) : E_x[\tau] < \infty\}$ and so $w_\beta(x) \leq w_{\varepsilon, \beta}(x) \leq w_\beta(x) + \varepsilon \|f - A \Psi_\beta\|$. Therefore

$$w_{\varepsilon, \beta} \downarrow w_\beta(x).$$

It is a standard fact from semigroup theory that

$$\lim_{\beta \rightarrow \infty} \|\Psi_\beta - \bar{G}\Psi\| = 0.$$

Moreover since

$$\|w_\beta - u\| \leq \|\Psi_\beta - \bar{G}\Psi\| \quad \text{and} \quad \|w_{\varepsilon, \beta} - u_\varepsilon\| \leq \|\Psi_\beta - \bar{G}\Psi\|$$

it follows that $u_\varepsilon(x) \downarrow u(x)$ as $\varepsilon \rightarrow 0$.

LEMMA 4.4. $u_\alpha \uparrow u$.

Proof. Let

$$w_{\alpha, \beta}(x) = \inf_\tau E_x \left[\int_0^\tau \bar{e}^{\alpha t} \bar{G}(X_t) f(X_t) dt + \bar{e}^{\alpha \tau} \Psi_\beta(X_\tau) \right]$$

and $w_{\varepsilon, \alpha, \beta}(x)$ be analogously defined. Then as before

$$\|w_{\varepsilon, \alpha, \beta} - w_{\alpha, \beta}\| \leq \varepsilon \|f - A \Psi_\beta\|.$$

Letting $w_{\varepsilon, \beta}$ and w_β be as in the proof of Lemma 4.3,

$$\begin{aligned} |w_{\alpha, \beta}(x) - w_\beta(x)| &\leq \|w_{\alpha, \beta} - w_{\varepsilon, \alpha, \beta}\| + |w_{\varepsilon, \alpha, \beta}(x) - w_{\varepsilon, \beta}(x)| + \|w_{\alpha, \beta} - w_\beta\| \\ &\leq 2\varepsilon \|f - A \Psi_\beta\| + |w_{\varepsilon, \alpha, \beta}(x) - w_{\varepsilon, \beta}(x)|. \end{aligned}$$

By Remark 4.4 and Lemma 4.2 $w_{\varepsilon, \alpha, \beta} \uparrow w_{\varepsilon, \beta}$ as $\varepsilon \rightarrow 0$ and therefore $w_{\alpha, \beta}(x) \rightarrow w_\beta(x)$ as $\alpha \rightarrow 0$.

Now

$$\begin{aligned} |u_\alpha(x) - u(x)| &\leq |w_{\alpha, \beta}(x) - u_\alpha(x)| + |w_{\alpha, \beta}(x) - w_\beta(x)| + |w_\beta(x) - u(x)| \\ &\leq 2\|\Psi_\beta - \bar{G}\Psi\| + |w_{\alpha, \beta}(x) - w_\beta(x)|. \end{aligned}$$

Letting $\alpha \rightarrow 0$, we have that $\beta \rightarrow \infty$ yields $\lim_{\alpha \rightarrow 0} u_\alpha(x) = u(x)$. It is then clear that $u_\alpha(x) \uparrow u(x)$ since for all stopping times τ , $J_x^\alpha(\tau) \uparrow J_x(\tau)$.

Proof of Theorem 4.1. From Lemmas 4.2 and 4.4 we have $u_\varepsilon(x) \downarrow u(x)$ and $u_\alpha(x) \uparrow u(x)$. Hence $u \in C_b(R^+)$ and $u_\alpha \rightarrow u$ uniformly on compact sets.

Since we know from the general theory (see [3] or [10]) that $u_\alpha(x)$ is the maximal solution of the problem

$$h \in C_b(R^+), \quad h \leq \bar{G}\Psi$$

and

$$h(x) \leq \bar{e}^{\alpha t} T_t h(x) + \int_0^t \bar{e}^{\alpha s} T_s(\bar{G}f)(x) ds,$$

it follows by letting $\alpha \rightarrow 0$ that $u(x)$ is a solution of (4.5). It is also standard that if h is any solution of (4.5), $h(x) \leq J_x(\tau)$ for any stopping time τ and therefore $h(x) \leq u(x)$. What is left to show is that

$$\bar{\tau} = \inf \{t: u(X_t) = \bar{G}(X_t)\Psi(X_t)\};$$

then $u(x) = J_x(\hat{\tau})$.

To this end let $B = \{x: u(x) = \bar{G}(x)\Psi(x)\}$. If $x \in B$, then $\tau \equiv 0$ and there is nothing to prove. If $x \notin B$, find $\delta > 0$ so that

$$u(x) + \delta < \bar{G}(x)\Psi(x).$$

Define for $R > 0$, $\tau_R = \inf \{t: |X_t - x| \geq R\}$ and $\tau_\delta = \inf \{t: u(X_t) + \delta \geq \bar{G}(X_t)\Psi(X_t)\}$. Since $u \in C_b(R^+)$, Lemma 4.3 yields that $u_\varepsilon(x) \rightarrow u(x)$ uniformly on compact sets. Choose $\varepsilon_{\delta,R}$ so that if $0 < \varepsilon < \varepsilon_{\delta,R}$

$$\sup_{|x-y| < R} |u_\varepsilon(y) - u(y)| < \delta/2.$$

Thus for $t \in [0, \tau_\delta \wedge \tau_R)$

$$u_\varepsilon(X_t) \leq u(X_t) + \frac{\delta}{2} \leq \bar{G}(X_t)\Psi(X_t) - \frac{\delta}{2}.$$

So from Lemma 4.2(iii),

$$u_\varepsilon(x) = E_x \left[u_\varepsilon(X_{\tau_\delta \wedge \tau_R}) + \int_0^{\tau_\delta \wedge \tau_R} \bar{G}(X_t)f(X_t) dt \right].$$

Letting $\varepsilon \rightarrow 0$, we obtain

$$(4.14) \quad u(x) = E_x \left[u(X_{\tau_\delta \wedge \tau_R}) + \int_0^{\tau_\delta \wedge \tau_R} \bar{G}(X_t)f(X_t) dt \right].$$

Since $0 \leq u(x) \leq \bar{G}(x)\Psi(x)$, $\lim_{x \rightarrow \infty} u(x) = 0$ and so by (3.6) $\lim_{t \rightarrow \infty} u(X_t) = 0$. Moreover by (2.2), we can then let $R \rightarrow \infty$ in (4.14) to obtain

$$u(x) = E_x \left[u(X_{\tau_\delta}) + \int_0^{\tau_\delta} \bar{G}(X_t)f(X_t) dt \right].$$

Last, we note that $\tau_\delta \uparrow \hat{\tau}$. This is so because $\tau_\delta \uparrow \hat{\tau}$ as $\delta \rightarrow 0$, and by (2.2), $X_{\tau_\delta} \rightarrow X_{\hat{\tau}}$ on $(\hat{\tau} < \infty)$ as $\delta \rightarrow 0$. Thus $u(X_{\hat{\tau}}) \leq \bar{G}(X_{\hat{\tau}})\Psi(X_{\hat{\tau}})$ on $(\hat{\tau} < \infty)$. Since $\hat{\tau} \leq \hat{\tau}$, we have $\hat{\tau} = \hat{\tau}$

on $(\tilde{\tau} < \infty)$ because of the definition of $\hat{\tau}$. Since $\tilde{\tau} \leq \hat{\tau}$, in general, $\tilde{\tau} = \hat{\tau}$ on $(\tilde{\tau} = \infty)$. Hence

$$\begin{aligned} u(x) &= \lim_{\delta \downarrow 0} E_x \left[u(X_{\tau_\delta}) + \int_0^{\tau_\delta} \bar{G}(X_t) f(X_t) dt \right] \\ &= E \left[u(X_{\hat{\tau}}) + \int_0^{\hat{\tau}} \bar{G}(X_t) f(X_t) dt \right] = J_x(\hat{\tau}), \end{aligned}$$

which finishes the argument.

5. The long-run average cost problem. Let $V^\alpha(x)$ be the value function for the Discounted Replacement Problem of Theorem 2.1. Recall that

$$\begin{aligned} (5.1) \quad V^\alpha(x) &= \inf_{\tau} E_x \left[\int_0^{\tau} \bar{e}^{\alpha t} (\bar{G}(X_t) f(X_t) + \alpha(V^\alpha(0) + c_0)G(X_t)) dt \right. \\ &\quad \left. + \bar{e}^{\alpha \tau} (V^\alpha(0) + \bar{G}(X_\tau) c(X_\tau) + c_0 G(X_\tau)) \right]. \end{aligned}$$

Let

$$r(x) = \frac{AG(x)}{\bar{G}(x)}.$$

Since $X_t: t \geq 0$ and $G(x)$ are both nondecreasing, it follows that $r(x) \geq 0$.

Define

$$(5.2) \quad \mu = \frac{E_0[\int_0^\infty \bar{G}(X_t)(f(X_t) + c_0 r(X_t)) dt]}{E_0[\int_0^\infty \bar{G}(X_t) dt]}.$$

Remark 5.1. By Dynkin's formula

$$0 = E_0 \left[G(X_t) - \int_0^t AG(X_s) ds \right].$$

By (3.6) we obtain, letting $t \rightarrow \infty$,

$$1 = E_0 \left[\int_0^\infty \bar{G}(X_t) r(X_t) dt \right].$$

Hence (5.2) has the form

$$(5.3) \quad \mu = \frac{E_0[\int_0^\infty \bar{G}(X_t) f(X_t) dt] + c_0}{E_0[\int_0^\infty \bar{G}(X_t) dt]}.$$

Remark 5.2. We claim

$$(5.4) \quad 0 < \lambda = \overline{\lim} \alpha V^\alpha(0) \leq \mu.$$

First recall that

$$H(t) = P_0(\sigma \leq t) = P_0(X_t \geq Y) = E_0[G(X_t)]$$

where $\sigma = \inf \{t: X_t \geq Y\}$. Hence

$$\hat{H}(\alpha) = E_0[\bar{e}^{\alpha \sigma}] = E_0 \left[\int_0^\infty \alpha \bar{e}^{\alpha t} G(X_t) dt \right].$$

From (5.1), letting $\tau \equiv \infty$, we obtain

$$0 \leq V^\alpha(0) \leq E_0 \left[\int_0^\infty \bar{e}^{\alpha t} \bar{G}(X_t) f(X_t) dt \right] + (V^\alpha(0) + c_0) \hat{H}(\alpha)$$

and so

$$(5.5) \quad 0 \leq \alpha V^\alpha(0) \leq \frac{E_0[\int_0^\infty \bar{e}^{\alpha t} \bar{G}(X_t) f(X_t) dt] + c_0 \hat{H}(\alpha)}{(1 - \hat{H}(\alpha))/\alpha}.$$

Since

$$\lim_{\alpha \rightarrow 0} \frac{1 - \hat{H}(\alpha)}{\alpha} = E_0[\sigma] = E_0 \left[\int_0^\infty \bar{G}(X_t) dt \right],$$

(5.4) follows from (5.5).

Remark 5.3. Define

$$\mu_0(x) = \frac{E_x[\int_0^\infty \bar{G}(X_t)(f(X_t) + c_0 r(X_t) - \mu) dt]}{\bar{G}(x)}$$

and let

$$\rho = \sup \mu_0(x) \quad \text{and} \quad \mu_\rho(x) = \mu_0(x) - \rho.$$

Now $\mu_\rho(x) \leq 0$, and since $\mu_0(0) = 0$, $\rho \geq 0$. By Theorem 3.2, $\mu_\rho(x)$ is the unique solution of

$$-\tilde{A}\mu_\rho(x) = f(x) + c_0 r(x) - \mu$$

with $\mu_\rho(0) = -\rho$. Also

$$\tilde{A}\mu_\rho(x) = \frac{A(\bar{G}\mu_\rho)(x)}{\bar{G}(x)} + \mu_\rho(0) \frac{AG(x)}{\bar{G}(x)} = \frac{A(\bar{G}\mu_\rho)(x)}{\bar{G}(x)} - \rho r(x).$$

Therefore we have both

$$(5.6) \quad \begin{aligned} & \bar{e}^{\alpha t} \bar{G}(X_t) \mu_\rho(X_t) + \int_0^t \bar{e}^{\alpha s} (\bar{G}(X_s)(f(X_s) + c_0 r(X_s) \\ & \quad - (\mu + \rho r(X_s))) + \alpha \bar{G}(X_s) \mu_\rho(X_s)) ds \end{aligned}$$

and

$$(5.7) \quad \bar{G}(X_t) \mu_\rho(X_t) + \int_0^t \bar{G}(X_s)(f(X_s) + c_0 r(X_s) - (\mu + \rho r(X_s))) ds$$

are martingales.

Remark 5.4. By Dynkin's formula

$$\bar{e}^{\alpha t} G(X_t) - \int_0^t \bar{e}^{\alpha s} (\bar{G}(X_s) r(X_s) - \alpha \bar{G}(X_s)) ds$$

and

$$G(X_t) - \int_0^t \bar{G}(X_s) r(X_s) ds$$

are both martingales.

Define

$$(5.8) \quad \bar{V}^\alpha(x) = V^\alpha(x) - V^\alpha(0).$$

From the identity

$$(1 - \bar{e}^{\alpha\tau}) V^\alpha(0) = \int_0^\tau \alpha \bar{e}^{\alpha t} dt V^\alpha(0)$$

we see from (5.1) that

$$(5.9) \quad \bar{V}^\alpha(x) = \inf_\tau E_x \left[\int_0^\tau \bar{e}^{\alpha t} (\bar{G}(X_t)(f(X_t) - \alpha V^\alpha(0)) + \alpha c_0 G(X_t)) dt + \bar{e}^{\alpha\tau} (\bar{G}(X_\tau)c(X_\tau) + c_0 G(X_\tau)) \right].$$

Let λ be as in (5.4) and

$$(5.10) \quad \bar{V}(x) = \inf_\tau E_x \left[\int_0^\tau \bar{G}(X_t)(f(X_t) - \lambda) dt + \bar{G}(X_\tau)c(X_\tau) + c_0 G(X_\tau) \right].$$

Suppose

$$(5.11) \quad \mu + \rho r(x) - \lambda \geq \delta > 0.$$

Remark 5.5. Since $\rho \geq 0$ and $r(x) \geq 0$, (5.10) is implied by

$$(5.12) \quad \mu - \lambda > 0.$$

LEMMA 5.1. Let $\alpha_n \rightarrow 0$ so that

$$\lim_{n \rightarrow \infty} \alpha_n V^{\alpha_n}(0) = \lambda.$$

Under the assumption (2.1)–(2.4), (3.3), (3.6), (3.7), (5.11), and $c(0) > 0$ and $f(x) \geq \beta > 0$.

(i) $\lim_{n \rightarrow \infty} \bar{V}^{\alpha_n}(x) = \bar{V}(x)$ uniformly on compact sets.

Moreover $\bar{V}(x)$ is the maximal solution of

$$(5.13) \quad \begin{aligned} h &\in C_b(R^+), \quad h(x) \leq \bar{G}(x)c(x) + c_0 G(x), \\ h(x) &\leq T_t h(x) + \int_0^t T_s(\bar{G}(f - \lambda)) ds. \end{aligned}$$

(ii) $\bar{V}(0) = 0$.

(iii) $\hat{\tau} = \inf \{t: \bar{V}(X_t) = \bar{G}(X_t)c(X_t) + c_0 G(X_t)\}$ is optimal, that is,

$$\bar{V}(x) = E_x \left[\int_0^{\hat{\tau}} \bar{G}(X_t)(f(X_t) - \lambda) dt + \bar{G}(X_{\hat{\tau}})c(X_{\hat{\tau}}) + c_0 G(X_{\hat{\tau}}) \right].$$

Proof. Define $w^\alpha(x) = \bar{V}^\alpha(x) - \bar{G}(x)\mu_\rho(x) - c_0 G(x)$ and $w(x) = \bar{V}(x) - \bar{G}(x)\mu_\rho(x) - c_0 G(x)$. From (5.9) and Remarks 5.3 and 5.4,

$$\begin{aligned} w^\alpha(x) = \inf_\tau E_x \left[\int_0^\tau \bar{e}^{\alpha t} \bar{G}(X_t)(\mu + \rho r(X_t) - \alpha V^\alpha(0)) dt \right. \\ \left. + \bar{e}^{\alpha\tau} \bar{G}(X_\tau)(c(X_\tau) - \mu_\rho(X_\tau)) \right]. \end{aligned}$$

From Remarks 5.3 and 5.4,

$$(5.14) \quad w(x) = \inf_\tau E_x \left[\int_0^\tau \bar{G}(X_t)(\mu + \rho r(X_t) - \lambda) dt + \bar{G}(X_\tau)(c(X_\tau) - \mu_\rho(X_\tau)) \right].$$

Note that (2.2) and (3.6) are required when τ is not finite almost surely P_x in (5.14).

Next define

$$\bar{w}^\alpha(x) = \inf_{\tau} E_x \left[\int_0^{\tau} \bar{e}^{\alpha t} \bar{G}(X_t) (\mu + \rho r(X_t) - \lambda) dt + \bar{e}^{\alpha \tau} \bar{G}(X_{\tau}) (c(X_{\tau}) - \mu_{\rho}(X_{\tau})) \right].$$

Now

$$|\bar{w}^\alpha(x) - w^\alpha| \leq (|\lambda - \alpha V^\alpha(0)| + \alpha \|\mu_{\rho}\|) E_x \left[\int_0^{\infty} \bar{G}(X_t) dt \right].$$

By (3.7), $E_x[\int_0^{\infty} \bar{G}(X_t) dt] \leq C$ independent of x and so $\lim_{n \rightarrow \infty} \bar{w}^{\alpha_n}(x) - w^{\alpha_n}(x) = 0$ uniformly in x . From (5.10) and the facts $\rho \geq 0$ and $c(x) - \mu_{\rho}(x) \geq 0$, we obtain from Theorem 4.1 that $\lim_{\alpha \rightarrow 0} \bar{w}^\alpha(x) = w(x)$ uniformly on compact sets. Hence

$$|\bar{V}^{\alpha_n}(x) - \bar{V}(x)| = |w^{\alpha_n}(x) - w(x)| \leq |w^{\alpha_n}(x) - \bar{w}^{\alpha_n}(x)| + |\bar{w}^{\alpha_n}(x) - w(x)|$$

and therefore $\lim_{n \rightarrow \infty} \bar{V}^{\alpha_n}(x) = \bar{V}(x)$ uniformly on compact sets.

Theorem 4.1 also yields that $w(x)$ is the maximal solution of

$$\begin{aligned} h &\in C_v(R^+), \quad h(x) \leq \bar{G}(x)(c(x) - \mu_{\rho}(x)), \\ h(x) &\leq T_t h(x) + \int_0^t T_s(\bar{G}(\mu + \rho r - \lambda))(x) ds. \end{aligned}$$

From Remarks 5.3 and 5.4 we conclude that $\bar{V}(x)$ is the maximal solution of (5.13) which finishes the proof of (i).

By (5.8), $\bar{V}^\alpha(0) = 0$, and hence $\bar{V}(0) = 0$ showing (ii). Theorem 4.1 implies $\hat{\tau} = \inf \{t: w(X_t) = \bar{G}(X_t)(c(X_t) - \mu_{\rho}(X_t))\}$ is optimal for (5.14) and therefore optimal for $\bar{V}(x)$ since $w(x) + \bar{G}(x)\mu_{\rho}(x) + c_0 G(x) = \bar{V}(x)$.

THEOREM 5.2. *Under the assumptions of Lemma 5.1, we have the following:*

- (i) $\lim_{\alpha \rightarrow 0} \alpha V^\alpha(0) = \lambda$.
- (ii) $\lim_{\alpha \rightarrow 0} V^\alpha(x) - V^\alpha(0) = \bar{V}(x)$ uniformly on compact sets where $\bar{V}(x)$ is given by (5.10) and satisfies (5.13).
- (iii) $\bar{V}(0) = 0$ and letting $\sigma = \inf \{t: X_t \geq Y\}$

$$(5.15) \quad \lambda = \inf_{\tau} \frac{E_0[\int_0^{\tau} f(X_t) dt + I_{(\tau < \sigma)} c(X_{\tau}) + c_0 I_{(\tau \geq \sigma)}]}{E_0[\tau \wedge \sigma]}.$$

- (iv) If $\hat{\tau} = \inf \{t: \bar{V}(X_t) = \bar{G}(X_t)c(X_t) + c_0 G(X_t)\}$ then

$$(5.16) \quad \lambda = \frac{E_0[\int_0^{\hat{\tau} \wedge \sigma} f(X_t) dt + I_{(\hat{\tau} < \sigma)} c(X_{\hat{\tau}}) + c_0 I_{(\hat{\tau} \geq \sigma)}]}{E_0[\hat{\tau} \wedge \sigma]}.$$

Proof. We first prove (iii) and (iv). By Lemma 5.1, $\bar{V}(0) = 0$, and if τ is any stopping time

$$(5.17) \quad 0 \leq E_0 \left[\int_0^{\tau} \bar{G}(X_t) (f(X_t) - \lambda) dt + \bar{G}(X_{\tau}) c(X_{\tau}) + c_0 G(X_{\tau}) \right]$$

and so

$$(5.18) \quad \lambda \leq \frac{E_0[\int_0^{\tau} \bar{G}(X_t) f(X_t) dt + \bar{G}(X_{\tau}) c(X_{\tau}) + c_0 G(X_{\tau})]}{E_0[\int_0^{\tau} \bar{G}(X_t) dt]}.$$

Since $\hat{\tau}$ is optimal we will have equality in (5.17) and (5.18) when $\hat{\tau}$ is used. From Remark 2.1 we see that (5.17) and (5.18) have forms (5.15) and (5.16).

For (i) and (ii), let $\alpha'_n \rightarrow 0$ with $\lim_{n \rightarrow \infty} \alpha'_n V^{\alpha'_n}(0) = \bar{\lambda}$. By repeating the argument for Lemma 5.1 we see that $\lim_{n \rightarrow \infty} \bar{V}^{\alpha'_n}(x) = \bar{V}(x)$, where $\bar{\lambda}$ is substituted for λ in the definition of $\bar{V}(x)$. By parts (iii) and (iv) we must have $\lambda = \bar{\lambda}$. Hence $\lim_{\alpha \rightarrow 0} \alpha V^\alpha(0) = \lambda$ and $\lim_{\alpha \rightarrow 0} \bar{V}^\alpha(x) = \bar{V}(x)$ uniformly on compact sets.

Remark 5.6. The do-nothing policy is to take $\tau \equiv \infty$. The cost of the do-nothing policy is then

$$\mu = \frac{E_0[\int_0^\infty \bar{G}(X_t) f(X_t) dt] + c_0}{E_0[\int_0^\infty \bar{G}(X_t) dt]}.$$

Therefore the do-nothing policy is optimal if and only if $\lambda = \mu$.

THEOREM 5.3. *Suppose $X_t: t \geq 0$ has the property that for any open set $U \subset R^+$, U not containing a neighborhood of the origin, and if $\tau_U = \inf\{t: X_t \in U\}$, then $P_0(\tau_U < \infty)$. Under the assumptions of Theorem 5.2 if $\lambda < \mu$ then $\{x: c(x) < \mu_0(x)\} \neq \emptyset$. Moreover if $r(x) \geq \beta > 0$ and $\{x: c(x) < \mu_0(x)\} \neq \emptyset$ then $\lambda < \mu$.*

Proof. Suppose $\lambda < \mu$ and let $\hat{\tau}$ be optimal. From Remark 5.6 $P_0(\hat{\tau} < \infty) > 0$. Since

$$0 = \bar{V}(0) = E_0 \left[\int_0^{\hat{\tau}} \bar{G}(X_t) (f(X_t) + c_0 r(X_t) - \lambda) dt + \bar{G}(X_{\hat{\tau}}) c(X_{\hat{\tau}}) \right]$$

and also

$$0 = \mu_0(0) = E_0 \left[\bar{G}(X_{\hat{\tau}}) \mu_0(X_{\hat{\tau}}) + \int_0^{\hat{\tau}} \bar{G}(X_t) (f(X_t) + c_0 r(X_t) - \mu) dt \right]$$

we obtain

$$0 < E_0 \left[\int_0^{\hat{\tau}} \bar{G}(X_t) (\mu - \lambda) dt \right] = E_0 [\bar{G}(X_{\hat{\tau}}) (\mu_0(X_{\hat{\tau}}) - c(X_{\hat{\tau}}))].$$

Hence $\{x: c(x) < \mu_0(x)\} \neq \emptyset$.

Suppose $r(x) \geq \beta > 0$ and $\{x: c(x) < \mu_0(x)\} \neq \emptyset$. Note that $\rho > 0$ in this case. Choose $\delta > 0$ so that $U = \{x: c(x) + \delta < \mu_0(x)\} \neq \emptyset$ and since $\mu_0(0) = 0$, U does not contain a neighborhood of the origin. Letting $\tau_U = \inf\{t: X_t \in U\}$, we have $P_0(0 < \tau_U < \infty) = 1$. Theorem 5.2 then applies and

$$0 \leq E_0 \left[\int_0^{\tau_U} \bar{G}(X_t) (\mu - \lambda) dt + \bar{G}(X_{\tau_U}) (c(X_{\tau_U}) - \mu_0(X_{\tau_U})) \right].$$

Therefore $0 < \delta E_0[\bar{G}(X_{\tau_U})] \leq E_0[\int_0^{\tau_U} \bar{G}(X_t) (\mu - \lambda) dt]$, and so $\lambda < \mu$.

REFERENCES

- [1] M. ABDEL-HAMEED, *Life distribution properties of devices subject to pure jump damage processes*, J. Appl. Probab., 21 (1984), pp. 816-825.
- [2] ———, *Life distribution properties of devices subject to Lévy wear processes*, Math. Oper. Res., 9 (1984), pp. 606-614.
- [3] A. BENSOUSSAN, *Stochastic Control by Functional Analysis Methods*, North-Holland, Amsterdam, New York, 1982.
- [4] A. BENSOUSSAN AND J. L. LIONS, *Impulse Control and Quasi-Variational Inequalities*, Gauthier-Villars, Paris, 1984.
- [5] J. W. DROSEN, *Failure times and optimal stopping rules of generalized shock models*, Ph.D. thesis, Northwestern Univ., Evanston, IL, 1983.
- [6a] E. B. DYNKIN, *Markov Processes*, Vol. I, Springer-Verlag, New York, 1965.
- [6b] ———, *Markov Processes*, Vol. II, Springer-Verlag, New York, 1965.
- [7] J. D. ESARY, A. W. MARSHALL, AND F. PROSCHAN, *Shock models and wear processes*, Ann. Probab., 1 (1973), pp. 627-649.

- [8] W. FELLER, *An Introduction to Probability Theory and Its Applications*, Vol. II, John Wiley, New York, 1966.
- [9] M. ROBIN, *On some impulse control problems with long run average cost*, SIAM J. Control Optim., 19 (1981), pp. 333–358.
- [10] ———, *Contrôle impulsif des processus de Markov*, Thèse d'état, Université de Paris IX, 1976.
- [11] ———, *Long-term average cost problems for continuous time Markov processes: a survey*, Acta Appl. Math., 1 (1983), pp. 281–299.
- [12] H. TAYLOR, *Optimal replacement under additive damage and other failure models*, Naval Res. Logist. Quart., 22 (1975), pp. 1–18.
- [13] D. ZUCKERMAN, *Optimal replacement policy for the case where the damage process is a one-sided Lévy process*, Stochastic Process. Appl., 7 (1978), pp. 141–151.

AN EVASION GAME WITH BARRIERS*

V. J. BASTON† AND F. A. BOSTOCK†

Abstract. This paper considers the following two-person zero-sum multistage game. An evader starts at some given point A_s in the ordered set of points A_0, A_1, \dots, A_n and at discrete intervals of time $t = 1, 2, \dots$ chooses to move to one of the points adjacent to him or stay where he is. A gunner with a single bullet may at each of the same discrete intervals of time either fire the bullet at any of the points A_r or hold his fire. The gunner always hits the point at which he aims and the bullet takes one unit of time to reach its target. The payoff to the gunner is 1 if he hits the evader, μ (where $|\mu| < 1$) if he fires and misses, and 0 if he never fires. It is shown that, whatever point A_s the evader starts at, the value of the game is $(1 + \mu)/2$.

Key words. two-person game, zero-sum game, time lag system, recursive matrix game

AMS(MOS) subject classifications. 90D05, 90D20

1. Introduction. Let A_r denote the point $(r, 0)$ of the x -axis, where r is any integer, and consider the following multistage zero sum game. An evader starts at some given point A_s and at discrete intervals of time $t = 1, 2, \dots$ chooses to move to one of the points adjacent to him or to stay where he is. A gunner with a single bullet may at each of the same discrete intervals of time either fire the bullet at any one of the points A_r or hold his fire. It is assumed that the gunner always hits the point at which he aims and that the bullet takes one unit of time to reach its target. The payoff to the gunner is 1 if he hits the evader, μ (where $|\mu| < 1$) if he fires and misses, and zero if he never fires. The solution of this game is of course very easy. If $\mu \geq -\frac{1}{2}$ suppose that, as soon as the game starts with the evader at the point A_s , the gunner employs the strategy whereby he fires at the points A_{s-1}, A_s, A_{s+1} each with probability $\frac{1}{3}$. Then clearly the gunner's expectation is at least $(1 + 2\mu)/3$; if $\mu < -\frac{1}{2}$ the gunner never fires and has an expectation of zero. This shows the value of the game is at least $\max\{0, (1 + 2\mu)/3\}$. On the other hand, if the evader employs the strategy whereby at each interval of time he chooses to move to the right, to stay where he is, and to move to the left with equal probability, then clearly the gunner's expectation will be at most $\max\{0, (1 + 2\mu)/3\}$. Thus the value of the game is $\max\{0, (1 + 2\mu)/3\}$ and is independent of the starting point A_s . In his paper [6] Lee has considered a variation of this trivial game (with j bullets and $\mu = 0$) in which one of the points, say A_0 , is designated as a safe point, that is, if the evader at some time reaches this point then the game is deemed to end. The payoff to the gunner in Lee's game is the number of times he hits the evader.

In this paper we will consider the game which, instead of having a safe point, has two barriers, say at the points A_0 and A_N , and the gunner has one bullet. The game starts at one of the points A_0, A_1, \dots, A_N and all of the future play takes place only on these points. So if the evader at some time arrives at a barrier, then at the next time interval, instead of having the usual three choices he has now only two, namely staying where he is or moving to the single permitted adjacent point. Note that if any game starts with the evader actually at a barrier, then there are obvious strategies for the players which show the value of such a game to be $(1 + \mu)/2$. The game with barriers is in marked contrast to the game with a safe point; in the latter there is a

* Received by the editors March 18, 1987; accepted for publication (in revised form) November 20, 1987.

† Faculty of Mathematical Sciences, University of Southampton, Southampton, SO9 5NH, United Kingdom.

point to which the evader is essentially attracted, while in the former there are points to be avoided. This fundamental difference in the games is forcefully reflected in the nature of optimal strategies for the gunner. With a safe point, as Lee shows, there is an optimal strategy for the gunner where he does not need the option to hold his fire; in contrast, as we will see, in the game with two barriers he needs very much to avail himself of this option.

We will solve our game with barriers by modeling it as a recursive matrix game. The theory of recursive games is not that straightforward and the interested reader may wish to consult [3] and [7], or at an elementary level [8]. We will have occasion to use a result which as far as we can determine is not readily available in the literature; mainly for this reason, but also for the sake of completeness, we will outline the relevant details of the general theory in § 2. This will essentially follow Everett [3], who does not permit the player to use a strategy which depends on the past history of the game.

The problem can also be modeled as a stochastic game with absorbing states and a long-run average reward criterion; such games have aroused considerable interest and there is now an extensive body of literature on them [1], [2], [4], [5], [9]. In this approach the players can use strategies which depend on the past history of the game, whereas in our model they cannot. This could be a serious objection to our model, as Blackwell and Ferguson [2] and Thuijsman and Vrieze [9] have constructed history-remembering strategies for the Big Match and the Bad Match, respectively, which are better than any history-forgetting strategy. However, Orkin [7] has shown that, for recursive matrix games, stationary optimal, or ε -optimal strategies (which always exist) remain so when the strategy spaces of the players are enlarged to include history remembering ones (see our comments just after Lemma 1). This enables a recursive matrix game to be regarded as a special case of a stochastic game, and implies that any stationary optimal (ε -optimal) strategy resulting from the recursive matrix game formulation remains optimal (ε -optimal) in the stochastic game formulation.

2. Recursive matrix games. A recursive matrix game Γ consists of a finite number $\Gamma_1, \Gamma_2, \dots, \Gamma_n$ of so-called game components. Each game component Γ_r is essentially a matrix game but with the difference that the entries of the matrix need not be numbers but may be probabilities of playing other game components. The two players P_1 and P_2 start with one of the game components at time $t=1$. They choose a row and a column, respectively. If the corresponding entry is a number, then that number is the payoff and play terminates. If the entry is not a number, a game component Γ_r is chosen according to the given probabilities and the players then play Γ_r at time $t=2$. Should a number turn up at this stage, that number is the payoff and play terminates; otherwise play continues in the obvious manner. In the event of an infinite play the payoff is defined to be zero. A strategy X for P_1 in Γ is defined as a sequence $X = X^1, X^2, \dots, X^t, \dots$, where each term X^t is a sequence X^t_1, \dots, X^t_n with the significance that X^t_r is a mixed strategy to be used at time t in the game component Γ_r . Note that X^1 covers all possible starts at $t=1$. If, for a particular r , X^t_r is independent of t , X is said to be stationary in the r th component, and a strategy which is stationary in all components is simply described as stationary. A strategy Y for P_2 is defined similarly.

Now suppose the players use the strategies X and Y . Let q_{ij}^t denote the conditional probability when in Γ_i of going to Γ_j under X^t_i, Y^t_i , and denote the matrix (q_{ij}^t) by $Q^t(X, Y)$. Let $E^t_i(X, Y)$ be the contribution to the expectation at time t conditional upon being in Γ_i , and denote the corresponding vector by $E^t(X, Y)$. At $t=1$ play starts in Γ_r ; let $E_r^{(t)}(X, Y)$ denote P_1 's expectation up to and including time t , and

denote the corresponding vector by $E^{(t)}(X, Y)$. It is then clear that

$$(1) \quad E^{(t)}(X, Y) = \sum_{j=0}^{t-1} \left(\prod_{i=0}^j Q^i(X, Y) \right) E^{j+1}(X, Y)$$

where $Q^0(X, Y)$ denotes the unit matrix I of order n . As $t \rightarrow \infty$ it is not difficult to see that $E^{(t)}(X, Y)$ converges to, say, $E(X, Y)$, and because a zero payoff is assigned to an unending play, this limit is the expectation which results when the players use the strategies X and Y . For each real n -vector $W = (w_1, \dots, w_n)$ let $\Gamma_r(W)$ denote the real matrix obtained by substituting w_i for each game component Γ_i which occurs in the matrix for Γ_r . Regarded as an ordinary matrix game, the value of $\Gamma_r(W)$ is denoted by $\text{val } \Gamma_r$, and the value map V from real n -vectors to real n -vectors is defined by taking the r th component of $V(W)$ as $\text{val } \Gamma_r$. We let $E'_i(X, Y; W)$ signify the expectation pertaining to the use of the strategies X'_i and Y'_i in the ordinary matrix game $\Gamma_i(W)$, and we denote the corresponding vector by $E'(X, Y; W)$.

Now suppose that X and W are such that for all Y and t , $E'(X, Y; W) \geq W$. Then since clearly $E'(X, Y; W) = E'(X, Y) + Q'(X, Y)W$, it follows that, for all Y and t , $E'(X, Y) \geq (I - Q'(X, Y))W$. It is then an easy deduction from 1 that, for all Y and t ,

$$E^{(t)}(X, Y) \geq W - \left(\prod_{i=1}^t Q^i(X, Y) \right) W.$$

This result essentially proves the following lemma.

LEMMA 1. *In the recursive matrix game $\Gamma = (\Gamma_1, \dots, \Gamma_n)$ suppose X is a strategy for P_1 and W is an n -vector satisfying the following:*

- (i) *For all strategies Y for P_2 and for all t , $E'(X, Y; W) \geq W$;*
- (ii) *For all strategies Y for P_2 , $\prod_{i=1}^t Q^i(X, Y) \rightarrow 0$ as $t \rightarrow \infty$;*

then for all strategies Y for P_2 , $E(X, Y) \geq W$.

We observe that if W is an n -vector which satisfies $V(W) \geq W$ and $*X$ is the stationary strategy for P_1 where, for all t , $*X'_t$ is defined as an optimal strategy for player one in the ordinary matrix game $\Gamma_i(W)$, then condition (i) of Lemma 1 would clearly be satisfied. Since $*X$ is stationary, the conclusion of Lemma 1 regarding $*X$ is valid when the strategy space for P_2 is augmented by the set of history remembering strategies (see Orkin [7]). Later in our investigation of the game with barriers we will consider a particular strategy $*X$ which we will also see satisfies condition (ii).

3. The double barrier game. We take barriers at the points A_0 and A_N ($N \geq 2$) and let Γ_r ($r = 0, 1, \dots, N$) represent the game where the evader starts at the point A_r . It has already been remarked that the games Γ_0 and Γ_N each have the value $(1 + \mu)/2$. Also by symmetry we note that the value of Γ_r is equal to the value of Γ_{N-r} for $r = 1, 2, \dots, [N/2]$, where $[x]$ denotes the integer part of x . Although the analysis for N odd and N even turns out to be generally the same, there are some differences which we feel merit a separate treatment of the cases. We find it convenient to treat N even in detail, this being the somewhat harder case, and at the end indicate the variations to be encountered when N is odd. It also turns out that a solution of the game Γ_r for $|\mu| < 1$ may be deduced from the single case $\mu = 0$.

In view of the above remarks, we let $N = 2M$, and regard the problem in terms of the recursive matrix game $\Gamma = (\Gamma_1, \Gamma_2, \dots, \Gamma_M)$, where

$$\Gamma_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & \Gamma_1 & \Gamma_2 \end{bmatrix},$$

$$\Gamma_r = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \Gamma_{r-1} & \Gamma_r & \Gamma_{r+1} \end{bmatrix} \quad \text{for } r=2, 3, \dots, M-1,$$

$$\Gamma_M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \Gamma_{M-1} & \Gamma_M & \Gamma_{M-1} \end{bmatrix}.$$

In Γ_r the columns 1, 2, and 3, respectively, represent the pure strategies for the evader in which he moves to the points A_{r-1} , A_r , A_{r+1} ; the first three rows, respectively, represent the pure strategies for the gunner in which he fires at the points A_{r-1} , A_r , A_{r+1} ; the fourth row represents that of not firing. Let W_0 denote the zero M -vector and for $k=1, 2, \dots$ define $W_k = (w_1^k, \dots, w_M^k)$ by $W_k = V(W_{k-1})$, that is, W_k is the k th iterate of the zero vector under the value map V . Since all the number entries in all the game component matrices are nonnegative, and because the value of an ordinary matrix game is an increasing function of its entries it is easy to see by induction that the sequence W_k is increasing. We will see that each sequence w_r^k converges to the value of the game where the players are considered to start in Γ_r . The next three lemmas enable us to find the limits of the sequences w_r^k and to find optimal or ε -optimal strategies for the players.

LEMMA 2. Let a, b, c be real numbers satisfying $\frac{1}{3} \leq a, b, c \leq \frac{1}{2}$, $a \geq b$, and $a \geq c$. Then the value of the two-person zero-sum matrix game

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ a & b & c \end{bmatrix}$$

is given by $v = a/(1+2a-b-c)$.

Proof. Consider strategies X, Y for players one and two given by $X = (0, a-b, a-c, 1)/D$ and $Y = (1-b-c, a, a)/D$, where $D = 1+2a-b-c$. It is easy to verify that, using X , player one's expectation against each of columns 1, 2, and 3 is precisely v ; also when player two uses Y player one's expectation against rows two, three, and four is precisely v and against row one is $(1-b-c)/D \leq a/D = v$. Thus the value of the matrix game is v and the lemma is proved. \square

LEMMA 3. We have the following:

- (i) For $r=1, \dots, M$; $k \geq 1$, $\frac{1}{3} \leq w_r^k < \frac{1}{2}$.
- (ii) For $r=1, \dots, M-1$; $k \geq 1$, $w_r^k \geq w_{r+1}^k$.
- (iii) For $r=1, \dots, M$, $k \geq 2$, $w_r^k = w_{r-1}^{k-1}/D_r^{k-1}$, where $D_r^{k-1} = 1+2w_{r-1}^{k-1}-w_r^{k-1}-w_{r+1}^{k-1}$, and for all $k \geq 2$ we define $w_0^{k-1} = \frac{1}{2}$ and $w_{M+1}^{k-1} = w_{M-1}^{k-1}$.

Proof. The strategies $(1, 1, 1, 0)/3$ for player one and $(1, 1, 1)/3$ for player two show that for $r=1, \dots, M$ $w_r^1 = \frac{1}{3}$, and also show that the assumption $\frac{1}{3} \leq w_r^k < \frac{1}{2}$ for $r=1, \dots, M$ implies $\frac{1}{3} \leq w_r^{k+1} < \frac{1}{2}$ for $r=1, \dots, M$. This proves (i).

Since $w_r^1 = \frac{1}{3}$ for $r=1, \dots, M$ we have $w_r^1 \geq w_{r+1}^1$ for $r=1, \dots, M-1$. Assume as induction hypothesis that $w_r^k \geq w_{r+1}^k$ for $r=1, \dots, M-1$ and for all $k \geq 1$ define $w_0^k = \frac{1}{2}$ and $w_{M+1}^k = w_{M-1}^k$. By (i) above, all of the conditions of Lemma 2 with $a = w_{r-1}^k$, $b = w_r^k$ and $c = w_{r+1}^k$ are seen to hold concerning the component game Γ_r . Thus $w_r^{k+1} = w_{r-1}^k/D_r^k$ so that (iii) holds at $k+1$.

We now establish (ii) at $k+1$; this will complete the proof of the lemma. For $r=1, \dots, M-1$ we have

$$\begin{aligned} w_r^{k+1} - w_{r+1}^{k+1} &= (w_{r-1}^k / D_r^k) - (w_r^k / D_{r+1}^k) \\ &= \{w_{r-1}^k(1 - w_{r+1}^k - w_{r+2}^k) - w_r^k(1 - w_r^k - w_{r+1}^k)\} / D_r^k D_{r+1}^k. \end{aligned}$$

From (i), the definitions of w_0^k and w_{M+1}^k , and the induction hypothesis, it follows that $D_r^k D_{r+1}^k > 0$, $w_{r-1}^k \geq w_r^k$, and $0 < 1 - w_r^k - w_{r+1}^k \leq 1 - w_{r+1}^k - w_{r+2}^k$ for $r=1, \dots, M-1$. Thus $w_r^{k+1} \geq w_{r+1}^{k+1}$; this concludes the proof of the lemma. \square

LEMMA 4. For $r=0, 1, \dots, M-1$ and $k \geq M$, $w_r^k > w_{r+1}^k$.

Proof. Recall in the proof of Lemma 3 that for $r=1, \dots, M-1$,

$$w_r^{k+1} - w_{r+1}^{k+1} = \{w_{r-1}^k(1 - w_{r+1}^k - w_{r+2}^k) - w_r^k(1 - w_r^k - w_{r+1}^k)\} / D_r^k D_{r+1}^k$$

where $D_r^k D_{r+1}^k > 0$ and $0 < 1 - w_r^k - w_{r+1}^k \leq 1 - w_{r+1}^k - w_{r+2}^k$. Thus if $w_{r-1}^k > w_r^k$ then $w_r^{k+1} > w_{r+1}^{k+1}$. However, by Lemma 3(i) and the definition of w_0^k , we have $w_0^k > w_1^k$ for all $k \geq 1$. This clearly proves the lemma. \square

Since the sequence W_k is increasing and is bounded above (by Lemma 3(i)), each sequence w_r^k , $k=1, 2, \dots$ converges, say to w_r . From Lemma 3 we immediately deduce that

$$(a) \quad \frac{1}{2} \geq w_1 \geq w_2 \geq \dots \geq w_M \geq \frac{1}{3};$$

$$(b) \quad \text{For } r=1, \dots, M \quad w_r = w_{r-1} / (1 + 2w_{r-1} - w_r - w_{r+1}),$$

whence $w_{r+1} = 1 + 2w_{r-1} - w_r - (w_{r-1}/w_r)$, where $w_0 = \frac{1}{2}$ and $w_{M+1} = w_{M-1}$.

It follows easily from (b) that if $w_{r-1} = w_r = \frac{1}{2}$ then $w_{r+1} = \frac{1}{2}$. Thus it follows from (a) that if for some $r=1, \dots, M$, $w_r = \frac{1}{2}$ then, for all $r=1, \dots, M$, $w_r = \frac{1}{2}$.

The next lemma shows, perhaps rather surprisingly, that w_r is independent of r .

LEMMA 5. For each $r=1, \dots, M$, $w_r = \frac{1}{2}$.

Proof. As we have just seen, for $r=1, \dots, M$, $w_r = w_{r-1} / (1 + 2w_{r-1} - w_r - w_{r+1})$, where $w_0 = \frac{1}{2}$ and $w_{M+1} = w_{M-1}$. Thus, if we use $r=M$, $w_M = w_{M-1} / (1 + w_{M-1} - w_M)$, whence $(w_{M-1} - w_M)(w_M - 1) = 0$. Therefore, $w_{M-1} = w_M$ since $w_M \leq \frac{1}{2}$. Using $r=M-1$, $w_{M-1} = w_{M-2} / (1 + 2w_{M-2} - w_{M-1} - w_M) = w_{M-2} / (1 + 2w_{M-2} - 2w_{M-1})$, whence $(w_{M-2} - w_{M-1})(2w_{M-1} - 1) = 0$. If $w_{M-1} = \frac{1}{2}$ then the result follows by our observations immediately preceding the theorem, so we may assume that $w_{M-2} = w_{M-1}$. Using $r=M-2$, $w_{M-2} = w_{M-3} / (1 + 2w_{M-3} - w_{M-2} - w_{M-1}) = w_{M-3} / (1 + 2w_{M-3} - 2w_{M-2})$, whence $(w_{M-3} - w_{M-2})(2w_{M-2} - 1) = 0$. As before we must have $w_{M-3} = w_{M-2}$. Continuing in this manner as far as $r=1$, we see that $w_0 = w_1$, whence $w_1 = \frac{1}{2}$ and the theorem is proved. \square

At this point we consider briefly the situation when $N=2M+1$ is odd, that is, with barriers at the points A_0 and A_{2M+1} . The only variation in the definition of the recursive game $\Gamma = (\Gamma_1, \Gamma_2, \dots, \Gamma_M)$ is that now

$$\Gamma_M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \Gamma_{M-1} & \Gamma_M & \Gamma_M \end{bmatrix};$$

this has the effect that w_{M+1}^k is defined as w_M^k . Lemma 2 of course still stands. Lemma 3 holds so long as we now define, in (iii), $w_{M+1}^{k-1} = w_M^{k-1}$, and the proof holds with only very minor changes concerning the game component Γ_M . Both the statement and the proof of Lemma 4 are unchanged. The statement of Lemma 5 remains true, the proof being somewhat easier, since for $r=M$ we have $w_M = w_{M-1} / (1 + 2w_{M-1} - w_M - w_{M+1}) = w_{M-1} / (1 + 2w_{M-1} - 2w_M)$, and there is no exceptional first step in the inductive process such as we encountered when $N(=2M)$ was even.

The following theorem solves the double barrier game when $\mu = 0$.

THEOREM. *In the double barrier game with $\mu = 0$ the value is $\frac{1}{2}$, independent of the point at which the game starts.*

Proof. We do not need the theory of recursive games to see a strategy for the evader which will hold the gunner's expectation down to $\frac{1}{2}$. He may when at a point A_r at time t simply choose two of his alternatives, each with equal probability $\frac{1}{2}$. If the evader so wishes, it is obviously possible to choose a strategy of this kind which is stationary.

The situation for the gunner is far more complex. Unless the game starts at a barrier, the gunner must be satisfied with a strategy, which, although stationary, is only ε -optimal. Given $\varepsilon > 0$ we look to the recursive game; by Lemma 5 choose k so that $k \geq M$ and, for $r = 1, \dots, M$, $\frac{1}{2} - w_r^k < \varepsilon$. In the matrix game $\Gamma_r(W_k)$ let X_r be the optimal strategy for player one which is used in the proof of Lemma 3, namely

$$(0, w_{r-1}^k - w_r^k, w_{r-1}^k - w_{r+1}^k, 1)/(1 + 2w_{r-1}^k - w_r^k - w_{r+1}^k),$$

and define the stationary strategy $*X$ for P_1 by taking $*X_r^t = X_r$ for all t . Since $V(W_k) \geq W_k$, condition (i) in Lemma 1 regarding $*X$ and W_k is satisfied. By Lemma 4 $2w_{r-1}^k - w_r^k - w_{r+1}^k > 0$ so that there exists $\alpha < 1$ such that, for all $r = 1, \dots, M$, $1/(1 + 2w_{r-1}^k - w_r^k - w_{r+1}^k) \leq \alpha$. Thus for all strategies Y for P_2 and for all t each matrix $Q'(*X, Y)$ clearly has each of its row sums at most α . It is an easy exercise to see that this implies that condition (ii) of Lemma 1 is also satisfied. Now define a stationary strategy $X(\varepsilon)$ for the gunner as follows. Suppose the evader is at A_r (at time t); then we have the following:

- (a) For $r = 1, \dots, M$ the gunner uses the strategy X_r .
- (b) For $r = M + 1, \dots, N - 1$ the gunner uses the strategy stemming from X_{N-r} .
- (c) For $r = 0$ and N , when the evader is at a barrier the gunner fires at the two possible positions for the evader at time $t + 1$ each with probability $\frac{1}{2}$.

By Lemma 1, and the remarks following it regarding history remembering strategies, it is clear that when the gunner uses $X(\varepsilon)$, then irrespective of what the evader does, his expectation is at least $\frac{1}{2} - \varepsilon$. This concludes the proof of the theorem. \square

The following two notes will be of use when we next show how we may deduce a solution of our game with $|\mu| < 1$ from that with $\mu = 0$.

Note 1. If we wish we may interpret the gunner's expectation (with $\mu = 0$) as the probability of his hitting the evader; thus the use of $X(\varepsilon)$ ensures that the gunner hits the evader with probability at least $\frac{1}{2} - \varepsilon$.

Note 2. In the proof of the theorem we see that, when the gunner uses $X(\varepsilon)$, the probability that he has not fired by time $t = n$ is at most α^n which has limit zero as $n \rightarrow \infty$.

We now prove that for $|\mu| < 1$ the value of our game is $(1 + \mu)/2$ no matter where the game starts. It is clear the evader has no difficulty in holding down the gunner's expectation to $(1 + \mu)/2$. Suppose the gunner uses the strategy $X(\varepsilon)$. Let the evader adopt a strategy Y and let the game start at the point A_s , which we may suppose is not A_0 or A_N . Let $H_n(Y, s)$, $M_n(Y, s)$, and $B_n(Y, s)$, respectively, be the probabilities that up to time $t = n$ the gunner has hit the evader, fired and missed, or not fired. If $E_n(Y, s)$ denotes the accumulated expectation up to $t = n$, then $E_n = H_n + \mu M_n$. Since $H_n + M_n + B_n = 1$ we have $M_n \leq 1 - H_n$ so that, when $\mu \leq 0$, $E_n \geq H_n + \mu(1 - H_n) = (1 - \mu)H_n + \mu$. By Note 1 the gunner hits the evader with probability at least $\frac{1}{2} - \varepsilon$, so we may choose n_1 so that, for all $n \geq n_1$, $H_n \geq \frac{1}{2} - 2\varepsilon$, whence $E_n \geq (1/2 - 2\varepsilon)(1 - \mu) + \mu = (1 + \mu)/2 - 2\varepsilon(1 - \mu)$, and it follows that $X(\varepsilon)$ is $2\varepsilon(1 - \mu)$ -optimal when $\mu \leq 0$. On the other hand, assume that $\mu \geq 0$. We have $E_n = H_n + (1 - H_n - B_n)\mu = (1 - \mu)H_n + (1 - B_n)\mu$, and since, by Note 2, $B_n \rightarrow 0$, we may

choose n_2 so that, for all $n \geq n_2$, $1 - B_n \geq 1 - \varepsilon$. Let $n_3 = \max \{n_1, n_2\}$ then, for all $n \geq n_3$, $E_n \geq (\frac{1}{2} - 2\varepsilon)(1 - \mu) + (1 - \varepsilon)\mu = (1 + \mu)/2 - \varepsilon(2 - \mu)$ and it follows that $X(\varepsilon)$ is $\varepsilon(2 - \mu)$ -optimal. This concludes the extension of our solution to the range $|\mu| < 1$.

4. General remarks. When the gunner has j bullets there are two obvious generalizations of the game; the gunner may try to maximize the number of hits (payoff one for each hit, μ for each miss) or he may maximize his chances of a single hit. The value of the former game is clearly $(1 + \mu)j/2$; that of the latter is $1 - [(1 + \mu)/2]^j$.

An interesting generalization of the game is to play it on a finite tree. An intuitive appreciation of the gunner's technique in the double barrier game leads us to conjecture that with one bullet the value would again be $(1 + \mu)/2$, but this does not seem easy to prove.

Another game of significance is that with a single barrier; in this game there is a barrier only at the point A_0 , and all of the play takes place only on the points A_0, A_1, \dots . This game has an infinite number of game components, and in general it is an open question whether or not such games always have a solution (see [7]). However, it is of interest to observe that our Lemma 1 essentially remains valid. We have been able to solve the single barrier game with one bullet, but the solution with j bullets eludes us.

A referee has drawn our attention to the (discrete) Rabbit and Hunter Game [1], which has a number of features in common with ours. The hunter (gunner) has an unlimited supply of bullets and the payoff is given by the time to a hit. In another version the payoff is the probability of a hit when the hunter has only a limited time T to fire, but still has an unlimited supply of bullets. In both versions, the hunter would clearly choose to fire at each stage and, in this respect, the games differ fundamentally from ours.

REFERENCES

- [1] P. BERNHARD, A. L. COLOMB, AND G. P. PAPAVASSILOPOULOS, *Rabbit and Hunter Game: two discrete stochastic formulations*, Comput. Math. Appl., 13 (1987), pp. 205-225.
- [2] D. BLACKWELL AND T. S. FERGUSON, *The big match*, Ann. Math. Statist., 39 (1968), pp. 159-163.
- [3] H. EVERETT, *Recursive games*, Ann. of Math. Stud., 39 (1957), pp. 47-78.
- [4] J. A. FILAR, *Single loop stochastic game which one player can terminate*, Opsearch, 18 (1981), pp. 185-203.
- [5] E. KOHLBERG, *Repeated games with absorbing states*, Ann. Statist., 2 (1974), pp. 724-738.
- [6] K. T. LEE, *A firing game with time lag*, J. Optim. Theory Appl., 41 (1983), pp. 547-558.
- [7] M. ORKIN, *Recursive matrix games*, J. Appl. Probab., 9 (1972), pp. 813-820.
- [8] L. C. THOMAS, *Games, Theory and Applications*, Ellis Horwood, Chichester, England, 1984.
- [9] F. THUIJSMAN AND O. J. VRIEZE, *The bad match; a total reward stochastic game*, OR Spektrum, 9 (1987), pp. 93-99.

CONTROLLABILITY IS HARDER TO DECIDE THAN ACCESSIBILITY*

EDUARDO D. SONTAG†

Abstract. This article compares the difficulties of deciding controllability and accessibility. These are standard properties of control systems, but complete algebraic characterizations of controllability have proved elusive. The article shows in particular that, for subsystems of bilinear systems, accessibility can be decided in polynomial time, but controllability is NP-hard.

Key words. accessibility, controllability, nonlinear control, complexity

AMS(MOS) subject classifications. 93B05, 93C60, 68C25

1. Introduction. One of the most important and basic outstanding problems in control theory is that of finding necessary and sufficient conditions for deciding when a continuous-time analytic nonlinear system is (locally or globally) controllable. The goal is to provide some sort of generalization of the classical Kalman controllability rank condition. An early success of this line of research was achieved with the characterization of the *accessibility property*: there is a Lie-algebraic rank condition for deciding if it is possible to reach an open set from a given initial state. When this accessibility rank condition does not hold, all trajectories must remain in a lower-dimensional submanifold of the state space. See for instance [HK], [Su1], or [I] for a discussion of this and related results. It is known that local controllability can also be *in principle* checked in terms of linear relations between Lie brackets of the vector fields defining the system [Su1], and recent research has succeeded in isolating a number of necessary as well as a number of sufficient explicit conditions for controllability. The literature regarding this question is very large; see, for instance, [Su2] and the references therein. No complete characterization is yet available, however.

The purpose of this note is to point out that, whatever necessary and sufficient conditions are eventually found, these are likely to be rather hard to check. One way to quantify this difficulty is in terms of complexity of computation. There has been previous work dealing with difficulty of computation in the context of control and system theory. For instance, [So1] showed the undecidability of the realization problem, and more recently [PT] (and references therein) dealt with the study of complexity of decentralized control problems, while [So2] characterized the complexity of decision problems for an algebra used to study piecewise linear control systems. For more in the spirit of this paper, see [BW].

We shall show that the existence of easily verifiable conditions for controllability—local or global, and even several “small-time” variants—would imply solutions to problems known to be hard. The relative difficulty of controllability vis-a-vis the already understood accessibility problem is clarified in the case of the class of systems that can appear as subsystems of bilinear ones. This is a large class of nonlinear systems, including, for instance, all minimal realizations of finite Volterra series, and of course all linear systems. In the context of this class, we can make the precise statement that the accessibility question can be decided in polynomial time, while controllability is

* Received by the editors August 3, 1987; accepted for publication (in revised form) November 20, 1987. This research was supported in part by U.S. Air Force grant 0247.

† Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903. Electronic Mail address: sontag@math.rutgers.edu.

(at least) NP-hard. Recall that NP-hard problems are widely believed to be intractable, and one of the main open problems in theoretical computer science is that of establishing rigorously this intractability, the famous “ $P \neq NP$ ” question [GJ], [PS]. It could be argued that, by proving that controllability is NP-hard, we are not in fact establishing precisely that this is harder than accessibility, only that it is true provided the above open question in computer science is resolved. This is, however, the standard way in which we “prove” that a problem is hard in combinatorics, operations research, theoretical computer science, or in a control-theoretic framework [PT]. In any case, we conjecture that, even for the class of bilinear subsystems, it must be possible to establish exponential-time lower bounds, as has been done in the area of decision methods for logical theories and certain problems in language theory (see, e.g., [AHU, Chap. 11]). We have not yet been able to prove this stronger fact, however.

2. A few preliminaries. The systems we shall deal with have equations

$$\dot{x}(t) = f(x(t), u(t)),$$

where the state $x(t)$ is in a differentiable manifold M for each t , and the control values $u(t) = (u_1(t), \dots, u_m(t))$ belong to a Euclidean space \mathbb{R}^m at each time t . We assume that the dynamics f are real analytic. Generalizations to more arbitrary control value sets and to nonanalytic systems could be made, but since our purpose is mainly to provide negative results, we shall make these results stronger by restricting to even simpler kinds of systems below.

Given any fixed state $x_0 \in \mathbb{R}^n$, we can pose several types of problems relative to x_0 : reachability from x_0 , controllability to x_0 , controllability in any fixed time T . We may also consider the property of complete controllability, being able to find controls that transfer any desired state to any other state. We use the notation

$$A^T(x)$$

for the set of states that can be reached from x in time exactly T ; when T is negative, we mean states from which x can be reached in time $-T$. We may take any reasonable family of controls: all measurable locally essentially bounded controls or piecewise continuous controls; the results will be the same. The union of all the sets $A^T(x)$ over all nonnegative T is denoted

$$A^+(x);$$

this is the set of states reachable from x . Similarly,

$$A^-(x)$$

is the union over $T \leq 0$, the set of states controllable to x . With this notation, for instance, controllability from x_0 means that $A^+(x_0) = M$, controllability to x_0 means that $A^-(x_0) = M$, and local reachability in small time means for each $T > 0$, x_0 is in the interior of the union of the sets $A^\varepsilon(x_0)$, $0 \leq \varepsilon \leq T$.

Two issues which must be clarified are the meanings of the words “given” (a system, and possibly also an initial state x_0) and “decide” (if the system is controllable from x_0 , reachable, etc.). In its weakest sense, *given* could be taken to mean “given a recursive description” of the system, that is, we should provide a *computable* real function f , as well as a *computable* vector x_0 if a fixed initial state is of interest. (See [A] for a discussion of computable analysis, as well as [K] for an alternative viewpoint.) *Decide* should mean *provide a computer algorithm* which, when presented as an input with the description of f (and x_0), will answer “yes” or “no” after a finite number of steps. At this level, controllability is undecidable for trivial reasons, even for linear

systems. For example, the one-dimensional system

$$\dot{x} = bu$$

is controllable if and only if b is nonzero. But it is impossible to decide if a “given” real number is zero or not (see [A, Thm. 6.1]). We obviously want to avoid such logical traps, which have to do with the fact that a recursive description of the dynamics is not necessarily in what we would intuitively call “explicit form.” For linear systems, the simplest way to get around this difficulty is to restrict ourselves to systems with rational coefficients, explicitly given in some notation, for instance, in binary. More generally, we could look, for instance, at a class like that of systems with polynomial or rational functions f , again requiring rational coefficients.

In order to avoid such trivial counterexamples, and to give a stronger negative result, we shall restrict ourselves to bilinear subsystems. These are systems with a finite-dimensional Lie algebra, specified as follows. Given are integers N , m , and l , and $m+2$ matrices

$$A, G_1, \dots, G_m, B$$

over the rational numbers. Each of A, G_1, \dots, G_m is square of size $N \times N$, and B is of size $N \times m$. Also given is a set of l polynomials with rational coefficients

$$\phi_i(x_1, \dots, x_N), \quad i = 1, \dots, l$$

with $\phi_i(0) = 0$ and such that the Jacobian of $(\phi_1, \dots, \phi_l)'$ (prime indicates transpose) has constant rank, say equal to $N - n$. Further, we assume that the n -dimensional manifold M , where all the ϕ_i simultaneously vanish, is invariant for the differential equation

$$(2.1) \quad \dot{x} = \left(A + \sum_{i=1}^m u_i G_i \right) x + Bu,$$

no matter what the control $u(\cdot)$ is. The latter can be expressed algebraically by the requirement that the Lie derivatives

$$(2.2) \quad L_X \phi_i$$

vanish identically on M , for each vector field X of the type $(A + \sum \alpha_i G_i)x + B\alpha$, $\alpha \in \mathbb{R}^m$. Then to the data

$$(2.3) \quad (A, G_1, \dots, G_m, B, \phi_1, \dots, \phi_l)$$

we associate the system Σ whose state space is

$$M = \{x \mid \forall i, \phi_i(x) = 0\}$$

and whose dynamics are given by the restriction of (2.1). We shall call a system of this type a *bilinear subsystem*.

The above definition is meant to capture the idea of a system whose dynamics can be embedded algebraically into a bilinear system. This is a rich enough class of systems for the purposes of this note, and in fact includes many subclasses of interest. For instance, *bilinear* systems result when we take all the $\phi_i \equiv 0$ (so $n = N$, $M = \mathbb{R}^N$), and in particular *linear* systems result when also all the G_i are zero. Further, minimal realizations of finite Volterra series are always of this type [Cr].

In order to express difficulty of computation, we associate to each Σ as in (2.3) a *size*. This is the total number of bits needed in order to store the data (2.3). We assume a fixed data structure for the matrices, say that they are listed by row, and that

each entry is listed as a quotient of integers by giving sign, and the numerator and denominator in binary. Similarly, each of the polynomials ϕ_i may be given by specifying (again in binary) all coefficients in a fixed order. We denote by

size Σ

the resulting integer. When we say that a certain property can be *decided in polynomial time* for such systems, we mean that there is a (fixed) polynomial P and an algorithm which, when given the data (2.3), will answer correctly in time at most

$$P(\text{size } \Sigma)$$

whether this property holds or not. The precise definition of “algorithm” is not very critical in this context; it may be, for instance, a multitape Turing machine as in [AHU], or one of several types of abstract computer models. For this and other related notions, we refer the reader to the standard literature in complexity theory, which we shall not repeat here.

Remark 2.1. A somewhat subtle point: note that when presented with a bilinear subsystem we assume that the Jacobians have constant rank and that the derivatives (2.2) vanish on M , and we shall only be interested in answering questions related to controllability. Checking the consistency of the data, for instance, via the Tarski–Seidenberg theory, could require a large computational effort, and we do not wish to make the problem even harder due to such reasons; we want to show that controllability is hard to check even if the data is reliable.

3. Accessibility. As an illustration of the terminology, we now restate in complexity terms the simplicity of the controllability problem for linear systems. Consider the following property:

The linear system (A, B) is controllable.

The classical condition is that the rank of the $n \times nm$ Kalman block matrix

$$(3.1) \quad (B, AB, A^2B, \dots, A^{n-1}B)$$

must equal the dimension n of the state space. Without loss of generality, we may assume that A and B are integer matrices; if they are not, we can multiply by a common denominator, which increases the total size of the data at most polynomially and does not affect controllability. Whether the rank of the Kalman matrix is n can be checked by Gaussian elimination, which (see, e.g., [PS, Proof of Thm. 8.2]) requires a number of algebraic operations and is polynomial in n , m , and the size of the integers appearing in the composite matrix (3.1). The size of these integers is, in turn, polynomial in the size of the original data; more generally, the size in binary of each entry of a product matrix

$$A = A_1 \cdots A_k$$

is bounded by a polynomial in k and in the size of the integer matrices A_i .

The analogue of the above for nonlinear systems will be obtained, as may be expected, for the accessibility problem. It turns out indeed that accessibility can be also decided in polynomial time for the class of bilinear subsystems, as we shall prove next.

In general, a system Σ is said to be *accessible from* the state x_0 if and only if the reachable set from x_0 has full dimension, that is, if

$$(3.2) \quad \text{int } A^+(x_0) \neq \emptyset.$$

For bilinear subsystems (2.3), we shall take $x_0 := 0$ and say only that Σ is *accessible*. Note that the state space is M , so in (3.2) we mean, of course, the interior with respect to M . When Σ is linear in particular, accessibility is equivalent to controllability, but these concepts are in general different.

Assume now a given bilinear subsystem Σ . Consider the $m + 1$ affine vector fields

$$X_0(x) := Ax$$

and

$$X_i(x) := G_i x + b_i$$

for each $i = 1, \dots, m$, where b_i denotes the i th column of B . The set \mathcal{A} of all affine vector fields on \mathbb{R}^N is a Lie algebra of dimension

$$k := N^2 + N,$$

with multiplication

$$[Ax + b, Cx + d] := (CA - AC)x + (Cb - Ad).$$

Let \mathcal{L}_i , $i \geq 1$, be the sequence of linear subspaces of \mathcal{A} defined as follows:

$$\mathcal{L}_1 := \text{span} \{X_0, X_1, \dots, X_m\}$$

and inductively,

$$\mathcal{L}_{i+1} := \mathcal{L}_i + \text{span} \{[X_i, X] | i = 0, \dots, m, X \in \mathcal{L}_i\}.$$

Let \mathcal{L} be the union of all the \mathcal{L}_i . It follows from the definition that if $\mathcal{L}_i = \mathcal{L}_{i+1}$ for some integer i , then also $\mathcal{L}_i = \mathcal{L}$. By dimensionality we then conclude that

$$\mathcal{L}_k := \mathcal{L}.$$

For any subspace $L \subseteq \mathcal{A}$, denote

$$L(0) := \{b | Ax + b \in L \text{ for some } A\}.$$

This is the tangent space at the state $x_0 = 0$, corresponding to the distribution L . With this notation, we can state the (by now) classical characterization of accessibility (see, e.g., [I, Thm. 6.15]).

PROPOSITION 3.1. *The system Σ is accessible if and only if $\mathcal{L}_k(0)$ has dimension n .*

Note that the rank at the origin of the Jacobian matrix of (ϕ_1, \dots, ϕ_l) is $N - n$; this Jacobian can be computed in polynomial time, and its rank can be obtained again by Gaussian elimination. Thus n can be computed in this form, and it is only necessary to find the dimension of $\mathcal{L}_k(0)$. We now show how to compute a basis of \mathcal{L}_k in polynomial time.

First of all, the problem is not changed by multiplying all the matrices in the description of Σ by the product of all the denominators of all the entries. This increases the size of Σ at most polynomially, so we assume from now on that A, G_1, \dots, G_m, B are matrices of integers.

We shall represent elements $X = Ax + b$ of \mathcal{A} as vectors of size k , listing first the entries of A in some fixed order and then those of b . For any such element, we let $\mu(X)$ denote the maximum of the absolute values of its entries. Also, we take $\bar{\mu}$ to be the largest of the values of the $\mu(X_i)$, $i = 0, \dots, m$, for the generators of \mathcal{L}_1 . Directly from the definition of matrix product, we obtain the formula

$$\mu([X, Y]) \leq 2N\mu(X)\mu(Y)$$

for any $X, Y \in \mathcal{A}$.

Next we show how to build in polynomial time, for each $i = 1, \dots, k$, a basis

$$\{Y_1, \dots, Y_{n_i}\}$$

(note that $n_i \leq k$) of \mathcal{L}_i such that

$$\mu(Y_j) \leq (2N\bar{\mu})^i$$

for any $j = 1, \dots, n_i$. The case $i = 1$ is clear by definition: start with the X_i and use Gaussian elimination to take a subset which forms a basis. By induction, it is necessary to consider now all Lie products

$$(3.3) \quad [X_j, Y_l]$$

for $j = 0, \dots, m$ and $l = 1, \dots, n_i$. There are at most k^2 of these. Each of them has entries of largest magnitude

$$\mu([X_j, Y_l]) \leq 2N\mu(X_j)(2N\bar{\mu})^i \leq (2N\bar{\mu})^{i+1}.$$

Let B be the matrix that lists all these generators (3.3). Each entry of B , expressed in binary, has length at most equal to

$$(i+1) \log_2 (2N\bar{\mu})$$

(plus a bit for the sign). Since we may assume that $i+1 \leq k = N^2 + N$, this quantity is bounded by a polynomial

$$a + bN^3 + cN^2 \log \bar{\mu},$$

which is in turn bounded by an expression of order $O(M^3)$, where M is the total size of the original data (A, G_1, \dots, G_m, B) . Thus Gaussian elimination can be performed in polynomial time to select a subset of (3.3) which forms a basis. Note that it is essential that this elimination be performed at each step to the algorithm: otherwise we end up with an exponential number, $(m+1)^k$, of generators for the space \mathcal{L} . After at most k steps we have a basis for $\mathcal{L}_k = \mathcal{L}$, and hence also by evaluation at zero and one last elimination step, we can determine the dimension of $\mathcal{L}(0)$. This establishes the following fact.

THEOREM 1. *For bilinear subsystems, the accessibility property can be decided in polynomial time.*

Remark 3.2. We set the convention that $x_0 = 0$ only for notational simplicity. It is equivalent to studying accessibility (and later controllability) from an *equilibrium* state. The results in the case of more general x_0 are entirely analogous, with accessibility as well as *strong* (fixed-time) accessibility both verifiable in polynomial time.

Remark 3.3. Another property that is sometimes of interest is that of *span reachability*, meaning that the linear span of the states reachable from the origin should be the entire space. This can also be checked in polynomial time, by an argument like the one above. In fact, the accessibility property is basically the same as a span-reachability property at the Lie algebra level.

4. A controllability remark. In this section we shall restrict our attention to systems of the following very special type:

$$(4.1) \quad \begin{aligned} \dot{y} &= w^2 f(z), \\ \dot{z} &= Az + bu_1, \\ \dot{w} &= u_2. \end{aligned}$$

These are systems with state space

$$M = \mathbb{R} \times \mathbb{R}^k \times \mathbb{R}$$

of dimension $k+2$, and states partitioned as $x = (y, z, w)$,¹ with the block of variables z evolving as a linear system of dimension k . The control value set is \mathbb{R}^2 , and we assume that the single-input system (A, b) is controllable. We shall show that for systems (4.1), several variants of the notion of controllability are all equivalent to the indefiniteness of the mapping $f: \mathbb{R}^k \rightarrow \mathbb{R}$.

DEFINITION 4.1. The mapping f is *definite* if and only if $f(z) \geq 0$ for all z or $f(z) \leq 0$ for all z . Otherwise, it is *indefinite*.

This allows us to reduce the problem of deciding if a polynomial is definite to a controllability question, and we conclude that controllability is at least as hard to decide as definiteness. Further, systems of the special form (4.1) with f polynomial can be rewritten as bilinear subsystems, and this rewriting can be done in polynomial time, relative to any class of polynomials f of fixed degree. Together with the NP-hardness of the problem of deciding definiteness (next section), the desired negative result will follow.

It is clear that indefiniteness is necessary for any type of controllability: if f were definite, say $f(z) \geq 0$ for all z , then

$$y \geq y_0$$

whenever $x = (y, z, w) \in A^+(x_0)$, for each $x_0 = (y_0, z_0, w_0) \in M$. Thus it is impossible in that case for any x_0 to be in the interior of $A^+(x_0)$ or of $A^-(x_0)$.

Assume then that f is indefinite. We show now that, for each $\delta > 0$, and for each two states x_0 and \bar{x} , there is a control $u(\cdot)$ which steers x_0 into \bar{x} in time 4δ . We build u in five steps.

Step 1. Apply the control $u_1 \equiv 0$, $u_2(t) \equiv -w_0/\delta$ on the interval $[0, \delta]$. There results a state of the form $x_1 = (y_1, z_1, 0)$.

Now consider the number

$$y^* := \bar{y} - \left(\frac{\bar{w}}{\delta}\right)^2 \int_0^\delta s^2 f(e^{(s-\delta)A} \bar{z}) ds.$$

Either (a) $y_1 = y^*$ or (b) $y_1 \neq y^*$. Assume first that (b) holds. We know by indefiniteness of f that there exists some vector z_2 such that

$$\text{sign } f(z_2) = \text{sign } (y^* - y_1).$$

Pick any such z_2 . By continuity of the exponential, there is some $0 < \varepsilon < \delta$ such that

$$(4.2) \quad \text{sign } f(e^{sA} z_2) = \text{sign } (y^* - y_1) \quad \text{for all } s \in [0, \varepsilon].$$

Finally, pick any solution α of the equation

$$\alpha^2 \left[\int_0^{\varepsilon/2} s^2 f(e^{sA} z_2) ds + \int_{\varepsilon/2}^\varepsilon (\varepsilon - s)^2 f(e^{sA} z_2) ds \right] = y^* - y_1.$$

There is some such α because of (4.2). In case (a), make an arbitrary choice, say $z_2 = z_1$, let $0 < \varepsilon < \delta$ be also arbitrary, and let $\alpha := 0$.

Step 2. Apply a control with $u_2 \equiv 0$, on the interval $[0, \delta - \varepsilon]$, that takes x_1 into the state $x_2 = (y_1, z_2, 0)$. A suitable $u_1(\cdot)$ exists because of the assumed controllability of the pair (A, b) .

¹ For simplicity, we write (y, z, w) instead of, more accurately, $(y, z', w)'$.

Step 3. Apply a control on the interval $[0, \varepsilon]$ as follows. The u_1 component is identically zero, and

$$u_2(t) := \begin{cases} \alpha & \text{if } t < \varepsilon/2, \\ -\alpha & \text{if } t \geq \varepsilon/2. \end{cases}$$

There results the state $x_3 = (y^*, e^{\varepsilon A} z_2, 0)$; the total time elapsed is 2δ .

Step 4. On the interval $[0, \delta]$, let $u_2 \equiv 0$ and let $u_1(\cdot)$ be a control steering $e^{\varepsilon A} z_2$ into $e^{-\delta A} \bar{z}$. Again such a control exists by the controllability of the linear system (A, b) . The resulting state is $x_4 = (y^*, e^{-\delta A} \bar{z}, 0)$.

Step 5. Finally, in one last interval of length δ , use $u_1 \equiv 0$ and $u_2 \equiv \bar{w}/\delta$. The result is the desired state \bar{x} .

We can summarize the discussion above.

PROPOSITION 4.2. *Let Σ be a system as in (4.1), and pick any fixed $x_0 \in M$. The following properties are then equivalent:*

- (i) $A(x) = M$ for each $x \in M$ (complete controllability);
- (ii) $A^T(x) = M$ for each T and each $x \in M$;
- (iii) $x_0 \in \text{int } A^+(x_0)$;
- (iv) $x_0 \in \text{int } A^-(x_0)$;
- (v) f is indefinite.

It follows that other intermediate properties are also equivalent to the above, for instance *local small-time reachability* from x_0 :

$$x_0 \in \text{int } \bigcup_{t=0}^{\varepsilon} A^t(x_0) \quad \text{for each } \varepsilon > 0,$$

as well as local controllability to x_0 in small time. Thus checking either of these properties is equivalent to checking the indefiniteness of f .

For accessibility, it is sufficient only that f not be identically zero, which illustrates in this particular case the gap between the two concepts. Note that, even for the very simple case in which f is a homogeneous quadratic form, checking definiteness already requires some computational effort.

5. Deciding definiteness. The previous section shows how, at least for some systems, controllability is no easier to check than definiteness of a map. This latter property can be checked for polynomials via decision methods for real closed fields (see, e.g., [Co]) in doubly-exponential time; however, it is not clear if there are faster algorithms. We remark here that the problem is NP-hard, and we do this by polynomial time reduction of the classical NP-complete problem, 3-SAT, to the definiteness question. Thus deciding definiteness is at least as hard as any problem in NP. The remark is not at all surprising, but it is the best lower bound that we have been able to obtain until now.

Recall the definition of the 3-SAT problem [GJ, p. 48]. A *clause* $c(x, y, z)$ in the three (distinct) variables x, y, z is an expression of the type

$$(5.1) \quad \phi_1(x) \vee \phi_2(y) \vee \phi_3(z),$$

where each “literal” ϕ_i is of the form

$$\phi_i(a) = a$$

or

$$\phi_i(a) = 1 - a$$

and the binary variables x, y, z can take values in $\{0, 1\}$. We interpret the values 1 and 0 as “true” and “false,” respectively. For any assignment (x^*, y^*, z^*) of values $\{0, 1\}$

to x, y, z , we say that $c(x^*, y^*, z^*)$ is *true* if at least one of $\phi_1(x^*)$, $\phi_2(y^*)$ or $\phi_3(z^*)$ is 1, and *false* otherwise. Equivalently, $c(x^*, y^*, z^*)$ is true if and only if the real polynomial

$$\tilde{c}(x, y, z) := \phi_1(x)^2 + \phi_2(y)^2 + \phi_3(z)^2$$

does not vanish at (x^*, y^*, z^*) . A set

$$\mathcal{S} = \{c_i(t_{i,1}, t_{i,2}, t_{i,3}), i = 1, \dots, L\}$$

of L clauses in the variables (t_1, \dots, t_n) , with each $t_{i,j} \in \{t_1, \dots, t_n\}$, is *satisfiable* if and only if there is some binary assignment $t^* = (t_1^*, \dots, t_n^*)$ to the variables (t_1, \dots, t_n) such that the clauses $c_i(t^*)$ become all simultaneously true. The 3-SAT problem is that of finding an algorithm for checking satisfiability. It was the first problem to be shown to be NP-complete, in the sense that if there were such an algorithm, which would run in time polynomial in L , then every other problem in the wide class NP, which includes many, if not most, combinatorial problems of interest, would also be decidable in polynomial time. It is a long-standing conjecture (“P \neq NP”) in theoretical computer science, widely believed to be true, that indeed none of these problems can be solved in polynomial time.

It is easy to reduce 3-SAT to the problem of deciding if a polynomial has any real zeros, and hence to establish that the latter problem is NP-hard. We first show how to do that, and then modify the construction to deal with the definiteness problem instead. Let \mathcal{S} be as above. Consider first the polynomial

$$(5.2) \quad \theta(t) = \theta(t_1, \dots, t_n) = \sum_{i=1}^n t_i^2(1-t_i)^2.$$

Denote by B_n the set of binary n -vectors, $\{t = (t_1, \dots, t_n) \mid \text{for all } i, t_i \in \{0, 1\}\}$, and note that B_n is the set of zeros of θ . Now let $u = (u_1, \dots, u_n)$ be L new variables, and introduce

$$(5.3) \quad \psi(t, u) := \sum_{i=1}^L (u_i \tilde{c}_i(t) - 1)^2 + \theta(t).$$

If $\psi(t^*, u^*) = 0$ then the last term in the sum vanishes, so t^* is binary, while the vanishing of the other terms implies that $\tilde{c}_i(t^*) \neq 0$ for all i . Conversely, if $t^* \in B_n$ is such that all $\tilde{c}_i(t^*) \neq 0$, there is some vector u^* such that $\psi(t^*, u^*) = 0$. We conclude that \mathcal{S} is satisfiable if and only if ψ has a real zero.

We next modify ψ in order to reduce to definiteness instead. Now let

$$(5.4) \quad \psi(t, u) := \sum_{i=1}^L (2u_i \tilde{c}_i(t) - u_i^2 - 1)^2 + \theta(t).$$

It is again true that \mathcal{S} is satisfiable if ψ has a real zero. This is because an expression of the type $2u\tilde{c} - u^2 - 1$ is strictly negative unless $\tilde{c} \neq 0$. Conversely, assume that \mathcal{S} is satisfiable, and let $t^* \in B_n$ be such that all $\tilde{c}_i(t^*) \neq 0$. Consider each $2u\tilde{c}_i(t^*) - u^2 - 1 = 0$ as an equation on $u \in \mathbb{R}$. Writing this as

$$\tilde{c}_i(t^*) = \frac{u^2 + 1}{2u}$$

and using the fact that, since $t^* \in B_n$ and $\tilde{c}_i(t^*) \neq 0$, $\tilde{c}_i(t^*) \in \{1, 2, 3\}$, and that

$$\alpha : (0, \infty) \rightarrow [1, \infty) : u \mapsto \frac{u^2 + 1}{2u}$$

is onto, we conclude that (5.4) has a zero.

We show below that when \mathcal{S} is not satisfiable, not only is ψ always positive, but in fact it is bounded away from zero. Note that, in general, it is false that a positive polynomial must be bounded below by a positive constant, as evidenced by the example $x^2 + (1 - xy)^2$, so some care is required. Moreover, we need an explicit value for this lower bound, which in our case will turn out to be $1/4n^2$.

Assume then that \mathcal{S} is not satisfiable. Pick any element $t^* \in B_n$. There is some clause $c_i(t^*)$ which is false. Relabeling variables if necessary, we may assume that c_i involves the polynomials $\phi_1(t_1)$, $\phi_2(t_2)$, $\phi_3(t_3)$. Since these all vanish at t^* , we conclude that

$$\phi_j(t)^2 = (t_j^* - t)^2$$

for $j = 1, 2, 3$. In particular, using Euclidean norm, it holds that

$$(5.5) \quad \|t^* - t\|^2 \geq \tilde{c}_i(t)$$

for each $t \in \mathbb{R}^n$. Now consider any fixed element $(t, u) \in \mathbb{R}^{n+L}$. Either (1) $\|t^* - t\|^2 \leq \frac{1}{2}$ for some $t^* \in B_n$, or (2) the distance from t to B_n is at least $1/\sqrt{2}$. If there is any t^* as in (1), pick a clause c_i as above. Then, for this fixed i , using (5.5),

$$\left| \left(\frac{2u_i}{u_i^2 + 1} \right) \tilde{c}_i(t) \right| \leq \tilde{c}_i(t) \leq \frac{1}{2},$$

so

$$|2u_i \tilde{c}_i(t) - u_i^2 - 1| = (u_i^2 + 1) \left| \left(\frac{2u_i}{u_i^2 + 1} \right) \tilde{c}_i(t) - 1 \right| \geq \frac{u_i^2 + 1}{2} \geq \frac{1}{2}.$$

Hence, $\psi(t, u) \geq \frac{1}{4} \geq 1/4n^2$. Suppose that (2) holds instead. Then necessarily

$$t_j^2(1 - t_j^2) \geq \frac{1}{4n^2}$$

for at least one j , and therefore again $\psi(t, u) \geq 1/4n^2$. Indeed, if this were not the case, then it would hold for each $j = 1, \dots, n$ that either

$$(5.6) \quad |t_j| < \frac{1}{\sqrt{2n}} \quad \text{or} \quad |t_j - 1| < \frac{1}{\sqrt{2n}}.$$

Choose $t^* \in B_n$ with $t_j^* = 0$ if the first case in (5.6) holds, and 1 otherwise. Then this particular t^* would satisfy that $\|t^* - t\|^2 \leq \frac{1}{2}$, contradicting case (2).

The conclusion from the above discussion is that \mathcal{S} is satisfiable if and only if there is some pair (t, u) such that

$$f(t, u) := 4n^2\psi(t, u) - 1 < 0.$$

On the other hand, since f certainly admits positive values, for instance

$$f((2, 2, \dots, 2), u) \geq 16n^3 - 1$$

for any u , it follows that f is indefinite if and only if it takes any negative values, that is, if and only if \mathcal{S} is satisfiable. The construction, including expanding f into a standard polynomial form, can be done in polynomial time, and we summarize.

LEMMA 5.1. *For each set \mathcal{S} of L clauses in n variables, there is a polynomial f of degree 6 in $n + L$ variables, whose coefficients are integers of magnitude less than or equal to cLn^2 , such that \mathcal{S} is satisfiable if and only if f is indefinite. The polynomial f can be obtained in polynomial time from \mathcal{S} , and c is a constant independent of \mathcal{S} .*

6. A reduction to bilinear subsystems. The idea behind the construction to be given in this section is basically a classical one in the field of bilinear systems, and can be traced at least as far back as the paper [B]. The point of giving the details is to keep track of the computational effort required and of the size of the numbers appearing.

LEMMA 6.1. *Let Σ be a system of the form*

$$(6.1) \quad \dot{\xi}_0 = \Psi(\xi), \quad \dot{\xi} = P\xi + Qu,$$

where Ψ is a polynomial of degree d in the r variables $\xi = (\xi_1, \xi_2, \dots, \xi_r)$ with integer coefficients and with $\Psi(0) = 0$, P is an integer matrix of dimensions $r \times r$, and Q is an integer matrix of dimensions $r \times m$. Assume that the coefficients of Ψ, P, Q are all of magnitude bounded by ρ . Then there is a bilinear system

$$\Sigma_b = (A, G_1, \dots, G_m, B, \phi_1, \dots, \phi_l)$$

with $N = \binom{r+d}{r}$, $l = N - r - 1$, each coefficient of the matrices A, G_1, \dots, G_m, B of magnitude less than or equal to $d\rho^2$, and the polynomials ϕ_i of degree less than or equal to d and with each coefficient equal to 0, 1, or -1 , such that each of the following properties holds for the system Σ if and only if it holds for the system Σ_b :

- (a) $0 \in \text{int } A^+(0)$,
- (b) $0 \in \text{int } A^-(0)$,
- (c) $A^+(x) = M$ for all $x \in M$.

Further, the system Σ_b can be constructed in polynomial time from the data Ψ, P, Q .

Proof. Note that N is the number of possible monomials of degree less than or equal to d in the variables $\xi = \xi_1, \dots, \xi_r$. We shall use multi-indices $\alpha = (\alpha_1, \dots, \alpha_r)$ with weight $|\alpha| := \sum \alpha_i \leq d$, and

$$\xi^\alpha := \xi_1^{\alpha_1} \cdots \xi_r^{\alpha_r}$$

to denote these monomials. The coordinates of vectors in \mathbb{R}^N will be denoted as η_α , for such indices α , ordered lexicographically. In particular, we let $e_i := (0, 0, \dots, 0, 1, 0, \dots, 0)$ (1 in i th position), and write η_{e_i} just as η_i . For each of the $l = N - r - 1$ indices α with weight $|\alpha| \geq 2$, we introduce the polynomial in N variables

$$\phi_\alpha(\eta) := \eta_\alpha - \eta_1^{\alpha_1} \cdots \eta_r^{\alpha_r}.$$

Note that the Jacobian matrix of the ϕ_α 's has constant rank $N - r - 1$. The idea of the construction is to introduce a variable for each monomial in Ψ (the η_α 's) in such a manner that the equation for $\dot{\xi}_0$ becomes linear:

$$\dot{\eta}_0 = \sum_\beta \psi_\beta \eta_\beta,$$

where $\Psi(\xi) = \sum_\beta \psi_\beta \xi^\beta$, and to introduce a differential equation for each of the monomials ξ^α , thought of now as new variables η_α . Note that if $\xi(\cdot)$ is a solution of

$$\dot{\xi} = P\xi + Qu$$

and if $|\alpha| > 0$, then

$$(6.2) \quad \frac{d\xi^\alpha}{dt} = \sum_{|\beta|=|\alpha|} a_\beta^\alpha \xi^\beta + \sum_{|\beta|=|\alpha|-1} (g_\beta^{\alpha,1} u_1 + \cdots + g_\beta^{\alpha,m} u_m) \xi^\beta,$$

where the coefficients a_β^α, \dots are obtained as follows. Let $P = (p_{ij}), Q = (q_{ij})$; with this notation,

$$(6.3) \quad a_\beta^\alpha = \sum_{i,j} \alpha_i p_{ij},$$

the sum over all those indices $1 \leq i, j \leq r$ for which $\beta + e_i = \alpha + e_j$, and for each j ,

$$(6.4) \quad g_{\beta}^{\alpha, j} = \sum_i \alpha_i q_{ij},$$

the sum over all those indices $1 \leq i \leq r$ for which $\beta + e_i = \alpha$. We also denote, for the case $\alpha = (0, \dots, 0)$, each β , and each $j = 1, \dots, m$, $\alpha_{\beta}^{\alpha} := \psi_{\beta}$ and $g_{\beta}^{\alpha, j} := 0$. Finally, let A and G_j , $j = 1, \dots, m$ be the matrices (a_{β}^{α}) and $(g_{\beta}^{\alpha, j})$ respectively, and let B be the block matrix

$$\begin{pmatrix} 0 \\ Q \\ 0 \end{pmatrix},$$

where the first block is a row of size $1 \times m$ and the last one is of size $N - r - 1$ by m .

Since there are at most r^2 terms in each of (6.3) and (6.4), the claimed estimates for the magnitudes of the entries of these matrices do indeed hold. Further, the constructions can be clearly carried out in polynomial time.

Given any vector $x = (x_0, \dots, x_r) \in \mathbb{R}^{r+1}$, let $h(x) \in \mathbb{R}^N$ be defined as follows: $h_{(0, \dots, 0)}(x) := x_0$, and

$$h_{\alpha}(x) := x_1^{\alpha_1} \cdots x_r^{\alpha_r}$$

for $|\alpha| > 0$. Note that h is a diffeomorphism $\mathbb{R}^r \simeq M$, and $h(0) = 0$. Now assume that $x(\cdot) = (x_0(\cdot), x_1(\cdot), \dots, x_r(\cdot))$ solves (6.1) with respect to a given control $u(\cdot)$. It follows by construction that $\eta(t) := h(\xi(t))$ satisfies

$$\dot{\eta}(t) = (A + \sum u_i(t) G_i) \eta(t) + Bu(t).$$

Therefore $x(0)$ can be steered into $x(T)$ if and only if $h(x(0))$ can be steered into $h(x(T))$, and this establishes properties (a)–(c). \square

7. Controllability is NP-hard. We are only left to put together all the pieces from the previous sections. Assume that \mathcal{S} is any set of L clauses. Note that it can involve at most $n \leq 3L$ variables. By Lemma 5.1 we may build in polynomial time an integer polynomial f of degree 6 in $n + L$ variables, with coefficients of magnitude bounded by cL^3 , such that f is indefinite if and only if \mathcal{S} is satisfiable. For this f , we now consider the system (4.1), where (A, b) is a cascade of integrators

$$\dot{z}_1 = z_2, \quad \dot{z}_2 = z_3, \quad \dots, \quad \dot{z}_k = u_1$$

and $k = n + L$. By Proposition 4.2, the system is controllable, in either of the senses there, if and only if \mathcal{S} is satisfiable. We now apply Lemma 6.1, with $r = k + 1$, $m = 2$,

$$\Psi(\xi_1, \dots, \xi_{k+1}) := \xi_{k+1}^2 f(\xi_1, \dots, \xi_r),$$

and P, Q found from A, b plus the last equation $\dot{w} = u_2$. Note that $d = 8$, and that we may take $\rho = cL^3$. We thus obtain, in polynomial time, the bilinear subsystem Σ_b in Lemma 6.1. Listing all entries of A, \dots, ϕ_i results in a size of order at most $N^2 \log_2(r^2 \rho)$, which is bounded by a polynomial in L . A polynomial time decision method for controllability of Σ_b would thus imply one for 3-SAT. Thus our problem is at least as hard as that one.

THEOREM 2. *Each of the following decision problems is NP-hard, for bilinear subsystems:*

- (a) $0 \in \text{int } A^+(0)$ (local reachability at 0).
- (b) $0 \in \text{int } A^-(0)$ (local controllability at 0).
- (c) $A^+(x) = M$ for all $x \in M$ (complete controllability). \square

As remarked earlier, many other questions, such as local small-time reachability, are shown to be NP-hard by the same argument. As directions for further research, we suggest looking for a similar result using only single-input systems—the proof above shows that it is hard to decide controllability if at least two controls are allowed—and also for the case of controls constrained to compact sets. Alternatively, it would be interesting to establish better lower bounds for the problem studied here.

REFERENCES

- [A] O. ABERTH, *Computable Analysis*, McGraw-Hill, New York, 1980.
- [AHU] A. V. AHO, J. HOPCROFT, AND J. D. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.
- [BW] W. M. BOOTHBY AND E. N. WILSON, *Determination of the transitivity of bilinear systems*, SIAM J. Control Optim., 17 (1979), pp. 212–221.
- [B] R. BROCKETT, *On the algebraic structure of bilinear systems*, in Theory and Applications of Variable Structure Systems, R. Mohler and A. Ruberti, eds., Academic Press, New York, 1972, pp. 153–168.
- [Co] G. COLLINS, *Quantifier elimination for real-closed fields by cylindrical algebraic decomposition*, Lecture Notes in Computer Science 35, Springer-Verlag, Berlin, New York, 1975, pp. 134–183.
- [Cr] P. E. CROUCH, *Dynamical realizations of finite Volterra series*, Ph.D. thesis, Harvard Univ., Cambridge, MA, 1977.
- [GJ] M. R. GAREY AND D. JOHNSON, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, 1979.
- [HK] R. HERMANN AND A. J. KRENER, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, 22 (1977), pp. 728–740.
- [I] A. ISIDORI, *Nonlinear Control Systems: An Introduction*, Springer-Verlag, Berlin, New York, 1985.
- [K] K. KO, *Applying techniques of discrete complexity theory to numerical computation*, in Studies in Complexity Theory, R. V. Book, ed., Pitman, London, 1986.
- [PS] C. H. PAPADIMITRIOU AND K. STEIGLITZ, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, NJ, 1982.
- [PT] C. H. PAPADIMITRIOU AND J. TSITSIKLIS, *Intractable problems in control theory*, SIAM J. Control Optim., 24 (1986), pp. 639–654.
- [So1] E. D. SONTAG, *On certain questions of rationality and decidability*, J. Comput. System. Sci., 11 (1975), pp. 375–381.
- [So2] ———, *Real addition and the polynomial hierarchy*, Inform. Process. Lett., 20 (1985), pp. 115–120.
- [Su1] H. J. SUSSMANN, *Lie brackets, real analyticity, and geometric control*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhauser, Boston, 1983.
- [Su2] ———, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.

GLOBAL (f, g) -INVARIANCE OF NONLINEAR SYSTEMS*

W. DAYAWANSA†, D. CHENG‡, W. M. BOOTHBY‡, AND T. J. TARN§

Abstract. Consider a nonlinear system $\dot{x} = f(x) + \sum_{j=1}^m g_j(x)u_j$ and a nonsingular involutive distribution Δ defined on a state space M which is a C^∞ manifold. The authors investigate the existence of nonlinear feedback (α, β) such that $[f + g\alpha, \Delta] \subset \Delta$ and $[g\beta, \Delta] \subset \Delta$, i.e., (f, g) -invariance. A geometric interpretation of weak (f, g) -invariance, i.e., $[f, \Delta], [G, \Delta] \subset \Delta + G$ is given in terms of a transverse $GL(q, R)$ -structure to the codimension q foliation \mathcal{F} defined by Δ . This interpretation allows a simplified proof of the Quaker Lemma—proved by other means in [15], [16], and [21]—and helps to somewhat clarify the global picture. As an application the case $M = S^3$ is considered in detail. The introduction contains a discussion of the relation of this work to other studies of global (f, g) -invariance and to relevant foliation theory literature. Resolution of this problem is important for extension of results using the concept of (A, B) -invariant subspaces to nonlinear systems.

Key words. (f, g) -invariance, nonlinear disturbance decoupling, normal bundle of a foliation, basic connections, transverse e -structures, local absolute parallelism along leaves

AMS(MOS) subject classifications. 93C10, 93C15, 93C25

1. Introduction. Consider the linear system

$$(1.1) \quad \dot{x} = Ax + Bu, \quad x \in R^n, \quad u \in R^m.$$

A subspace V of R^n is said to be (A, B) -invariant if $AV \subset V + \mathcal{B}$, where \mathcal{B} denotes the column span of B . It is well known (see, e.g., [20], [30]) that V is (A, B) -invariant if and only if there exists a matrix F such that $(A + BF)V \subset V$. This has been called the Quaker Lemma [4], a terminology which we will adopt here. In linear system theory, (A, B) -invariance plays a crucial role in many important problems such as disturbance localization and decoupling (see [1], [20], [30]). In the basic papers [12], [15], [21], [22], and especially the seminal paper of Isidori et al. [14], the above concept was extended to the nonlinear setting as follows.

Consider the affine nonlinear system,

$$(1.2) \quad \dot{x} = f(x) + \sum_{i=1}^m g_i(x)u_i,$$

where $x \in M$, a smooth manifold, the u_i 's are input functions from a suitable function space (generally measurable R -valued functions defined on closed intervals of the form $[0, T]$), and f, g_1, \dots, g_m are smooth vector fields on M with g_1, \dots, g_m linearly independent everywhere. The following definitions were taken from [3], [14], and [15] although the terminology is changed slightly.

DEFINITION 1.1. Let U be an open subset of M . A pair of smooth functions (α, β) , where $\alpha: U \rightarrow R^m$ and $\beta: U \rightarrow GL(m, R)$ are called *feedback functions on U* . When $U = M$ we refer to feedback functions as *global feedback*. If $\alpha = (\alpha_1, \dots, \alpha_m)$ and $\beta = (\beta_{ij}), 1 \leq i, j \leq m$, are feedback functions on U , then the new system obtained by feedback,

$$(1.3) \quad \dot{x} = \left(f + \sum_{i=1}^m g_i \alpha_i \right) + \sum_{i=1}^m \left(\sum_{j=1}^m \beta_{ij} g_j \right) v_i = \hat{f}(x) + \sum_{i=1}^m v_i \hat{g}_i(x),$$

* Received by the editors December 1, 1986; accepted for publication (in revised form) November 27, 1987. This research was supported in part by National Science Foundation grants ECS-8515899, ECS-8518832, DMC-8615963, and INT-8519654.

† Department of Mathematics, Texas Tech University, Lubbock, Texas 79409.

‡ Department of Mathematics, Washington University, St. Louis, Missouri 63130.

§ Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130.

is said to be *feedback equivalent* or *feedback related* to (1.2).

Remark 1.2. We use the following shorter notation: g denotes (g_1, \dots, g_m) , $g\alpha$ denotes $\sum_{i=1}^m g_i \alpha_i$, $g\beta$ denotes $\{\sum_{j=1}^m \beta_{ij} g_j\}$, $1 \leq i \leq m$, and $u = (u_1, \dots, u_m)$, etc. Then the feedback law can be obtained from $u = \alpha + v\beta$ with $v\beta = (v_1, \dots, v_m)\beta$.

DEFINITION 1.3. Let G denote the distribution spanned by g_1, \dots, g_m . A distribution Δ on M is said to be *weakly (f, g) -invariant at $x \in M$* , if there exists a neighborhood U of x such that on U ,

$$(1.4) \quad [f, \Delta] \subset \Delta + G,$$

$$(1.5) \quad [G, \Delta] \subset \Delta + G.$$

Δ is *(globally) weakly (f, g) -invariant* if Δ is weakly (f, g) -invariant at x for all $x \in M$.

DEFINITION 1.4. A distribution Δ on M is *(f, g) -invariant at $x \in M$* if there exists a neighborhood U of x , and smooth feedback functions (α, β) on U such that

$$(1.6) \quad [f + g\alpha, \Delta] \subset \Delta,$$

$$(1.7) \quad [g\beta, \Delta] \subset \Delta.$$

Similarly, Δ is *(globally) (f, g) -invariant* if there exists a global feedback (α, β) such that (1.6) and (1.7) are satisfied.

In the particular case, when Δ is a flat distribution on R^n (i.e., Δ is spanned by a finite set of constant vectors of R^n) and the system under consideration is linear (1.1), then all three notions of (f, g) -invariance defined above reduce to the notion of (A, B) -invariance. In particular no distinction between weak (f, g) -invariance and global (f, g) -invariance is necessary. However the nonlinear situation is much more difficult. First, the distribution Δ may give rise to a far more complicated foliation than parallel planes. Second, we must distinguish between local and global properties for even though (f, g) -invariance implies weak (f, g) -invariance, the reverse implication need not be true. The following does hold, however.

LEMMA 1.5 (Quaker Lemma [3], [14]–[16], [21]). *If Δ is involutive and nonsingular in a neighborhood of x and if $G + \Delta/\Delta$ has constant dimension around x , then Δ is weakly (f, g) -invariant at x if and only if Δ is (f, g) -invariant at x .*

This lemma is a powerful tool that can be used to study the *local* behavior of nonlinear systems. It has been used in various nonlinear problems such as disturbance localization, decoupling, etc. (see [15]). However, for studying *global* properties of nonlinear systems we need global (f, g) -invariance. It is well known that (f, g) -invariance at every $x \in M$ does not necessarily imply global (f, g) -invariance (see § 3 of Byrnes [4], which is closely related to this paper, or Example 5.6 below). In the literature, the proofs of Lemma 1.5 make use of direct arguments using existence theorems for the differential equations involved (see, for instance, [16, pp. 126–134], which gives a very careful proof based on that found in [15]).

Our main purpose here is to apply fairly standard foliation theory and vector bundle machinery, especially the related notions of parallelism along leaves and transverse structures to foliations. These notions were introduced in the very beginning of foliation theory by Reeb [33] and Ehresmann [34] and have been studied by Reinhart [24], Pasternak [35], and especially (for our purposes) by Conlon [7] and his students. For a complete discussion of transverse structures see Godbillon [31, Chap. III].

This approach stems from our observation that the conditions $[f, \Delta] \subset \Delta + G$ and $[G, \Delta] \subset \Delta + G$, $G = \text{span}\{g_1, \dots, g_m\}$ for weak (f, g) -invariance can be interpreted geometrically as a statement that G defines a transverse $GL(q, R)$ -structure to the

foliation \mathcal{F} defined by Δ , or, equivalently, that a subspace of the normal bundle to the foliation is parallel along the leaves. This is Lemma 3.1. We have proved this lemma using a Bott basic connection associated to \mathcal{F} , although it is quite unnecessary to do so. However, this connection is a useful tool in studying transverse structures (see Conlon [7]). This connection was also used in the papers of Krener [18] and of Byrnes and Krener [3], but in a quite different and more fundamental way, namely, to study characteristic classes of Δ following the idea of the Bott Vanishing Theorem in order to find characteristic classes whose vanishing was necessary for Δ to be (f, g) -invariant.

To the best of our knowledge it has not been previously recognized that weak (f, g) -invariance is equivalent to a transverse structure, although it is implicit, of course, in the geometry. We use this fact extensively in what follows and also in the paper [32] written at the same time. Among the consequences is what seems to us a simpler and more intuitive proof of the Quaker Lemma (1.5). It avoids any overt use of partial differential equations and seems better adapted to globalization questions.

As we have mentioned, the problem of uncovering global obstructions to global (f, g) -invariance has been studied in Byrnes and Krener [3] and Byrnes [4], [5] (and in Krener [18] discussed above). In [4] and [5] Byrnes discusses a vector bundle denoted V_β over the leaf space M/\mathcal{F} (or M/Δ) and goes on to define a cohomology class η_α and to indicate the definition of a second class η_β . These ingredients are combined into Theorem (3.1) of [4], which asserts a necessary and sufficient condition for (f, g) -invariance of Δ . No proof of the theorem is given and no formal definition of V_β is provided. The latter is discussed in terms of an example in which several very special features are present: foremost, the fact that the distribution spanned by $\{g_1, \dots, g_m\}$ and the distribution Δ intersect only along the zero section. In both [4] and [5] the reader is referred for all proofs to a paper of Byrnes and Krener entitled "Global Methods for Nonlinear Geometric Control Theory I." To the best of our knowledge this paper has not appeared. When it does, it will no doubt clarify the overlap between some sections of [4] and [5] and this paper.

We conclude this section by listing notation and assumptions used in the sequel.

(1.8) Notation

- M is a smooth manifold of dimension n ;
- Δ is an involutive $(n - q)$ -plane distribution;
- E is the subbundle of $T(M)$ corresponding to Δ ;
- \mathcal{F} is the foliation of codimension q defined by Δ (or E);
- Q is the "normal bundle" of E , formally Q is the vector bundle $T(M)/E$;
- G is the distribution spanned by g_1, \dots, g_m , and also the vector subbundle of $T(M)$ it determines.

(1.9) Assumptions

- (i) We assume throughout that the vector fields f and g_1, \dots, g_m are globally defined and smooth, i.e., C^2 at least, and that g_1, \dots, g_m are independent at every $x \in M$. Thus G is a *nonsingular* distribution, i.e., of constant dimension. Of course, unlike Δ it is not assumed to be involutive.
- (ii) It is further assumed throughout that Δ is globally weakly (f, g) -invariant, as in Definition 1.3.

2. The normal bundle to a foliation. We will assume that the concept of a (smooth) vector bundle on a manifold M is known to the reader (see [2], [13], [24], and [25]). The simplest examples are the tangent bundle $T(M)$ and any nonsingular distribution E of dimension p , which of course, at each point x determines a p -dimensional subspace

E_x of $T_x(M)$ and thus a p -dimensional subbundle. The subbundle E determines a quotient bundle $Q = T(M)/E$, having $T_x(M)/E_x$ as the fiber over x , a vector space of dimension $q = n - p$ attached to x . There is no canonical way to identify Q with a subbundle of $T(M)$, but if a Riemannian metric is chosen on M , then at each point Q_x has a natural identification with E_x^\perp . If E determines an *involutive* distribution Δ on M , then a very special situation arises. Through each $x \in M$ there is a p -dimensional maximal connected integral manifold L_x , the *leaf* of the foliation Δ , which has E_x as its tangent bundle at x . If Q is identified with E^\perp via a Riemannian metric on M , then it is indeed a normal bundle to each immersed submanifold L_x in the usual sense.

A C^∞ section σ of a vector bundle E over M is a C^∞ mapping $\sigma: M \rightarrow E$ such that $\pi \cdot \sigma = id_M$. In particular, a section of $T(M)$ is a vector field on M . The *zero section* maps $x \in M$ to O_x , the zero vector of E_x . In general, sections which are nowhere zero may not exist. Let $m = \dim E_x$. If m sections exist which are everywhere independent, then $E = M \times \mathbb{R}^m$ and is said to be *trivial*. If M is contractible to a point (e.g., $M = \mathbb{R}^n$), then every vector bundle over M is trivial. (Remark on notation: Whenever X is a section of E we will usually simply write $X \in E$.)

It was observed as soon as foliations were defined that there is a natural way, associated with the foliation, to “parallel translate” vectors of the bundle Q along a leaf. Let (U, ϕ) be a “Frobenius” chart on M relative to the foliation, i.e., we assume that $\phi(U) = C^n$ the unit cube in \mathbb{R}^n and that the intersection of any leaf L with U maps to points in C^n whose first q coordinates are constant. If $\sigma: C^n \rightarrow C^q$ maps (x^1, \dots, x^n) onto (x_1, \dots, x_q) , then $\tilde{\phi} = \sigma \cdot \phi: U \rightarrow C^q \subset \mathbb{R}^q$ is a submersion (i.e., has rank q “at every point of U ”). If $x \in U$ and L'_x is the connected component through x of $L_x \cap U$, then $\tilde{\phi}(L'_x) = y$, a single point of C^q . Moreover, $E_x = \ker(\tilde{\phi}_*)_x$ and $(\tilde{\phi}_*)_x: Q_x \rightarrow T_y(\mathbb{R}^q)$ is a vector space isomorphism. If $\tilde{X}_x = X_x + E_x \in Q_x$, then for any $y \in L'_x$ we define \tilde{X}_y uniquely by $(\tilde{\phi}_*)_y(\tilde{X}_y) = (\tilde{\phi}_*)_x \tilde{X}_x$. M is covered by an atlas $\{U_\alpha, \phi_\alpha\}$ of such neighborhoods. If $U_\alpha \cap U_\beta \neq \emptyset$, then $\phi_\beta \cdot \phi_\alpha^{-1}$ is constant on a component of $L \cap U_\alpha \cap U_\beta$ for any leaf L . Hence this parallelism is the same for (U_α, ϕ_α) as for (U_β, ϕ_β) and can be extended from neighborhood to neighborhood along any continuous curve $s(t)$, $a \leq t \leq b$, lying on a leaf L . If x^0, x^1 are on the same leaf L , then parallel translation from x^0 to x^1 depends only on the homotopy class of the curve $s(t)$ from x^0 to x^1 . Thus if L is simply connected, the parallelism along leaves is absolute and the bundle Q restricted to L is trivial, having a basis of (parallel) frames.

The *basic connection* ∇ defined by Bott is closely related: first identify Q with a subbundle Q' of $T(M)$, say by use of a Riemannian metric. A vector field Y on M then decomposes uniquely into $Y = Y_E + Y_{Q'}$, vector fields (sections) with values in E, Q' , respectively. A connection ∇ on Q' is called *basic* if for X a vector field with values in E (tangent to leaves) and Y a vector field with values in Q' , we have $\nabla_X Y = [X, Y]_{Q'}$. Such connections are easily shown to exist and to define a connection in Q which is independent of the choice of Q' (see [2]). Thus if \tilde{Y} is a section in Q then define $\nabla_X \tilde{Y} = \nabla_X Y'$, where Y' is the unique vector field in $Q' \subset T(M)$ which projects onto \tilde{Y} at each point x by the natural projection $T_x(M) \rightarrow T_x(M)/E_x$, i.e., $\tilde{Y}_x = Y'_x + E_x$. We see that \tilde{Y} is parallel if and only if $\nabla_X Y' = [X, Y']_{Q'} = 0$, i.e., $[X, Y']$ is tangent to the leaf, in other words $[\Delta, Y'] \subset \Delta$, where Δ is the involutive distribution coinciding with E . This parallelism, and that described above without the use of connections, are equivalent.

Comparison at this point of this latter condition on Y' , with the condition that a vector field Y' leaves an involutive distribution Δ invariant, as found in [3], [6], and [16], makes it clear that these are essentially the same concept. Precisely, Δ is Y' -invariant if and only if $Y = Y' + E$ is parallel along leaves. This was noted but not

really used in [5]. In [6] this idea was used to discuss the relation of accessibility and strong (exact time) accessibility. The advantage of using Q and parallelism along leaves is that it allows immediate use of standard tools from foliation theory.

Remark 2.1. An easy consequence of these definitions is the following. Let Y be a vector field on M , i.e., $Y \in \Gamma(T(M))$, and let \bar{Y} be its projection into Q . The \bar{Y} is parallel (with respect to any basic connection) if and only if $[Y, \Delta] \subset \Delta$.

3. Global weak (f, g) -invariance. We have assumed that on M there is a nonsingular involutive distribution $\Delta (= E)$ which is globally weakly (f, g) -invariant (with respect to the system (1.2)); \mathcal{F} denotes the corresponding foliation. We will interpret this in the light of the previous section. Given any vector field on M , we denote its projection into Q by placing a bar over it. Thus $\bar{f}, \bar{g}_1, \dots, \bar{g}_m$ are the projections of f, g_1, \dots, g_m into Q . Conversely, a section \bar{Y} in Q determines at each $x \in M$ only an equivalence class $E_x + Y_x = \Delta_x + Y_x$ in $T_x(M)$. Let $\bar{G} = (\Delta + G)/\Delta$ denote the subbundle of Q spanned by $\bar{g}_1, \dots, \bar{g}_m$ and let ∇ be a basic connection, which defines parallel transport in Q along curves lying on leaves of \mathcal{F} .

LEMMA 3.1. *Given any $x^0 \in M$ and curve $s(t), 0 \leq t \leq 1, s(0) = x^0$ on L_{x^0} , the leaf of \mathcal{F} through x^0 , then the unique parallel vectors $\bar{Z}_i(s(t)), i = 1, \dots, m$, in Q such that $\bar{Z}_i(x^0) = \bar{g}_i(x^0)$ lie in $\bar{G}(s(t))$ for $0 \leq t \leq 1$, and $\dim(\bar{G})$ is constant on L_{x^0} . Hence $\dim(G \cap \Delta)$ is invariant on leaves of \mathcal{F} .*

Remark 3.2. We should interpret this lemma as saying that \bar{G} is invariant under parallel translation along the leaves, i.e., if (U_α, ϕ_α) is a Frobenius chart and $\tilde{\phi}_\alpha$ is ϕ_α followed by projection to R^q , then at every point x of a component of any leaf L intersected with U_α , $(G + \Delta)_x$ projects to the same subspace of $T_{\tilde{\phi}_\alpha(x)}(R^q)$. Equivalently, \bar{G} defines a parallel subbundle of the restriction of Q to L . If the foliation \mathcal{F} is regular, i.e., $N = M/\mathcal{F}$ is a C^∞ manifold, and if $\pi: M \rightarrow N$ denotes the natural projection, then $\pi(x) = \pi(y)$ implies $\pi_*(G + \Delta)_x = \pi_*(G + \Delta)_y$.

Proof of Lemma 3.1. It is sufficient to prove that the lemma holds in a neighborhood of any point p on the leaf L through x^0 . Let U be a simply connected neighborhood of p in L and suppose $\text{rank}(\bar{g}_{1p}, \dots, \bar{g}_{mp}) = d$. Let $(\bar{Z}_1, \dots, \bar{Z}_d)$ be sections of $Q|_U$ such that they are parallel and span $\{\bar{Z}_{1p}, \dots, \bar{Z}_{dp}\} = \text{span}(\bar{g}_{1p}, \dots, \bar{g}_{mp})$. (Such sections exist since U is simply connected and thus parallel transport relative to ∇ is absolute on U .) Extend $\bar{Z}_1, \dots, \bar{Z}_d$ to a parallel frame field $(\bar{Z}_1, \dots, \bar{Z}_q)$ of $Q|_U$. Let $\bar{g}_i = \sum_{j=1}^q \gamma_{ij} \bar{Z}_j$ for $1 \leq i \leq m$, where $\gamma_{ij} \in C^\infty(U)$. Note that $\gamma_{ij}(p) = 0$ for $j > d$.

Without loss of generality, we may assume that the image of the curve s lies entirely in U . Let $t \in [0, 1]$ and take covariant derivatives of \bar{g}_i along s . Since $D\bar{Z}/dc = 0$, we have

$$(3.1) \quad \frac{d\bar{g}_i}{dt} = \sum_{j=1}^q \frac{D}{dt}(\gamma_{ij}(s(t)))\bar{Z}_j = \sum_{j=1}^q \frac{d}{dt}(\gamma_{ij}(s(t)))\bar{Z}_j.$$

On the other hand, by definition of D/dt ,

$$\frac{D\bar{g}_i}{dt} = \nabla_{\dot{s}(t)} \bar{g}_{is(t)}.$$

Let X be a vector field in U such that $X_{s(t)} = \dot{s}(t)$ and such that $X \in \Delta$, i.e., X is tangent to the leaves of \mathcal{F} , then $\nabla_{\dot{s}(t)}(\bar{g}_{is(t)}) = \nabla_{X_{s(t)}}(\bar{g}_{is(t)})$. By the definition of ∇ ,

$$\nabla_X \bar{g}_i = [\overline{X, g_i}]_{Q'} = [X, g_i] \bmod \Delta.$$

Since $[X, g_i] \in \Delta + G$ we see that

$$\nabla_X \bar{g}_i = \sum_{k=1}^m \theta_{ik} \bar{g}_k \quad \text{for some smooth } \theta_{ik} \text{'s on } U.$$

Therefore,

$$\nabla_X \bar{g}_i = \sum_{k=1}^m \theta_{ik} \sum_{j=1}^m \gamma_{kj} \bar{Z}_j.$$

Hence, restricting to $s(t)$ for $0 \leq t \leq 1$, we have

$$(3.2) \quad \frac{d}{dt} \gamma_{ij} = \sum_{k=1}^m \theta_{ik} \gamma_{kj}, \quad 1 \leq i, \quad j \leq q.$$

Let $r_j = [\gamma_{ij}, \dots, \gamma_{mj}]^T$ and $\theta = [\theta_{ij}]_{m \times m}$. Then

$$\frac{d}{dt} r_j = \theta r_j, \quad j = 1, \dots, q,$$

a system of linear equations. Now $r_j(s(0)) = 0$ for $j > d$ and thus $r_j(s(t)) = 0$ for $j > d$, showing that $\bar{Z}_j(s(t)), j = 1, \dots, d$ span \bar{G} for $0 \leq t \leq 1$. Also $(r_1(s(0)), \dots, r_d(s(0)))$ are linearly independent. Thus $(r_1(s(t)), \dots, r_d(s(t)))$ are linearly independent for all t in the domain of definition. Therefore $\text{rank}(\bar{g}_1, \dots, \bar{g}_m) = d$ at all points on s . Constancy of the dimension of \bar{G} on L_{x^0} follows at once since any two points of L_{x^0} can be joined by a smooth curve.

Remark 3.3. We can interpret the above lemma as saying that weak (f, g) -invariance is equivalent to the statement " \bar{G} is parallel along leaves." Now suppose that Δ is globally (f, g) -invariant. For the sake of explanation we will assume that $[g_i, \Delta] \subset \Delta$ and $[f, \Delta] \subset \Delta$, i.e., that suitable feedback has been applied. In terms of a basic connection, this says that $\nabla_X \bar{g}_i = 0 = \nabla_X \bar{f}$ for all $X \in \Delta$ (where overbar indicates equivalence classes). In other words \bar{g}_i and \bar{f} are parallel along leaves. This is a stronger statement than saying that \bar{G} is invariant under parallel transport along leaves.

4. Smooth versus continuous feedback. In geometric control theory it is typical to assume that all vector fields, foliations, etc. are smooth. However in certain situations such as disturbance decoupling, we can relax the smoothness conditions. For example, consider the following disturbance decoupling problem:

$$(4.1) \quad \dot{x} = f(x) + \sum_{i=1}^m g_i(x) u_i + p(x) w, \quad y = h(x),$$

where (u_1, \dots, u_m) are the inputs, y is the output and w is a disturbance. It is well known that if (i) there exists a nonsingular distribution Δ such that $p \in \Delta$, (ii) (f, g) are smooth, and (iii) Δ is globally (f, g) -invariant, and (iv) $\Delta \subset \ker dh$, then the disturbance w does not affect the output. Below we relax (ii) (i.e., smoothness of (f, g)).

Let $x \in M$ and let (U, ϕ) be a Frobenius chart around x . Let (x_1, \dots, x_{n-q}) be the coordinates tangential to leaves and (y_1, \dots, y_q) the coordinates transverse to leaves, then $f = \sum_{i=1}^{n-q} \lambda_i(x, y) \partial / \partial x_i + \sum_{j=1}^q \mu_j(x, y) \partial / \partial y_j$. We say that f is smooth along leaves in U if $\lambda_j(\cdot, y)$ and $\mu_j(\cdot, y)$ are smooth (similarly for the g_i 's). It is easy to see that this definition does not depend on the particular choice of (U, ϕ) , and hence we can define the notion of smoothness along leaves in M . For the remainder of the section we assume that f, g are continuous vector fields with uniquely determined flows and are smooth along leaves of \mathcal{F} . By considering the local coordinate expressions, we see that $[X, g_i]$ and $[X, f]$, etc. still make sense for all $X \in \Delta$. Thus, even in this situation we can define the notion of weak (f, g) -invariance and (f, g) -invariance.

LEMMA 4.1. *Suppose that Δ is globally (f, g) -invariant and $p \in \Delta \subset \ker dh$. Then the disturbance w does not affect the output $y = h(x)$.*

Proof. The same proof is valid as in the case when (f, g) are smooth (see [15]).

Because of this we will examine the existence of both smooth (α, β) and continuous (α, β) in our search for feedback functions to render Δ globally (f, g) -invariant. In the continuous case we will assume that (f, g) are continuous and smooth along leaves. But unless otherwise mentioned we are considering the smooth case.

5. Global (f, g) -invariance when $G \cap \Delta = \{0\}$. In addition to our fixed assumption that g_1, \dots, g_m are everywhere independent we now suppose they are independent mod Δ , i.e., that $G \cap \Delta = 0$ at all points.

Our reasons for paying attention to this special case are the following:

(i) In the single input case, if $\dim G$ and $\dim G \cap \Delta$ are constant, then either $G \subset \Delta$ or $G + \Delta / \Delta = G = \text{span}\{g\}$. In the first case Δ is obviously globally (f, g) -invariant. In the second case G satisfies our assumption.

(ii) As we will see in the next section, under a suitable hypothesis the general case can be reduced to this case.

In view of our assumption G is isomorphic to the subbundle \bar{G} of Q . By Lemma 3.1, \bar{G} is invariant under parallel transport along leaves.

Our first result is under the hypothesis that M/\mathcal{F} is a smooth but not necessarily Hausdorff manifold. In particular, when the foliation arises as (i) the connected components of the level sets of a submersion or (ii) cosets of a closed subgroup of a Lie group, our results are applicable. Let $\pi: M \rightarrow M/\mathcal{F}$ be the projection map. We can assign a subspace of $T_{\pi(p)}(M/\mathcal{F})$ to each $p \in M$ by $\hat{G}(\pi(p)) = \{\pi_{*p}X \mid X \in G_p\}$. Invariance of \bar{G} under parallel transport along leaves implies that $\hat{G}(\pi(p))$ does not depend on the choice of p on a leaf of \mathcal{F} . In this case, this is exactly the meaning of Lemma 3.1. Since π is a submersion and since local cross sections to \mathcal{F} always exist, it follows that \hat{G} is a smooth distribution on M/\mathcal{F} . Let \hat{G} denote the corresponding subbundle of $T(M/\mathcal{F})$.

PROPOSITION 5.1. *If \hat{G} is trivial and if there exists a complementary subbundle \hat{H} to \hat{G} in $T(M/\mathcal{F})$, then Δ is globally (f, g) -invariant.*

Proof. Let (Y_1, \dots, Y_m) be a smooth frame field of \hat{G} . Let $p \in M$ be arbitrary. Since $\Delta \cap G = \{0\}$, it follows that $\pi_{*p}: \bar{G}_p \rightarrow T_{\pi(p)}(M/\mathcal{F})$ is one-to-one with image equal to $\hat{G}_{\pi(p)}$. Hence we can define smooth, parallel frame fields $(\bar{g}'_1, \dots, \bar{g}'_m)$ on \bar{G} in Q by setting $\bar{g}'_{ip} = (\pi_{*p})^{-1}(Y_{i\pi(p)})$ and corresponding smooth frame fields (g'_1, \dots, g'_m) in $G \subset T(M)$. This determines a smooth $\beta: M \rightarrow GL(m, R)$ such that $(g'_1, \dots, g'_m) = (g_1, \dots, g_m)\beta$. By Remark 2.1, $[g'_i, \Delta] \subset \Delta$. Noting that $Q = \pi^*T(M/\mathcal{F})$, (π^* denotes pull back by π), we can pull back \bar{H} by π to define a complementary subbundle \bar{H} to \bar{G} in Q . At this point we note that if a Riemannian metric is chosen on M then, as before, we can identify Q with a subbundle Q' of $T(M)$. Q' can be chosen to contain G : let $Q' = G + (G + \Delta)^\perp$. Then $Q' = G + H$ an orthogonal decomposition with $G \approx \bar{H}$ under the projection of $T(M) \rightarrow Q$. Now f can be written as $f = f_\Delta + f_G + f_H$ with $f_\Delta \in \Delta$; $f_G \in G$ and $f_H \in H$. But by construction H is invariant under parallel transport along leaves. Hence $[f_H, \Delta] \subset \Delta + H$. Thus if $[f, \Delta] \subset \Delta + G$ we must have $[f_H, \Delta] \subset \Delta$. Write $f_G = \sum -\theta_i g'_i$ and set $\alpha = \beta[\theta_1, \dots, \theta_m]^T$, then $[f + \alpha g, \Delta] \subset \Delta$. This defines (α, β) as required.

PROPOSITION 5.2. *Suppose that M/\mathcal{F} is a Hausdorff manifold. Then Δ is globally (f, g) -invariant if and only if \hat{G} is trivial. In particular when M is compact global (f, g) -invariance implies that the Euler characteristic of M/\mathcal{F} is zero.*

Proof. Sufficiency. M/\mathcal{F} has a Riemannian metric in this case. Let $\hat{H} = \hat{G}^\perp$ be the desired complementary bundle and use Proposition 5.1.

Necessity. Let (α, β) be feedback rendering Δ globally (f, g) -invariant. Let $(\tilde{g}_1, \dots, \tilde{g}_m) := (g_1, \dots, g_m)\beta$. Then \tilde{g}_i , the projection of \tilde{g}_i into Q , is parallel along leaves. Hence $(\tilde{g}_1, \dots, \tilde{g}_m)$ can be pushed forward by π to define a smooth frame

field of \hat{G} . Hence \hat{G} is trivial. The last remark follows at once since triviality of \hat{G} implies the existence of a nowhere zero vector field on M/\mathcal{F} and thus vanishing Euler characteristic (see [11]) if M is compact.

COROLLARY 5.3. *Suppose that $\dim G = q = \text{codim } \mathcal{F}$ and that M/\mathcal{F} is a smooth (not necessarily Hausdorff) manifold. Then Δ is globally (f, g) -invariant if and only if M/\mathcal{F} is parallelizable.*

Proof. Sufficiency follows from Proposition 5.1. A repetition of the necessity argument of Proposition 5.2 completes the proof.

Example 5.4. Let $M = U(2)/O(2)$. Then $(\det)^2: M \rightarrow S^1$ is a smooth fiber bundle [3]. Let Δ be the foliation with fibers as leaves $f \in \Delta$ and g an arbitrary transverse vector field to Δ . This system satisfies the hypotheses of Corollary 5.3, and hence Δ is globally (f, g) -invariant.

Alternative construction. Let ω be a nowhere zero, closed one form on S^1 . Let $\eta = (\det^2)^*\omega$. Now η is closed, nowhere zero and $\Delta = \ker \eta$. Define β by $\beta = 1/\eta(g)$. Now if $X \in \Delta$, then $0 = d\eta(X, \beta g) = \eta[X, \beta g]$. Therefore $[\beta g, \Delta] \subset \Delta$. Write $f = f_\Delta + f_G$ with $f_\Delta \in \Delta$ and $f_G \in G$. Find α by $f_G = -\alpha g$. Clearly $[f + \alpha g, \Delta] \subset \Delta$. Hence (α, β) satisfy the requirements.

Remark 5.5. This example was presented in [3], but it was concluded that Δ is not globally (f, g) -invariant. Our results are also at variance with the converse part of Proposition 2.3 of [3]. More precisely we believe that what is required for global (f, g) -invariance in this situation is the parallelizability of M/\mathcal{F} and not the triviality of $\pi: M \rightarrow M/\mathcal{F}$, as was stated there.

Example 5.6. Let $M = S^1 \times S^1$, and consider $\mathfrak{so}(n)$ as a subalgebra of $\mathfrak{so}(n+1)$. Let Δ be the distribution generated by $\mathfrak{so}(n)$. Suppose there exists a transverse vector field g to Δ such that $[g, \Delta] \subset \Delta + \text{span}\{g\}$. (Note that according to our theory existence of such g implies and is implied by the existence of a one-dimensional distribution on $S^n \simeq S^1 \times S^1/S^1$. Hence n cannot be even.) Let f be such that Δ is globally weakly (f, g) -invariant. Then in view of Proposition 5.2, Δ is globally (f, g) -invariant. More generally, in the single input case if M/\mathcal{F} is Hausdorff and if g is transverse to Δ , then global weak (f, g) -invariance and global (f, g) -invariance are equivalent.

Example 5.7. Let M and Δ be as in Example 5.6. Suppose that g_1, \dots, g_n are everywhere linearly independent vector fields such that $G \cap \Delta = \{0\}$. (This situation occurs, for example, when g_1, \dots, g_n are elements of $\mathfrak{so}(n+1)$ which span a subspace complementary to $\mathfrak{so}(n)$.) Let f be arbitrary. Then Δ is globally weakly (f, g) -invariant. However, $M/\mathcal{F} \simeq S^n$; hence (by Corollary 5.3) Δ is globally (f, g) -invariant if and only if $n = 1, 3$ or 7 , since these are the only dimensions for which S^n is parallelizable.

PROPOSITION 5.8. *Suppose that $\text{codim } \Delta = 1$ and g is transverse to Δ . Let f be arbitrary. Then Δ is globally (f, g) -invariant if and only if there exists a nowhere zero closed one form ω such that $\Delta = \ker \omega$.*

Proof. Sufficiency. Let $\beta = 1/\omega(g)$. Define α by $f + \alpha g = 0 \pmod{\Delta}$. Clearly $[f + \alpha g, \Delta] \subset \Delta$, $[\beta g, \Delta] \subset \Delta$ by the alternative argument of Example 5.4.

Necessity. Let (α, β) be feedback as required. Define a one form ω by $\omega(\beta g) = 1$ and $\ker \omega = \Delta$. Then for all $X \in \Delta$, $d\omega(X, \beta g) = X(\omega(\beta g)) - \beta g(\omega(X)) - \omega([X, \beta g]) = 0$; hence $d\omega = 0$.

COROLLARY 5.9. *Assume M is compact, $\text{codim } \Delta = 1$, and f, g are vector fields satisfying the assumption of (5.8), i.e., $\text{span}\{g\} \cap \Delta = 0$ and g is nowhere zero. If Δ is globally (f, g) -invariant, then M admits the structure of a fiber bundle over S^1 . In particular M is not simply connected.*

Proof. Tischler [29] proved that any compact manifold admitting a nowhere zero, closed one form, fibers over S^1 .

Example 5.10. Let $M = T^2$ and let Δ be the one-dimensional foliation on T^2 given by the Kronecker flow; thus the leaves (integral curves) are all dense in M . Let g be a transverse vector field to Δ and let f be arbitrary. It is well known that Kronecker flow is determined by a closed one form. Hence by Proposition 5.8, Δ is globally (f, g) -invariant. Note that in this case the foliation is *not* regular since the leaves are dense on T^2 .

PROPOSITION 5.11. *Suppose that M is compact and Δ is of codimension 1. If Δ is globally (f, g) -invariant with feedback (α, β) which is continuous transverse to the leaves, then Δ does not have limit cycles.*

Proof. Suppose that \tilde{g} is continuous, transverse to the leaves, and is such that $\dot{x} = \tilde{g}(x)$ has unique solutions for all initial conditions. Since $[\tilde{g}, \Delta] \subset \Delta$, it follows that \tilde{g} leaves the foliation invariant. Hence Δ does not have limit cycles (see [11], [17], [28] for details).

In the case when M is compact and $\dim G = q = \text{codim } \Delta$, global (f, g) -invariance is equivalent to saying that the normal bundle Q admits a trivialization by parallel sections. Conlon [7] describes this situation by saying that \mathcal{F} admits a transverse e -structure. Among other things he proves that all leaves of \mathcal{F} are diffeomorphic and the foliation does not have limit cycles. Moreover if \mathcal{F} admits a closed leaf, then there is a fiber bundle $p: M \rightarrow N$, where N is parallelizable and \mathcal{F} is the foliation of M by fibers of p . The reader is referred to [7] for many other interesting details.

6. Global (f, g) -invariance when Δ separates controls and $\dim(G + \Delta/\Delta)$ is constant. In view of the assumption that (g_1, \dots, g_m) are linearly independent and $\dim(G + \Delta/\Delta)$ is constant, Krener's definition [19] of " Δ separates controls" reduces to Definition 6.1 below.

DEFINITION 6.1. Δ *separates controls* if there exists feedback $(0, \beta_1)$ and $0 \leq d \leq m$ such that $\text{span}(\tilde{g}_1, \dots, \tilde{g}_d) \cap \Delta = \{0\}$ and $\text{span}(\tilde{g}_{d+1}, \dots, \tilde{g}_m) \subset \Delta$, where $(\tilde{g}_1, \dots, \tilde{g}_m) = (g_1, \dots, g_m)\beta_1$.

Throughout this section, we assume that Δ separates controls. This assumption is equivalent to $\dim(G \cap \Delta)$ is constant (possibly zero) and both $(G \cap \Delta)$ and $G + \Delta/\Delta$ are trivial bundles. In particular this holds when $\dim(G \cap \Delta)$ is constant and M is contractible.

If Δ separates controls, then we may identify $(G + \Delta/\Delta)$ with $\text{span}\{\tilde{g}_1, \dots, \tilde{g}_d\}$. We will denote $(\tilde{g}_1, \dots, \tilde{g}_d)$ by \tilde{g} and $\text{span}\{\tilde{g}_1, \dots, \tilde{g}_d\}$ by \tilde{G} .

LEMMA 6.2. Δ is globally (f, g) -invariant if Δ is globally (f, \tilde{g}) -invariant.

Proof. Let (α_2, β_2) be feedback which renders Δ , globally (f, \tilde{g}) invariant. Then set

$$\beta = \beta_1 \begin{bmatrix} \beta_2 & 0 \\ 0 & I \end{bmatrix} \quad \text{and} \quad \alpha = \beta_1 \begin{bmatrix} 0 \\ \alpha_2 \end{bmatrix}.$$

Clearly (α, β) makes Δ globally (f, g) -invariant.

Remark 6.3. In view of Lemma 6.2, all of our results in §6 yield sufficient conditions for global (f, g) -invariance when G is replaced by \tilde{G} , which satisfies $\tilde{G} \cap \Delta = \{0\}$.

PROPOSITION 6.4 (Global Quaker Lemma). *Let Δ be weakly (f, g) -invariant on M and separate controls. Suppose that either of the following conditions holds:*

(i) *There exists a smooth manifold N , which is transverse to the leaves of \mathcal{F} and intersects each leaf exactly once.*

(ii) *M/\mathcal{F} is a Hausdorff manifold and the subbundle \hat{G} of M/\mathcal{F} obtained by pushing forward $(G + \Delta/\Delta)$ is trivial. Then Δ is globally (f, g) -invariant.*

Proof. It suffices to show that Δ is globally (f, \tilde{g}) -invariant. We will first reduce (i) to (ii). Let $i: N \rightarrow M$ be the inclusion. Since N is transverse to \mathcal{F} , it follows that $T(N) = i^*Q$. Moreover $i^*\tilde{G}$ is a trivial subbundle of $T(N)$. Since $\tilde{G} := \{\pi_{*p} Y \mid Y \in \tilde{G}, p \in M\}$ is isomorphic to $i^*\tilde{G}$ and since M/\mathcal{F} is diffeomorphic to N , (i) implies (ii). Sufficiency of (ii) for global (f, \tilde{g}) -invariance follows at once from Proposition 5.2.

COROLLARY 6.5. Δ is (f, \tilde{g}) -invariant on all cubical Frobenius charts on M .

Proof. The proof follows at once from Proposition 6.4.

Remark. This appears to supply a proof, independent of Theorem 3.1 of [4] for Corollary (3.4) of that theorem—at least if we interpret the chart U , referred to in its hypothesis, to be a “Frobenius” chart. Our Corollary 6.5 and (i) of Proposition 6.4 are also results obtained independently in the Ph.D. dissertation of Cheng [8].

7. Global (f, g) -invariance on S^3 . In this section we give a rather detailed description of global (f, g) -invariance on S^3 . Reasons for our choice of S^3 are twofold. First, many rotational problems have $SO(3)$ as the state space. Since S^3 is the simply connected covering of $SO(3)$, these problems can be lifted to S^3 . Second, topological properties of foliations on S^3 are better understood than on many other manifolds. It is well known that any codimension 1 foliation on S^3 has a Reeb component (see [11], [23], etc.). We show below that a Reeb component presents a formidable obstruction to global (f, g) -invariance.

The inspiration for this section came from Example (3.1) of Byrnes [5] as follows. On $SU(2)$ take $\{X, Y, Z\}$ as the standard basis of $SU(2)$, the Lie algebra of left invariant vector fields on $SU(2)$ (i.e., X, Y , and Z are the infinitesimal rotations around the x, y , and z axes, respectively). Take $\Delta = \text{span}\{X\}$, $f = 0$, $g_1 = Y$, $g_2 = Z$. The key observation made by Byrnes is that Novikov’s Theorem on the existence of a Reeb component of codimension 1 foliation on $SU(2)$ and the Poincaré–Bendixson Theorem together imply that Δ is not globally (f, g) invariant. Here we carry this idea further.

Throughout we assume that $[G, \Delta] \subset G + \Delta$, $\dim G = \text{constant}$ and that $\tilde{G} = G + \Delta/G$ has constant, nonzero rank. Unless otherwise specified \tilde{G} may be a nontrivial bundle. With these assumptions fixed we now consider the possibilities.

(i) $\text{Codim } \Delta = 1$; $\dim G = 1, 2$, or 3 . We claim that Δ is not globally (f, g) -invariant. For other cases, there exists a vector field W such that \bar{W} , the equivalence class of W in Q is nonzero at a point p on the toral leaf of a Reeb component and such that $[W, \Delta] \subset \Delta$. By continuity \bar{W} is nonzero on a neighborhood of p . But $[W, \Delta] \subset \Delta$ implies that \bar{W} is parallel along leaves and thus \bar{W} is nonzero on a saturated neighborhood of p . But this and completeness of W imply that the toral leaf is stable (i.e., all leaves in the vicinity of a toral leaf are toral leaves themselves). But the toral leaf of a Reeb component is known to be an *isolated* toral leaf.

(ii) $\text{Codim } \Delta = 2$; $\dim G = 2$; $G \cap \Delta = \{0\}$. We claim that Δ is not globally (f, g) -invariant. This is a direct consequence of Corollary A of [7, p. 80], which states that on any compact manifold with finite fundamental group there does not exist a foliation of codimension less than or equal to 2 which admits a trivialization of the normal bundle respecting the parallelism along leaves.

(iii) $\text{Codim } \Delta = 2$; $\dim G = 1$ or 2 ; $\dim \tilde{G} = 1$. We claim that Δ is not globally (f, g) -invariant. For other cases, let $\tilde{\mathcal{F}}$ denote the codim 1 foliation generated by $\Delta + G$. Let L be the toral leaf of a Reeb component of $\tilde{\mathcal{F}}$. Since S^3 is simply connected, the line bundle E defined by Δ is orientable, and hence there exists a vector field X such that $\Delta = \text{span}\{X\}$. We show below that this leads to a contradiction.

Let $\mathcal{F}|L$ denote the foliation of Δ on L . $\mathcal{F}|L$ cannot have limit cycles. Otherwise, for the same reason as in (i), the limiting closed leaves are stable, which is a contradic-

tion. Hence there exists a homeomorphism $\phi: L \rightarrow L$ which maps $\mathcal{F}|_L$ onto a foliation on L , induced by parallel lines in R^2 under the identification $L = R^2/Z^2$ (see [11, p. 63]). Without any loss of generality we may assume that ϕ induces the identity homomorphism at the fundamental group level. Let $\hat{\mathcal{F}}$ denote the image foliation. We consider two possible cases as shown in Figs. 1 and 2.

Suppose that a “crosswise band to L ” (call this \hat{C}) is transverse to $\hat{\mathcal{F}}$. Let $C = \phi^{-1} \cdot \hat{C}$. Now C generates trivial holonomy on \hat{F} , and hence by radial projection C can be pushed inward to construct closed curves on planar leaves in the Reeb component. Let L_1 be a planar leaf in the Reeb component. Then the procedure above determines an infinite number of closed curves $\{C_n\}_{n=1}^\infty$ in L_1 , and there are closed curves arbitrarily close to C in the Reeb component. Hence by continuity there exists n such that X is transverse to C_n . But now the Poincaré–Bendixson Theorem states that X has an equilibrium point in L_1 contrary to our assumption of nonsingularity of X .

Case 1 exhausts all possibilities except for the one in which $\hat{\mathcal{F}}$ itself consists of “crosswise bands.” In this case pick \hat{C} as the lengthwise band. Let $C = \phi^{-1} \cdot \hat{C}$. Once again push C radially inward. Instead of closed curves we now get infinite curves. Let L_1 be a planar leaf of the Reeb component and let $\{C_n\}_{n=1}^\infty$ be the collection of curves produced on L_1 as above. By continuity, close enough to L the curves C_n are transverse to \mathcal{F}/L_1 . Moreover, since all X trajectories on L are closed curves on L inducing trivial holonomy, then by continuity, X trajectories on L_1 close enough to L , are either closed curves or intersect C_n twice for some n . But by the Poincaré–Bendixson Theorem either possibility implies the existence of a singular point of X on L_1 , which is again a contradiction.

(iv) Codim $\Delta = 2$, $\dim G = 3$. We do not have a general theory for this situation. However the following example shows that global (f, g) -invariance is possible.

Let $\pi: S^3 \rightarrow S^2$ be the Hopf fibration and let \mathcal{F} be the foliation with fibers of π as leaves. Let X be a nowhere zero vector field tangential to \mathcal{F} . Since X is in the

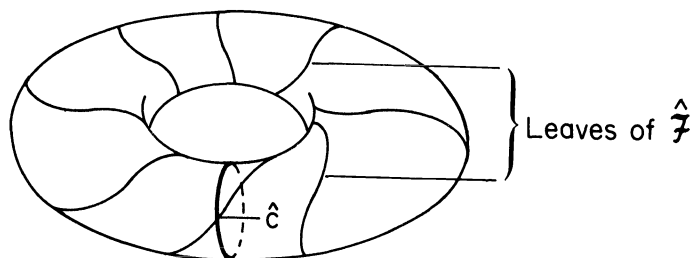


FIG. 1. Illustration of Case I: \hat{C} is a crosswise band.

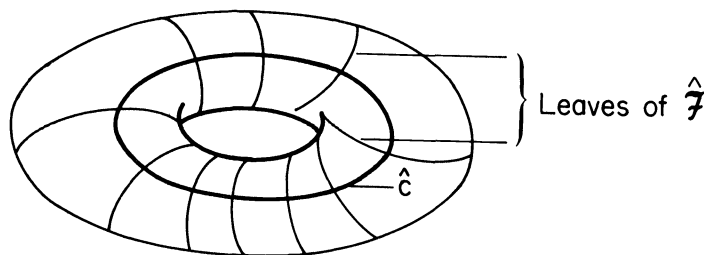


FIG. 2. Illustration of Case II: \hat{C} is a lengthwise band.

direction of a vector field in the Lie algebra of the Lie group S^3 , we can find vector fields Y_1, Y_2 such that (X, Y_1, Y_2) are linearly independent everywhere. Now identify the normal bundle Q to \mathcal{F} with $\text{span}\{Y_1, Y_2\}$. Pick three vector fields Z_1, Z_2, Z_3 on S^2 such that at each point on S^2 , two of the vector fields are linearly independent. Using π , pull back Z_1, Z_2, Z_3 to define sections $\hat{g}_1, \hat{g}_2, \hat{g}_3$ of Q (and hence vector fields on S^3). Let $\hat{g}_i = \sum_{j=1}^2 \gamma_{ij} Y_j$. Now the subbundle H of $T(S^3)$ spanned by $W_1 = \gamma_{11}X + \gamma_{21}Y_1 + \gamma_{31}Y_2$ and $W_2 = \gamma_{12}X + \gamma_{22}Y_1 + \gamma_{32}Y_2$ is a two-dimensional subbundle of S^3 . Pick a complementary one-dimensional subbundle and then, using simple connectedness of S^3 , pick a nowhere zero vector field W_3 such that $\text{span}\{W_1, W_2, W_3\} = T(S^3)$. Let $W_3 = \gamma_{13}X + \gamma_{23}Y_1 + \gamma_{33}Y_2$. Now set $g_i = \hat{g}_i + \gamma_{i3}X$; $i = 1, 2, 3$. By construction g_1, g_2 , and g_3 are linearly independent and $[g_i, \Delta] \subset \Delta$.

Remark. This example shows that when \mathcal{F} is the foliation as described above and g_1, g_2, g_3 are linearly independent (arbitrary) vector fields and f is arbitrary, then Δ is globally (f, g) -invariant.

8. Applications. This section includes some results which describe the global behavior of nonlinear systems using the global Quaker Lemma (Proposition 6.4). For the sake of convenience, we will consider nonlinear systems defined over R^n .

8.1. Cascade decomposition problem with global coordinates.

DEFINITION 8.1. The following is termed the *cascade decomposition problem with global coordinates*: Find $\alpha: M \rightarrow R^n$ and $\beta: M \rightarrow GL(m, R)$ (α, β smooth) and a diffeomorphism $T: R^n \rightarrow R^n$, such that after using feedback control $u = \alpha(x) + \beta(x)v$, and state space transformation T , the new system has the form

$$\begin{aligned} \dot{z}^1 &= \bar{f}_1(z^1, z^2) + \sum_1^m \bar{g}_i^1(z^1, z^2)v_i, \\ \dot{z}^2 &= \bar{f}_2(z^2) + \sum_1^m \bar{g}_i^2(z^2)v_i. \end{aligned} \quad (8.1)$$

PROPOSITION 8.2. *The cascade decomposition problem with global coordinates is solvable if there exists a nonsingular involutive distribution Δ which is globally weakly (f, g) -invariant and such that $G \cap \Delta$ is nonsingular, and there exists a Frobenius chart (R^n, T) with $T(R^n) = R^n$.*

Proof. The proof is immediate from Proposition 6.4.

8.2. Global parallel decomposition problem.

DEFINITION 8.3. Let $I := (I_1, \dots, I_k)$ be a partition of $\{1, \dots, m\}$ and $J := (J_1, \dots, J_k)$ be a partition of $\{1, \dots, n\}$. Let $x^j = \{x_s | s \in J_j\}$. The following is termed the *global parallel decomposition problem* with associated partitions I and J . Find $\alpha: M \rightarrow R^n$ and $\beta: M \rightarrow GL(m, R)$ (α, β smooth) and a diffeomorphism $T: R^n \rightarrow R^n$ such that by using the feedback control law $u = \alpha + \beta v$, the system in new coordinates becomes

$$\begin{aligned} \dot{z}^1 &= \bar{f}^1(z^1) + \sum_{j \in I_1} \bar{g}_j(z^1)v_j \\ &\vdots \\ \dot{z}^k &= \bar{f}^k(z^k) + \sum_{j \in I_k} \bar{g}_j(z^k)v_j. \end{aligned} \quad (8.2)$$

DEFINITION 8.4. Let $\Delta_1, \dots, \Delta_k$ be a set of simultaneously integrable distributions with $\Delta_1 \oplus \dots \oplus \Delta_k = T(R^n)$. A coordinate chart (V, T) is called a *Frobenius chart with respect to $\Delta_1, \dots, \Delta_k$* if in these coordinates $\Delta_j = \text{sp}\{\partial/\partial x_s | s \in J_j\}$.

PROPOSITION 8.5. *The global parallel decomposition problem is solvable if the following hold:*

(i) *There exist distributions $\Delta_1, \dots, \Delta_k$ which are simultaneously integrable, globally weakly (f, g) -invariant and $\Delta_i \cap G$ has constant rank.*

(ii) *There exists a Frobenius chart (R^n, T) with respect to $\Delta_1, \dots, \Delta_k$ such that $T(R^n) = R^n$.*

(iii) $G = G \cap \Delta_1 \oplus G \cap \Delta_2 \oplus \dots \oplus G \cap \Delta_k$.

Proof. Since R^n is contractible and since $\Delta_i \cap G$ has constant rank, we can find separating feedback $(0, \beta_1)$ such that $\tilde{g} = (\tilde{g}_1, \dots, \tilde{g}_m) := g\beta_1$ has the property that $\text{span}\{\tilde{g}_j | j \in I_i\} = G \cap \Delta_i$. Denote $\Delta^i := \sum_{j \neq i} \Delta_j$, $i = 1, \dots, k$, $\tilde{g}^i = \{\tilde{g}_j | j \in I_i\}$ and $\tilde{G}^i := \text{span}\{\tilde{g}_j | j \in I_i\}$.

By (i) it follows easily that Δ^i is globally weakly (f, \tilde{g}^i) -invariant. Hence by Proposition 6.4, there exist α^i and β^i making Δ^i globally (f, \tilde{g}^i) -invariant. We set

$$\beta = \beta_1 \begin{bmatrix} \beta^1 & 0 \\ & \ddots \\ 0 & \beta^k \end{bmatrix}, \quad \alpha = \beta_1 \begin{bmatrix} \alpha^1 \\ \vdots \\ \alpha^k \end{bmatrix}.$$

Now we can see easily that T as in (ii) and (α, β) above satisfy the requirements.

8.3. Global output decoupling problem. Consider the system (1.2) with $M = R^n$ and with output functions $y = (y_1, \dots, y_r) = h(x)$. Let (y^1, \dots, y^k) be a partition of the outputs.

DEFINITION 8.6. The following problem is termed the *output decoupling problem with global coordinates*: Find feedback $u = \alpha + \beta v$ with α, β as usual and a diffeomorphism $T: R^n \rightarrow R^n$ such that in T -coordinates the closed-loop system state equation has the form (8.2) and the output equation has the form $y^i = y^i(z^i)$, $i = 1, 2, \dots, k$.

PROPOSITION 8.7. *The output decoupling problem with global coordinates is solvable if (i)–(iii) of Proposition 8.5 are satisfied and*

$$\Delta_i \subset \bigcap_{j \neq i} \ker dh^j, \quad i = 1, \dots, k.$$

Proof. The proof follows at once from Proposition 8.5.

REFERENCES

- [1] G. BASILE AND G. MARRO, *Controlled and conditioned invariant subspaces in linear system theory*, J. Optim. Theory Appl., 3 (1969), pp. 306–315.
- [2] R. BOTT, *Lectures on Characteristic Classes of Foliations* (Notes by L. Conlon), Lecture Notes in Math. 279, Springer-Verlag, New York, Berlin, 1972, pp. 1–80.
- [3] C. I. BYRNES AND A. J. KRENER, *On the existence of globally (f, g) -invariant distributions*, Differential Geometric Control Theory Conference, Cambridge, MA, Birkhauser, Boston, 1983.
- [4] C. I. BYRNES, *Toward a global theory of (f, g) -invariant distributions with singularities*, in Proc. Mathematical Theory of Network and Systems Conference, Beer-Sheba, Israel, 1983.
- [5] ———, *Feedback decoupling of rotationed disturbances for spherically constrained systems*, in Proc. IEEE Conference on Decision and Control, December 1984.
- [6] W. M. BOOTHBY, *Transversely complete e -foliations of codimension one and accessibility property of nonlinear systems*, in Proc. Conference on Geometric Control Theory, Math-Sci Press, Brookline, MA, 1976, pp. 361–384.
- [7] L. CONLON, *Transversally parallelizable foliations of codimension two*, Trans. Amer. Math. Soc., 194 (1974), pp. 79–102.
- [8] D. CHENG, *On Linearization and decoupling problems of nonlinear systems*, D.Sc. thesis, Washington Univ., St. Louis, MO, 1985.

- [9] N. FLIESS, *Cascade decompositions, invariant foliations and ideals of Lie Algebras*, in Proc. IEEE Conference on Decision and Control, December, 1985.
- [10] I. C. GASPARINI, *Global reduction of a dynamical system on a foliated manifold and controlled projectability*, preprint.
- [11] G. HECTOR AND U. HIRSCH, *Introduction to the geometry of foliations*, Parts A and B, Vieweg, Braunschweig, 1983.
- [12] R. HIRSCHORN, *(A, B)-invariant distributions and disturbance decoupling of nonlinear systems*, SIAM J. Control Optim., 19 (1981), pp. 1–19.
- [13] D. HUSEMOLLER, *Fiber bundles*, 2nd edition, Springer-Verlag, New York, Berlin, 1975.
- [14] A. ISIDORI, A. J. KRENER, C. GORI-GIORI, AND S. MONACO, *Nonlinear decoupling via feedback. A differential geometric approach*, IEEE Trans. Automat. Control AC, 26 (1981), pp. 331–345.
- [15] ———, *Locally (f, g)-invariant distributions*, Systems Control Lett., 1 (1981), pp. 12–15.
- [16] A. ISIDORI, *Nonlinear Control Systems: An Introduction*, Lecture Notes in Control and Information Sci., 71, Springer-Verlag, New York, Berlin, 1985.
- [17] H. IMANISHI, *On the theorem of Denjoy–Sacksteder for codimension one foliations without holonomy*, J. Math. Kyoto Univ., 14 (1974), pp. 607–634.
- [18] A. KRENER, *(f, g)-invariant distributions, connections and Pontryagin classes*, in Proc. IEEE Conference on Decision and Control, 1981.
- [19] ———, *(ad_f, g), (ad_f, g) and locally (ad_f, g) invariant and controllability distributions*, SIAM J. Control and Optim., 23 (1985), pp. 523–549.
- [20] A. S. MORSE AND W. M. WONHAM, *Status of noninteracting control*, IEEE Trans. Automat. Control, 16 (1971), pp. 568–581.
- [21] H. NIJMEIJER, *Controlled invariance for affine control systems*, Internat. J. Control, 24 (1981), pp. 825–833.
- [22] S. H. NIJMEIJER AND A. VAN DER SCHAFT, *Controlled invariance for nonlinear systems*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 904–914.
- [23] S. P. NOVIKOV, *Topology of foliations*, Trans. Moskow Math. Soc., 14 (1965), pp. 248–278, Trans. Amer. Math. Soc. 14 (1967), pp. 268–304.
- [24] B. L. REINHART, *Foliated manifolds with bundle-like metrics*, Ann. of Math., Vol. 69, No. 1, January 1959.
- [25] ———, *Differential Geometry of Foliations*, Springer-Verlag, New York, 1983.
- [26] W. RESPONDEK, *On decomposition of nonlinear control systems*, System Control Lett., 1 (1982), pp. 301–308.
- [27] ———, *Global aspects of linearization, equivalence to polynomial forms and decomposition of nonlinear control systems*, in Proc. Conference on the Algebraic and Geometric Methods in Nonlinear Control, Centre National de la Recherche Scientifique, Paris, 1985.
- [28] R. SACKSTEDER, *Foliations and pseudogroups*, Ann. of Math., 87 (1965), pp. 79–102.
- [29] D. TISCHLER, *On fibering certain foliated manifolds over S^1* , Topology, 9 (1970), pp. 79–102.
- [30] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1979.
- [31] C. GODBILLON, *Feuilletages: Etudes Géométriques I*, Université Louis Pasteur, Strasbourg, France, 1985.
- [32] W. DAYAWANSA, D. CHENG, W. M. BOOTHBY, AND T. J. TARN, *On the global (f, g)-invariance of a class of nonlinear systems*, in Proc. 25th IEEE Conference on Decision and Control, Athens, Greece, December 10–12, 1986, pp. 2065–2068.
- [33] G. REEB, *Sur certaines propriétés topologiques des variétés feuilletées*, Actualités Sci. Ind. 1183, Hermann, Paris, 1952, pp. 91–154, 157–158.
- [34] C. EHRESMANN, *Les connexions infinitésimales dans un espace fibré différentiable*, Colloque de topologie (espaces fibrés), Brussels, G. Thone, ed., Masson, Paris, 1950, pp. 29–55.
- [35] J. S. PASTERNAK, *Foliations and compact Lie group actions*, Comment. Math. Helv., 46 (1972), pp. 55–65.

EXIT TIME PROBLEMS IN OPTIMAL CONTROL AND VANISHING VISCOSITY METHOD*

G. BARLES† AND B. PERTHAME†

Abstract. The authors study the connections between deterministic exit time control problems and possibly discontinuous viscosity solutions of a first-order Hamilton–Jacobi (HJ) equation up to the boundary. This equation admits a maximum and a minimum solution that are the value functions associated to stopping time problems on the boundary. When these solutions are equal, they can be obtained through the vanishing viscosity method. Finally, when the HJ equation has a continuous solution, it is proved to be the value function for the first exit time of the domain. It is also the vanishing viscosity limit arising, in particular, in some large deviations problems.

Résumé. Les auteurs étudient les liens entre les problèmes de contrôle déterministe avec temps de sortie et les solutions de viscosité, généralement discontinues, d’une équation d’Hamilton–Jacobi du premier ordre posée jusqu’au bord. Cette équation admet une solution maximale et une minimale qui sont des fonctions valeur de problèmes de temps d’arrêt sur le bord. Quand ces solutions sont égales, elles peuvent être obtenues grâce à la méthode de viscosité évanescence. Enfin, quand l’équation de Hamilton–Jacobi a une solution continue, on découvre que cela est la fonction valeur pour le premier temps de sortie de l’ouvert. C’est aussi la limite par viscosité évanescence qui apparaît, en particulier, dans certains problèmes de grandes déviations.

Key words. deterministic optimal control, exit time, Hamilton–Jacobi equations, viscosity solutions, vanishing viscosity, large deviations

Mots-clé. contrôle optimal déterministe, temps de sortie, équations de Hamilton–Jacobi, solutions de viscosité, viscosité évanescence, grandes déviations

AMS(MOS) subject classifications. primary 35F30, 49C20; secondary 35B25

Introduction. In this work, we are interested in a systematic understanding of the relations between Hamilton–Jacobi (HJ) equations and value functions that can be considered in deterministic exit time control problems, even (and mainly) when these functions are discontinuous. As a consequence of this study, we obtain that no uniqueness holds for discontinuous solutions to the HJ equation under consideration, since the control problem leads naturally to very different value functions solving it. Thus the question is, what happens in the vanishing viscosity method? Does it converge (as it can be thought intuitively) to the value function for the first exit time of the domain? We answer this question only for two cases: when the value functions are not too different, and when the HJ equation admits a continuous solution. Then we obtain a uniqueness result which allows us to conclude, even for Hamiltonians more general than the one appearing in optimal control theory (nonconvex Hamiltonians). As a byproduct of this result, we can simplify the proof (and weaken the assumptions) of convergence theorems of Wentzell–Freidlin type [29] following the new approach, initiated in Evans and Ishii [10], to the method of Fleming [11], [12].

This work is based on the notion of viscosity solutions introduced by Crandall and Lions [7] (see also Crandall, Evans, and Lions [8] or Lions [22]) and extended to different forms of boundary conditions by Lions [23], Perthame and Sanders [25], Soner [27], and Capuzzo-Dolcetta and Lions [6]. This notion has also been extended to discontinuous solutions by Ishii [16], [17] and Barles and Perthame [4].

* Received by the editors June 29, 1987; accepted for publication (in revised form) December 8, 1987.

† Université Paris Dauphine, Place de Lattre de Tassigny, 75775 Paris Cedex 16, France.

This paper is organized as follows. Section I is devoted to the study of the exit time problem and its relation to an HJ equation. In § II we prove a new comparison result between a continuous and a discontinuous solution and we apply it to some vanishing viscosity problem. Finally, the Appendix is devoted to a technical result on the two-sided obstacle problem. Sections I and II are nearly independent, although § I shows the pathology of discontinuous solutions.

I. Exit time problems in optimal control. In this section we consider a system, the state of which is given by the solution of

$$(1) \quad dy_x(t) + b(y_x(t), v(t)) dt = 0, \quad y_x(0) = x \in \bar{\Omega},$$

or, in the case of *relaxed controls*, by

$$(1') \quad d\hat{y}_x(t) + \left[\int_V b(\hat{y}_x(t), v(t)) d\mu_t \right] dt = 0, \quad \hat{y}_x(0) = x \in \bar{\Omega},$$

and the basic cost functions are

$$J(x, v, \theta) = \int_0^\theta f(y_x(s), v(s)) e^{-\lambda s} ds + \varphi(y_x(\theta)) e^{-\lambda \theta}$$

and

$$\hat{J}(x, \mu, \theta) = \int_0^\theta \int_V f(\hat{y}_x(s), v) e^{-\lambda s} d\mu_s ds + \varphi(\hat{y}_x(\theta)) e^{-\lambda \theta}.$$

Ω is a bounded domain in \mathbb{R}^N , f, b are given functions, $\lambda > 0$, θ is a nonnegative number, and φ is a given function on the complementary set of Ω . $v(\cdot) \in L^\infty(\mathbb{R}^+, V)$, $\mu_s \in L^\infty(\mathbb{R}^+, P(V))$ are, respectively, the control and the relaxed control; V is a compact metric space and $P(V)$ is the set of probability measures on V . We assume that

$$(2) \quad \begin{aligned} &\text{For } \phi = f, \quad b_i \quad (1 \leq i \leq N), \quad \phi \in C(\mathbb{R}^N \times V), \quad \phi(\cdot, v) \in W^{1,\infty}(\mathbb{R}^N), \\ &\|\phi(\cdot, v)\|_{W^{1,\infty}} \leq C \quad (\text{independent of } v) \end{aligned}$$

(so that (1) and (1') have a unique solution).

$$(3) \quad \varphi \in \text{BUC}(\Omega^C)$$

(BUC denotes the set of bounded uniformly continuous functions).

Since we are interested in the exit time problem of Ω , let us denote

$$(4) \quad \tau = \inf \{t \geq 0, y_x(t) \notin \Omega\},$$

$$(5) \quad \bar{\tau} = \inf \{t \geq 0, y_x(t) \notin \bar{\Omega}\}$$

(and $\hat{\tau}, \hat{\bar{\tau}}$ are defined by (4) or (5) using \hat{y}_x). In particular we will use the following three value functions:

$$(6) \quad u^+(x) = \inf_{v(\cdot) \in L^\infty(\mathbb{R}^+; V)} [\text{Sup} \{J(x, v, \theta); \tau \leq \theta \leq \bar{\tau}, y_x(\theta) \in \partial\Omega\}], \quad x \in \bar{\Omega},$$

$$(7) \quad u_-(x) = \inf \{\hat{J}(x, \mu, \theta); \hat{\tau} \leq \theta \leq \hat{\bar{\tau}}, \hat{y}_x(\theta) \in \partial\Omega, \mu \in L^\infty(\mathbb{R}^+, P(V))\}, \quad x \in \bar{\Omega},$$

$$(8) \quad \begin{aligned} u(x) &= \inf \{J(x, v, \tau); v(\cdot) \in L^\infty(\mathbb{R}^+, V)\} \quad \text{for } x \in \Omega, \\ u(x) &= \liminf_{y \rightarrow x, y \in \Omega} u(y) \quad \text{for } x \in \partial\Omega. \end{aligned}$$

The reason we introduce a different value for u on $\partial\Omega$ is that u_- is lower semicontinuous (l.s.c.) on $\bar{\Omega}$ and u^+ is upper semicontinuous (u.s.c.) on $\bar{\Omega}$, but u has no such property

and this definition will simplify the notation. Let us mention that u_- has been studied by Quadrat [26] and that this type of value function has already been studied by Ishii [16] and the authors [4]. These works show the connections between the above control problems and the following HJ equation:

$$(9) \quad \begin{aligned} H(x, u, Du) &= 0 \quad \text{in } \Omega, \\ \text{Min } \{H(x, u, Du); u - \varphi\} &\leq 0 \quad \text{on } \partial\Omega, \\ \text{Max } \{H(x, u, Du); u - \varphi\} &\geq 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where

$$(10) \quad H(x, t, p) = \sup_{v \in V} \{b(x, v) \cdot p + \lambda t - f(x, v)\}.$$

Of course, these equations have to be understood in the viscosity sense of the following definition.

DEFINITION. A u.s.c. function u (on $\bar{\Omega}$) is called a viscosity subsolution (or simply a subsolution) of (9) if for any $\phi \in C^1(\bar{\Omega})$ and any $x_0 \in \bar{\Omega}$ such that $\text{Max}_{\bar{\Omega}} (u - \phi) = (u - \phi)(x_0)$,

$$\begin{aligned} H(x_0, u(x_0), D\phi(x_0)) &\leq 0 \quad \text{if } x_0 \in \Omega, \\ \text{Min } (H(x_0, u(x_0), D\phi(x_0)), u(x_0) - \varphi(x_0)) &\leq 0 \quad \text{if } x_0 \in \partial\Omega. \end{aligned}$$

We refer to the references in the Introduction for the motivation of this definition and for variants and properties of viscosity solutions. Let us note only that a supersolution may be defined in a similar fashion for l.s.c. functions. A function u is said to be a solution if u^* is a subsolution and u_* is a supersolution, where

$$(11) \quad \begin{aligned} u^*(x) &= \limsup_{y \in \bar{\Omega}, y \rightarrow x} u(y), \\ u_*(x) &= \liminf_{y \in \bar{\Omega}, y \rightarrow x} u(y). \end{aligned}$$

Our main results are the following. First, we prove that u_- , u_+ , u are viscosity solutions of (9) and an example shows that they may be very different. Thus, no general uniqueness result holds for solutions that are discontinuous on the boundary (see however [16], [4]). Since u^+ (respectively, u_-) is the maximum (respectively, minimum) solution of (9), if we assume

$$(12) \quad (u^+)_* = u_- \quad \text{in } \Omega,$$

then the solution is, in some sense, unique and it can be recovered by vanishing viscosity. Let u_ε satisfy

$$(13) \quad -\varepsilon \Delta u_\varepsilon + H(x, u_\varepsilon, Du_\varepsilon) = 0 \quad \text{in } \Omega, \quad u_\varepsilon|_{\partial\Omega} = \varphi;$$

then

$$u_-(x) = \liminf_{y \rightarrow x, \varepsilon \rightarrow 0} u_\varepsilon(y) \quad \text{in } \Omega.$$

Equation (12) holds, in particular, if $u_- = \varphi$ on $\partial\Omega$. This is a compatibility condition close to that of Lions [22]. Finally, we notice that u has a remarkable regularity property:

$$(14) \quad [(u_*)^*]_* = u_* \quad \text{in } \bar{\Omega}.$$

Our conclusion is that such a property could be a criterion for uniqueness, as can be seen from the uniqueness theorem of § II, but that a general result is still lacking.

The section is divided into three smaller sections. In the first we study u_+ and u_- . The second is devoted to the properties of u , and the last gives examples which show that our results are nearly optimal and that u and u_- can be very different.

I.1. Properties of u_+ and u_- . In this section, we characterize u_+ and u_- as the maximal and minimal solutions of (9). We also prove the “regularity” property $(u^+)_* = u_*$ and that the vanishing viscosity method converges to u in the case when $u_* = u_-$. To do so, we need the following assumptions:

$$(15) \quad \Omega \text{ is a bounded open set satisfying } \overset{\circ}{\Omega} = \Omega.$$

We assume that φ satisfies (3) and we define

$$\begin{aligned} \psi_1 &= C \quad \text{in } \Omega, & \psi_1 &= \varphi \quad \text{in } \Omega^c, \\ \psi_2 &= -C \quad \text{in } \Omega, & \psi_2 &= \varphi \quad \text{in } \Omega^c, \end{aligned}$$

where C is large enough ($C \geq \sup |f|/\lambda + \sup |\varphi|$).

We have the following theorem.

THEOREM I.1. *Under assumptions (2), (3), (15), let v be defined in $\bar{\Omega}$, and let us denote by \hat{v} the extension of v by φ outside $\bar{\Omega}$. Then v is a viscosity subsolution (respectively, supersolution) of (9) if and only if \hat{v} is a viscosity subsolution (respectively, supersolution) of*

$$(16) \quad \text{Max} (u - \psi_1, \text{Min} (u - \psi_2, H(x, u, Du))) = 0 \quad \text{in } \mathbb{R}^N.$$

The definition of solutions of (16) and its main properties are given in the Appendix. In particular, (16) admits a maximal subsolution \bar{u} and a minimal supersolution \underline{u} (explicit formulas are given for \underline{u} and \bar{u}) and \bar{u} and \underline{u} are solutions of (16). Thus, we have the following corollary.

COROLLARY I.2. *$u^+ = \bar{u}|_{\bar{\Omega}}$ and $u_- = \underline{u}|_{\bar{\Omega}}$ are, respectively, the maximal subsolution and the minimal supersolution of (9), and u^+ and u_- are solutions of (9). u^+ is u.s.c. and u_- is l.s.c. in $\bar{\Omega}$.*

The following theorem shows that u^+ and u are not very different.

THEOREM I.3. *Under assumptions (2), (3), (15), we have*

$$(u^+)_* = u_* \quad \text{in } \bar{\Omega}.$$

COROLLARY I.4. *If $u_- = u_*$ in Ω , then there exists a unique l.s.c. (in $\bar{\Omega}$) viscosity solution of (9) satisfying (14). This is the case, for example, if $u_- = \varphi$ on $\partial\Omega$.*

COROLLARY I.5. *Assume that (13) has a solution $u^\varepsilon \in C^2(\Omega) \cap C(\bar{\Omega})$ for ε small enough. Then*

$$u_-(x) \leq \liminf_{\varepsilon \rightarrow 0, y \rightarrow x} u^\varepsilon(y) \leq \limsup_{\varepsilon \rightarrow 0, y \rightarrow x} u^\varepsilon(y) \leq u^+(x),$$

and the function defined above by the \liminf (respectively, the \limsup) is a supersolution (respectively, subsolution) of (9) and if $u_- = u_$ in Ω*

$$u_-(x) = \liminf_{\varepsilon \rightarrow 0, y \rightarrow x} u^\varepsilon(y) \quad \text{in } \Omega.$$

Most of these results are mere adaptations of the method of [4] and Proposition A.1 in the Appendix and we will skip the proofs. Theorem I.1 is an easy consequence of the definition of viscosity solutions. Corollary I.2 can be proved without any difficulty using the Appendix. The only new point is Theorem I.3, which we prove below. Then, Corollary I.4 and I.5 are clear. Let us mention that Corollary I.5 is a consequence of the stability results of [16] and [4], which show that $\limsup_{\varepsilon \rightarrow 0, y \rightarrow x} u^\varepsilon(y)$ and

$\liminf_{\varepsilon \rightarrow 0, y \rightarrow x} u^\varepsilon(y)$ are, respectively, viscosity sub- and supersolutions of (9), together with Corollary I.2.

Proof of Theorem I.3. Since $u^+ \geq u$ in Ω , it is clear that $(u^+)_* \geq u_*$ and it remains to prove the other inequality. We will need the following lemma.

LEMMA I.6. *Let x be a point of Ω , $v(\cdot)$ a control, τ the exit time from Ω of the trajectory associated to x and $v(\cdot)$. There exists a sequence $x_n \in \Omega \rightarrow x$ such that*

$$\liminf_n u^+(x_n) \leq J(x, v, \tau).$$

First, we prove Theorem I.3 by using Lemma I.6.

Let $x \in \Omega$. Then there exists a sequence $x_n \rightarrow x$ such that

$$\lim_n u(x_n) = u_*(x).$$

Now, we take $v_n(\cdot)$ such that

$$u(x_n) + \frac{1}{n} \geq J(x_n, v_n, \tau_n),$$

τ_n being the exit time from Ω of the trajectory y_{x_n} associated to $v_n(\cdot)$. By Lemma I.6, there exists $x_n^p \rightarrow x_n$ such that

$$J(x_n, v_n, \tau_n) \geq u^+(x_n^p) - \frac{1}{p}.$$

Therefore, by a diagonal procedure

$$u_*(x) = \lim_n u(x_n) \geq \liminf_n \left(u^+(x_n^n) - \frac{2}{n} \right) \geq (u^+)_*(x)$$

and the result is proved. \square

Next, we prove Lemma I.6.

Proof of Lemma I.6. We treat only the case when $\tau < \infty$. If $\tau = +\infty$, then $u^+(x) \leq J(x, v(\cdot), \tau)$ and we are done. We are going to build a sequence x_n converging to x such that the trajectory y_{x_n} associated to v and x_n satisfies

$$\tau - 1/n \leq \tau_n \leq \bar{\tau}_n \leq \tau,$$

where τ_n is the first exit time of y_{x_n} from Ω and $\bar{\tau}_n$ from $\bar{\Omega}$. Since the map $Y: z \rightarrow y_z(\tau)$ is an homeomorphism from a neighborhood of x onto some neighborhood of $y_x(\tau)$, and since Ω satisfies (15) for ε small enough, the set $A = Y^{-1}(B_\varepsilon(y_x(\tau)) \cap \bar{\Omega}^C)$ —where $B_\varepsilon(y_x(\tau))$ is the ball of center $y_x(\tau)$ and of radius ε —is not empty and is open, and $x \in \bar{A}$. Let us remark that for $z \in A$, the exit time of y_z from $\bar{\Omega}$ is less than τ . Now, since $\inf \{d(y_x(s), \partial\Omega), 0 \leq s \leq \tau - 1/n\} > 0$ and since $z \rightarrow y_z(s)$ is uniformly Lipschitz in z for $0 \leq s \leq \tau$, there exists $\eta > 0$ such that for $|z - x| \leq \eta$, we have

$$(i) \quad y_z(s) \in \Omega \quad \text{for } 0 \leq s \leq \tau - 1/n,$$

$$(ii) \quad |y_z(s) - y_x(s)| \leq 1/n.$$

Therefore, if we take $x_n \in B_\eta(x) \cap A$, we have

$$(iii) \quad \tau - 1/n \leq \tau_n \leq \bar{\tau}_n \leq \tau$$

and

$$u^+(x_n) \leq \sup \{J(x_n, v, \theta), \theta \in [\tau_n, \bar{\tau}_n], y_{x_n}(\theta) \in \partial\Omega\}.$$

But using the Lipschitz properties for b and f and the continuity of φ we easily deduce

$$u^+(x_n) \leq J(x, v, \tau) + \alpha(n),$$

where $\alpha(n) \rightarrow 0$ when $n \rightarrow \infty$, and the lemma is proved. \square

I.2. Some properties of u . In this section, we study some properties of the function u defined by (8). Let us begin by commenting on the value of u on $\partial\Omega$. If we define \tilde{u} by

$$\begin{aligned}\tilde{u}(x) &= u(x), & x \in \Omega, \\ \tilde{u}(x) &= \varphi(x), & x \in \partial\Omega,\end{aligned}$$

then it is easy to check that \tilde{u}^* (respectively, \tilde{u}_*) is a viscosity subsolution (respectively, supersolution) of (9) if and only if u^* (respectively, u_*) is. We are going to prove that u and u_* are solutions of (9) and some regularity property of u . Finally, let us note that Ishii [16] has already proved that \tilde{u} is a viscosity solution of (9).

THEOREM I.7. Assume (2), (3), (15); then u and u_* are viscosity solutions of (9). Moreover, u_* satisfies

$$(14) \quad [(u_*)^*]_* = u_* \quad \text{in } \bar{\Omega}.$$

Remark I.8. It is known that the functions satisfying (14) are continuous almost everywhere in the Baire sense; this justifies the term regularity. Let us mention that u^+ satisfies (14) since it is u.s.c. The new point here is that u_* is a viscosity solution with our definition of u on $\partial\Omega$. In general, this would be wrong for u_- as we will see in the next section.

Proof of Theorem I.7. Let us first show the regularity property. Since $u \leq u^+$ in Ω and u^+ is u.s.c., we have

$$u_* \leq (u_*)^* \leq u^+ \quad \text{in } \bar{\Omega};$$

therefore, since u_* is l.s.c.

$$u_* \leq [(u_*)^*]_* \leq (u^+)_* \quad \text{in } \bar{\Omega}.$$

But, using Theorem I.3, we have

$$(u^+)_* = u_* \quad \text{in } \bar{\Omega}.$$

Thus (14) is proved. To prove that u is a solution of (9), we give a more direct proof than that of Ishii [16]. We prove only that u is a supersolution on $\partial\Omega$. The other properties may be obtained by the same method. Let x be a point of $\partial\Omega$ and x_n a sequence of points of Ω such that $x_n \rightarrow x$ and $u(x_n) \rightarrow u_*(x)$. Using the Dynamic Programming Principle (cf. Fleming and Rishel [13], Lions [22], Ishii [16]), we know that

$$\begin{aligned}u(x_n) &= \inf_{v(\cdot)} \left(\int_0^{T \wedge \tau} f(y_{x_n}(t), v(t)) e^{-\lambda t} dt + \varphi(y_{x_n}(\tau)) e^{-\lambda \tau} \cdot 1_{\{\tau \leq T\}} \right. \\ &\quad \left. + u(y_{x_n}(T)) e^{-\lambda T} \cdot 1_{\{T < \tau\}} \right).\end{aligned}$$

If $u_*(x) \geq \varphi(x)$, we have nothing to prove. In the other case, we choose T such that

$$(17) \quad T \|f\|_\infty + \rho_\varphi(\|b\|_\infty T) \leq [\varphi(x) - u_*(x)]/2,$$

where ρ_φ is the modulus of continuity of φ . Now, we choose $v_n(\cdot)$ such that

$$\begin{aligned}u(x_n) + 1/n &\geq \int_0^{T \wedge \tau_n} f(y_{x_n}(t), v_n(t)) e^{-\lambda t} dt + \varphi(y_{x_n}(\tau_n)) e^{-\lambda \tau_n} 1_{\{\tau_n \leq T\}} \\ &\quad + u(y_{x_n}(T)) e^{-\lambda T} 1_{\{T < \tau_n\}}.\end{aligned}$$

It is easy to check that for n large enough, $\tau_n > T$, and using the compactness of relaxed controls (cf. [5], [28]) we get

$$u_*(x) \cong \int_0^T \int_V f(\hat{y}_x(t), v) e^{-\lambda t} d\mu_t dt + u_*(\hat{y}_x(T)) e^{-\lambda T},$$

where \hat{y}_x is defined by (1'). We have used the fact that

$$y_{x_n}(T) \rightarrow \hat{y}_x(T)$$

and $\liminf_n u(y_{x_n}(T)) \cong u_*(\hat{y}_x(T))$ by definition. So, we finally obtain that, for T small enough,

$$u_*(x) \cong \inf_{\mu} \left(\int_0^T \int_V f(\hat{y}_x(t), v) e^{-\lambda t} d\mu_t dt + u_*(\hat{y}_x(T)) e^{-\lambda T} \right).$$

This is a superoptimality principle of dynamic programming (cf. Lions and Souganidis [24]) and when it is used, classical arguments lead to the result. Therefore, we skip the end of the proof. \square

We can use the same kind of ideas to prove that $(u_*)^*$ is a subsolution of (9). Again we consider only the case of a point $x \in \partial\Omega$ and we choose a sequence $x_n \in \Omega$ such that $x_n \rightarrow x$ and $u_*(x_n) \rightarrow (u_*)^*(x)$. For any control $v(\cdot)$ and $y \in \Omega$ we have, for some modulus of continuity ρ_n ,

$$\begin{aligned} u_*(x_n) - \rho_n(|x_n - z|) &\leq u(z) \leq \int_0^{T \wedge \tau} f(y_z(s), v(s)) e^{-\lambda s} ds \\ &\quad + \varphi(y_z(\tau)) e^{-\lambda \tau} \cdot 1_{\{\tau \leq T\}} + u(y_z(T)) e^{-\lambda T} \cdot 1_{\{T < \tau\}}. \end{aligned}$$

As before, we must consider the case when $(u_*)^*(x) > \varphi(x)$ and we choose T such that

$$(17') \quad T \|f\|_{\infty} + \rho_{\varphi}(\|b\|_{\infty} T) \leq \frac{(u_*)^*(x) - \varphi(x)}{2}.$$

Then, for n large enough and $|z - x_n|$ small enough (depending on n), we easily obtain $T < \tau(y)$, and thus

$$u_*(x_n) - \rho_n(|x_n - z|) \leq \int_0^T f(y_z(s), v(s)) e^{-\lambda s} ds + u(y_z(T)) e^{-\lambda T}.$$

Since the range of a small neighbourhood of x_n by the application $z \rightarrow y_z(T)$ is a neighbourhood of $y_{x_n}(T)$, we may choose a sequence z_p such that $z_p \rightarrow x_n$ as $p \rightarrow \infty$ and $u(y_{z_p}(T)) \rightarrow u_*(y_{x_n}(T))$. This gives

$$u_*(x_n) \leq \int_0^T f(y_{x_n}(s), v(s)) e^{-\lambda s} ds + u_*(y_{x_n}(T)) e^{-\lambda T},$$

and, letting n go to infinity, we obtain

$$(u_*)^*(x) \leq \int_0^T f(y_x(s), v(s)) e^{-\lambda s} ds + (u_*)^*(y_x(T)) e^{-\lambda T}.$$

And since this holds for any control $v(\cdot)$ we have proved a suboptimality of dynamic programming. This is enough to conclude that $(u_*)^*$ is a subsolution by classical arguments. \square

I.3. An example. The aim of this section is to describe a situation where u_- is very different from u . For (9), it is the only real nonuniqueness feature, since $(u^+)_* = u_*$

in $\bar{\Omega}$. We shall make some comments concerning the connections between the boundary values, the boundary conditions in (9), and the nonuniqueness feature for (9). Let us describe our example. With our notation, we take the following:

$$\Omega = \{(x, y) \in [-1, 1]^2 / x < 0 \text{ or } y < 0\},$$

$$f \equiv 0, \quad \lambda = 0 \text{ (but is not relevant for our example),}$$

$$b(x, v) = -v \quad \text{where } v \in V = \{(v_1, v_2) \in \mathbb{R}^2 / v_1^2 + v_2^2 \leq 1, v_1 \geq 0, v_2 \leq 0\}.$$

Finally, we take $\varphi \equiv 0$ on $\partial\Omega$, except on $[0, 1] \times \{0\}$, on which we define φ by

$$\varphi(x, y) = -4x(1-x).$$

It is now easy to compute u^+ , u , and u_- , and we obtain

$$u^+ \equiv u \equiv 0 \quad \text{in } \bar{\Omega}$$

and

$$u_-(x, y) = \begin{cases} 0 & \text{if } y < 0, \\ -1 & \text{if } x \leq \frac{1}{2}, \quad y \geq 0, \\ \varphi & \text{if } x > \frac{1}{2}, \quad y = 0. \end{cases}$$

Of course, u and u_- are viscosity solutions of (3) in $\bar{\Omega}$. However, we can remark that the values of u_- on $[0, 1] \times \{0\}$ are very different from the behavior of u_- on $[0, 1] \times]-\varepsilon, 0[$. This remark leads to two comments.

(i) It does not seem possible to use the idea of Soner [27] (and used for this type of problem by Ishii [16]) to prove uniqueness results for (9) with functions such as u_- since the behavior at the boundary is completely different from the behavior of the interior points; this will motivate the uniqueness result of § II.

(ii) If we define the l.s.c. and u.s.c. envelopes of u_- in the same way as for u in (8), we must consider the function ω defined by

$$\omega(x, y) = \begin{cases} -1 & \text{if } x \leq 0, \quad y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

$\omega(x) = \liminf_{y \in \Omega, y \rightarrow x} u_-(y)$, and it is easy to check that ω is *not* a viscosity supersolution at the point $(0, 0)$. So, for u_- , the boundary values play an essential role for the viscosity solution property. This is a striking difference from u . It seems that u_- is a singular solution and that the “good” solutions are those which have the regularity property (14).

Remark I.9. In this example, u_- has the regularity property (14) in Ω . But this is not the case in general. (In the above example replace b by \tilde{b} defined by $\tilde{b}(x, v) \equiv (1, 0)$.)

Remark I.10. In this example, u_* is the maximum continuous viscosity subsolution of (9). It is easy to construct examples in which the supremum of the continuous subsolution is u_- , and u_* is different from u_- at some points of Ω . Finally, let us point out that Theorem II.1 below may be applied to this example and provides a uniqueness theorem for u .

II. Uniqueness and vanishing viscosity method for HJ equations. In this section, we prove an extension of the uniqueness result of Ishii [16] and we apply it to simplify the proof and weaken the assumptions for asymptotic theorems of large deviations type.

II.1. A uniqueness result. We prove the uniqueness result for a general HJ equation for which we need the following assumptions:

$$(18) \quad H(x, t, p) \text{ is uniformly continuous in } p \text{ for } x \in \bar{\Omega}, t \in \mathbb{R};$$

(19) There exists a continuous nondecreasing function $m: (0, \infty) \rightarrow [0, \infty)$ such that $m(0) = 0$ and $|H(x, t, p) - H(y, t, p)| \leq m(|x - y|(1 + |p|))$;

(20) $\exists \gamma \geq 0, \forall x \in \bar{\Omega}, p \in \mathbb{R}^N, s \leq t, H(x, t, p) - H(x, s, p) \geq \gamma(t - s)$.

Moreover, we need an assumption for the bounded set Ω , which is a classical assumption introduced by Soner [27]:

(21) There exists a continuous function η defined on a neighborhood of $\partial\Omega$ with values in \mathbb{R}^N and a constant $b > 0$ such that $B(x + t\eta(x), bt) \subset \Omega$ for $x \in \bar{\Omega}$, $0 < t < b$.

THEOREM II.1. Assume (18)–(21), and either $\gamma > 0$ in (20) and $a = 0$, or $\gamma = 0$ and $a > 0$. Let v , an l.s.c. (in $\bar{\Omega}$) function, be a supersolution of (9). Let $w \in C(\bar{\Omega})$ satisfy (in the viscosity sense)

$$(22) \quad \begin{aligned} H(x, w, Dw) &\leq -a \quad \text{in } \Omega, \\ \text{Min}(w - \varphi, H(x, w, Dw) + a) &\leq 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Finally, assume that

$$(23) \quad \forall x \in \partial\Omega \quad \text{if } w(x) > \varphi(x) \quad \text{then } \liminf_{y \rightarrow x, y \in \Omega} v(y) = v(x).$$

Then $w \leq v$ in $\bar{\Omega}$. \square

Remark II.2. Of course, the same result holds for a continuous supersolution and a u.s.c. subsolution. It is also possible to weaken slightly the assumption of continuity of the subsolution by assuming that it is continuous only at the points of $\partial\Omega$. But when we deal with solutions this is of no use, since a solution that is continuous at the points of $\partial\Omega$ is continuous on $\bar{\Omega}$. Indeed, let w be a solution that is continuous at the points of $\partial\Omega$, choose two continuous functions w_1, w_2 such that

$$\begin{aligned} w_1 &\leq w \leq w_2 \quad \text{in } \bar{\Omega}, \\ w_1 &= w_2 = w \quad \text{on } \partial\Omega, \end{aligned}$$

and extend w, w_1, w_2 in \mathbb{R}^N such that

$$w, w_1, w_2 \in \text{BUC}(\mathbb{R}^N), \quad w = w_1 = w_2 \quad \text{in } \Omega^C$$

(BUC denotes the space of uniformly bounded continuous functions). It is easy to check that w is a solution of

$$(24) \quad u + \inf\{-w_1; \sup(H(x, u, Du) - u, -w_2)\} = 0 \quad \text{in } \mathbb{R}^N,$$

and the unique solution of (24) belongs to $\text{BUC}(\mathbb{R}^N)$ (cf. [9], [2]). Therefore, w is continuous in $\bar{\Omega}$.

Thus, the main difference between Theorem II.1 and Theorem 2.1 of [16] is that w is not assumed to satisfy $w \leq \varphi$ on $\partial\Omega$.

Proof of Theorem II.1. The proof is based on the idea of Soner [27] and is almost the same as that of Ishii [16]. So, we just indicate the adaptation required for this proof and we treat the case $\gamma > 0, a = 0$. We are interested in $M = \max_{x \in \bar{\Omega}} (w(x) - v(x))$, which is achieved, say, at \bar{x} . The only new argument is used when $\bar{x} \in \partial\Omega$, $w(\bar{x}) > \varphi(\bar{x})$, and $v(\bar{x}) \geq \varphi(\bar{x})$. It is contained in the proof of the following lemma.

LEMMA II.3. Let $\bar{x}_n \in \Omega$ satisfy (as in (23)), $\bar{x}_n \rightarrow \bar{x}$, $v(x_n) \rightarrow v(\bar{x})$ as $n \rightarrow \infty$ and let ε_n and z_n be defined by

$$\begin{aligned}\varepsilon_n &= \inf \{t \geq 0, \bar{x}_n - t\eta(\bar{x}) \notin \Omega\}, \\ z_n &= \bar{x}_n - \varepsilon_n \eta(\bar{x}).\end{aligned}$$

Then $\varepsilon_n \rightarrow 0$ and if ψ_n is defined by

$$\psi_n(x, y) = w(x) - v(y) - \left| \frac{y - x}{\varepsilon_n} - \eta(\bar{x}) \right|^2 - |x - \bar{x}|^2;$$

then, denoting (x_n, y_n) a maximum point of ψ_n , we have

$$|x_n - \bar{x}| \rightarrow 0, \quad \left| \frac{y_n - x_n}{\varepsilon_n} - \eta(\bar{x}) \right| \rightarrow 0, \quad v(y_n) \rightarrow v(\bar{x}). \quad \square$$

Proof of Lemma II.3. Using (21), it is not difficult to check that $\varepsilon_n \rightarrow 0$. Therefore, for n large enough, $\varepsilon_n \leq b$. First, it is easy to check that

$$|y_n - x_n| \leq C \cdot \varepsilon_n.$$

And so

$$\psi_n(x_n, y_n) \leq (w(x_n) - w(y_n)) + (w(y_n) - v(y_n))$$

which gives

$$(25) \quad \psi_n(x_n, y_n) \leq \rho_w(C \cdot \varepsilon_n) + M,$$

where ρ_w is the modulus of continuity of w . Moreover,

$$\psi_n(x_n, y_n) \geq \psi(z_n, \bar{x}_n)$$

since (x_n, y_n) is a maximum point of ψ_n . We obtain

$$\psi_n(x_n, y_n) \geq w(z_n) - v(\bar{x}_n) - |z_n - \bar{x}|^2.$$

But $z_n \rightarrow \bar{x}$ and $v(\bar{x}_n) \rightarrow v(\bar{x})$, so there exists $\alpha(n) \rightarrow 0$ such that

$$(26) \quad \psi_n(x_n, y_n) \geq M - \alpha(n).$$

From (25) and (26), we deduce that $\psi_n(x_n, y_n) \rightarrow M$. Moreover, since Ω is bounded, we can consider subsequences—still denoted by x_n, y_n —such that $x_n, y_n \rightarrow z \in \bar{\Omega}$. Hence

$$M \leq \limsup_n \psi_n(x_n, y_n) \leq w(z) - \liminf_n v(y_n) - \liminf_n \left| \frac{x_n - y_n}{\varepsilon_n} - \eta(\bar{x}) \right|^2 - |z - \bar{x}|^2.$$

The right-hand side is estimated by

$$w(z) - v(z) \leq M.$$

From the above we deduce the following:

- (i) $z = \bar{x}$,
- (ii) $\liminf_n v(y_n) = v(\bar{x})$,
- (iii) $\liminf_n \left| \frac{x_n - y_n}{\varepsilon_n} - \eta(\bar{x}) \right|^2 = 0$.

Since (i)–(iii) are true for every converging subsequence, the proof is complete.

Using this lemma, we easily conclude the proof, as in Ishii [16]. \square

II.2. A general case of uniqueness. In this section we consider the case when H satisfies the assumption

$$(27) \quad \forall R > 0, H(x, t, p) \rightarrow +\infty \text{ as } |p| \rightarrow +\infty, \text{ uniformly for } x \in \bar{\Omega}, |t| \leq R.$$

Our purpose is to show that Theorem II.1 applies to getting a general uniqueness theorem. Indeed, we will prove that any subsolution is continuous in Ω . We refer to Ishii [16] for other continuity criteria.

THEOREM II.4. *Under assumptions (18)–(21) and (27), any bounded u.s.c. (in $\bar{\Omega}$) subsolution v of (9) is uniformly Lipschitz continuous in Ω and satisfies $v \leq \varphi$ on $\partial\Omega$.*

Proof of Theorem II.4. The proof of this theorem uses classical arguments and we only sketch it. First, we prove that $v \leq \varphi$ on $\partial\Omega$. Take a sequence of smooth functions φ_n on \mathbb{R}^N , $\varphi + 2/n \geq \varphi_n \geq \varphi + 1/n$ on $\partial\Omega$, and define for $\lambda > 0$,

$$w(x) = \lambda d(x, \partial\Omega) + \varphi_n(x).$$

On $E_\alpha = \{x/d(x, \partial\Omega) \geq \alpha\}$, $w(x)$ is continuous and satisfies, for a proper choice of α, λ ,

$$(28) \quad \begin{aligned} H(x, u, Du) &\geq 1/n \quad \text{in } E_\alpha, \\ u &\geq \varphi + 1/n \quad \text{on } \partial\Omega, \\ u &\geq v + 1/n \quad \text{on } \partial E_\alpha \setminus \partial\Omega. \end{aligned}$$

Indeed, for any $\phi \in C^1(\bar{\Omega})$ and any minimum point $x_0 \in \Omega$ of $w - \phi$, we have, for $y_0 \in \partial\Omega$, $d(x_0, \partial\Omega) = |x_0 - u_0|$,

$$\begin{aligned} \phi(x_0) - \phi(x) &\geq w(x_0) - w(x) \\ &\geq -|x_0 - x| \cdot \|D\varphi_n\|_\infty + \lambda(d(x_0, \partial\Omega) - d(x, \partial\Omega)) \\ &\geq -|x_0 - x| \cdot \|D\varphi_n\|_\infty + \lambda(|x_0 - y_0| - |x - y_0|) \\ &\geq -|x_0 - x| \cdot \|D\varphi_n\|_\infty + \lambda|x - x_0|, \end{aligned}$$

choosing $x \in [x_0, y_0]$, and thus $D\phi(x_0) \geq \lambda - \|D\varphi_n\|_\infty$. Now, we choose α (depending upon λ) such that $\alpha\lambda - \|\varphi_n\|_\infty \geq \max v + 1$ so that the third inequality of (28) holds and w remains bounded. Then we choose λ such that (27), the bound on w and the estimate on $D\phi(x_0)$, implies the first equation of (28). Finally, the condition $w \geq \varphi + 1/n$ on $\partial\Omega$ is clear enough. Thus, we may apply Theorem II.1 to compare v and w (notice that (21) is not needed on $\partial E_\alpha/\partial\Omega$, since on this part of the boundary we have a pure Dirichlet condition). Therefore, we obtain $v(x) \leq \varphi_n \leq 2/n + \varphi(x)$ on $\partial\Omega$ and we have proved that $v(x) \leq \varphi(x)$ on $\partial\Omega$.

Now, we prove that v is uniformly Lipschitz continuous in Ω . We introduce v_ε defined by

$$v_\varepsilon(x) = \text{Max}_{y \in \bar{\Omega}} \left\{ v(y) - \frac{|x - y|^2}{\varepsilon} \right\}.$$

(This operation is called sup-convolution and is studied in Lasry and Lions [20].) v_ε is Lipschitzian, is nondecreasing to v , and satisfies, for any strict open subset of Ω and ε small enough,

$$H(x, v_\varepsilon, Dv_\varepsilon) \leq C \quad (\text{independent of } \varepsilon).$$

Since v_ε remains uniformly bounded, (27) gives a uniform bound on Dv_ε and Theorem II.4 is proved.

Remark II.5. With the notation of Theorem II.4, setting $v^0(x) = \limsup_{y \in \Omega, y \rightarrow x} v(y)$, we define a Lipschitzian function on $\bar{\Omega}$ and v^0 is still a subsolution.

II.3. Application to large deviations problems. Assumption (27) is naturally satisfied when we look at the asymptotic behavior for exponentially small probabilities and expectations associated with processes with small diffusion term. We refer the reader to Evans and Ishii [10], Fleming and Souganidis [14], or Bardi [1] for the motivations and the main results obtained recently from the method initiated by Fleming [11], [12].

In order to illustrate how the theory above may simplify the approach to these problems, we have, for ε small enough, a solution $u_\varepsilon \in C^1(\Omega) \cap C(\bar{\Omega})$ of

$$(29) \quad -\varepsilon a_{ij}(x) \frac{\partial^2 u_\varepsilon}{\partial x_i \partial x_j} + H(x, u_\varepsilon, Du_\varepsilon) = 0, \quad u_\varepsilon|_{\partial\Omega} = \varphi,$$

where $a_{ij}(x)$ is a positive symmetric continuous matrix.

(Here we assume that u_ε exists. Generally, it is explicitly given by the logarithmic transformation of a "cost function." Above, it is possible to assume that $u_\varepsilon \in C(\bar{\Omega})$ is a viscosity solution of (29) in the sense of Lions [21]).

In the context of large deviations the following Hamiltonian is relevant:

$$(30) \quad H(x, p) = p' \cdot a(x) \cdot p + b(x) \cdot p.$$

Generally, the question of studying the behavior of u_ε is performed as follows (see [10]):

- (1) Uniform bounds in L^∞ are proved on u_ε ;
- (2) Uniform bounds on the first derivatives of u_ε are proved;
- (3) Then, it is possible to pass to the limit in (29);
- (4) Identify the limit of u_ε by a uniqueness theorem for (9).

We remark that step (2) may be hard to establish (and either wrong if, as below, a boundary layer appears in (29) and a part of the boundary condition is lost). The following proposition shows that this step is not necessary to obtain uniform convergence of u_ε inside Ω . Therefore, it simplifies the program described above, weakens the assumptions on H , and makes the program applicable to more general situations.

PROPOSITION II.6. *Let u_ε be a viscosity solution of (29) in $C(\bar{\Omega})$ satisfying $\|u_\varepsilon\|_\infty \leq C$ (independent of ε) and assume that H satisfies (18)–(21).*

(i) *Then the functions*

$$\bar{u}(x) = \limsup_{y \rightarrow x, \varepsilon \rightarrow 0} u_\varepsilon(y), \quad \underline{u}(x) = \liminf_{y \rightarrow x, \varepsilon \rightarrow 0} u_\varepsilon(y),$$

are, respectively, viscosity sub- and supersolutions of (9).

(ii) *If (27) holds, then $\tilde{u}(x) = \limsup_{y \in \Omega, y \rightarrow x} \bar{u}(y)$ is a Lipschitzian subsolution of (9) and $\tilde{u} \leq \varphi$ on $\partial\Omega$.*

(iii) *Assume (27). Then, assume $\gamma > 0$ in (20), or that $H(x, t, p)$ is convex in p and that (9) admits a strict subsolution in the sense of (31) below; then $\underline{u} = \bar{u}$ in Ω and u_ε converges uniformly on every compact subset of Ω to \bar{u} (or \underline{u}). \square*

(By a strict subsolution we mean a function $w \in C^1(\bar{\Omega})$, such that, for all $t \in \mathbb{R}$,

$$(31) \quad H(x, t, Dw) \leq -a < 0 \quad \text{in } \bar{\Omega}.$$

Proof of Proposition II.6. Much of this proposition is an adaptation of the above results and classical ones. Item (i) is proved in [16] and is an adaptation of the stability result of [4]. Item (ii) is nothing but Theorem II.4 and Remark II.5. Item (iii) is an adaptation to discontinuous solutions of a result of Kruskov [19], Crandall and Lions [7], Lions [22], Ishii [15], Lasry and Lions [20], and Barles [3]. Let us just recall the proof of [15] (we consider the case $\gamma = 0$ only). Let $\theta \in (0, 1)$ and set $u_\theta = \theta \tilde{u} + (1 - \theta)w$.

Thanks to the convexity of H in p and to (31), we get

$$H(x, u_\theta, Du_\theta) \leq -(1-\theta)a < 0 \quad \text{in } \Omega.$$

Since this still holds if we replace w by $w - M$ we get also

$$\text{Min}(u_\theta - \varphi, H(x, u_\theta, Du_\theta)) \leq -(1-\theta)a \quad \text{on } \partial\Omega.$$

Thus, we may apply Theorem II.1 and obtain

$$u_\theta \leq \underline{u},$$

and we conclude the uniqueness statement by letting θ go to 1. Finally, we have obtained

$$\limsup_{\varepsilon \rightarrow 0, y \rightarrow x} u_\varepsilon(y) = \liminf_{\varepsilon \rightarrow 0, y \rightarrow x} u_\varepsilon(y) \quad \forall x \in \Omega,$$

and this implies the uniform convergence of u_ε on every compact subset of Ω . \square

Remark II.7. For the particular Hamiltonian given by (30), the existence of a strict subsolution is known to hold if

$$\forall x(s) \in \bar{\Omega}, x(\cdot) \in H^1_{\text{loc}}([0, \infty); \mathbb{R}^N) \quad \int_0^\infty |\dot{x}(s) - b(x(s))|^2 dx = \infty.$$

Proposition II.6 may be applied to this example as soon as b and a are continuous (and a is uniformly positive). Indeed, it is well known that (18)–(20) must hold only for bounded p since (27) gives the Lipschitz continuity of the subsolutions.

Appendix. On the two-sided obstacle problem. In this Appendix, we consider the problem

$$(16) \quad \text{Max}(u - \psi_1, \text{Min}(u - \psi_2, H(x, u, Du))) = 0 \quad \text{in } \mathbb{R}^N$$

for discontinuous obstacles ψ_1, ψ_2 . We prove that it has a maximum subsolution and a minimum supersolution which are solutions. These results have been used in § I, since the exit time problem is a particular case of (16).

Let us recall what we mean by a solution of (16).

DEFINITION. A u.s.c. u on \mathbb{R}^N is called a viscosity subsolution of (16) if, for any $\phi \in C^1(\mathbb{R}^N)$ and any point x_0 such that $\text{Max}(u - \phi) = (u - \phi)(x_0)$, then

$$H^*(x_0, u(x_0), D\phi(x_0)) \leq 0.$$

The definition of supersolutions or solutions is obtained as usual (see § I).

In the following, H is given by (10) with b and f satisfying (2).

To state our result, we need the following assumption:

$$(32) \quad \psi_1 \text{ and } \psi_2 \text{ are bounded and satisfy } (z^*)_* = z_*, (z_*)^* = z^* \text{ (where } z = \psi_1 \text{ or } \psi_2 \text{)}.$$

PROPOSITION A.1. Under assumptions (2), (32) the functions \bar{u} and \underline{u} defined by

$$\begin{aligned} \bar{u}(x) = & \inf_{v(\cdot), \theta_1} \sup_{\theta_2} \left[\int_0^{\theta_1 \wedge \theta_2} f(y_x(t), v(t)) e^{-\lambda t} dt \right. \\ & \left. + \psi_1^*(y_x(\theta_1)) e^{-\lambda \theta_1} 1_{\{\theta_1 < \theta_2\}} + \psi_2^*(y_x(\theta_2)) e^{-\lambda \theta_2} 1_{\{\theta_2 \leq \theta_1\}} \right], \\ \underline{u}(x) = & \sup_{\theta_2} \inf_{(\mu, \theta_1)} \left[\int_0^{\theta_1 \wedge \theta_2} f(\hat{y}_x(t), v(t)) e^{-\lambda t} dt \right. \\ & \left. + \psi_{1*}(\hat{y}_x(\theta_1)) e^{-\lambda \theta_1} 1_{\{\theta_1 < \theta_2\}} + \psi_{2*}(\hat{y}_x(\theta_2)) e^{-\lambda \theta_2} 1_{\{\theta_2 \leq \theta_1\}} \right] \end{aligned}$$

(where y_x and \hat{y}_x are, respectively, defined by (1) and (1') in \mathbb{R}^N) are, respectively, the maximum viscosity subsolution and solution of (16) and the minimum viscosity supersolution and solution of (16).

Proof of Proposition A.1. Since the arguments are routine adaptations of the arguments of [4], we only sketch the proof for \bar{u} . Let $(\psi_1^n)_n$ and $(\psi_2^n)_n$ be two nonincreasing sequences of functions of $BUC(\mathbb{R}^N)$ such that

$$\psi_1^* = \inf_n \psi_1^n, \quad \psi_2^* = \inf_n \psi_2^n.$$

Let v be a bounded u.s.c. viscosity subsolution of (16). By classical results (see [2], [7]–[9], [22]), we know that there exists a unique bounded and continuous solution u^n of

$$(33) \quad \text{Max} [(u - \psi_1^n), \text{Min} ((u - \psi_2^n), H(x, u, Du))] = 0 \quad \text{in } \mathbb{R}^N.$$

Moreover, since (33) satisfies Isaac's condition [24], u^n is given by

$$u^n(x) = \inf_{(v(\cdot), \theta_1)} \sup_{\theta_2} \left[\int_0^{\theta_1 \wedge \theta_2} f(y_x(t), v(t)) e^{-\lambda t} + \psi_1^n(y_x(\theta_1)) e^{-\lambda \theta_1} 1_{\{\theta_1 < \theta_2\}} + \psi_2^n(y_x(\theta_2)) e^{-\lambda \theta_2} 1_{\{\theta_2 \leq \theta_1\}} \right].$$

Let us only note that, in this particular case, u belongs to $BUC(\mathbb{R}^N)$ and that we could invert the Sup and Inf in the above formula. Moreover, the comparison results for viscosity sub- and supersolution give

$$(34) \quad v \leq u^{n+1} \leq u^n \quad \text{in } \mathbb{R}^N,$$

since v is still a subsolution of (33). So, it is enough to identify the function

$$w = \inf_n u^n.$$

It is clear enough that

$$w \geq \bar{u}.$$

To prove the other inequality we denote by

$$J^n(v, \theta_1, \theta_2) = \int_0^{\theta_1 \wedge \theta_2} f(y_x(t), v(t)) e^{-\lambda t} + \psi_1^n(y_x(\theta_1)) e^{-\lambda \theta_1} 1_{\{\theta_1 < \theta_2\}} + \psi_2^n(y_x(\theta_2)) e^{-\lambda \theta_2} 1_{\{\theta_2 \leq \theta_1\}}$$

and by $J(v, \theta_1, \theta_2)$ the same expression, where ψ_1^n and ψ_2^n are replaced by ψ_1^* and ψ_2^* . We have

$$w(x) = \inf_{(v, \theta_1)} \inf_n \sup_{\theta_2} J^n(v, \theta_1, \theta_2).$$

So, it is enough to compute $\lim_n \sup_{\theta_2} J^n(v, \theta_1, \theta_2)$, v, θ_1 being fixed with $\theta_1 < +\infty$. Let θ_2^n be a maximum point of $J^n(v, \theta_1, \theta_2)$; we may assume that $\theta_2^n \leq \theta_1 + 1$ and therefore we can take a subsequence, still denoted by θ_2^n , which converges to θ_2 :

$$\inf_n \sup_{\theta_2} J^n(v, \theta_1, \theta_2) = \lim_{n \rightarrow +\infty} J^n(v, \theta_1, \theta_2^n) \leq \lim_{\substack{n \rightarrow \infty \\ \theta \rightarrow \theta_2}} \sup J^n(v, \theta_1, \theta)$$

and it is easy to see that the right-hand side of the inequality is exactly $J(v, \theta_1, \theta_2)$. Finally

$$w(x) \leq \inf_{(v, \theta_1)} \sup_{\theta_2} J(x, \theta_1, \theta_2) = \bar{u}(x)$$

which is the result we want. Let us just note that this proof shows that \bar{u} is u.s.c. and it is easy to check that

$$\begin{aligned} \bar{u}(x) &= \limsup_{n \rightarrow \infty, y \rightarrow x} u^n(y), \\ (\bar{u})_*(x) &= \liminf_{n \rightarrow \infty, y \rightarrow x} u^n(y). \end{aligned}$$

So, by the stability result of [4], we conclude that \bar{u} is the viscosity solution of (16). Moreover, by (34), we see that \bar{u} is the maximum viscosity subsolution of (6). Assumption (32) is used to identify the limits of the Hamiltonians and to ensure \bar{u} and \underline{u} are solutions of the same Hamiltonian. The proof is complete. \square

REFERENCES

- [1] M. BARDI, *An asymptotic formula for the Green's function of an elliptic operator*, Ann. Scuola Norm. Sup. Pisa, to appear.
- [2] G. BARLES, *Existence results for first-order Hamilton-Jacobi equations*, Ann. Inst. H. Poincaré, Anal. Non Linéaire, 1 (1984), pp. 325-340.
- [3] ———, *An approach of deterministic unbounded control problems and of first-order Hamilton-Jacobi equations with gradient constraints*, manuscript.
- [4] G. BARLES AND B. PERTHAME, *Discontinuous solutions of deterministic optimal stopping time problems*, Math. Methods Numer. Anal., to appear.
- [5] I. CAPUZZO-DOLCETTA AND H. ISHII, *Approximate solutions of the Bellman equations of deterministic control theory*, Appl. Math. Optim., 11 (1984), pp. 161-181.
- [6] I. CAPUZZO-DOLCETTA AND P. L. LIONS, *Hamilton-Jacobi equations and state constraint problems*, to appear.
- [7] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.
- [8] M. G. CRANDALL, L. C. EVANS, AND P. L. LIONS, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487-502.
- [9] M. G. CRANDALL, H. ISHII, AND P. L. LIONS, *Uniqueness of viscosity solutions revisited*, J. Math. Soc. Japan, 39 (1987).
- [10] L. C. EVANS AND H. ISHII, *A PDE approach to some asymptotic problems concerning random differential equations with small noise intensities*, Ann. Inst. H. Poincaré, 2 (1985), pp. 1-20.
- [11] W. H. FLEMING, *Logarithmic transformations and stochastic control*, in Advances in Filtering and Stochastic Control, W. F. Fleming and L. G. Gorostiza, eds., Springer-Verlag, Berlin, New York, 1983.
- [12] ———, *A stochastic control approach to some large deviations problems*, Lecture Notes in Math. 1119, Springer-Verlag, Berlin, New York, 1984.
- [13] W. H. FLEMING AND R. W. RISHEL, *Deterministic and stochastic optimal control*, Lecture Notes in Math. 1119, Springer-Verlag, Berlin, New York, 1975.
- [14] W. H. FLEMING AND P. E. SOUGANIDIS, *A P.D.E. approach to asymptotic estimates for optimal exit probabilities*, Ann. Scuola Norm. Sup. Pisa, 1986.
- [15] H. ISHII, *A simple, direct proof of uniqueness for solutions of the Hamilton-Jacobi equations of eikonal type*, manuscript.
- [16] ———, *A boundary value problem of the Dirichlet type for Hamilton-Jacobi equations*, manuscript.
- [17] ———, *Perron's method for Hamilton-Jacobi equations*, Duke Math. J., to appear.
- [18] S. KAMIN, *Elliptic perturbation for linear and nonlinear equations with a singular point*, to appear.
- [19] S. N. KRUSKOV, *Generalized solutions of the Hamilton-Jacobi equations of eikonal type. I*, Math. USSR-Sb., 27 (1975), pp. 406-446.
- [20] J. M. LASRY AND P. L. LIONS, *A remark on regularization in Hilbert spaces*, Israel J. Math., 55 (1986), pp. 257-266.

- [21] P. L. LIONS, *Optimal control of diffusion processes and Hamilton-Jacobi equations. Part II: viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1229–1276.
- [22] ———, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, 1982.
- [23] ———, *Neumann type boundary conditions for Hamilton-Jacobi equations*, Duke Math. J., 52 (1985), pp. 793–820.
- [24] P. L. LIONS AND P. E. SOUGANIDIS, *Differential games, optimal control and directional derivatives of viscosity solutions of Bellman's and Isaac's equations*, SIAM J. Control and Optim., 23 (1985), pp. 566–583.
- [25] B. PERTHAME AND R. SANDERS, *The Neumann problem for nonlinear second order singular perturbation problems*, SIAM J. Math. Anal., 19 (1988), pp. 295–311.
- [26] J. P. QUADRAT, *Thèse d'état*, Université Paris IX-Dauphine, 1981.
- [27] H. M. SONER, *Optimal control problems with state space constraints*, SIAM J. Control and Optim., 24 (1986), pp. 552–561.
- [28] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [29] A. D. WENTZELL AND M. I. FREIDLIN, *Random Perturbations of Dynamical Systems*, Springer-Verlag, New York, 1984.

DIFFERENTIAL GEOMETRIC METHODS IN HYBRID PARAMETRIZATION OF LINEAR DYNAMICAL SYSTEMS*

B. K. GHOSH† AND W. P. DAYAWANSA‡

Abstract. This paper introduces a non-Euclidean parametrization of linear dynamical systems both in the open loop and in the closed loop. In particular, the various parametrizations introduced have a bundle structure which appears to be of relevance in system identification. Furthermore a quotient topology is obtained on the space of all systems and compared with the well-known graph topology. The article describes many desirable features of the quotient topology. Applications of vector bundle and fiber bundle theory in parametrization of control systems in the closed loop introduced in this paper are new.

Key words. vector bundle, parametrization, characteristic class

AMS(MOS) subject classifications. 14, 30, 93

1. Introduction. In feedback control system design, there has been recent interest in considering collections of linear multi-input multi-output dynamical systems, rather than fixed ones, and recursive schemes to update the parameters of compensators in order to satisfy specific sets of design constraints (e.g., sensitivity minimization, stabilization, etc.) in the closed loop. As has already been noted in the literature [1]–[3], the collection of plants might arise as a result of parameter variation or of our ignorance about the true value of a set of parameters and wherein only a bound on the parameters is known. Likewise the compensator parameters might have to be updated or adjusted in view of more recent information about the plant parameters or as a result of a change in the design requirements.

As a first step in the design of a compensator, it is important to select a suitable parametrized space for systems. Understanding the topology of this space is important since the convergence of algorithms may depend on it. Brockett [27] has studied the space of single-input single-output systems of McMillan degree n (called $\text{Rat}(n)$) and has paved the way for interesting research work by many others. In addition to this, we are clearly interested in obtaining a suitable parametrization of a class of plants and compensators for which recursive adjustment of the parameters is feasible. At the very least, we are interested in both off-line and on-line variations of the parameters of the plant-compensator pair in the closed loop that achieve robust stability. In this paper, we initiate such a program and parametrize plants and compensators that are robustly stable in the closed loop. An advantage of our parametrization is that the McMillan degrees of the family of plants and compensators considered are not assumed to be fixed. This is a very desirable feature, since it is extremely difficult to estimate the McMillan degree in parameter identification. As a special case, we parametrize the space of strictly proper multi-input multi-output systems and show that the associated space is graded (i.e., there is a sequence of subsets $\Omega_1 \subset \Omega_2 \subset \cdots \subset \Omega_n \subset \cdots$ such that $\bigcup_{n=1}^{\infty} \Omega_i$ is the given space) and each of the graded spaces (i.e., Ω_i 's) is diffeomorphic to a Euclidean space. This property is particularly desirable in system identification, as has already been argued in [4], if the identification algorithm is defined by a locally and globally convergent vector field. This fact is justified due to the following theorem of Milnor (see Appendix I and [5]): "If a manifold admits a locally

* Received by the editors November 6, 1986; accepted for publication (in revised form) October 12, 1987. This research was partly supported by National Science Foundation grant ECS-8414220.

† Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130.

‡ Department of Mathematics, Texas Tech University, Lubbock, Texas 79409.

and globally asymptotically stable vector field then it is diffeomorphic to a Euclidean space." Thus a non-Euclidean space does not admit a globally convergent vector field. On the other hand, because of the Euclidean structure of the graded spaces, it may now be possible to globally generalize many known parameter identification algorithms described in [10], where the associated parameter space is assumed to be Euclidean.

In general however, i.e., when the plants and the compensators are not necessarily strictly proper, we show that the proposed graded space is not Euclidean. However, it has the structure of a vector bundle [6], [7]. A vector bundle has the desirable feature that its fibers are isomorphic to a Euclidean space. Thus a point on the fiber may be identified via globally convergent identification schemes defined on a Euclidean space. Identifying points on the non-Euclidean base space may be difficult in general. However, in our parametrization the base space is sufficiently lower-dimensional compared to the dimension of the total space. Existence of separate identification algorithms on the base space and on the fibers is currently under study.

The main results of this paper are now described. In § 3 we introduce the notion of a lag and parametrize the space of $p \times m$ strictly proper systems of lag $\leq n$. In § 4 we study the vector bundle structure of the space of $p \times m$ proper or improper systems of lag $\leq n$. In § 5 we parametrize plant-compensator pairs of lag $\leq n$ and $\leq q$, respectively, that are stable in the closed loop and show that the parametrized space has the structure of a vector bundle. In § 6, we construct a fiber bundle of the space of proper or improper systems with unbounded lag and describe a quotient topology on the space of systems. This topology is compared with the well-known graph topology [8] and many desirable features of the proposed topology are described. Finally, in § 7 we parametrize plant-compensator pairs of lag $\leq n$ and $\leq q$, respectively, that are not necessarily stable in the closed loop and show that such a space has the structure of a fiber bundle.

In this paper we introduce the application of non-Euclidean geometry in parametrization problems and therefore cover some of the work of Byrnes [11], [12], Delchamps [13], and Helmke [14], [15]. The main point of this paper, however, is that, via a suitable over-parametrization, the parameter space we obtain is endowed with the structure of either a vector bundle or a fiber bundle. The base of the bundle so obtained is generally not a Euclidean space. Thus we might consider a set of charts which cover the base. This would agree with some of the works of Hazewinkel [16] on the numerical aspect of the parametrization problem as a consequence of "chart changes."

Here we assume that the reader is familiar with the "characteristic class" theory at the level of [7]. For a good introduction of this subject and its relevance to system theory we refer the reader to [13].

2. Notation.

- $LS_{m,p}^n$: Space of $p \times m$ strictly proper systems of lag n .
- $LP_{m,p}^n$: Space of $p \times m$ proper systems of lag n .
- $L\Omega_{m,p}^n$: A parametrization of the set of $p \times m$ strictly proper systems of lag $\leq n$.
- $L\Pi_{m,p}^n$: A parametrization of the set of $p \times m$ proper systems of lag $\leq n$.
- $L\Omega_{m,p}^{n,q}$: A parametrization of the set of $p \times m$ strictly proper systems of lag $\leq n$ that can be stabilized by a compensator of lag $\leq q$.
- $L\Pi_{m,p}^{n,q}$: A parametrization of the set of $p \times m$ proper systems of lag $\leq n$ that can be stabilized by a compensator of lag $\leq q$.
- $LFB_{m,p}^{n,q}$: A parametrization of $p \times m$ systems $G(s)$ of lag $\leq n$ and $m \times p$ compensators $K(s)$ of lag $\leq q$ such that the closed-loop system $G(s)[I + K(s)G(s)]^{-1}$ is proper and stable.

- \mathbb{R} : The real line.
 \mathbb{RP}^n : The n -dimensional real projective space.
 \mathbb{C} : The complex plane.
 $\bar{\mathbb{C}}$: The complex plane together with the point at infinity.
 \mathbb{D}_s : The open interior of the unit disc.
 \mathbb{D}_u : $\bar{\mathbb{C}} - \mathbb{D}_s$.
 $L_{m,p}^n$: A parametrization of $p \times m$ proper or improper autoregression moving average (ARMA) systems having a lag of at most n .
 $L_{m,p}^{n,q}$: The space of ARMA systems in $L_{m,p}^n$ that can be stabilized by an $m \times p$ ARMA system in $L_{p,m}^q$.
 $L_{m,p}^\infty$: A parametrization of $p \times m$ proper or improper ARMA systems having an arbitrary large lag.
 $\tilde{L}_{m,p}^\infty$: A quotient space of all $p \times m$ ARMA systems.
 H : The ring of rational functions with poles in \mathbb{D}_s .
 J : Set of invertible elements in H .
 $H^{p \times m}$: A $p \times m$ matrix with elements in H .
 $LF_{m,p}^{n,q}$: A parametrization of m -input p -output feedback control systems of plants of lag $\leq n$ and compensators of lag $\leq q$.

3. A parametrization of the space of $p \times m$ strictly proper systems of lag $\leq n$. In this section we consider the space $LS_{m,p}^n$ of $p \times m$ strictly proper systems of lag n . The notion of "lag" rather than the notion of "McMillan degree" is frequently considered in economics and statistics in the modeling of autoregressive moving average (ARMA) systems. In particular, we refer to Deistler and Hannan [9] and consider the following difference equation:

$$(3.1) \quad D_0 y(t) + D_1 y(t-1) + \cdots + D_n y(t-n) = N_0 u(t) + N_1 u(t-1) + \cdots + N_n u(t-n)$$

where $y(t)$ is the output p vector, $u(t)$ is the input m vector, and where D_i, N_i for $i=0, \dots, n$ are, respectively, $p \times p$ and $p \times m$ real matrices. The above difference equation, (3.1), can be represented in the frequency domain as follows:

$$(3.2) \quad D(z)y(z) = N(z)u(z)$$

where

$$(3.3) \quad \begin{aligned} D(z) &= D_0 z^n + D_1 z^{n-1} + \cdots + D_n, \\ N(z) &= N_0 z^n + N_1 z^{n-1} + \cdots + N_n. \end{aligned}$$

Let us now consider the following definitions.

DEFINITION 3.1. The pair $D(z), N(z)$ is coprime if

$$(3.4) \quad \text{rank} [D(z), N(z)] = p$$

for all $z \in \bar{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$.

DEFINITION 3.2. The difference equation (3.1) is said to represent an ARMA model if $D(z), N(z)$ are coprime.

DEFINITION 3.3. An ARMA model (3.1) is said to be proper if D_0 is nonsingular. Otherwise it is called improper.

DEFINITION 3.4. An ARMA model (3.1) is said to be strictly proper if D_0 is nonsingular and $N_0 = 0$.

Let us now describe the space $LS_{m,p}^n$ as follows:

$$(3.5) \quad LS_{m,p}^n \triangleq \{(D_0, \dots, D_n, N_0, \dots, N_n) : D_0 = I, N_0 = 0 \text{ and } D(z), N(z) \text{ are coprime}\}.$$

From (3.5), we clearly have the following proposition which we state without proof.

PROPOSITION 3.5. $LS_{m,p}^n$ is an open and dense subset of $\mathbb{R}^{np(m+p)}$.

We now describe the following parametrization problem originally considered in an earlier paper [4].

Problem 3.6. Assume that $D_0 = I$, $N_0 = 0$. Parametrize the set of points $\xi \triangleq (D_1, \dots, D_n, N_1, \dots, N_n)$ in $\mathbb{R}^{np(m+p)}$ with the property that there exists a neighborhood $N(\xi)$ of ξ in $\mathbb{R}^{np(m+p)}$ such that $N(\xi) \cap LS_{m,p}^n$ is simultaneously stabilizable by a dynamic compensator.

Throughout this paper, we shall assume that the region of stability is \mathbb{D}_s , the open interior of the unit disc. Elementary arguments would show that the set of points ξ , satisfying the property described in Problem 3.6, give rise to the space $L\Omega_{m,p}^n$ defined as follows:

$$(3.6) \quad L\Omega_{m,p}^n \triangleq \{(D_0, D_1, \dots, D_n, N_0, N_1, \dots, N_n) : D_0 = I, N_0 = 0 \text{ and rank } [D(z), N(z)] = p \text{ for all } z \text{ in } \mathbb{D}_u\}.$$

Clearly the space $L\Omega_{m,p}^n$ parametrizes the set of ARMA systems with $\text{lag} \leq n$. In this parametrization, however, a given system is represented in a nonunique way. To see why (3.6) indeed describes the set of points defined in Problem 3.6 let $N_c(z)D_c(z)^{-1}$ be the transfer function of a dynamic compensator. The return difference polynomial is given by $\det [D(z)D_c(z) + N(z)N_c(z)]$. Thus if there exists a point ξ in $\mathbb{R}^{np(m+p)}$ for which $\text{rank } [D(z)N(z)] < p$ for some z_0 in \mathbb{D}_u , then every neighbor $N(\xi)$ of ξ would have a point ξ' in $LS_{m,p}^n$ with the property that if $D(z)^{-1}N'(z)$ is the corresponding transfer function, then $\det [D'(z)D_c(z) + N'(z)N_c(z)]$ vanishes arbitrary close to z_0 . Thus $\xi' \notin L\Omega_{m,p}^n$. The sufficiency condition, on the other hand, is clear.

We now generalize the spaces $LS_{m,p}^n$ and $L\Omega_{m,p}^n$ and consider proper systems of lag n and $\text{lag} \leq n$, respectively, as follows:

$$(3.7) \quad LP_{m,p}^n \triangleq \{(D_0, \dots, D_n, N_0, \dots, N_n) : D_0 = I \text{ and } D(z), N(z) \text{ are coprime}\},$$

$$(3.8) \quad L\Pi_{m,p}^n \triangleq \{(D_0, \dots, D_n, N_0, \dots, N_n) : D_0 = I \text{ and rank } (D(z), N(z)) = p \text{ for all } z \text{ in } ID_u\}.$$

We shall say more about the spaces $LP_{m,p}^n$ and $L\Pi_{m,p}^n$ in subsequent sections. Presently however we need the space $L\Pi_{m,p}^n$ in order to define the stabilizability problem.

In order to study the stabilizability properties of a control system we now propose to restrict the space $L\Omega_{m,p}^n$ and consider the space $L\Omega_{m,p}^{n,q}$ of ARMA systems in $L\Omega_{m,p}^n$ that can be stabilized by an $m \times p$ proper ARMA system of $\text{lag} \leq q$ in $L\Pi_{p,m}^q$. Likewise we consider the space $L\Pi_{m,p}^{n,q}$ of proper ARMA systems in $L\Pi_{m,p}^n$ that can be stabilized by a compensator in $L\Pi_{p,m}^q$. The main result of this section is a generalization of Theorem 2.3 in [4] described as follows.

THEOREM 3.7. The space $L\Omega_{m,p}^{n,q}$ is diffeomorphic to $\mathbb{R}^{np(m+p)}$ for all $q = 0, \dots, \infty$.

Remark 3.8. Using a different argument we shall show subsequently that $L\Pi_{m,p}^{n,q}$ is diffeomorphic to $\mathbb{R}^{np(m+p)+mp}$ for all $q = 0, \dots, \infty$. It may be noted however that for $q = \infty$, the spaces $L\Omega_{m,p}^{n,\infty}$, $L\Omega_{m,p}^n$ and $L\Pi_{m,p}^{n,\infty}$, $L\Pi_{m,p}^n$ are identical.

Proof of Theorem 3.7. Consider the map

$$(3.9) \quad \phi : \mathbb{R}^{np(m+p)} \times \mathbb{R} \rightarrow \mathbb{R}^{np(m+p)},$$

$$(3.10) \quad \phi(D_1, \dots, D_n, N_1, \dots, N_n, t) = (e^{-t}D_1, \dots, e^{-nt}D_n, e^{-t}N_1, \dots, e^{-nt}N_n).$$

The map ϕ defines a flow on $\mathbb{R}^{np(m+p)}$. Let X be the corresponding vector field. Clearly X has a unique equilibrium point at the origin of $\mathbb{R}^{np(m+p)}$. Moreover it may be seen that X restricts to a globally asymptotically stable vector field on $L\Omega_{m,p}^{n,q}$. Therefore

by Milnor's Theorem (see Appendix I), it follows that $L\Omega_{m,p}^{n,q}$ is diffeomorphic to $\mathbb{R}^{np(m+p)}$. \square

In view of Remark 3.8, we have the following corollary.

COROLLARY 3.9. *The space $L\Omega_{m,p}^n$ is diffeomorphic to $\mathbb{R}^{np(m+p)}$.*

Remark 3.10. Note that for a parameter identification algorithm defined by a locally and globally convergent vector field, the associated space must be diffeomorphic to a Euclidean space. As a consequence of Theorem 3.7 it may now be possible to generalize many known parameter identification algorithms to $L\Omega_{m,p}^{n,q}$ (see [10]), where the associated space is assumed to be Euclidean. On the basis of a referee's comment, however, we remark that parameter identification algorithms may not be necessarily defined by a locally and globally convergent vector field. Thus the exact correspondence between the existence of convergent on-line identification algorithms and the structure of the associated parameter space is not entirely clear and is a subject of future investigation.

The main result of this section is to obtain a parametrization of the space of strictly proper systems. In particular we show that the parameter space is diffeomorphic to a Euclidean space.

4. A vector bundle of the space of $p \times m$ systems of lag $\leq n$. In this section, we generalize the notion of a strictly proper and a proper ARMA system and consider the space $L_{m,p}^n$ of ARMA systems with m inputs and p outputs having a lag of at most n described as follows:

$$(4.1) \quad L_{m,p}^n \triangleq \{[D_0, \dots, D_n, N_0, \dots, N_n] \in \text{Grass}(p, (n+1)(p+m)) : \\ \text{rank}(D(z), N(z)) = p \text{ for all } z \in \mathbb{D}_u\}$$

where $[D_0, \dots, D_n, N_0, \dots, N_n]$ denotes the subspace spanned by the rows of the matrix (D_0, \dots, N_n) . We will use this notation throughout the paper. Note that left-multiplying (3.1) by a nonsingular $p \times p$ matrix does not change the corresponding input-output relationship. Thus we may view the ARMA system (3.1) as a subset of a Grassmannian.

Of course the space $L_{m,p}^n$ consists of ARMA systems that are not necessarily proper since the matrix D_0 can be singular and in fact can be zero. The improper ARMA systems may be viewed as "infinite objects" that are of interest in the study of system degeneration and high gain compensation (see [12], [17]). Some of the infinite points in $L_{m,p}^n$ are the well-known generalized dynamical systems [18] and may be viewed in $L_{m,p}^n$ as limits of regular systems.

As noted in our earlier paper [4], $L_{m,p}^n$ is an over-parametrization of the space of $p \times m$ ARMA systems of lag $\leq n$. In fact we have the following definition.

DEFINITION 4.1. Two points $[D_0^1, \dots, D_n^1, N_0^1, \dots, N_n^1]$ and $[D_0^2, \dots, D_n^2, N_0^2, \dots, N_n^2]$ in $L_{m,p}^n$ are said to be equivalent if there exist matrices $K_1(z)$ and $K_2(z)$ such that

$$(4.2) \quad K_2(z)D_2(z) = K_1(z)D_1(z),$$

$$(4.3) \quad K_2(z)N_2(z) = K_1(z)N_1(z)$$

and where $\det K_1(z)$ and $\det K_2(z)$ vanish in \mathbb{D}_s and where

$$(4.4) \quad D_i(z) = \sum_{j=0}^n D_j^i z^{n-j}, \quad N_i(z) = \sum_{j=0}^n N_j^i z^{n-j}, \quad i = 1, 2.$$

In view of Definition 4.1 we can define the quotient space $L_{m,p}^n / \sim$ and a quotient topology on such a space. We shall describe these in detail in § 6.

In order to study the properties of $L_{m,p}^n$ we begin with the following preliminary result.

LEMMA 4.2. $L_{m,p}^n$ is an open subset of $\text{Grass}(p, (m+p)(n+1))$.

Proof. Let $\lambda = [D_0, \dots, D_n, N_0, \dots, N_n]$ be a point in $L_{m,p}^n$. Since λ is of full row rank, there exists a $p \times p$ submatrix of λ which can be scaled to identity. Assume without loss of generality that λ can be written as

$$(4.5) \quad \lambda \triangleq [I, D_1, \dots, D_n, N_0, \dots, N_n].$$

The argument will be similar for any other structure of λ . Since $\lambda \in L_{m,p}^n$ it follows that

$$(4.6) \quad \text{rank}[D(z), N(z)] = p$$

for all $z \in \mathbb{D}_u$. Let $\psi_1(z), \dots, \psi_t(z)$ be the set of principal minors of the matrix $[D(z), N(z)]$. It follows that $\psi_1(z), \dots, \psi_t(z)$ do not have a common zero in \mathbb{D}_u . Let Ω be a neighborhood of $D_1, \dots, D_n, N_0, \dots, N_n$ in $\mathbb{R}^{(m+p)n+m}$. For Ω sufficiently small, corresponding to Ω there exists a neighborhood N of λ in $\text{Grass}(p, (m+p) \times (n+1))$ such that $N \in L_{m,p}^n$. \square

By analogous arguments we state and prove the subsequent results of this section for the space $L_{m,p}^{n,q}$ defined as follows.

DEFINITION 4.3. Let $L_{m,p}^{n,q}$ be the space of ARMA systems in $L_{m,p}^n$ that can be stabilized by an $m \times p$ ARMA system in $L_{p,m}^q$. The following trivial result is now stated without proof.

LEMMA 4.4. $L_{m,p}^{n,q}$ is an open subset of $\text{Grass}(p, (m+p)(n+1))$ for all $q = 0, \dots, \infty$.

Note 4.5. $L_{m,p}^{n,\infty}$ is clearly identical to the space $L_{m,p}^n$.

In the subsequent part of this section, we show that $L_{m,p}^{n,q}$ is a vector bundle. As we have argued in § 1, it is this property of $L_{m,p}^{n,q}$ which is of great importance in parameter identification. Thus the parameter identification problem can now be broken up into two distinct parts. The first involves identifying a point on the base; the second involves identifying the corresponding point on the fiber.

We would now detail the structure of $L_{m,p}^{n,q}$ as a vector bundle. Let us remind the reader at this point that if

$$(4.7) \quad [D_0, \dots, D_n, N_0, \dots, N_n] \in L_{m,p}^{n,q}$$

then

$$(4.8) \quad [D_0, N_0] \in \text{Grass}(p, m+p).$$

This is because

$$(4.9) \quad \text{rank}[D_0, N_0] = \text{rank}[D(\infty), N(\infty)] = p.$$

Hence we have a well-defined map

$$(4.10) \quad \psi: L_{m,p}^{n,q} \rightarrow \text{Grass}(p, m+p),$$

$$(4.11) \quad \psi([D_0, \dots, D_n, N_0, \dots, N_n]) = [D_0, N_0].$$

Since for an arbitrary element $[D_0, N_0] \in \text{Grass}(p, m+p)$, the corresponding point $[D_0, 0, \dots, 0, N_0, 0, \dots, 0]$ is an element in $L_{m,p}^n$ and $\psi([D_0, 0, \dots, 0, N_0, \dots, 0]) = [D_0, N_0]$, it follows that ψ is a surjection. Furthermore it is clear that ψ is a smooth mapping. The main result in this section is the following theorem.

THEOREM 4.6.

$$\psi: L_{m,p}^{n,q} \rightarrow \text{Grass}(p, m+p)$$

can be endowed with the structure of a smooth vector bundle for every $q = 0, 1, \dots, \infty$.

The proof of Theorem 4.6 is lengthy and will therefore be broken up into several lemmas.

LEMMA 4.7. *For every $[D_0, N_0] \in \text{Grass}(p, m+p)$, $\psi^{-1}([D_0, N_0])$ is an embedded submanifold of $L_{m,p}^{n,q}$ which is diffeomorphic to $\mathbb{R}^{(m+p)np}$.*

Proof. Since $L_{m,p}^{n,q}$ is an open subset of $\text{Grass}(p, (m+p)(n+1))$ it follows that ψ is a submersion. Hence $\psi^{-1}([D_0, N_0])$ is an embedded submanifold of $L_{m,p}^{n,q}$ of dimension $\mathbb{R}^{(m+p)(np)}$. Let us now consider the local flow

$$(4.12) \quad \textcircled{H} : L_{m,p}^{n,q} \times [0, \infty) \rightarrow L_{m,p}^{n,q}$$

$$(4.13) \quad \textcircled{H}([D(z), N(z)], t) = D(e^t z), N(e^t z).$$

We leave it to the reader to verify that \textcircled{H} is indeed a local flow on $L_{m,p}^{n,q}$, i.e., to verify the smoothness, the semigroup property, and the fact that $\textcircled{H}([D(z), N(z)], t) \in L_{m,p}^{n,q}$ for all $t \in (-\varepsilon, \infty)$ for some $\varepsilon > 0$ which depends on $[D(z), N(z)]$. Let Y denote the vector field on $L_{m,p}^{n,q}$ generated by this local flow. The key observation is that \textcircled{H} preserves the fiber $\psi^{-1}([D_0, N_0])$ for all $[D_0, N_0] \in \text{Grass}(p, m+p)$. Thus the vector field Y is tangential to the fibers of the submersion ψ . Moreover when restricted to the fiber $\psi^{-1}([D_0, N_0])$, Y has a unique equilibrium point $[z^n D_0, z^n N_0]$ which is a local and global attractor. Thus by Milnor's theorem (Appendix I) it follows that each fiber is diffeomorphic to $\mathbb{R}^{(m+p)np}$. \square

Remark 4.8. In order to prove Theorem 4.6 we need to do some more work. In fact we need to show that the diffeomorphisms of Lemma 4.7 can be chosen in such a way that when we endow the fibers $\psi^{-1}([D_0, N_0])$ with vector space structures by using our diffeomorphisms, local triviality of the bundle follows.

Let us now describe the following vector bundle. Denote by Σ the subset of $\text{Grass}(p, (m+p)(n+1))$ defined by

$$(4.14) \quad \Sigma \triangleq \{[D_0, \dots, D_n, N_0, \dots, N_n] : [D_0, N_0] \in \text{Grass}(p, m+p)\}.$$

Let us define

$$(4.15) \quad \Sigma \rightarrow \text{Grass}(p, m+p),$$

$$(4.16) \quad \pi([D(z), N(z)]) = [D_0, N_0].$$

Clearly π is a smooth submersion. Moreover each fiber $\pi^{-1}([D_0, N_0])$ has a natural vector space structure obtained by defining

$$(4.17) \quad \begin{aligned} & [D_0, D_1, \dots, D_n, N_0, N_1, \dots, N_n] + [D_0, \bar{D}_1, \dots, \bar{D}_n, N_0, \bar{N}_1, \dots, \bar{N}_n] \\ &= [D_0, D_1 + \bar{D}_1, \dots, D_n + \bar{D}_n, N_0, N_1 + \bar{N}_1, \dots, N_n + \bar{N}_n], \end{aligned}$$

$$(4.18) \quad \begin{aligned} & \lambda[D_0, D_1, \dots, D_n, N_0, N_1, \dots, N_n] \\ &= [D_0, \lambda D_1, \dots, \lambda D_n, N_0, \lambda N_1, \dots, \lambda N_n]. \end{aligned}$$

Clearly Σ is a vector bundle under the above vector space operations on fibers.

We now proceed to put a smooth Riemannian metric on the vector bundle Σ . In order to do this, when we write an element $[D_0, N_0]$ in $\text{Grass}(p, m+p)$, we shall assume that the representative matrix $[D_0, N_0]$ has orthonormal rows. Hence, if $[D_0, N_0] = [\tilde{D}_0, \tilde{N}_0]$, then there exists an orthogonal matrix g such that

$$(4.19) \quad gD_0 = \tilde{D}_0, \quad gN_0 = \tilde{N}_0.$$

Now if $[D_0, D_1, \dots, D_n, N_0, N_1, \dots, N_n] \in \Sigma$ let us denote the columns of D_i as D_{ij} , $j = 1, \dots, p$ and the columns of N_i as N_{ij} , $j = 1, \dots, m$. Define a Riemannian metric on the vector bundle Σ by specifying

$$(4.20) \quad \langle [D_0, D_1, \dots, D_n, N_0, \dots, N_n], [D_0, \bar{D}_1, \dots, \bar{D}_n, N_0, \bar{N}_1, \dots, \bar{N}_n] \rangle \\ = \sum_{i=1}^n \sum_{j=1}^p D_{ij} \cdot \bar{D}_{ij} + \sum_{j=1}^m N_{ij} \cdot \bar{N}_{ij}$$

where $D_{ij} \cdot \bar{D}_{ij}$ and $N_{ij} \cdot \bar{N}_{ij}$ denote the usual dot product in \mathbb{R}^p . Note that, as a consequence of the restriction on $[D_0, N_0]$ to having orthonormal rows, it follows that the definition of the metric does not depend on the representative element.

Let S_μ denote the sphere bundle [23] over Grass $(p, m+p)$ consisting of vectors in Σ of magnitude equal to $1/\mu$. We now prove the following.

LEMMA 4.9. *For μ sufficiently large, $S_\mu \subseteq L_{m,p}^{n,q}$ and the flow (\mathcal{H}) is transverse to S_μ .*

Proof. Let W be the space of $p \times (m+p)$ matrices with the property that

$$(4.21) \quad Q \in W \quad \text{iff} \quad QQ^T = I.$$

Elementary arguments show that W is a compact subset in $\mathbb{R}^{p \times (m+p)}$. Hence there exists $\varepsilon > 0$ such that if R is a $p \times m+p$ matrix with columns having magnitude $< \varepsilon$ then $R + Q$ has rank p for all $Q \in W$. Thus for μ large enough we see that if

$$[D_0, D_1, \dots, D_n, N_0, N_1, \dots, N_n] \in S_\mu$$

then the matrix

$$(4.22) \quad \left[\frac{1}{z} D_1 + \frac{1}{z^2} D_2 + \dots + \frac{1}{z^n} D_n, \frac{1}{z} N_1 + \frac{1}{z^2} N_2 + \dots + \frac{1}{z^n} N_n \right]$$

has columns whose magnitude are less than ε for all $z \in \mathbb{D}_u$ and for all representative elements. Hence it follows that

$$(4.23) \quad \left[D_0 + \frac{1}{z} D_1 + \dots + \frac{1}{z^n} D_n, N_0 + \frac{1}{z} N_1 + \dots + \frac{1}{z^n} N_n \right]$$

has rank p for all $z \in \mathbb{D}_u$. This shows that for large μ , $S_\mu \subseteq L_{m,p}^n$.

We now show that for a given q , there exists μ large enough such that $S_\mu \subseteq L_{m,p}^{n,q}$. Notice that since $[D_0, N_0]$ is of rank p there exist matrices D_c, N_c such that

$$(4.24) \quad D_0 D_c + N_0 N_c = I_p$$

where D_c is invertible. In other words there exists a compensator $N_c D_c^{-1}$ which stabilizes the plant $D_0^{-1} N_0$. Moreover for μ large enough the plant

$$(4.25) \quad \left[D_0 + \frac{1}{z} D_1 + \dots + \frac{1}{z^n} D_n \right]^{-1} \left[N_0 + \frac{1}{z} N_1 + \dots + \frac{1}{z^n} N_n \right]$$

is stabilizable by the compensator $N_c D_c^{-1}$. It now follows by the compactness of S_μ that for large μ , $S_\mu \subseteq L_{m,p}^{n,q}$. Moreover S_μ is an embedded submanifold of Σ , and hence of $L_{m,p}^{n,q}$ since $L_{m,p}^{n,q} \subseteq_{\text{open}} \Sigma$.

To show that the flow (\mathcal{H}) is transverse to S_μ let us define

$$(4.26) \quad [D(z), N(z)] = [D_0, D_1, \dots, D_n, N_0, N_1, \dots, N_n] \in S_\mu$$

where we still assume that $[D_0, N_0]$ has orthonormal rows. It follows that

$$(4.27) \quad (\mathcal{H})([D(z), N(z), t]) = [D(e^t z), N(e^t z)] \\ = [D_0, e^{-t} D_1, \dots, e^{-nt} D_n, N_0, e^{-t} N_1, \dots, e^{-nt} N_n].$$

Hence

$$\begin{aligned}
 \|D(e'z), N(e'z)\|^2 &< \sum_{j=1}^p \sum_{i=1}^n D_{ij} \cdot D_{ij} + \sum_{j=1}^m \sum_{i=1}^n N_{ij} \cdot N_{ij} \\
 (4.28) \qquad \qquad \qquad &= \|D(s), N(s)\|^2 = \frac{1}{\mu^2}.
 \end{aligned}$$

Thus if $t > 0$ then

$$(4.29) \qquad \qquad \qquad \|\mathcal{H}([D(s), N(s)], t)\| < \frac{1}{\mu},$$

thus proving the assertion. \square

Finally, before we prove Theorem 4.6, we need the following technical result from the literature which we state without proof (see [28, p. 85]).

LEMMA 4.10. *Suppose that M is a paracompact C^∞ manifold and X is a smooth vector field. Then there exist a nowhere zero smooth function f such that fX is a complete vector field.*

We now prove Theorem 4.6.

Proof of Theorem 4.6. Let Y be the vector field on $L_{m,p}^{n,q}$ which has been defined earlier and let f be a strictly positive smooth function on $L_{m,p}^{n,q}$ such that fY is complete. Let μ be large enough such that Y is transversal to S_μ . Clearly fY is transversal to S_μ also. Moreover fY restricts to a vector field on $\psi^{-1}([D_0, N_0])$ for each $[D_0, N_0] \in \text{Grass}(p, m+p)$. Since fY is transverse to S_μ for large μ and since fY is globally attracting to $[D_0, N_0]$, it follows that $[D_0, N_0]$ is the unique globally asymptotically stable equilibrium point on $\psi^{-1}([D_0, N_0])$ of the vector field fY . Let us now define the map

$$(4.30) \qquad \qquad \qquad \Phi: \Sigma \rightarrow L_{m,p}^{n,q},$$

$$\begin{aligned}
 (4.31) \qquad \qquad \qquad \Phi(\xi) &= \xi \quad \text{if } \xi \in \text{Grass}(p, m+p) \\
 &= \phi_{-\log \mu \|\xi\|}^Y \left(\frac{1}{\mu \|\xi\|} \xi \right).
 \end{aligned}$$

It is easy to check that

- (4.32) (1) Φ is a diffeomorphism;
 (2) the diagram

$$(4.33) \qquad \begin{array}{ccc} \Sigma & \xrightarrow{\Phi} & L_{m,p}^{n,q} \\ \pi \downarrow & & \downarrow \psi \\ \text{Grass}(p, m+p) & \xrightarrow{id} & \text{Grass}(p, m+p) \end{array}$$

commutes.

We now use Φ to endow a vector space structure on each fiber of

$$(4.34) \qquad \qquad \qquad \psi: L_{m,p}^{n,q} \rightarrow \text{Grass}(p, m+p).$$

Thus we have $L_{m,p}^{n,q}$ as a vector bundle isomorphic to Σ . \square

Remark 4.11. The vector bundle structure endowed on $L_{m,p}^{n,q}$ is independent of the choice of the function f up to bundle isomorphism.

An important corollary of Theorem 4.6 is described as follows.

COROLLARY 4.12. $L\Pi_{m,p}^{n,q}$ is isomorphic to $\mathbb{R}^{p[(m+p)n+m]}$.

Proof. Consider the subspace B of $\text{Grass}(p, m+p)$ defined as follows:

$$(4.35) \quad B \triangleq \{[D_0, N_0] \in \text{Grass}(p, m+p) : D_0 = I\}.$$

It is clear that B is a chart of $\text{Grass}(p, m+p)$ and is diffeomorphic to \mathbb{R}^{mp} . We now consider the pullback of the vector bundle (4.34) as follows:

$$\begin{array}{ccc} f^*L_{m,p}^{n,q} & & L_{m,p}^{n,q} \\ f^*\psi \downarrow & & \downarrow \psi \\ B & \xrightarrow{f} & \text{Grass}(p, m+p) \end{array}$$

where f is the inclusion of B in $\text{Grass}(p, m+p)$. Of course, since B is contractible to a point, the vector bundle $f^*\psi : f^*L_{m,p}^{n,q} \rightarrow B$ is trivial [23]. Thus $f^*L_{m,p}^{n,q} \cong B \times \mathbb{R}^{(m+p)np}$. Finally from the definition of $L\Pi_{m,p}^{n,q}$ in § 3, it is clear that $f^*L_{m,p}^{n,q}$ and $L\Pi_{m,p}^{n,q}$ are identical. Thus $L\Pi_{m,p}^{n,q} \cong \mathbb{R}^{mp} \times \mathbb{R}^{(m+p)mp}$. \square

We now proceed to describe the structure of the vector bundle $\psi : L_{m,p}^{n,q} \rightarrow \text{Grass}(p, m+p)$. Define a canonical vector bundle $\gamma^p(\mathbb{R}^{m+p})$ over $\text{Grass}(p, m+p)$ as follows. Let

$$(4.36) \quad E = E(\gamma^p(\mathbb{R}^{m+p}))$$

be the set of all pairs

$$(p \text{ plane in } \mathbb{R}^{m+p}, \text{ vector in that } p\text{-plane}).$$

This is to be topologized as a subset of $\text{Grass}(p, m+p) \times \mathbb{R}^{m+p}$. The projection map $\pi_1 : E \rightarrow \text{Grass}(p, m+p)$ is defined by $\pi_1(X, x) = X$ and the vector space structure in the fiber over X is defined by

$$t_1(X, x_1) + t_2(X, x_2) = (X, t_1x_1 + t_2x_2).$$

The main result on the structure of $L_{m,p}^{n,q}$ is described as follows.

THEOREM 4.13. *The vector bundle*

$$(4.37) \quad \psi : L_{m,p}^{n,q} \rightarrow \text{Grass}(p, m+p)$$

is isomorphic to $(m+p)n$ Whitney sums of canonical vector bundles $\gamma^p(\mathbb{R}^{m+p})$.

$$(4.38) \quad L_{m,p}^{n,q} \cong \gamma^p(\mathbb{R}^{m+p}) \oplus \cdots \oplus \gamma^p(\mathbb{R}^{m+p}).$$

Proof. Let us consider Σ as defined in (4.14), together with the map π given by (4.15), (4.16). In view of the proof of Theorem 4.6, it suffices to show that this map

$$(4.39) \quad \pi : \Sigma \rightarrow \text{Grass}(p, m+p)$$

has the structure of a vector bundle which is isomorphic to $n(m+p)$ copies of canonical vector bundle $\gamma^p(\mathbb{R}^{m+p})$.

We now proceed to prove the above claim. Recall from (4.17), (4.18) that π has the structure of a vector bundle.

Let D_{ij} denote the j th column of D_i and let N_{ij} denote the j th column of N_i . Restrict the representative elements $[D_0, D_1, \dots, N_0, N_1, \dots, N_n]$ to have the

property that the rows of $[D_0, N_0]$ are orthonormal. Let us now define the map σ as follows:

$$\begin{array}{ccc}
 \Sigma & \xrightarrow{\sigma} & \gamma^p \oplus \cdots \oplus \gamma^p (n(m+p) \text{ copies}) \\
 \pi \downarrow & & \downarrow \\
 \text{Grass}(p, m+p) & \xleftarrow{id} & \text{Grass}(p, m+p)
 \end{array}$$

(4.40)

$$\begin{aligned}
 \sigma: \Sigma &\rightarrow \gamma^p \oplus \cdots \oplus \gamma^p, \\
 \sigma[D_0 D_1 \cdots D_n N_0 N_1 \cdots N_n] \\
 &= \left(\sum_{j=1}^p D_{0j} (D_0, N_0)^j, \sum_{j=1}^p D_{1j} (D_0, N_0)^j, \cdots, \sum_{j=1}^m D_{nj} (D_0, N_0)^j \right)
 \end{aligned}$$

where $(D_0, N_0)^j$ denotes the j th row of the matrix (D_0, N_0) considered as a vector in \mathbb{R}^{m+p} . We need to check, however, that σ is well defined; i.e., if K is an orthogonal $p \times p$ matrix then

$$\begin{aligned}
 (4.41) \quad &\sigma([KD_0, KD_1, \cdots, KD_n, KN_0, \cdots, KN_n]) \\
 &= \sigma([D_0, D_1, \cdots, D_n, N_0, \cdots, N_n]).
 \end{aligned}$$

The above fact, however, follows at once since

$$(4.42) \quad (KD_0, KN_0)^T K V = (D_0, N_0)^T K^T K V = (D_0, N_0)^T V.$$

The rest of the proof that σ is a diffeomorphism and the diagram commutes is clear and is omitted. \square

When $p = 1$, the vector bundle (4.37) reduces to a vector bundle over a projective space. In this situation, it is possible in many cases to decide whether or not the bundle (4.37) is trivial.

Let us assume $p = 1$ and write

$$(4.43) \quad m+1 = 2^r t \quad \text{and} \quad n = 2^{r_1} t_1$$

where t, t_1 are odd. We now have the following interesting corollaries.

COROLLARY 4.14. $2^{r_1} < t \Rightarrow L_{m,1}^{n,q}$ is nontrivial.

COROLLARY 4.15. $L_{1,1}^{n,q}, L_{3,1}^{n,q}, L_{7,1}^{n,q}$ are trivial vector bundles.

COROLLARY 4.16. $L_{15,1}^{1,q}, L_{31,1}^{1,q}, L_{63,1}^{1,q} \cdots$ are nontrivial vector bundles.

Before we state and prove the corollaries above, we need certain notions from the theory of “characteristic classes,” for which we refer to [7]. We would like to summarize some of the main ideas here as follows.

Let $H^i(B, G)$ denote the i th singular cohomology group of B with coefficients in G . We have the following important result.

FACT 4.17. The group $H^i(\mathbb{R}P^n; \mathbb{Z}/2)$ is cyclic of order two for $0 \leq i \leq n$ and is zero for higher values of i . Furthermore, if a denotes the nonzero element of $H^1(\mathbb{R}P^n; \mathbb{Z}/2)$, then each $H^i(\mathbb{R}P^n; \mathbb{Z}/2)$ is generated by the i -fold cap product a^i .

Thus $H^*(\mathbb{R}P^n; \mathbb{Z}/2)$ can be described as the algebra with unit over $\mathbb{Z}/2$ having one generator a and one relation $a^{n+1} = 0$. Moreover the total Stiefel–Whitney class of the canonical line-bundle $\gamma^1(\mathbb{R}^{m+1})$ over $\mathbb{R}P^m$ is given by

$$(4.44) \quad w(\gamma^1(\mathbb{R}^{m+1})) = 1 + a.$$

FACT 4.18. Let $m+1 = 2^r t$, $n = 2^{r_1} t_1$, where t, t_1 are odd. Then

$$(4.45) \quad 2^{r_1} < t \Leftrightarrow (1+a)^{n(m+1)} \neq 1$$

where $a^{m+1} = 0$.

Proof. Assume $2^{r_1} < t$. Clearly we have

$$(4.46) \quad \begin{aligned} (1+a)^{n(m+1)} &= (1+a^{2^{r_1}})^{n_1} \\ &= 1 + nt_1 a^{2^{r_1}} + (\text{higher-order terms}). \end{aligned}$$

Since $2^{r_1} < m+1$, it follows that

$$(1+a)^{n(m+1)} \neq 1.$$

Conversely, let $2^{r_1} \geq t$. It follows that

$$(1+a)^{n(m+1)} = (1+a^{2^{r_1}})^{n_1} = 1. \quad \square$$

We now proceed to prove the corollaries.

Proof of Corollary 4.14. The Stiefel–Whitney class $w(L_{m,1}^{n,q})$ is given by $(1+a)^{n(m+1)}$, which is not equal to 1 from Fact 4.18. Hence $w(L_{m,1}^{n,q})$ is a nontrivial vector bundle. \square

Proof of Corollary 4.15. Let $m \in \{1, 3, 7\}$. We have

$$(4.47) \quad w(L_{m,1}^{n,q}) = (1+a)^{n(m+1)} = 1.$$

Let $n = 1$; then

$$L_{m,1}^{1,q} \cong \gamma^1(\mathbb{R}^{m+1}) \oplus \cdots \oplus \gamma^1(\mathbb{R}^{m+1}) (m+1 \text{ copies}).$$

By a theorem due to Stiefel [19], $m+1$ Whitney sums of $\gamma^1(\mathbb{R}^{m+1})$ are a trivial vector bundle if $m = 1, 3, 7$. Thus $L_{m,1}^{1,q}$ is a trivial vector bundle. Moreover $L_{m,1}^{n,q}$ is an n -fold Whitney sum of trivial vector bundles and is therefore trivial. \square

Proof of Corollary 4.16. From Theorem 4.13 we may conclude that if $m = 15, 31, 63, \dots$, $L_{m,1}^{1,q}$ is not trivial. This follows from a theorem due to Bott and Milnor [20], Kervaire [21], and Adams [22] that $m+1$ Whitney sums of $\gamma^1(\mathbb{R}^{m+1})$ are nontrivial if $m = 15, 31, 63, \dots$. \square

Example 4.19. Assume $p = 2$, $m = 3$. In this example we show that $L_{3,2}^{n,q}$ is nontrivial if n is not a multiple of 8.

First of all, note that $L_{3,2}^{n,q}$ is isomorphic to $5n$ Whitney sums of canonical vector bundles $\gamma^2(\mathbb{R}^5)$. In fact, it follows from [7, Problem 7-B, p. 87] that $H^*(G_2(\mathbb{R}^5))$ over $\mathbb{Z}/2$ is generated by the Stiefel–Whitney classes w_1, w_2 of $\gamma^2(\mathbb{R}^5)$ and the dual classes $\bar{w}_1, \bar{w}_2, \bar{w}_3$ subject only to the $n+k$ defining relations

$$(4.48) \quad (1+w_1+w_2)(1+\bar{w}_1+\bar{w}_2+\bar{w}_3) = 1.$$

The restriction (4.48) defines the relation

$$(4.49) \quad w_2 w_1^3 = 0$$

and

$$(4.50) \quad w_2^2 + w_1^4 + w_2 w_1^2 = 0.$$

The total Stiefel–Whitney class of $\gamma^2(\mathbb{R}^5)$ is given by $1 + w_1 + w_2$. Hence

$$(4.51) \quad w(L_{3,2}^{n,q}) = (1 + w_1 + w_2)^{5n}.$$

It can be checked easily that the minimum positive integer such that $(1 + w_1 + w_2)^j = 1$ is given by $j = 8$, i.e.,

$$(4.52) \quad (1 + w_1 + w_2)^8 = 1.$$

Thus if n is not a multiple of 8,

$$(4.53) \quad (1 + w_1 + w_2)^{5n} \neq 1$$

so that $L_{3,2}^{n,q}$ is nontrivial if n is not a multiple of 8.

To summarize, in this section we show that the space $L_{m,p}^{n,q}$, of $p \times m$ proper or improper ARMA systems of lag $\leq n$ that can be stabilized by an $m \times p$ ARMA system of lag $\leq q$, has the structure of a vector bundle over Grass $(p, m+p)$ for all $n = 0, 1, \dots, q = 0, 1, \dots$. The vector bundle above is isomorphic to an $n(m+p)$ -fold Whitney sum of canonical vector bundle γ^p over Grass $(p, m+p)$. Finally the vector bundles $L_{m,1}^{n,q}$ are trivial for $m = 1, 3, 7$.

5. A parametrization of stable feedback control systems. In this section we parametrize the space of $p \times m$ ARMA systems $G(z)$ of lag $\leq n$, and $m \times p$ ARMA systems $K(z)$ of lag $\leq q$, such that the closed-loop system $G(z)[I + K(z)G(z)]^{-1}$ is stable. Of course in the above definition $G(z)$ is assumed to represent the plant, $K(z)$ is assumed to represent the compensator, and the region of stability is assumed to be \mathbb{D}_s , the open interior of the unit disc.

Thus we define the space

$$(5.1) \quad LFB_{m,p}^{n,q} \triangleq \{([D_0, \dots, D_n, N_0, \dots, N_n], [L_0, \dots, L_q, M_0, \dots, M_q]) \in L_{m,p}^n \times L_{p,m}^q \mid \text{the pair of plants } D(z)^{-1}N(z) \text{ and } L(z)^{-1}M(z) \text{ is a stable pair}\}.$$

In (5.1) we define $D(z)$, $N(z)$ as in (3.3) and define

$$\begin{aligned} L(z) &= L_0 z^q + L_1 z^{q-1} + \dots + L_q, \\ M(z) &= M_0 z^q + M_1 z^{q-1} + \dots + M_q. \end{aligned}$$

In order to justify the need to consider the space $LFB_{m,p}^{n,q}$ let us consider the following. In system design problems wherein we are designing a compensator for an unknown plant, frequently we assume an initial value of the plant and its stabilizing compensator. Subsequently both the plant and the compensator parameters are updated and the design objective is to ensure that the plant parameters converge to its true value. In this technique of identifying the parameters of a closed loop system, we define a recursive algorithm on the space $LFB_{m,p}^{n,q}$. Likewise, in adaptive control problems we are frequently interested in updating the parameters of an unknown plant and compensator in real time, with the added hypothesis that if the adaptive algorithm is switched off, the corresponding plant-compensator pair is stable. Thus we are interested in adaptively updating points on $LFB_{m,p}^{n,q}$ in real time.

The main result we now show in this section is described as follows.

MAIN THEOREM 5.1. $LFB_{m,p}^{n,q}$ has the structure of a vector bundle over the base space Grass $(p, m+p)$.

The ideas behind the proof of the main theorem can be most easily explained by considering the following preliminary lemmas. First, however, we must set up the following notation.

Define W to be

$$(5.2) \quad W \triangleq \{(A, B) \in \text{Grass}(p, m+p) \times \text{Grass}(m, m+p) \mid A \oplus B = \mathbb{R}^{m+p}\}.$$

Let ψ denote the projection

$$(5.3) \quad \psi_{m,p}^n : L_{m,p}^n \rightarrow \text{Grass}(p, m+p)$$

as in (4.34). We now state and prove the following lemmas.

LEMMA 5.2. The map

$$(5.4) \quad \phi : LFB_{m,p}^{n,q} \rightarrow \text{Grass}(p, m+p) \times \text{Grass}(m, m+p)$$

defined by

$$(5.5) \quad \phi([D(z), N(z)], [L(z), M(z)]) = (\psi_{m,p}^n[D(z), N(z)], \psi_{p,m}^q[L(z), M(z)])$$

has W as its image and ϕ is a submersion onto W .

Proof. Suppose that $([D(z), N(z)], [L(z), M(z)])$ is in $LFB_{m,p}^{n,q}$. It now follows that

$$(5.6) \quad \begin{bmatrix} z^{-n}D(z) & z^{-n}N(z) \\ z^{-q}M(z) & z^{-q}L(z) \end{bmatrix}$$

has full rank for all $z \in \mathbb{D}_u$. In particular, since $\infty \in \mathbb{D}_u$ it follows that the matrix

$$(5.7) \quad \begin{bmatrix} D_0 & N_0 \\ M_0 & L_0 \end{bmatrix}$$

has full rank. This shows that $\text{im}(\phi) \subseteq W$. Since for each pair $([D_0 \ N_0], [L_0 \ M_0]) \in W$, the corresponding closed loop system is stable, it now follows that $\text{im}(\phi) = W$. Finally since $LFB_{m,p}^{n,q}$ is an open subset of $L_{m,p}^n \times L_{p,m}^q$ and since

$$(5.8) \quad \psi_{m,p}^n \times \psi_{p,m}^q : L_{m,p}^n \times L_{p,m}^q \rightarrow \text{Grass}(p, m+p) \times \text{Grass}(m, m+p)$$

is a submersion, it follows that (5.4) is also a submersion. \square

LEMMA 5.3. *The map (5.4) has the structure of a fiber bundle with fibers diffeomorphic to $\mathbb{R}^{(m+p)(np+mq)}$.*

Proof. Consider the local flow

$$(5.9) \quad \mathcal{H}_{m,p}^{n,q} : LFB_{m,p}^{n,q} \times [0, \infty) \rightarrow LFB_{m,p}^{n,q}$$

defined by

$$(5.10) \quad \mathcal{H}_{m,p}^{n,q}([D(z), N(z)], [L(z), M(z)]; t)([D(e^t z), N(e^t z)], [L(e^t z), M(e^t z)]).$$

It is easy to check that the above local flow is well defined and smooth. Clearly, for each $(A, B) \in W$, the fiber $\phi^{-1}(A, B)$ of the submersion (5.4) at (A, B) is invariant under the flow $\mathcal{H}_{m,p}^{n,q}$. Moreover, restricted to $\phi^{-1}(A, B)$, this flow has exactly one equilibrium point which is globally and locally attracting. Since ϕ is a submersion, it follows that $\phi^{-1}(A, B)$ is a smooth Hausdorff manifold of dimension $(m+p)(mq+np)$. Thus using Milnor's theorem [5] we conclude that $\phi^{-1}(A, B)$ is diffeomorphic to $\mathbb{R}^{(m+p)(np+mq)}$.

The proof of local triviality of $LFB_{m,p}^{n,q}$ is similar to the proof sketched in the Theorem 4.6. Let $X_{m,p}^{n,q}$ be the vector field generated by the local flow $\mathcal{H}_{m,p}^{n,q}$ and let g be a nowhere zero C^∞ function on $LFB_{m,p}^{n,q}$ such that the vector field $Y \triangleq gX_{m,p}^{n,q}$ is complete. Let us define $\Sigma_{m,p}^n$ as the space defined by (4.14) and consider the vector bundle

$$(5.11) \quad \psi_{m,p}^n : \Sigma_{m,p}^n \rightarrow \text{Grass}(p, m+p)$$

defined via equations (4.15), (4.16). Endow a Riemannian metric on this vector space via (4.20). Similarly consider the vector bundle

$$(5.12) \quad \psi_{p,m}^q : \Sigma_{p,m}^q \rightarrow \text{Grass}(m, m+p)$$

and let

$$(5.13) \quad \psi : \Sigma \rightarrow \text{Grass}(p, m+p) \times \text{Grass}(m, m+p)$$

denote the Cartesian product of the two vector bundles (5.11) and (5.12). Endow the vector bundle Σ in (5.13) with the Cartesian product of the metrics on $\Sigma_{m,p}^n$ and $\Sigma_{p,m}^q$. Let S_μ denote the sphere bundle over $\text{Grass}(p, m+p) \times \text{Grass}(m, m+p)$ consisting of vectors of Σ of magnitude μ . If U is an open subset of $\text{Grass}(p, m+p) \times \text{Grass}(m, m+p)$, we will denote the restrictions of Σ and S_μ to U by $\Sigma|U$ and $S_\mu|U$, respectively.

Now let (A, B) be in W , and let U be a relatively compact open neighborhood of (A, B) in W . It follows easily that for μ sufficiently small, $S_\mu|U \subseteq LFB_{m,p}^{n,q}$. Moreover the vector field $X_{m,p}^{n,q}$ (and hence the vector field Y also) is transverse to $S_\mu|U$. Now by an argument analogous to that used in the proof of Theorem 4.6, we conclude that $LFB_{m,p}^{n,q}|U$ is a trivial fiber bundle with fibers diffeomorphic to $\mathbb{R}^{(m+p)(np+mq)}$. This proves that local triviality of

$$(5.14) \quad \phi : LFB_{m,p}^{n,q} \rightarrow W.$$

Hence $LFB_{m,p}^{n,q}$ is a fiber bundle over W with fibers diffeomorphic to $\mathbb{R}^{(m+p)(np+mq)}$. \square

The above argument is not powerful enough to show that $LFB_{m,p}^{n,q}$ is a vector bundle. Before proceeding to prove this stronger result, we need to set up some notation.

If $A \in \text{Grass}(p, m+p)$, we will denote the orthogonal projection of A in \mathbb{R}^{m+p} by A^\perp . Let

$$(5.15) \quad BW \triangleq \{(A, B) \in W \mid A \in \text{Grass}(p, m+p), B = A^\perp\}.$$

Let

$$(5.16) \quad i : BW \rightarrow W$$

be the inclusion map. We now consider the pullback of the fiber bundle (5.14) via the diagram

$$(5.17) \quad \begin{array}{ccc} i^*LFB_{m,p}^{n,q} & & LFB_{m,p}^{n,q} \\ i^*\phi \downarrow & & \downarrow \phi \\ BW & \xrightarrow{i} & W \end{array}$$

to obtain the locally trivial fiber bundle

$$(5.18) \quad i^*\phi : i^*LFB_{m,p}^{n,q} \rightarrow BW.$$

We now have the following two lemmas.

LEMMA 5.4. *$i^*LFB_{m,p}^{n,q}$ can be endowed with the structure of a vector bundle.*

Proof. Let Σ and S_μ be as in the proof of Lemma 5.3. Since BW is a compact subset of W , it follows that there exists a μ small enough such that $S_\mu|BW$ is contained in $i^*LFB_{m,p}^{n,q}$. Now by constructing a transverse and complete flow and using the argument in the proof of Theorem 4.6, we conclude that $i^*LFB_{m,p}^{n,q}$ can be endowed with the structure of a vector bundle which is isomorphic to $i^*\Sigma$. \square

LEMMA 5.5. *Let us consider the map*

$$(5.19) \quad \rho : W \rightarrow BW$$

defined by

$$(5.20) \quad \rho(A, B) = (A, A^\perp);$$

then W can be endowed with the structure of a vector bundle over BW .

Proof. It is clear that $\rho : W \rightarrow BW$ is a submersion. Identify BW with $\text{Grass}(p, m+p)$ via the mapping

$$(5.21) \quad (A, A^\perp) \rightarrow A.$$

Let us put coordinates on BW by using standard coordinate charts on $\text{Grass}(p, m+p)$, i.e., let i_1, \dots, i_p be integers such that $1 \leq i_1 < i_2 < \dots < i_p \leq m+p$. If V is in $\text{Grass}(p, m+p)$, denote by $V^{(i_1, \dots, i_p)}$ the subspace in \mathbb{R}^p obtained by projecting V along the i_1, \dots, i_p coordinate directions. Let

$$(5.22) \quad U^{(i_1, \dots, i_p)} \triangleq \{V \in \text{Grass}(p, m+p) \mid V^{(i_1, \dots, i_p)} = \mathbb{R}^p\}.$$

For each V in $U^{(i_1, \dots, i_p)}$ find the unique basis of V consisting of vectors $\{v_1, \dots, v_p\}$ with the property that

$$(5.23) \quad V_j = e_{i_j} + \sum_{k \notin \{i_1, \dots, i_p\}} x_{jk} e_k$$

where $\{e_1, \dots, e_{m+p}\}$ is the standard basis of \mathbb{R}^{m+p} . Now let us define a map

$$(5.24) \quad \phi^{(i_1, \dots, i_p)}: U^{(i_1, \dots, i_p)} \rightarrow \mathbb{R}^{mp}$$

by

$$(5.25) \quad V \mapsto (x_{jk}; 1 \leq j \leq p; k \notin \{i_1, \dots, i_p\}).$$

We can easily see that $(u^{(i_1, \dots, i_p)}, \phi^{(i_1, \dots, i_p)})$ is a coordinate chart of $\text{Grass}(p, m+p)$. Such charts will be referred to as standard charts.

Let us consider a standard chart $(U^{(i_1, \dots, i_p)}, \phi^{(i_1, \dots, i_p)})$. Let $(A, B) \in W$ be such that $A \in U^{(i_1, \dots, i_p)}$. With respect to this chart we can identify B with a unique element of \mathbb{R}^{mp} in the following way.

Let (v_1, \dots, v_p) be the unique basis associated with A as described above. Let j_1, \dots, j_m be integers such that $1 \leq j_1 < j_2 < \dots < j_m \leq m+p$ and $\{i_1, \dots, i_p, j_1, \dots, j_m\} = \{1, 2, \dots, m+p\}$. Now begin with the basis $\{v_1, v_2, \dots, v_p, e_{j_1}, \dots, e_{j_m}\}$ and use the Gram-Schmidt procedure to define an orthonormal basis $\{u_1, \dots, u_{m+p}\}$. Now B can be written uniquely as $\text{span}\{z_1, \dots, z_m\}$, where

$$(5.26) \quad z_i = u_{p+i} + \sum_{k=1}^p y_{ik} u_k.$$

Now identify B with the matrix $[y_{ik}]$. Let

$$(5.27) \quad \phi: \rho^{-1}(u^{(i_1, \dots, i_p)}) \rightarrow u^{(i_1, \dots, i_p)} \times \mathbb{R}^{mp}$$

be the map

$$(5.28) \quad (A, B) \mapsto (A, [y_{ik}])$$

where $[y_{ik}]$ is constructed as above. It is easily seen that (5.27) is a local trivialization of the map $\rho: W \rightarrow BW$ defined in (5.19), (5.20).

Finally, if A belongs to two standard coordinate neighborhoods, $U^{(i_1, \dots, i_p)}$ and $U^{(\hat{i}_1, \dots, \hat{i}_p)}$, and if ϕ and $\hat{\phi}$ are the local trivializations constructed as in (5.27), then the map

$$(5.29) \quad \phi \cdot \hat{\phi}^{-1}|_{A \times \mathbb{R}^{m+p}}: \mathbb{R}^{mp} \rightarrow \mathbb{R}^{mp}$$

is easily seen to be a linear map.

Thus $\rho: W \rightarrow BW$ has the structure of a vector bundle. \square

We now prove a theorem which may be considered as a generalization of Lemma 5.2.

THEOREM 5.6. *The map ϕ defined in (5.4) has the structure of a vector bundle.*

Proof. Refer to the following diagram:

$$(5.30) \quad \begin{array}{ccc} LFB_{m,p}^{n,q} & & i^* LFB_{m,p}^{n,q} \\ \phi \downarrow & & \downarrow i^* \phi \\ W & \xrightleftharpoons[\rho]{i} & BW \end{array}$$

By Lemma 5.3 we know that $\phi: LFB_{m,p}^{n,q} \rightarrow W$ is a fiber bundle with fibers diffeomorphic to $\mathbb{R}^{(m+p)(np+qm)}$. By Lemma 5.5 we know that $\rho: W \rightarrow BW$ is a homotopy equivalence.

Finally by Lemma 5.4 we know that $i^* \phi : i^* LFB_{m,p}^{n,q} \rightarrow BW$ is a vector bundle. It therefore follows from the homotopy property of vector bundles, that $\phi : LFB_{m,p}^{n,q} \rightarrow W$ is a vector bundle (see [23, p. 57]). \square

Finally we sketch the proof of Main Theorem 5.1.

Proof of Main Theorem 5.1. From the Lemma 5.5 and Theorem 5.6 it follows that $LFB_{m,p}^{n,q}$ has the structure of a vector bundle over W and W has the structure of a vector bundle over BW . It therefore follows from [24] that $LFB_{m,p}^{n,q}$ is a vector bundle over BW . Finally, since BW can be identified with $\text{Grass}(p, m+p)$, it follows that

$$(5.31) \quad \rho \cdot \phi : LFB_{m,p}^{n,q} \rightarrow \text{Grass}(p, m+p)$$

has the structure of a vector bundle. \square

The following corollary is immediate from the proof of the Main Theorem 5.1.

COROLLARY 5.7. *The vector bundle $LFB_{m,p}^{n,q}$ over BW is a Whitney sum of the two vector bundles (5.18) and (5.19).*

The proof of Corollary 5.7 follows immediately from [24] and is omitted.

In general, the structure of the vector bundle (5.31) is not known. The following theorem summarizes a partial result in this direction.

THEOREM 5.8. *Assume that $\min(m, p) = 1$. The vector bundle (5.31) is trivial if and only if $\max(m, p) = 1, 3$, or 7 .*

Before we sketch the proof of Theorem 5.8 we state and prove the following lemma.

LEMMA 5.9. *Assume that $\min(m, p) = 1$. The vector bundle $\rho : W \rightarrow BW$ is isomorphic to the tangent bundle of the projective space $\mathbb{RP}^{\max(m,p)}$.*

Proof. Assume that $p = 1$ without loss of generality. The vector bundle $\rho : W \rightarrow BW$ is therefore defined on the base space \mathbb{RP}^m . Let τ be the tangent bundle over \mathbb{RP}^m . We now consider the following diagram:

$$\begin{array}{ccc} D\mathbb{RP}^m & & W \\ \tau \downarrow & & \downarrow \rho \\ \mathbb{RP}^m & \xleftarrow{id} & \mathbb{RP}^m \end{array}$$

It is well known (see [7, p. 44]) that the tangent manifold $D\mathbb{RP}^m$ can be identified with the set of all pairs

$$(5.32) \quad \{(x, v), (-x, -v) : x, v \in \mathbb{R}^{m+1}, x \cdot x = 1, x \cdot v = 0\}.$$

Similarly the space W can be identified with the set of all pairs

$$(5.33) \quad \{(x, v), (-x, -v) : x, v \in \mathbb{R}^{m+1}, x \cdot x = 1, x \cdot v > 0\}.$$

The spaces W or $D\mathbb{RP}^m$ are endowed with the vector space structure in the following way. Let $[x, v]$ denote the pair $(x, v), (-x, -v)$; then

$$(5.34) \quad \alpha[x, v_1] + \beta[x, v_2] = [x, \alpha v_1 + \beta v_2].$$

We therefore consider an isomorphism between the spaces $D\mathbb{RP}^m$ and W via

$$(5.35) \quad (x, v) \mapsto (x, x + v).$$

This completes the proof. \square

We shall now prove Theorem 5.8.

Proof of Theorem 5.8. Consider the vector bundle (5.18). In Lemma 5.4 we have shown that this vector bundle is isomorphic to $i^* \Sigma$, where Σ is described in (5.13) and

the inclusion map i is given by (5.16). Let us now include W in $\text{Grass}(p, m+p) \times \text{Grass}(m, m+p)$ via the inclusion map α . We now have the following diagram:

$$\begin{array}{ccccc}
 i^*LFB_{m,p}^{n,q} & & LFB_{m,p}^{n,q} & & L_{m,p}^n \times L_{p,m}^q \\
 i^*\phi \downarrow & & \phi \downarrow & & \downarrow \psi_{m,p}^n \times \psi_{p,m}^q \\
 BW & \xrightarrow{i} & W & \xrightarrow{\alpha} & \text{Grass}(p, m+p) \times \text{Grass}(m, m+p)
 \end{array}$$

where the vector bundle

$$(5.36) \quad i^*LFB_{m,p}^{n,q} \rightarrow BW$$

is a pullback of the vector bundle

$$(5.37) \quad L_{m,p}^n \times L_{p,m}^q \rightarrow \text{Grass}(p, m+p) \times \text{Grass}(m, m+p).$$

From Corollaries 4.14–4.16 we know that the vector bundle (5.36) is trivial if $\min(m, p) = 1, 3$, or 7 . It therefore follows that the vector bundle (5.18) is trivial if $\min(m, p) = 1, 3$, or 7 . Finally from Theorem 5.8 it follows that the vector bundle (5.19) is trivial if and only if $\min(m, p) = 1, 3$, or 7 . Thus in view of Corollary 5.7 we have the proof. \square

6. A fiber bundle of the space of proper systems with unbounded lag: quotient topology. The various spaces that we have considered so far in this paper assume that the ARMA systems are of bounded lag. This assumption enabled us to parametrize spaces that are finite-dimensional vector bundles. Frequently, however, it is unreasonable to assume under the presence of high frequency parasitics, that the family of systems under consideration is of bounded lag. In this section we now parametrize the space $L_{m,p}^\infty$ of all $p \times m$ ARMA systems of arbitrary large lag.

Let \mathbb{R}^∞ denote the vector space consisting of those infinite sequences

$$(6.1) \quad x = (x_1, x_2, x_3, \dots)$$

of real numbers for which all but a finite number of x_i are zero. For fixed k , the subspace consisting of all

$$(6.2) \quad x = (x_1, x_2, \dots, x_k, 0, 0, 0, \dots)$$

will be identified with the coordinate space \mathbb{R}^k . Thus $\mathbb{R}^1 \subset \mathbb{R}^2 \subset \dots$ with union \mathbb{R}^∞ .

DEFINITION 6.1. The infinite Grassmann manifold

$$(6.3) \quad G_n = \text{Grass}(n, \infty)$$

is the set of all n -dimensional linear subspaces of \mathbb{R}^∞ , topologized as the direct limit of the sequence

$$(6.4) \quad \text{Grass}(n, n) \subset \text{Grass}(n, n+1) \subset \dots$$

In other words, a subset of G_n is open if and only if its intersection with $\text{Grass}(n, n+k)$ is open as a subset of $\text{Grass}(n, n+k)$ for each $k \geq n$. The following assertion is rather trivial to check and its proof is omitted.

PROPOSITION 6.2. *If $\text{Grass}(p, (m+p)(n+1))$ is included in $\text{Grass}(p, (m+p) \times (n+2))$ as follows:*

(6.5)

$$[D_0, D_1, \dots, D_n, N_0, N_1, \dots, N_n] \mapsto [D_0, D_1, \dots, D_n, 0, N_0, N_1, \dots, N_n, 0],$$

we have

$$(6.6) \quad \begin{array}{ccccc} \text{Grass}(p, (m+p)(n+1)) & \subset & \text{Grass}(p, (m+p)(n+2)) & \subset & \dots & \subset & \text{Grass}(p, \infty) \\ \cup & & \cup & & & & \cup \\ L_{m,p}^n & \subset & L_{m,p}^{n+1} & \subset & \dots & \subset & L_{m,p}^\infty \end{array}$$

where $L_{m,p}^\infty$ is defined to be the direct limit of the sequence

$$(6.7) \quad L_{m,p}^n \subset L_{m,p}^{n+1} \subset \dots$$

One of the results that we would now like to state and prove in this section is about the structure of $L_{m,p}^\infty$.

THEOREM 6.3. *$L_{m,p}^\infty$ can be endowed with the structure of a fiber bundle over the base $\text{Grass}(p, m+p)$ and with fibers homeomorphic to \mathbb{R}^∞ .*

Proof. We know from the Theorem 4.6 that $L_{m,p}^n$ can be endowed with the structure of a vector bundle over $\text{Grass}(p, m+p)$ with fibers isomorphic to $\mathbb{R}^{np[m+p]}$. Consider a trivializing chart U of $\text{Grass}(p, m+p)$. Let ψ_n be the bundle map

$$(6.8) \quad \psi_n: L_{m,p}^n \rightarrow \text{Grass}(p, m+p).$$

It follows that $\psi_n^{-1}(U)$ is isomorphic to $U \times \mathbb{R}^{np[m+p]}$. It is easy to check the validity of the following diagram:

$$(6.9) \quad \begin{array}{ccccccc} L_{m,p}^n & \subset & L_{m,p}^{n+1} & \subset & \dots & \subset & L_{m,p}^\infty \\ \cup & & \cup & & & & \cup \\ \psi_n^{-1}(U) & \subset & \psi_{n+1}^{-1}(U) & \subset & \dots & \subset & \psi_\infty^{-1}(U) \end{array}$$

where $\psi_n^{-1}(U)$ has been included in $\psi_{n+1}^{-1}(U)$ by including $\mathbb{R}^{np(m+p)}$ within $\mathbb{R}^{(n+1)(m+p)}$ via the construction of \mathbb{R}^∞ . It follows that the space $\psi_\infty^{-1}(U)$ obtained by considering the direct limit of $\psi_\infty^{-1}(U) \subset \psi_{n+1}^{-1}(U) \subset \dots$ is homeomorphic to $U \times \mathbb{R}^\infty$.

By considering various trivializing charts U_α which cover the base $\text{Grass}(p, m+p)$, it follows that $L_{m,p}^\infty$ can be endowed with the structure of a fiber bundle over $\text{Grass}(p, m+p)$. \square

We now define an equivalence relation \sim on $L_{m,p}^\infty$ as follows: Let $\omega_1, \omega_2 \in L_{m,p}^\infty$. Then $\omega_1 \sim \omega_2$ if and only if both ω_1 and ω_2 correspond to equivalent ARMA systems in the sense of Definition 4.1. Consider the quotient space $L_{m,p}^\infty / \sim \triangleq \tilde{L}_{m,p}^\infty$ together with the natural map

$$(6.10) \quad \psi: L_{m,p}^\infty \rightarrow \tilde{L}_{m,p}^\infty$$

which sends a point in $L_{m,p}^\infty$ to its equivalence class. We now endow $\tilde{L}_{m,p}^\infty$ with the quotient topology.

The space $\tilde{L}_{m,p}^\infty$ so constructed is a hybrid topology on the space of all ARMA systems. We now compare the space $\tilde{L}_{m,p}^\infty$ with the well-known graph topology [8] on the space of all systems. Let $P_{m,p}$ be the space of all multi-input multi-output ARMA systems endowed with the graph topology. We now consider the map

$$(6.11) \quad \phi_\infty: L_{m,p}^\infty \rightarrow P_{m,p}$$

which sends a point in $L_{m,p}^\infty$ to the corresponding ARMA system in $P_{m,p}$. Finally we consider the map ψ_∞ which makes the diagram

$$(6.12) \quad \begin{array}{ccc} L_{m,p}^\infty & \xrightarrow{\phi_\infty} & P_{m,p} \\ \psi \searrow & & \swarrow \psi_\infty \\ & \tilde{L}_{m,p}^\infty & \end{array}$$

commute. We now state and prove the following interesting results.

THEOREM 6.4. ϕ_∞ is a continuous function.

Proof. Clearly it is enough to show that

$$(6.13) \quad \phi_\infty: L_{m,p}^n \rightarrow P_{m,p}$$

is continuous for all n . Let \tilde{p} be a point in $P_{m,p}$ with the coprime factorization given by $\tilde{p}_1^{-1}\tilde{p}_2$, where \tilde{p}_1, \tilde{p}_2 are matrices with elements in H , the ring of stable and proper rational functions. Define an open neighborhood $N_{\varepsilon_1, \varepsilon_2}(\tilde{p}_1, \tilde{p}_2)$ of \tilde{p} in $P_{m,p}$ as follows:

$$(6.14) \quad N_{\varepsilon_1, \varepsilon_2}(\tilde{p}_1, \tilde{p}_2) = \{p \in P_{m,p}: p = p_1^{-1}p_2, p_1 \text{ and } p_2 \text{ are coprime,} \\ \|p_1 - \tilde{p}_1\|_\infty < \varepsilon_1, \|p_2 - \tilde{p}_2\|_\infty < \varepsilon_2\}.$$

The norm in (6.14) denotes the H^∞ norm.

Let p_0 be a point in $L_{m,p}^n$ such that

$$(6.15) \quad \phi_\infty(p_0) = \tilde{p}.$$

We would now show that there exists an open neighborhood $N(p_0)$ of p_0 in $L_{m,p}^n$ such that

$$(6.16) \quad \phi_\infty(N(p_0)) \subset N_{\varepsilon_1, \varepsilon_2}(\tilde{p}_1, \tilde{p}_2).$$

Let

$$(6.17) \quad p_0 = [\bar{D}_0 \bar{D}_1 \cdots \bar{D}_n \bar{N}_0 \cdots \bar{N}_n]$$

so that \tilde{p} may be written as

$$(6.18) \quad [\bar{D}_0 z^n + \bar{D}_1 z^{n-1} + \cdots + \bar{D}_n]^{-1} [\bar{N}_0 z^n + \bar{N}_1 z^{n-1} + \cdots + \bar{N}_n].$$

The coprime representation $\tilde{p}_2^{-1}\tilde{p}_1$ of \tilde{p} may be written as

$$(6.19) \quad \tilde{p}_1 = \Delta(z)[(\delta_0 z^n + \cdots + \delta_n)I]^{-1}[\bar{D}_0 z^n + \cdots + \bar{D}_n],$$

$$(6.20) \quad \tilde{p}_2 = \Delta(z)[(\delta_0 z^n + \cdots + \delta_n)I]^{-1}[\bar{N}_0 z^n + \cdots + \bar{N}_n]$$

where $\Delta(z) \in H^{p \times p}$ $\det \Delta(z) \in J$ and $\delta_0 z^n + \cdots + \delta_n$ is a Hurwitz polynomial of degree n .

Since $p_0 \in \text{Grass}(p, (n+1)(m+p))$, without loss of generality assume that \bar{D}_0 is nonsingular. In particular assume that $\bar{D}_0 = I$. For μ sufficiently small, we define $N_\mu(p_0)$ of p_0 in $L_{m,p}^n$ as follows:

$$(6.21) \quad N_\mu(p_0) = \{I, D_1, \dots, D_n, N_0, \dots, N_n\}: \sum (\bar{d}_{ijk} - d_{ijk})^2 + \sum (\bar{n}_{ijk} - n_{ijk})^2 < \mu\}$$

where \bar{d}_{ijk} is the ij th entry of \bar{D}_k and d_{ijk} is the ij th entry of D_k . We define \bar{n}_{ijk} and n_{ijk} similarly.

For any $p \in N_\mu(p_0)$, a coprime factorization of $\phi_\infty(p_0)$ is given by $p_2^{-1}p_1$, where

$$(6.22) \quad p_1 = \Delta(z)[(\delta_0 z^n + \cdots + \delta_n)I]^{-1}[Iz^n + D_1 z^{n-1} + \cdots + D_n],$$

$$(6.23) \quad p_2 = \Delta(z)[(\delta_0 z^n + \cdots + \delta_n)I]^{-1}[N_0 z^n + N_1 z^{n-1} + \cdots + N_n].$$

We now compute

$$(6.24) \quad \|p_1 - \tilde{p}_1\|_\infty \leq \|\Delta(z)\|_\infty \|[(\delta_0 z^n + \cdots + \delta_n)I]^{-1}[(D_1 - \bar{D}_1)z^{n-1} + \cdots + (D_n - \bar{D}_n)]\|_\infty$$

and

$$(6.25) \quad \|p_2 - \tilde{p}_2\|_\infty \leq \|\Delta(z)\|_\infty \|[(\delta_0 z^n + \cdots + \delta_n)I]^{-1}[(N_0 - \bar{N}_0)z^n + \cdots + (N_n - \bar{N}_n)]\|_\infty.$$

For μ sufficiently small, it is clear that

$$(6.26) \quad \|p_1 - \tilde{p}_1\|_\infty < \varepsilon_1 \quad \text{and} \quad \|p_2 - \tilde{p}_2\|_\infty < \varepsilon_2$$

implying that

$$(6.27) \quad \phi_\infty(N_\mu(p_0)) \subset N_{\varepsilon_1, \varepsilon_2}(\tilde{p}_1, \tilde{p}_2). \quad \square$$

Our next result is the following.

THEOREM 6.5. *Let $\tilde{\omega}_i, i = 1, 2, \dots$ be a sequence of points in $\tilde{L}_{m,p}^\infty$. Then the following two conditions are equivalent:*

- (1) *The sequence $\{\tilde{\omega}_i, i = 1, 2, \dots$ be a sequence of points in $\tilde{L}_{m,p}^\infty$. Then*
- (2) (a) *The sequence $\{\psi_\infty(\tilde{\omega}_i), i = 1, 2, \dots\}$ converges in $P_{m,p}$.*
 (b) *There exists n sufficiently large such that the lag of $\psi_\infty(\tilde{\omega}_i) \leq n$ for all $i = 1, 2, \dots$.*

Before we prove Theorem 6.5, we state the following interesting corollary.

COROLLARY 6.6. ψ_∞^{-1} is not a continuous function.

The proof of Corollary 6.6 is trivial and relies on the existence of a sequence $\delta_i, i = 1, 2, \dots$ of plants which converges in $P_{m,p}$ and is such that $\text{lag}(\delta_i) < \text{lag}(\delta_{i+1})$ for $i = 1, 2, \dots$.

Remark 6.7. The graph topology and the hybrid topology introduced in this paper are not identical. However, if we define $P_{m,p}^n$ to be the set of all $p \times m$ systems in $P_{m,p}$ of $\text{lag} \leq n$, then it follows from Theorem 6.5 that $\tilde{L}_{m,p}^n$ is homeomorphic to $P_{m,p}^n$, where $\tilde{L}_{m,p}^n = \psi(L_{m,p}^n)$. In general a set S is open in $\tilde{L}_{m,p}^\infty$ if $S \cap \tilde{L}_{m,p}^n$ is open in $\tilde{L}_{m,p}^n$ for all n . This property is unfortunately not true for the graph topology.

Proof of Theorem 6.5. ($2 \Rightarrow 1$) Let $\tilde{p}_1, \tilde{p}_2, \dots$ be a sequence of plants in $P_{m,p}$ which converges to \tilde{p}_0 in $P_{m,p}$. Let n be such that $\deg \tilde{p}_i \leq n$ for $i = 1, 2, \dots$. It follows that $\deg(\tilde{p}_0) \leq n$; otherwise there exists an open neighborhood U of \tilde{p}_0 in $P_{m,p}$ which does not contain any \tilde{p}_i . Let $\tilde{\omega}_i, i = 1, 2, \dots$ be the sequence in $\tilde{L}_{m,p}^\infty$ such that $\psi_\infty(\tilde{\omega}_i) = \tilde{p}_i, i = 1, 2, \dots$ and let $\tilde{\omega}_0$ be the point in $\tilde{L}_{m,p}^\infty$ such that $\psi_\infty(\tilde{\omega}_0) = \tilde{p}_0$. Let us write $\tilde{\omega}_i$ to be the equivalence class of

$$(6.28) \quad [D_{i0}, D_{i1}, \dots, D_{in}, N_{i0}, N_{i1}, \dots, N_{in}]$$

so that \tilde{p}_i may be written as

$$(6.29) \quad [D_{i0}z^n + \cdots + D_{in}]^{-1}[N_{i0}z^n + \cdots + N_{in}].$$

A coprime factorization of \tilde{p}_i may be written as $\tilde{p}_{i1}/\tilde{p}_{i2}$, where

$$(6.30) \quad \tilde{p}_{i1} = [D_{i0}z^n + \cdots + D_{in}]/z^n,$$

$$(6.31) \quad \tilde{p}_{i2} = [N_{i0}z^n + \cdots + N_{in}]/z^n.$$

By assumption, there exists a coprime factorization $\tilde{p}_{01}/\tilde{p}_{02}$ of \tilde{p}_0 in $P_{m,p}$, where

$$(6.32) \quad \tilde{p}_{01} = \Delta(z)[D_{00}z^n + \cdots + D_{0n}]/z^n,$$

$$(6.33) \quad \tilde{p}_{02} = \Delta(z)[N_{00}z^n + \cdots + N_{0n}]/z^n$$

such that

$$(6.34) \quad \lim_{i \rightarrow \infty} \|\tilde{p}_{i1} - \tilde{p}_{01}\| = 0,$$

$$(6.35) \quad \lim_{i \rightarrow \infty} \|\tilde{p}_{i2} - \tilde{p}_{02}\| = 0$$

where we assume that $\Delta(z) \in H^{p \times p}$, $\det \Delta(z) \in J$. Writing $\Delta(z) = \Delta_2(z)^{-1} \Delta_1(z)$, where Δ_1, Δ_2 are matrices with polynomial entries such that $\det \Delta_1(z)$ and $\det \Delta_2(z)$ are strictly Hurwitz polynomials. We may therefore conclude that in the space of polynomials of degree $n + d$ (for some d) with matrix coefficients, the polynomials

$$(6.36) \quad \Delta_2(z)[N_{i0}z^n + \cdots + N_{in}]$$

converge to the polynomial

$$(6.37) \quad \Delta_1(z)[N_{00}z^n + \cdots + N_{0n}]$$

and the polynomials

$$(6.38) \quad \Delta_2(z)[D_{i0}z^n + \cdots + D_{in}]$$

converge to the polynomial

$$(6.39) \quad \Delta_1(z)[N_{00}z^n + \cdots + N_{0n}]$$

as $i \rightarrow \infty$. It therefore follows that in $\tilde{L}_{m,p}^\infty$, the equivalence class of points which correspond to the function

$$(6.40) \quad [D_{i0}z^n + \cdots + D_{in}]^{-1}[N_{i0}z^n + \cdots + N_{in}]$$

converges to the function

$$(6.41) \quad [D_{00}z^n + \cdots + D_{0n}]^{-1}[N_{00}z^n + \cdots + N_{0n}].$$

Thus $\tilde{\omega}_i$ converges to $\tilde{\omega}_0$ in $\tilde{L}_{m,p}^\infty$ as $i \rightarrow \infty$.

(1 \Rightarrow 2) Since ϕ_∞ is continuous and $\tilde{L}_{m,p}^\infty$ has quotient topology, it follows that ψ_∞ is continuous. Thus 1 \Rightarrow 2(a). Finally, in order to show that 1 \Rightarrow 2(b), let p_i , $i = 1, 2, \dots$ be a sequence of plants in $P_{m,p}$ with the property that $\text{lag } p_{i+1} > \text{lag } p_i$ and $\text{lag } p_1 = 1$, $i = 1, 2, 3, \dots$. It now follows from the proof of Theorem 3.4 in [4] that $\psi_\infty^{-1}(p_i)$, $i = 1, 2, \dots$ does not converge in $\tilde{L}_{m,p}^\infty$. This is because if there exists $\tilde{\omega}_0$ in $\tilde{L}_{m,p}^\infty$ such that $\{\psi_\infty^{-1}(p_i)\}$ converges to $\tilde{\omega}_0$, define $\psi_\infty(\tilde{\omega}_0) = p_0$. Let U_1 be an open neighborhood of p_0 in $P_{m,p}$ which does not contain p_1 . Define

$$(6.42) \quad S_1 \triangleq \psi_\infty^{-1}(U_1) \cap \tilde{L}_{m,p}^1$$

and

$$(6.43) \quad V_1 \triangleq \psi_\infty(S_1).$$

Note that V_1 does not contain $\{p_i\}$. Assume that U_i, S_i, V_i are defined where V_i does not contain the above sequence $\{p_i, i = 1, 2, \dots\}$ in $P_{m,p}$ and where

$$(6.44) \quad S_i = \psi_\infty^{-1}(U_i) \cap \tilde{L}_{m,p}^i.$$

Define U_{i+1} to be an open neighborhood of V_i which does not contain p_1, \dots, p_{i+1} . Define

$$(6.45) \quad S_{i+1} \triangleq \psi_\infty^{-1}(U_{i+1}) \cap \tilde{L}_{m,p}^{i+1}$$

and

$$(6.46) \quad V_{i+1} \triangleq \psi_\infty(S_{i+1}).$$

It is clear that

$$(6.47) \quad S_i \subset S_{i+1} \quad \forall i = 1, 2, \dots$$

Define

$$(6.48) \quad S_\infty = \bigcup_{i=1}^{\infty} S_i.$$

It may be concluded that S_∞ is an open neighborhood of $\tilde{\omega}_0$ which does not contain the sequence $\{\psi_\infty^{-1}(p_i)\}$. \square

7. Some further remarks on the parametrization of feedback control systems. Continuing our earlier discussion in § 5, in this section we once again address the problem of parametrizing feedback control systems. However we do not assume that the closed loop system is stable. In particular we consider the space of $p \times m$ ARMA systems $G(z)$ of lag $\leq n$ and $m \times p$ ARMA systems $K(z)$ of lag $\leq q$ and parametrize in the product space $L_{m,p}^n \times L_{p,m}^q$ those pairs of plant/compensator that corresponds to a closed loop system in $L_{m,p}^{n+q}$.

Stated more precisely, let us consider the product space

$$(7.1) \quad L_{m,p}^n \times L_{p,m}^q$$

of plant/compensator pairs $G(z), K(z)$, where $G(z) \in L_{m,p}^n, K(z) \in L_{p,m}^q$. Suppose that $G(z)$ defines the input-output system

$$(7.2) \quad D_p(z)y = N_p(z)u$$

and $K(z)$ defines the input-output system

$$(7.3) \quad \tilde{D}_c(z) = \tilde{N}_c(z)u$$

where

$$(7.4) \quad \tilde{D}_c^{-1} \tilde{N}_c = N_c D_c^{-1}.$$

Here we assume that the pairs (D_p, N_p) , (D_c, N_c) and $(\tilde{D}_c, \tilde{N}_c)$ are coprime. We may define the space $LF_{m,p}^{n,q}$ as follows:

$$(7.5) \quad LF_{m,p}^{n,q} \triangleq \{[D_p, N_p], [\tilde{D}_c, \tilde{N}_c] \in L_{m,p}^n \times L_{p,m}^q : [D_p D_c + N_p N_c, N_p N_c] \in L_{m,p}^{n+q}\}.$$

Note that multiplying $[D_p, N_p]$ by a nonsingular matrix to the left and $[\tilde{D}_c, \tilde{N}_c]$ by a nonsingular matrix to the right does not change the condition in (7.5). Thus (7.5) is well defined.

The space $LF_{m,p}^{n,q}$ parametrizes plant/compensator pairs that define an ARMA system in the closed loop, and is clearly of interest in control system design. Furthermore, in off-line identification of parameters in $L_{m,p}^n \times L_{p,m}^q$, the closed-loop stability of the intermediate parameter values of the plant/compensator pair is not required (as opposed to an online identification problem and adaptive control). Thus whereas the space $LFB_{m,p}^{n,q}$ considered in § 5 is of importance in on-line recursive algorithms, the space $LF_{m,p}^{n,q}$ appears to be of importance in off-line recursive identification problems.

Remark 7.1. As an example of a plant-compensator pair that does not belong to $LF_{m,p}^{n,q}$ consider the pair

$$(7.6) \quad \frac{1}{z+1}, \frac{2z+3}{5}.$$

Define

$$(7.7) \quad D_p = \frac{z+1}{z}, \quad N_p = \frac{1}{z}, \quad \tilde{D}_c = \frac{5}{z}, \quad \tilde{N}_c = \frac{2z+3}{z}$$

and note that

$$(7.8) \quad D_p D_c + N_p N_c = (7z+8)/(z^2),$$

$$(7.9) \quad N_p N_c = (2z+3)/(z^2)$$

which vanishes at $z = \infty$.

The main result that we show in this section is described below.

Main Theorem 7.2. $LF_{m,p}^{n,q}$ is a fiber bundle over $L_{m,p}^{0,0}$.

Before we proceed to prove the Main Theorem it may be worthwhile to study the structure of $L_{1,1}^{0,0}$.

Example 7.3. In this example we assume that $m = p = 1$, $n = q = 0$, and describe $L_{1,1}^{0,0}$. Clearly, $L_{1,1}^{0,0}$ is a subset of $L_{1,1}^0 \times L_{1,1}^0$ described by

$$(7.10) \quad \{[a, b], [c, d]: [ac + bd, bd] \in \mathbb{RP}^1\}.$$

Equivalently, $L_{1,1}^{0,0}$ is described by

$$(7.11) \quad \mathbb{RP}^1 \times \mathbb{RP}^1 - \{([0, 1], [1, 0]), ([1, 0], [0, 1])\}.$$

Thus, $L_{1,1}^{0,0}$ is homeomorphic to a torus with two distinct points removed.

Remark 7.4. In view of Example 7.3, it appears that the structure of $LF_{m,p}^{0,0}$ is more complicated in comparison with the structure of $LFB_{m,p}^{0,0}$. Recall from § 5 that $LFN_{m,p}^{0,0}$ is unknown.

Proof of Main Theorem 7.3. The proof of this theorem is analogous to the technique of transverse and complete flow argument used in the proof of Theorem 4.6. Therefore we sketch only the main points here. First note that $L_{m,p}^{0,0}$ is a manifold since it is an open subset of $\text{Grass}(m, m+p) \times \text{Grass}(p, m+p)$. Consider the map

$$(7.12) \quad \psi: LF_{m,p}^{n,q} \rightarrow LF_{m,p}^{0,0}$$

given by

$$(7.13) \quad \psi([D_p(z), N_p(z)], [\tilde{D}_c(z), \tilde{N}_c(z)]) = ([D_p(\infty), N_p(\infty)], [\tilde{D}_c(\infty), \tilde{N}_c(\infty)]).$$

Consider the local flow

$$(7.14) \quad \mathcal{H}_{m,p}^{n,q}: LF_{m,p}^{n,q} \times [0, \infty) \rightarrow LF_{m,p}^{n,q}$$

defined by

$$(7.15) \quad \begin{aligned} & \mathcal{H}_{m,p}^{n,q}([D_p(z), N_p(z)], [\tilde{D}_c(z), \tilde{N}_c(z)], t) \\ &= ([D_p(e^t z), N_p(e^t z)], [\tilde{D}_c(e^t z), \tilde{N}_c(e^t z)]). \end{aligned}$$

It is easily seen that the above flow is well defined and smooth. Moreover if $Z \in LF_{m,p}^{0,0}$, the fiber $\psi^{-1}(Z)$ of the submersion (7.12) is invariant under the above flow. By argument similar to the proof of Lemma 5.3, $\psi^{-1}(Z)$ is diffeomorphic to $\mathbb{R}^{(np+qm)(m+p)}$. The proof of the local triviality of $LF_{m,p}^{n,q}$ is similar to the proof sketched in Lemma 5.3 and Theorem 4.6. \square

8. Conclusion. To conclude, in this paper we have studied the problem of parametrizing linear dynamical systems, both in the open loop and in the closed loop. Although we have restricted our attention to discrete-time ARMA systems, consideration of state-space, continuous-time systems would be analogous. The main result of

this paper is that the various parametrizations obtained have a bundle structure, which we believe would be particularly useful in system identification defined by a locally and globally convergent vector field wherein the structure of the fibers, diffeomorphic to a Euclidean space, would be exploited.

Appendix I. Milnor's theorem.

THEOREM. *Let M be an n -dimensional manifold. Suppose that M admits a vector field X with a unique locally and globally attracting singularity. Then M is diffeomorphic to \mathbb{R}^n .*

Remark. Among the geometric control theory community this is known as Milnor's theorem. (Perhaps this name is due to Chris Byrnes.) This is essentially contained in [5] and can be proved using several methods. We sketch a proof based on [5].

The following lemma is easy (see [5], [25], or [26] for a stronger version). In what follows, D_k is the closed disk of radius k in \mathbb{R}^n .

LEMMA 1. *Let N be an n -dimensional oriented manifold, and let $f_1, f_2: D_1 \rightarrow N$ be orientation-preserving embeddings into the interior of N . Then there exists a diffeomorphism $h: N \rightarrow N$ such that $h \cdot f_1 = f_2$.*

LEMMA 2 [5]. *Let N be an n -dimensional manifold such that each compact subset is contained in an open set diffeomorphic to \mathbb{R}^n . Then N is diffeomorphic to \mathbb{R}^n .*

Proof. We can easily construct a sequence $W_1 \subset W_2 \subset \cdots \subset N$ such that $\bigcup_{k=1}^{\infty} W_k = N$ and W_k is diffeomorphic to D_k . Say $g_k: D_k \rightarrow W_k$ is a diffeomorphism. We are now going to define a new sequence of diffeomorphisms $f_k: D_k \rightarrow W_k$ by induction. Set $f_1 = g_1$.

Now suppose that f_1, \dots, f_k have already been defined. Orient W_{k+1} and modify g_{k+1} if necessary such that

$$D_k \xrightarrow{f_k} W_k \xrightarrow{i_k} W_{k+1} \quad \text{and} \quad D_k \xrightarrow{j_k} D_{k+1} \xrightarrow{g_{k+1}} W_{k+1}$$

are orientation-preserving, where i_k, j_k are inclusions. By Lemma 1, find a diffeomorphism $h_{k+1}: W_{k+1} \rightarrow W_{k+1}$ such that $h_k \cdot g_k \cdot j_k = i_k \cdot f_k$ and define $f_{k+1} = h_{k+1} \cdot g_{k+1}$. Then $f_{k+1}|_{D_k} = f_k$, and hence pass to the direct limit to obtain the diffeomorphism $\lim_{\rightarrow} f_k: \mathbb{R}^n \rightarrow N$. \square

Proof of the theorem. By Lemma 4.10, without loss of generality we assume that X is complete. Let U be a neighborhood of p which is diffeomorphic to \mathbb{R}^n . Define $U_k = \phi_{-k}^X(U)$, $k = 1, 2, \dots$. Then each U_k is diffeomorphic to \mathbb{R}^n , and $M = \bigcup_{k=1}^{\infty} U_k$. Hence by Lemma 2, M is also diffeomorphic to \mathbb{R}^n .

REFERENCES

- [1] R. SAEKS AND J. J. MURRAY, *Fractional representation, algebraic geometry, and the simultaneous stabilization problem*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 895-903.
- [2] M. VIDYASAGAR AND N. VISWANADHAM, *Algebraic design techniques for reliable stabilization*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 1085-1095.
- [3] B. K. GHOSH AND C. I. BYRNES, *Simultaneous stabilization and simultaneous pole-placement by non-switching dynamic compensation*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 735-741.
- [4] B. K. GHOSH AND W. P. DAYAWANSA, *A hybrid parametrization of linear single-input single-output system*, Systems Control Lett., 8 (1987), pp. 231-239.
- [5] T. SAATY, ED., *Lectures in Modern Mathematics II*, John Wiley, New York, 1964.
- [6] C. CAMACHO AND A. L. NETO, *Geometric Theory of Foliations*, Birkhauser, Boston, 1985.
- [7] J. W. MILNOR AND J. D. STASHEFF, *Characteristic classes*, Ann. of Math. Stud., 76 (1974).
- [8] M. VIDYASAGAR, *Control Systems Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
- [9] M. DEISTLER AND E. J. HANNAN, *Some properties of the parametrization of ARMA systems with unknown order*, J. Multivariate Anal., 11 (1981), pp. 474-484.

- [10] G. C. GOODWIN AND K. S. SIN, *Adaptive Filtering, Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
- [11] C. I. BYRNES, *Geometric aspects of the convergence analysis of identification algorithms*, Nonlinear Stochastic Problems, R. S. Bucy and J. F. Moura, eds., D. Reidel, Dordrecht, the Netherlands, 1983.
- [12] ———, *On compactification of spaces of systems and dynamic compensation*, 22nd IEEE Conference on Decision and Control, San Antonio, TX, 1983, pp. 889–894.
- [13] D. F. DELCHAMPS, *Global structure of families of multivariable linear systems with an application to identification*, Math. Systems Theory, 18 (1985), pp. 329–380.
- [14] U. HELMKE, *The topology of a moduli space for linear dynamic systems*, Comment. Math. Helv., 60 (1985), pp. 630–655.
- [15] U. HELMKE AND W. MANTHEY, *A partial compactification for linear systems*, Systems Control Lett., 8 (1986), pp. 39–46.
- [16] M. HAZEWINKEL, *Parametrization of the space of all linear systems and numerical trouble*, preprint.
- [17] B. K. GHOSH, *An approach to simultaneous system design, part II: Nonswitching gain and dynamic feedback compensation by algebraic geometric methods*, SIAM J. Control Optim., 26(4), pp. 919–963.
- [18] D. J. COBB, *Descriptor variable and generalized singularly perturbed systems: A geometric approach*, Ph.D. thesis, Dept. of Electrical Engineering, Univ. of Illinois, Urbana, IL, 1980.
- [19] E. STIEFEL, *Richtungsfelder und Fernparallelismus in mannigfaltigkeiten*, Comment. Math. Helv., 8 (1936), pp. 3–51.
- [20] R. BOTT AND J. MILNOR, *On the parallelizability of spheres*, Bull. Amer. Math. Soc., 64 (1958), pp. 87–89.
- [21] M. Kervaire, *Nonparallelizability of the N -sphere for $n > 7$* , Proc. Nat. Acad. Sci., U.S.A., 44 (1958), pp. 280–283.
- [22] J. F. ADAMS, *On the non-existence of elements of Hopf invariant one*, Ann. of Math., 72 (1960), pp. 20–104.
- [23] R. BOTT AND L. W. TU, *Differential forms in Algebraic Topology*, Springer-Verlag, Berlin, New York, Graduate Texts in Math. 82, 1982.
- [24] W. POOR, *Differential Geometric Structures*, McGraw-Hill, New York, 1981.
- [25] R. PALAIS, *Extending diffeomorphisms*, Proc. Amer. Math. Soc., 11 (1960), pp. 274–277.
- [26] M. W. HIRSH, *Differential Topology*, Springer-Verlag, Berlin, New York, 1976.
- [27] R. W. BROCKETT, *Some geometric questions in the theory of linear systems*, IEEE Trans. Automat. Control, AC-21, (1976), pp. 449–464.
- [28] TH. BRÖCKER AND K. JÄNICH, *Introduction to Differential Topology*, Cambridge Univ. Press, Cambridge, U.K., 1982.

EQUIVALENCE BETWEEN OPEN-LOOP AND CLOSED-LOOP INVARIANCE FOR INFINITE-DIMENSIONAL SYSTEMS: A FREQUENCY DOMAIN APPROACH*

HANS ZWART†

Abstract. In this paper the relation between various system invariance concepts is considered, and it is proved that open- and closed-loop invariance are equivalent concepts. Using this equivalence the author investigates the existence of the largest closed-loop invariant subspace, and derives necessary and sufficient conditions such that the largest open-loop invariant subspace is equal to the largest closed-loop invariant subspace. If this equality holds, then the solvability of the disturbance decoupling problem (DDP) is equivalent to the solvability of a meromorphic matrix equation.

Key words. infinite-dimensional linear systems, controlled invariance, disturbance decoupling problem

AMS(MOS) subject classification. 93C25

1. Introduction. The theory of controlled invariance of a subspace has been investigated in detail in the case that the state space is finite-dimensional (see, e.g., [1], [9], [17], [19]). For the case where the state space is infinite-dimensional, while there have been some preliminary investigations in [2]–[4], [14], [16], and [22], many questions remain unanswered. In this paper we shall investigate this invariance for infinite-dimensional systems. We shall consider the following controlled system:

$$(1.1) \quad \dot{x} = Ax + Bu, \quad x \in X, \quad u \in U,$$

where X and U are Banach spaces, A is a generator of a C_0 -semigroup, $T(t)$, and furthermore we impose the conditions:

- (C1) B is a bounded linear operator with $\text{Im } B$ finite-dimensional.
- (C2) For all A -bounded (see Appendix A for definition) feedback laws F , $A + BF$ generates a C_0 -semigroup, $T_F(t)$, with the domain of $A + BF$ equal to the domain of A ; $D(A)$.

Let us remark that (C2) does not always hold, even if (C1) holds; for a counterexample see, e.g., Lasiecka and Triggiani [13]. However, if A generates an analytic semigroup, then (C2) holds if (C1) holds (see, e.g., Zabczyk [20]). The use of A -bounded feedback operators is naturally motivated by the geometric properties of invariance, as can be seen from Theorems 4.4 and 4.2. However they do necessitate the extra condition (C2) on (A, B) . If we restrict ourselves to bounded feedback laws, then $A + BF$ generates automatically a C_0 -semigroup and we can still give a version of an equivalence theorem (4.6).

For the system (1.1) we shall discuss various kinds of system invariance in § 2, and we shall pay particular attention to frequency invariance in § 3. This concept of system invariance was introduced by Hautus [9] for finite dimensions, and it turns out that this concept plays a key role in infinite dimensions. In § 4 the relation between the different invariance concepts is investigated, and there we show that, for closed

* Received by the editors February 17, 1987; accepted for publication (in revised form) November 30, 1987. This research was supported by the Netherlands Organization for the Advancement of Pure Scientific Research.

† Department of Mathematics, University of Groningen, P.O. Box 800, 9700 AV Groningen, the Netherlands.

subspaces, open- and closed-loop invariance are equivalent. Furthermore we prove that, for these subspaces, closed-loop invariance is equivalent to frequency invariance. Using this last equivalence we shall define the largest frequency invariant subspace contained in a given subspace, and show that the largest closed-loop invariant subspace in a given subspace need not necessarily exist. However, we shall give sufficient conditions such that this subspace exists and is equal to the largest frequency invariant subspace. If this equality holds, then we can give a frequency domain solution of the disturbance decoupling problem (§ 5).

Notation.

X	A general Banach space.
V	A subspace of X .
K	A closed subspace of X .
U	The input space, which is assumed to be finite-dimensional.
A	A generator of a C_0 -semigroup on X .
B	A bounded linear operator from U to X .
$\mathcal{B}^1(V)$	$\text{Im } B \cap V$.
$\mathcal{B}^0(V)$	A subspace of the image of B , $\text{Im } B$, with the properties that $\mathcal{B}^0(V) \cap V = \{0\}$ and $\mathcal{B}^0(V) + \mathcal{B}^1(V) = \text{Im } B$ (see (3.2)).
$\mathcal{V}^*(K)$	The largest closed-loop invariant subspace contained in K .
$\mathcal{V}_\Sigma(K)$	The largest frequency invariant subspace contained in K .
$*$	The convolution product.

2. Invariance concepts. The theory of system invariance entails many definitions. Here we shall summarize some of them and give some important properties. We shall start with the strongest.

Throughout this paper we consider system (1.1) under the standing assumptions (C1) and (C2).

DEFINITION 2.1. A subspace V of X is called closed-loop invariant if there exists an A -bounded feedback law F such that

$$(2.1) \quad T_F(t)V \subset V$$

for all t in $[0, \infty)$.

Remark. Under condition (C2) the semigroup $T_F(t)$ is well defined.

LEMMA 2.2. Assume that a closed linear subspace $V \subset X$ is $T_{F_1}(t)$ -invariant for a certain A -bounded operator F_1 . Then V is $T_{F_2}(t)$ -invariant for an A -bounded operator F_2 if and only if $\text{Im } B(F_1 - F_2)|_{V \cap D(A)} \subset V$.

Proof. See Appendix B for the proof. \square

Remark. Lemma 2.2 is a generalization of Lemma 4 in Curtain [3].

COROLLARY 2.3. Let \mathcal{B} be any subspace of $\text{Im } B$ such that $\mathcal{B} + (\text{Im } B \cap V) = \text{Im } B$. If a closed subspace V is closed-loop invariant, then there exists an A -bounded feedback law F such that V is $T_F(t)$ -invariant and $\text{Im } BF|_{V \cap D(A)} \subset \mathcal{B}$.

Proof. Assume that V is $T_F(t)$ -invariant; then from the facts that the range of \tilde{F} is finite-dimensional and \tilde{F} is A -bounded, $B\tilde{F}$ can be written as [11, pp. 195, 245] $\sum_{i=1}^q b_i \langle (A - \lambda) \cdot, f_i \rangle + \sum_{i=q+1}^{p_1} b_i \langle (A - \lambda) \cdot, f_i \rangle$, where λ is an arbitrary element of the resolvent set of A , $\text{span}_{i=1, \dots, q} \{b_i\} = \mathcal{B}$ and $\text{span}_{i=q+1, \dots, p_1} \{b_i\} \subset \text{Im } B \cap V$. Defining $F = (\langle (A - \lambda) \cdot, f_i \rangle)_{i=1}^q$, we obtain that $\text{Im } B(\tilde{F} - F)|_{V \cap D(A)} \subset V$ and $BF|_{V \cap D(A)} \subset \mathcal{B}$. Thus Lemma 2.2 concludes the assertion. \square

Before we can introduce the concept of open-loop invariance we need to define our input space.

DEFINITION 2.4. $C_e^{-1}([0, \infty); U)$ is the space of all distributional derivatives of continuous functions that are exponentially bounded and vanish at zero.

So $u(\cdot) \in C_e^{-1}([0, \infty); U)$ if $\int_0^t u(s) ds \in C([0, \infty); U)$, there are constants α and M in \mathbb{R} subject to

$$\left| \int_0^t u(s) ds \right| \leq M e^{\alpha t}, \text{ for all } t \geq 0, \text{ and } \lim_{t \downarrow 0} \int_0^t u(s) ds = 0.$$

Remark. By $C([0, \infty); U)$ we mean the space of all functions from $[0, \infty)$ to U which are continuous at all points in $[0, \infty)$; the functions need not be bounded on $[0, \infty)$.

DEFINITION 2.5. A subspace $V \subset X$ is said to be open-loop invariant if for every $x_0 \in V$ there exists a $u(\cdot) \in C_e^{-1}([0, \infty); U)$ such that the solution of (1.1) remains in V (see the following remark).

Remark. By the solution of (1.1) for $u(\cdot) \in C_e^{-1}([0, \infty); U)$ we mean the mild solution:

$$x(t) = T(t)x_0 + \{T(t)B * u(t)\}$$

where $*$ denotes the convolution product. So $\hat{x}(t) := \mathbb{1}_{[0, \infty)} * x(\cdot) = \int_0^t x(s) ds$ satisfies in the usual sense the following equation:

$$(2.2) \quad \hat{x}(t) = \int_0^t T(t-s) \left\{ x_0 + B \int_0^s u(\sigma) d\sigma \right\} ds.$$

Furthermore $x(t)$ will remain in V if $\hat{x}(t)$ remains in V .

It may seem more natural to define open-loop invariance as the existence of an input $u(t) \in \mathcal{L}_{oc}^1([0, \infty); U)$ such that the solution of (1.1) remains in V ; however, open-loop invariance should be a stronger concept than closed-loop invariance. Therefore we must allow distributional inputs of the form $FT_F(t)x_0$, where F is an A -bounded operator.

The input map $FT_F(t)x_0$ is not defined for all $t \geq 0$ and all $x_0 \in X$, but from Curtain and Pritchard [5] we have that $F(\int_0^t T_F(s)x_0 ds)$ is; it is even a continuous function with $\lim_{t \downarrow 0} F(\int_0^t T_F(s)x_0 ds) = 0$ and $F(\int_0^t T_F(s)x_0 ds)$ is exponentially bounded. So $FT_F(t)x_0$ is an element $C_e^{-1}([0, \infty); U)$, and this motivates our choice of the input space.

If we were to omit the condition $\lim_{t \downarrow 0} \int_0^t u(s) ds = 0$, then the Dirac distribution would be an element of our input space. For finite-dimensional X this reduces to the concept of asymptotic invariance (Stern [18]) which is stronger than closed-loop invariance. Since the aim of this paper is to generalize the known results for finite-dimensional systems, this is undesirable.

The condition that for open-loop invariance the input $u(\cdot)$ satisfies $|\int_0^t u(s) ds| \leq M e^{\alpha t}$ for some M and α in \mathbb{R} can be omitted if the state space is finite-dimensional. Thus we could hope that in the general case this condition can also be omitted. However a general proof for this assertion is at this moment still missing, and we shall need this condition in order to prove our main result; the equivalence between open- and closed-loop invariance for closed linear subspaces.

Let us remark that our definition of open-loop invariance is stronger than that given by Schmidt and Stern [16] in which they used the term "holdability subspace." However, it will be a direct consequence of Theorem 4.4 that the two definitions are equivalent for the case that A is a bounded operator. Thus if the state space X is finite-dimensional, then our concept of open-loop invariance is equivalent to the concept of controlled invariance, as defined by Basile and Marro [1].

DEFINITION 2.6. A subspace V of X is called (A, B) invariant if

$$(2.3) \quad A(V \cap D(A)) \subset V + \text{Im } B.$$

DEFINITION 2.7. A subspace V of X is called feedback invariant if there exists an A -bounded feedback law F such that

$$(2.4) \quad (A + BF)(V \cap D(A)) \subset V.$$

Remark. Let $\hat{\mathcal{B}}$ be any subspace of $\text{Im } B$ such that $\hat{\mathcal{B}} + (\text{Im } B \cap V) = \text{Im } B$. Then by a proof similar to that in Corollary 2.3, it can be shown that if a subspace is feedback invariant, it is also feedback invariant for an A -bounded feedback satisfying $\text{Im } BF|_{V \cap D(A)} \subset \hat{\mathcal{B}}$.

If X is finite-dimensional, then it is known that all these concepts are equivalent (Basile and Marro [1] and Wonham [19]). It is the aim of this paper to show that the same holds for infinite-dimensional systems that satisfy conditions (C1) and (C2). We shall show that for closed linear subspaces equivalence holds between open- and closed-loop invariance and between (A, B) - and feedback invariance. There is in general no equivalence between (A, B) - and closed-loop invariance even if we impose the extra condition that $V \cap D(A)$ is dense in V , as shown in Schmidt and Stern [16]. Furthermore we shall show that equivalence between open- and closed-loop invariance is lost if the subspace is not closed. Before we can prove these assertions we need a fifth concept of invariance which we shall investigate more deeply.

3. Frequency invariance. In Hautus [9] the concept of frequency invariance appeared for the first time. Only the finite-dimensional case was discussed there. In this section we shall discuss this concept for infinite-dimensional systems.

The next definition will generalize the concept of the space of all rational functions.

DEFINITION 3.1. Let Y be a subspace, not necessarily closed, of a Banach space Z . A function $f(s)$ from \mathbb{C} to Z is an element of $Y(s)$ if $f(s)$ is meromorphic on some right half-plane of \mathbb{C} and $f(s_0)$ is an element of Y for all s_0 in this right half-plane and s_0 is not a pole of $f(s)$. By $Y_+(s)$ we shall denote all functions in $Y(s)$ that are strictly proper, i.e.,

$$Y_+(s) = \left\{ f \in Y(s) \mid \lim_{\substack{s \rightarrow \infty \\ s \text{ real}}} f(s) = 0 \right\}.$$

DEFINITION 3.2. If $x_0 \in X$, $\xi(\cdot) \in D(A)(s)$ and $\omega(\cdot) \in U_+(s)$, then the expression

$$(3.1) \quad x_0 = (s - A)\xi(s) - B\omega(s)$$

is called a (ξ, ω) -representation of x_0 .

We remark that every x_0 in X has a (ξ, ω) -representation: take $\xi(s) = (s - A)^{-1}x_0$ and $\omega(s) = 0$.

The (ξ, ω) -representation was introduced by Hautus [9], and it has the following important property.

LEMMA 3.3. If $\xi(s)$, $\omega(s)$ is a (ξ, ω) -representation of x , then $\xi(s) \in D(A)_+(s)$ and $\lim_{s \rightarrow \infty, s \in \mathbb{R}} s\xi(s) = x$.

Proof. By the fact that A generates a C_0 -semigroup we have that $\lim_{s \rightarrow \infty, s \in \mathbb{R}} s(s - A)^{-1}x = x$ for all x in X . By the (ξ, ω) -representation

$$x = (s - A)\xi(s) - B\omega(s) \quad \text{or} \quad s\xi(s) = s(s - A)^{-1}x - s(s - A)^{-1}B\omega(s).$$

Thus

$$\lim_{\substack{s \rightarrow \infty \\ s \in \mathbb{R}}} s\xi(s) = \lim_{\substack{s \rightarrow \infty \\ s \in \mathbb{R}}} s(s - A)^{-1}x - \lim_{\substack{s \rightarrow \infty \\ s \in \mathbb{R}}} s(s - A)^{-1}B\omega(s) = x + 0,$$

since $s(s - A)^{-1}B$ is from the Hille-Yosida Theorem uniformly bounded on $[r, \infty)$ and $\lim_{s \rightarrow \infty, s \in \mathbb{R}} \omega(s) = 0$. \square

With this (ξ, ω) -representation we can introduce the concept of frequency invariance.

DEFINITION 3.4. Let V be a subspace, not necessarily closed, of X . V is said to be frequency invariant if every x_0 in V has a (ξ, ω) -representation with $\xi(\cdot) \in V_+(s)$.

Lemmas 3.5 and 3.6 will imply that the (ξ, ω) -representation is unique if $\text{Im } B \cap V = \{0\}$. First we must introduce some notation.

Let V be a subspace of X . By $\mathcal{B}^0(V)$ and $\mathcal{B}^1(V)$ we shall denote subspaces of X such that

$$\begin{aligned} \text{Im } B &= \mathcal{B}^0(V) \oplus \mathcal{B}^1(V), \\ \mathcal{B}^0(V) \cap V &= \{0\}, \\ \mathcal{B}^1(V) &\subset V. \end{aligned} \quad (3.2)$$

We remark that (3.2) is possible, since $\text{Im } B$ is finite-dimensional (condition (C1)).

So $\mathcal{B}^0(V)$ is a subspace contained in $\text{Im } B$ which has zero intersection with V and is of maximal dimension, and $\mathcal{B}^1(V)$ is $\text{Im } B \cap V$. Throughout this section we shall fix $\mathcal{B}^0(V)$ so if X is a Hilbert space, then we can take $\mathcal{B}^0(V) = \text{Im } B \cap (\text{Im } B \cap V)^\perp$. If there cannot be any doubt about V , then we shall simply use \mathcal{B}^0 and \mathcal{B}^1 .

LEMMA 3.5. *If a subspace V of X is frequency invariant, then every element of V has a (ξ, ω) -representation with $\xi(\cdot) \in V_+(s)$ and $B\omega(\cdot) \in \mathcal{B}_+^0(s)$.*

If in addition we assume that every element of V has a (ξ, ω) -representation such that $\lim_{s \rightarrow \infty, s \in \mathbb{R}} s\omega(s)$ exists, then it also has a (ξ, ω) -representation with $\xi(\cdot) \in V_+(s)$, $B\omega(\cdot) \in \mathcal{B}_+^0(s)$ and $\lim_{s \rightarrow \infty, s \in \mathbb{R}} s\omega(s)$ exists.

Proof. See Appendix B. For the proof. \square

LEMMA 3.6. *If a subspace V of X is frequency invariant and $\text{Im } B \cap V = \text{Im } B \cap \bar{V}$, then every x in V has a unique (ξ, ω) -representation with $B\omega(\cdot)$ in $\mathcal{B}_+^0(s)$ and $\xi(\cdot)$ in $V_+(s)$.*

Proof. Let b_1, \dots, b_p be a basis for \mathcal{B}^0 . Then by the Hahn-Banach Theorem and the fact that $\text{Im } B \cap V = \text{Im } B \cap \bar{V}$, there exist functionals $\{f_i \in X', i = 1, \dots, p\}$ such that $\langle f_i, b_j \rangle = \delta_{ij}$ and $f_i|_V = 0$.

Let x be an arbitrary element of V ; then by Lemma 3.5 it has a (ξ, ω) -representation with $B\omega(\cdot)$ contained in $\mathcal{B}_+^0(s)$. Thus x can be written as

$$(3.3) \quad x = (s - A)\xi(s) - \sum_{i=1}^p b_i \omega_i(s), \quad \xi(\cdot) \in V_+(s).$$

Equation (3.3) implies

$$(3.4) \quad (s - A)^{-1}x = \xi(s) - \sum_{i=1}^p (s - A)^{-1}b_i \omega_i(s).$$

Calculating $\langle f_i, (s - A)^{-1}x \rangle$; $i = 1, \dots, p$ gives

$$\begin{aligned} \langle f_i, (s - A)^{-1}x \rangle &= \langle f_i, \xi(s) \rangle - \sum_{j=1}^p \langle f_i, (s - A)^{-1}b_j \rangle \omega_j(s) \\ (3.5) \quad &= - \sum_{j=1}^p \langle f_i, (s - A)^{-1}b_j \rangle \omega_j(s) \quad \text{since } \xi(s) \text{ is in } V. \end{aligned}$$

If we premultiply (3.5) by s we obtain

$$(3.6) \quad \langle f_i, s(s - A)^{-1}x \rangle = \sum_{j=1}^p \langle f_i, s(s - A)^{-1}b_j \rangle \omega_j(s).$$

Or in matrix notation

$$(3.7) \quad \begin{pmatrix} \langle f_1, s(s-A)^{-1}x \rangle \\ \langle f_p, s(s-A)^{-1}x \rangle \end{pmatrix} = -S(s) \begin{pmatrix} \omega_1(s) \\ \omega_p(s) \end{pmatrix},$$

where $S_{ij}(s) = \langle f_i, s(s-A)^{-1}b_j \rangle$. Note that $S(s)$ and $\langle f_i, s(s-A)^{-1}x \rangle$ are analytic on the resolvent set of A , and since $\lim_{s \in \mathbb{R}, s \rightarrow \infty} s(s-A)^{-1}y = y$ for all y in X we have that $\lim_{s \in \mathbb{R}, s \rightarrow \infty} S_{ij}(s) := S_{ij}(\infty) = \langle f_i, b_j \rangle = \delta_{ij}$. So $S(\infty)$ is nonsingular. Then by the continuity of $S(s)$ in plus infinity there exists an interval $[\hat{s}, \infty)$ such that $S(s)$ is invertible on that interval.

$S(s)$ is an analytic function on the resolvent set of A , and so is certainly on a right half-plane. From (3.7) we get on this right half-plane

$$(3.8) \quad \begin{pmatrix} \omega_1(s) \\ \omega_p(s) \end{pmatrix} = -S^{-1}(s) \begin{pmatrix} \langle f_1, s(s-A)^{-1}x \rangle \\ \langle f_p, s(s-A)^{-1}x \rangle \end{pmatrix}.$$

The right-hand side of (3.8) depends only on x (for fixed A and B) and it is a meromorphic function. So the only possible choice for $\omega_i(s)$ is the i th row of the right-hand side of (3.8). By (3.4) we have only one choice for $\xi(s)$; that is,

$$(s-A)^{-1}x + (s-A)^{-1} \left(\sum_{j=1}^p b_j \omega_j(s) \right). \quad \square$$

COROLLARY 3.7. *If V is frequency invariant and $\text{Im } B \cap V = \text{Im } B \cap \bar{V}$, then every x in V has a unique (ξ, ω) representation with $\xi(\cdot) \in V_+(s)$, $B\omega(\cdot) \in \mathcal{B}_+^0(s)$ and there exists an interval $[\hat{s}, \infty)$, which is independent of x , such that $\xi(s)$ and $\omega(s)$ are continuous on this interval.*

Proof. Choose $\omega(s)$ to be the right-hand side of (3.8) and

$$\xi(s) = (s-A)^{-1}x + (s-A)^{-1} \left(\sum_{j=1}^p b_j \omega_j(s) \right).$$

These functions are meromorphic on the resolvent of A and by the proof of Lemma 3.6 continuous on an interval of the form $[\hat{s}, \infty)$. \square

COROLLARY 3.8. *If V is frequency invariant and $\text{Im } B \cap V = \text{Im } B \cap \bar{V}$, then there exists an $\hat{s} \in \mathbb{R}$ such that $(s-A)^{-1}\mathcal{B}^0 \cap \bar{V} = \{0\}$ for all s larger than \hat{s} .*

Proof. With the same notation as in the proof of Lemma 3.6, we have that $S(s) := (\langle f_i, s(s-A)^{-1}b_j \rangle)$ is nonsingular on an interval $[\hat{s}, \infty)$. On this interval we have that $(s-A)^{-1}\mathcal{B}^0 \cap \bar{V} = \{0\}$, since otherwise there would exist a vector $u^1 \neq 0$ and a $s_0 \geq \hat{s}$ such that $\sum_{j=1}^p (s_0-A)^{-1}b_j u_j^1$ is contained in \bar{V} , and so

$$\begin{aligned} S(s_0)u^1 &= \begin{pmatrix} \sum_{j=1}^p \langle f_1, s_0(s_0-A)^{-1}b_j u_j^1 \rangle \\ \sum_{j=1}^p \langle f_p, s_0(s_0-A)^{-1}b_j u_j^1 \rangle \end{pmatrix} = \begin{pmatrix} \left\langle f_1, s_0 \sum_{j=1}^p (s_0-A)^{-1}b_j u_j^1 \right\rangle \\ \left\langle f_p, s_0 \sum_{j=1}^p (s_0-A)^{-1}b_j u_j^1 \right\rangle \end{pmatrix} \\ &= s_0 \begin{pmatrix} \left\langle f_1, \sum_{j=1}^p (s_0-A)^{-1}b_j u_j^1 \right\rangle \\ \left\langle f_p, \sum_{j=1}^p (s_0-A)^{-1}b_j u_j^1 \right\rangle \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \end{aligned}$$

Thus $\det(S(s_0)) = 0$, providing the contradiction. \square

Let us remark that (3.8) implies that if X is finite-dimensional, and so A is a matrix, then $\omega_i(s)$ is a rational function, and with the last line of the proof of

Lemma 3.6, $\xi(s)$ is also a rational function. So if X is finite-dimensional, then Definition 3.1 is the same as if we were to restrict ourselves to strictly proper rational functions, as in Hautus [9].

Before we can prove the equivalence between frequency and closed-loop invariance, we need some properties of the set of all possible values of $\xi(s)$ for x in V .

DEFINITION 3.9. If V is a frequency invariant subspace, then Ξ_{s_1} consists of all $\xi_1 \in V \cap D(A)$ such that there exists an x in V with a (ξ, ω) -representation: $x = (s - A)\xi(s) - B\omega(s)$, $\xi(\cdot) \in V_+(s)$, $\omega(\cdot) \in U_+(s)$, such that $\xi(s_1) = \xi_1$.

LEMMA 3.10. If V is a closed subspace that is frequency-invariant and $\text{Im } B \cap V = \{0\}$, then we have that there exists a real \hat{s} such that for any $s_1, s_2 \geq \hat{s}$, the equalities

$$(3.9) \quad \begin{aligned} x &= (s_1 - A)\xi_1 - B\omega_1, \\ x &= (s_1 - A)\xi_2 - B\omega_2, \quad x, \xi_1, \text{ and } \xi_2 \text{ in } V, \end{aligned}$$

imply that $\xi_1 = \xi_2$.

Proof. Since V is a closed subspace of X and $\text{Im } B \cap V = \{0\}$, this lemma is a simple corollary of Corollary 3.8. \square

LEMMA 3.11. Let V be a closed-frequency invariant subspace with $\text{Im } B \cap V = \{0\}$; then

$$\Xi_{s_1} = \Xi_{s_2},$$

for all $s_1, s_2 \in [\hat{s}, \infty)$ (see Lemma 3.10).

Proof. Let ξ_1 be an element of Ξ_{s_1} ; then there exists an x in V with

$$(3.10) \quad \begin{aligned} x &= (s_1 - A)\xi_1 - B\omega(s_1), \quad \xi_1 = \xi(s_1) \\ \Rightarrow x &= (s_1 - s_2 + s_2 - A)\xi_1 - B\omega(s_1) \\ \Rightarrow (s_2 - s_1)\xi_1 + x &= (s_2 - A)\xi_1 - B\omega(s_1). \end{aligned}$$

$(s_2 - s_1)\xi_1 + x$ is an element of V , and thus it has a (ξ, ω) representation. So there exists a pair $(\hat{\xi}, \hat{\omega})$ such that

$$(3.11) \quad (s_2 - s_1)\xi_1 + x = (s - A)\hat{\xi}(s) - B\hat{\omega}(s).$$

Relations (3.10) and (3.11) with Lemma 3.10 imply that $\hat{\xi}(s_2) = \xi_1$. So $\Xi_{s_1} \subset \Xi_{s_2}$. By symmetry we conclude that $\Xi_{s_1} = \Xi_{s_2}$. \square

LEMMA 3.12. Let V be a closed-frequency invariant subspace with $\text{Im } B \cap V = \{0\}$; then $\Xi := \Xi_{\hat{s}}$ is closed in the graph norm of A , where \hat{s} is as defined in Lemma 3.10.

Proof. Let ξ_n be a sequence in Ξ such that $\xi_n \rightarrow y$ and $A\xi_n \rightarrow z$. Since A is a closed operator, we have that $y \in D(A)$ and $Ay = z$. Let $\{b_1, \dots, b_p\}$ be a basis for $\text{Im } B$; then since $\text{Im } B \cap V = \{0\}$ and V is a closed subspace, there exist $f_i \in X'$ such that $f_i|_V = 0$ and $\langle f_i, b_j \rangle = \delta_{ij}$. Since ξ_n is an element of Ξ , there exist x_n in V and ω_n in U such that

$$(3.12) \quad x_n = (\hat{s} - A)\xi_n - B\omega_n = (\hat{s} - A)\xi_n - \sum_{j=1}^p b_j \omega_{nj}.$$

Since $x_n \in V$, we have that

$$\begin{aligned} 0 &= \langle f_i, x_n \rangle = \langle f_i, (\hat{s} - A)\xi_n - B\omega_n \rangle = -\langle f_i, A\xi_n \rangle - \langle f_i, b_i \rangle \omega_{ni} \\ &\Rightarrow \langle f_i, A\xi_n \rangle = -\omega_{ni}. \end{aligned}$$

So ω_{ni} converges as $n \rightarrow \infty$, $i = 1, \dots, p$. Thus ω_n converges to, say, $\omega \in U$, and since $x_n = (\hat{s} - A)\xi_n - B\omega_n$ we have that x_n converges to x . Since V is closed we have that $x \in V$ and so there exist $\xi(s)$ and $\omega(s)$ such that

$$x = (\hat{s} - A)\xi(\hat{s}) - B\omega(\hat{s}).$$

By definition x is also equal to $(\hat{s} - A)y - B\omega$. From Lemma 3.10 we have that $y = \xi(\hat{s})$, and thus $y \in \Xi$. \square

4. Equivalence. In the theory of system invariant subspaces the equivalence between closed-loop and open-loop invariance is of great importance. This equivalence tells us that if we can find an input such that the trajectory stays in a subspace, then we can also find a feedback law such that the trajectory stays in this subspace. In Wonham [19] this equivalence was proved in the case that the state space is finite-dimensional. If the state space is infinite-dimensional, then the equivalence was proved by Schmidt and Stern [16], provided that A is bounded. However, the interesting case in infinite dimensions is when A is unbounded but generates a C_0 -semigroup; this is the focus of this section. We formulate and prove this equivalence under the conditions (C1) and (C2). Furthermore we shall prove that closed-loop invariance is equal to frequency invariance for closed linear subspaces.

We start by investigating the relation between (A, B) and feedback invariance, for which we need the following important lemma.

LEMMA 4.1. *If $V_1 \subset D(A)$ is a linear subspace, closed with respect to the graph norm of A , and $V_2 \subset X$ is a closed linear subspace with*

$$(4.1) \quad AV_1 \subset V_2 + \text{Im } B,$$

then there exists an A -bounded feedback law F such that

$$(4.2) \quad (A + BF)V_1 \subset V_2.$$

Proof. If X is finite-dimensional, then the proof can be found in Basile and Marro [1] and Wonham [19]. We refer the reader to Appendix B for the general proof; this proof is an adaptation of the proof given by Pandolfi [14]. \square

THEOREM 4.2. *If V is a closed subspace of X , then (A, B) and feedback invariance are equivalent.*

Proof. This is an easy corollary of Lemma 4.1, since if V is a closed subspace and A is a closed operator, then $V \cap D(A)$ is closed with respect to the graph norm of A . \square

Before we formulate our main equivalence theorem, we shall summarize the relation between frequency and closed-loop invariance if $B \equiv 0$.

THEOREM 4.3. *Let $\mathbb{C}_\beta := \{s \in \mathbb{C} | \text{Re}(s) > \beta\}$ be the largest right half-plane of \mathbb{C} that is contained in the resolvent set of A . Since A generates a C_0 -semigroup such a subset exists. Then the following four concepts of invariance are equivalent provided that V is a closed linear subspace:*

- (a) $T(t)V \subset V$ for all t in $[0, \infty)$.
- (b) There exists an s in \mathbb{C}_β such that $(s - A)^{-1}V \subset V$.
- (c) For all s in \mathbb{C}_β , $(s - A)^{-1}V \subset V$.
- (d) For all s in \mathbb{C}_β , $(s - A)(V \cap D(A)) = V$.

Proof. For the equivalence among (a), (b), and (c) see Pazy [15, p. 121], and for the equivalence between (a) and (d) see Kurtz [12]. \square

THEOREM 4.4. *Let V be a closed linear subspace of X ; then the following concepts of invariance are equivalent:*

- (a) V is closed-loop invariant.
- (b) V is open-loop invariant.
- (c) V is frequency invariant.
- (d) There exists an $s_0 \in \mathbb{R}$ such that for all s larger than s_0

$$(4.3) \quad (s - A)(V \cap D(A)) + \mathcal{B}^0 = V + \mathcal{B}^0,$$

where \mathcal{B}^0 is a subspace of $\text{Im } B$ which has zero intersection with V and has maximal dimension under this restriction.

It will be shown in the proof of the theorem that (4.3) is independent of the particular choice of \mathcal{B}^0 .

Let us remark that there is in general no equivalence between (a) and (d) if we use $\text{Im } B$ instead of \mathcal{B}^0 in (4.3). As a counterexample we can take

$$V = \text{Im } B = \text{span} \{b\} \quad \text{with } b \notin D(A);$$

then $(s - A)(V \cap D(A)) + \text{Im } B = \text{Im } B = V + \text{Im } B$, for all $s \in \mathbb{C}$. However, if $V = \text{span} \{b\}$ were closed-loop invariant, then b would be an eigenvector of $T_F(t)$ and thus an element of $D(A + BF) = D(A)$, providing the contradiction.

Presently it is unknown to the author whether (a) and (d), with \mathcal{B}^0 replaced by $\text{Im } B$, are equivalent if we assume in (d) the extra (and natural) condition that $V \cap D(A)$ is dense in V .

Relation (4.3) can be seen as a generalization of (A, B) invariance, and since in (4.3) only A , B , and V occur, we see (d) as the geometric reason for using A -bounded feedback laws instead of bounded feedback laws. If we impose stronger conditions on the input functions in Theorem 4.4 (b) and (c), then we have closed-loop invariance with bounded feedback laws as shall be shown in Theorem 4.6.

If X is finite-dimensional, then the equivalence between these concepts is known (see for (a) \Leftrightarrow (b), e.g., Basile and Marro [1], for (a) \Leftrightarrow (c) Hautus [9], and for (a) \Leftrightarrow (d) Schumacher [17]). Since we have equivalence between these invariance concepts we introduce a new concept that we shall use if a subspace satisfies (a), (b), (c), or (d) of Theorem 4.4.

DEFINITION 4.5. A closed linear subspace V of X is called controlled invariant if it satisfies (a), (b), (c) or, equivalently, (d) of Theorem 4.4.

THEOREM 4.6. Let V be a closed linear subspace of X ; then the following concepts of invariance are equivalent:

(a) V is closed-loop invariant for a bounded feedback law.

(b) For every x_0 in V there exists a $u(\cdot) \in C([0, \infty); U)$ with $\|u(t)\| \leq Me^{\alpha t}$; $t \geq 0$ for some M and α , such that the mild solution of the following equation remains in V :

$$(4.4) \quad \dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0.$$

(c) V is frequency invariant such that $\lim_{s \rightarrow \infty, s \in \mathbb{R}} s\omega(s)$ exists.

Remark. Theorem 4.6 is still valid if we do not impose condition (C2).

We shall prove Theorem 4.4 by showing (a) \Rightarrow (b) \Rightarrow (c) \Rightarrow (a) and (a) \Leftrightarrow (d).

Proof of Theorem 4.4. (a) \Rightarrow (b). Let F be the feedback such that $T_F(t)V \subset V$. Define $\hat{x}(t) = \int_0^t T_F(s)x_0 ds$ and $\hat{u}(t) = F \int_0^t T_F(s)x_0 ds$. From Curtain and Pritchard [5] we have that $\int_0^t T_F(s)x_0 ds \in D(A + BF) = D(A)$ for all $t \geq 0$, and

$$(4.5) \quad (A + BF) \int_0^t T_F(s)x_0 ds = T_F(t)x_0 - x_0.$$

Thus

$$\begin{aligned} \dot{\hat{x}}(t) &= x_0 + A\hat{x}(t) + B\hat{u}(t), \quad t \geq 0, \\ \hat{x}(0) &= 0 \end{aligned}$$

and so $\hat{x}(t)$ satisfies (2.2). Furthermore

$$\lim_{t \downarrow 0} \hat{u}(t) = \lim_{t \downarrow 0} F \int_0^t T_F(s)x_0 ds = \lim_{t \downarrow 0} F(\lambda - A - BF)^{-1}(\lambda - A - BF) \int_0^t T_F(s)x_0 ds,$$

where λ is an element of the resolvent set of $A + BF$. From Lemma A10, $F(\lambda - A - BF)^{-1}$ is a bounded operator; thus with (4.5) we have

$$\lim_{t \downarrow 0} F \int_0^t T_F(s)x_0 \, ds = F(\lambda - A - BF)^{-1} \lim_{t \downarrow 0} (\lambda - A - BF) \int_0^t T_F(s)x_0 \, ds = 0.$$

With a similar argument it can be proved that $\hat{u}(t) \in C([0, \infty); U)$ and $\|\hat{u}(t)\| \leq Me^{\alpha t}$. Defining $x(t) = T_F(t)x_0$ and $u(t)$ to be the distributional derivative of $\hat{u}(t)$ gives the desired result.

(b) \Rightarrow (c). Let $\hat{x}(t)$ and $\hat{u}(t)$ denote, respectively, $\mathbb{1}_{[0, \infty)} * x(\cdot) = \int_0^t x(s) \, ds$ and $\mathbb{1}_{[0, \infty)} * u(t)$, where $*$ denotes the convolution product and $\mathbb{1}_{[0, \infty)}$ is the unit step in zero. From the definition of open-loop invariance we have that $\|\hat{u}(t)\| \leq Me^{\alpha t}$, so we may take the Laplace transform of (2.2), which gives

$$(4.6) \quad \mathcal{L}(\hat{x})(s) = \frac{1}{s} (s - A)^{-1} x_0 + (s - A)^{-1} B \mathcal{L}(\hat{u})(s),$$

where $\mathcal{L}(\hat{x})$ denotes the Laplace transform of \hat{x} .

If we define $\xi(s) = s\mathcal{L}(\hat{x})(s)$ and $\omega(s) = s\mathcal{L}(\hat{u})(s)$, then since $\hat{x}(t)$ and $\hat{u}(t)$ are exponentially bounded, $\xi(\cdot)$ and $\omega(\cdot)$ are analytic on some right half-plane of \mathbb{C} and (4.6) implies that

$$(4.7) \quad x_0 = (s - A)\xi(s) - B\omega(s).$$

With the definition of the Laplace transform we have that $\xi(s) \in V$, for s sufficiently large, and with Theorem 8.6-1 of Zemanian [21] we have that $\lim_{s \rightarrow \infty} \omega(s) = \lim_{t \downarrow 0} \hat{u}(t) = 0$, and Lemma 3.3 with (4.7) implies that V is frequency-invariant.

(c) \Rightarrow (a). Since V is a closed subspace, we see, by Corollary 3.7, that we may restrict our input operator B to \tilde{B} , such that $\mathcal{B}^0(V) = \tilde{\mathcal{B}}^0(V)$ and $\tilde{\mathcal{B}}^1(V) = \text{Im } \tilde{B} \cap V = \{0\}$ (see (3.2)); then V is also frequency-invariant for the system (A, \tilde{B}) . So we may assume without loss of generality that $\text{Im } B \cap V = \{0\}$. For all x in V we have

$$(4.8) \quad x = (s - A)\xi(s) - B\omega(s).$$

If s is larger than \hat{s} (see Lemma 3.10 and Definition 3.9), then (4.8) implies that

$$(4.9) \quad A\Xi_s \subset V + \text{Im } B.$$

From Lemma 3.11 and Lemma 3.12 we have that $\Xi_s = \Xi$. So (4.9) implies that

$$(4.10) \quad A\Xi \subset V + \text{Im } B.$$

Ξ is closed in the graph norm of A , and so from Lemma 4.1 we have the existence of an A -bounded F such that

$$(A + BF)\Xi \subset V.$$

By rearranging (4.8) we have

$$x = (s - A)\xi(s) - B\omega(s) = (s - A - BF)\xi(s) - B(\omega(s) - F\xi(s)).$$

Thus $B(\omega(s) - F\xi(s)) \in V$, and so by the assumption made in the beginning of this proof $\omega(s) = F\xi(s)$, and

$$(4.11) \quad x = (s - A - BF)\xi(s), \quad s \geq \hat{s}.$$

So $(s - A - BF)^{-1}x = \xi(s) \in \Xi \subset V$, for all x in V , and Theorem 4.3 concludes this proof.

(a) \Rightarrow (d). Let F be an A -bounded feedback law such that V is $T_F(t)$ -invariant. By Corollary 2.3 we may assume that $\text{Im } BF|_{V \cap D(A)} \subset \mathcal{B}^0$. Since V is closed-loop invariant there exists, by Theorem 4.3(d), a real \hat{s} such that

$$(4.12) \quad (s - A - BF)(V \cap D(A + BF)) = V$$

for all $s \geq \hat{s}$. Since $D(A + BF) = D(A)$ and $\text{Im } BF|_{V \cap D(A)} \subset \mathcal{B}^0$, (4.12) implies (4.3).

(d) \Rightarrow (a). Suppose that (d) of Theorem 4.4 holds; then this implies that

$$(4.13) \quad A(V \cap D(A)) \subset V + \mathcal{B}^0 = V + \text{Im } B.$$

Since V is closed we may apply Theorem 4.2 to conclude that there exists an A -bounded feedback law F such that

$$(4.14) \quad (A + BF)(V \cap D(A)) \subset V$$

and from the remark made on Definition 2.7 we may assume that $\text{Im } BF|_{V \cap D(A)} \subset \mathcal{B}^0$, or $\text{Im } BF|_{V \cap D(A)} \cap V = \{0\}$. Let s be an element of \mathbb{R} , larger than s_0 (see (4.3)) such that $[s, \infty)$ is in the resolvent set of $A + BF$. Then by (4.3) every x in V can be written as

$$(4.15) \quad x = (s - A)v + Bu$$

for some $v \in V \cap D(A)$ and $Bu \in \mathcal{B}^0$. We rewrite (4.15) as

$$x = (s - A - BF)v + Bu + BFv$$

or $Bu + BFv = x - (s - A - BF)v$. So from (4.14) we see that $Bu + BFv$ is an element of V , but this is only possible if $Bu + BFv = 0$. Since $\text{Im } BF|_{V \cap D(A)} \subset \mathcal{B}^0$ and $Bu \in \mathcal{B}^0$, thus we have

$$(4.16) \quad x = (s - A - BF)v.$$

By premultiplying (4.16) with $(s - A - BF)^{-1}$ we obtain $(s - A - BF)^{-1}V \subset V$, and together with Theorem 4.3 we conclude that V is $T_F(t)$ -invariant. \square

Proof of Theorem 4.6. (a) \Rightarrow (b). Let F be the feedback law such that $T_F(t)V \subset V$. Defining $x(t)$ as $T_F(t)x_0$ and $u(t)$ as $FT_F(t)x_0$, then Theorem 2.31 of Curtain and Pritchard [5] gives the desired result.

(b) \Rightarrow (c). Define $\xi(\cdot)$ as the Laplace transform of $x(t)$ and $\omega(\cdot)$ as the Laplace transform of $u(t)$. Then $\xi(\cdot)$, $\omega(\cdot)$ is a (ξ, ω) representation of x_0 , with $\xi(s) \in V$, and the Theorem 8.6-1 of Zemanian [21] we have that $\lim_{s \rightarrow \infty} s\omega(s) = \lim_{t \downarrow 0} u(t) = u(0)$, and Lemma 3.3 implies that $\lim_{s \rightarrow \infty} \xi(s) = 0$.

(c) \Rightarrow (a). From Theorem 4.4 we have the existence of an A -bounded feedback operator, F , such that $(s - A - BF)^{-1}V \subset V$ and $\text{Im } BF|_{V \cap D(A)} \subset \mathcal{B}^0$, where \mathcal{B}^0 is the subspace of $\text{Im } B$ as defined in (3.2). If we define for $x_0 \in V$ $\xi(s) := (s - A - BF)^{-1}x_0$ and $\omega(s) := F(s - A - BF)^{-1}x_0$, then $\xi(s)$, $\omega(s)$ is a (ξ, ω) representation with $B\omega(s) \in \mathcal{B}_+^0(V)$. Thus from Lemmas 3.5 and 3.6 we have that $\lim_{s \rightarrow \infty, s \in \mathbb{R}} s\omega(s)$ exists, and so for every x in V $\lim_{s \rightarrow \infty, s \in \mathbb{R}} sF(s - A - BF)^{-1}x$ exists. Let $\tilde{F}x$ denote the limit value; then \tilde{F} is a linear operator defined on V and on $D(A) \cap V$, \tilde{F} is equal to F . By the Uniform Boundedness Theorem we have that \tilde{F} , and with that F , is a bounded operator on V . The Hahn-Banach Theorem gives that F can be extended as a bounded operator to the whole of X . Lemma 2.2 concludes the proof. \square

5. Disturbance decoupling problem. In this section we shall consider the disturbance decoupling problem (DDP):

Given the system

$$(5.1) \quad \dot{x}(t) = Ax(t) + Bu(t) + Eq(t), \quad z(t) = Dx(t),$$

where A and B are the same as in (1.1) and E and D are bounded linear operators from, respectively, Q to X and X to Z , find an A -bounded feedback law such that, in the closed-loop system, $z(\cdot)$ does not depend on $q(\cdot)$.

The next theorem gives the link between closed-loop invariance and DDP. By $\mathcal{V}^*(K)$ we shall denote the largest closed-loop invariant subspace contained in K , where K is a closed linear subspace of X .

THEOREM 5.1. *If $\mathcal{V}^*(\text{Ker } D)$ exists, then DDP is solvable if and only if*

$$(5.2) \quad \mathcal{V}^*(\text{Ker } D) \supset \text{Im } E.$$

Proof. See Curtain [2] for the proof. \square

In Curtain [2] the question of sufficient conditions for the existence of $\mathcal{V}^*(K)$ is posed. We shall show that $\mathcal{V}^*(K)$ need not always exist. Note that if $\mathcal{V}^*(K)$ exists, then it must be closed, since $T_F(t)V \subset V$ implies that $T_F(t)\bar{V} \subset \bar{V}$.

Keeping the equivalence between closed-loop and frequency invariance in mind we define the natural candidate for $\mathcal{V}^*(K)$.

DEFINITION 5.2. Let $\mathcal{V}_\Sigma(K)$ be the subset of X which contains all $x \in X$ with a (ξ, ω) -representation with $\xi(\cdot)$ in $K_+(s)$ (see Definitions 3.1 and 3.2).

The next lemma will show that $\mathcal{V}_\Sigma(K)$ is the supremal frequency invariant subspace in the closed subspace K .

LEMMA 5.3. (a) *Every frequency invariant subspace contained in K is contained in $\mathcal{V}_\Sigma(K)$.*

(b) *Every closed-loop invariant subspace contained in K is contained in $\mathcal{V}_\Sigma(K)$.*

(c) $\mathcal{V}_\Sigma(K) \subset K$.

(d) $\mathcal{V}_\Sigma(K)$ is the supremal frequency invariant subspace, contained in K .

Proof. (a) The proof is obvious with the definition of $\mathcal{V}_\Sigma(K)$ and the definition of frequency invariance.

(b) The proof is obvious with Theorem 4.4.

(c) Let $(\xi(s), \omega(s))$ be a (ξ, ω) representation of x in $\mathcal{V}_\Sigma(K)$; then $\xi(s) \in K$ and with Lemma 3.3 and the fact that K is closed we conclude that $x = \lim_{s \rightarrow \infty, s \in \mathbb{R}} s\xi(s)$ is an element of K .

(d) Let x be an element of $\mathcal{V}_\Sigma(K)$, so that there exist strictly proper meromorphic functions $\xi(s)$ and $\omega(s)$ such that $\xi(s)$ is in K and

$$(5.3) \quad x = (s - A)\xi(s) - B\omega(s) \quad \text{for } \text{Re}(s) > c_x$$

(see Definition 3.1). Let \hat{s} be an arbitrary, but fixed, point of \mathbb{C} with $\text{Re}(\hat{s}) > c_x$; then by using (5.3) we get

$$(5.4) \quad \begin{aligned} 0 &= (s - A)\xi(s) + (\hat{s} - A)(-\xi(\hat{s})) - B\omega(s) + B\omega(\hat{s}) \\ &= (s - A)(\xi(s) - \xi(\hat{s})) + (s - \hat{s})\xi(\hat{s}) - B(\omega(s) - \omega(\hat{s})). \end{aligned}$$

So if $s \neq \hat{s}$, then (5.4) implies that

$$(5.5) \quad \xi(\hat{s}) = (s - A) \left(\frac{(\xi(s) - \xi(\hat{s}))}{\hat{s} - s} \right) - B \left(\frac{(\omega(s) - \omega(\hat{s}))}{\hat{s} - s} \right).$$

Define $\xi_1(s) = ((\xi(s) - \xi(\hat{s})) / (\hat{s} - s))$ if $s \neq \hat{s}$ and $-\xi'(\hat{s})$ if $s = \hat{s}$. Define $\omega_1(s) = ((\omega(s) - \omega(\hat{s})) / (\hat{s} - s))$ if $s \neq \hat{s}$ and $-\omega'(\hat{s})$ if $s = \hat{s}$. It is not hard to show that (ξ_1, ω_1) is a strictly proper (ξ, ω) representation of $\hat{x} = \xi(\hat{s})$. Since $\xi(s)$ is in K , $\xi_1(s)$ is in K . Thus by the definition of $\mathcal{V}_\Sigma(K)$, $\xi(\hat{s})$ is in $\mathcal{V}_\Sigma(K)$. Since \hat{s} was an arbitrary element, we may conclude that $\mathcal{V}_\Sigma(K)$ is frequency invariant. By (c) we obtain $\mathcal{V}_\Sigma(K) \subset K$, and (a) implies that $\mathcal{V}_\Sigma(K)$ is the largest frequency invariant subspace with this property. \square

With this lemma we can easily prove the next theorem.

THEOREM 5.4. *Let K be a closed subspace of X . If $\mathcal{V}_\Sigma(K)$ is closed, then $\mathcal{V}^*(K)$ exists and is equal to $\mathcal{V}_\Sigma(K)$.*

Proof. If $\mathcal{V}_\Sigma(K)$ is closed, then by Lemma 5.3(d) and Theorem 4.4 we conclude that $\mathcal{V}_\Sigma(K)$ is closed-loop invariant. Furthermore, from Lemma 5.3(a) and (c), it must be the largest with this property. So $\mathcal{V}^*(K)$ exists and is equal to $\mathcal{V}_\Sigma(K)$. \square

The question which now arises is under what conditions $\mathcal{V}_\Sigma(K)$ is closed. It turns out that this condition can be formulated in terms of $\text{Im } B$ and $\mathcal{V}_\Sigma(K)$, as in the next lemma.

LEMMA 5.5. *$\mathcal{V}_\Sigma(K)$ is closed if and only if $\mathcal{V}_\Sigma(K) \cap \text{Im } B = \overline{\mathcal{V}_\Sigma(K)} \cap \text{Im } B$.*

Proof. (Only if.) The proof is obvious.

(If) $\mathcal{V}_\Sigma(K)$ is frequency invariant and $\mathcal{V}_\Sigma(K) \cap \text{Im } B = \overline{\mathcal{V}_\Sigma(K)} \cap \text{Im } B$. So, by Lemma 3.6, we obtain that every x in $\mathcal{V}_\Sigma(K)$ has a unique (ξ, ω) -representation with

$$B\omega(s) = \sum_{i=1}^p b_i \omega_i(s),$$

$\omega_i(s)$ given by (3.9) and $\xi(s) = (s - A)^{-1}x + (s - A)^{-1}(\sum_{i=1}^p b_i \omega_i(s))(b_i; i = 1, \dots, p)$ is a basis for \mathcal{B}^0 ; the subspace of $\text{Im } B$ that is defined in (3.2). So \mathcal{B}^0 has a zero intersection with $\mathcal{V}_\Sigma(K)$ and is of maximal dimension under this restriction.)

We shall recall (3.9). Let $f_i, i = 1, \dots, p$ be those elements of X' such that $f_i|_{\overline{\mathcal{V}_\Sigma(K)}} = 0$ and $\langle f_i, b_j \rangle = \delta_{ij}$, $S(s)$ is a $p \times p$ matrix with $S(s)_{ij} = \langle f_i, s(s - A)^{-1}b_j \rangle$, and

$$(5.6) \quad \begin{pmatrix} \omega_1(s) \\ \vdots \\ \omega_p(s) \end{pmatrix} = S(s)^{-1} \begin{pmatrix} \langle f_1, s(s - A)^{-1}x \rangle \\ \vdots \\ \langle f_p, s(s - A)^{-1}x \rangle \end{pmatrix}.$$

Let x be an element of $\overline{\mathcal{V}_\Sigma(K)}$; then there exists a sequence $\{x_n\}$ in $\mathcal{V}_\Sigma(K)$ which converges to x . x_n is an element of $\mathcal{V}_\Sigma(K)$, so it has a (ξ, ω) representation with

$$\omega_n(s) = \begin{pmatrix} \omega_1^n(s) \\ \vdots \\ \omega_p^n(s) \\ \vdots \\ \omega_{p_1}^n(s) \end{pmatrix},$$

where p_1 is the dimension of $\text{Im } B$. With the above we may choose

$$\omega_{p+1}^n(s) = \dots = \omega_{p_1}^n(s) = 0 \quad \text{and} \quad \begin{pmatrix} \omega_1^n(s) \\ \vdots \\ \omega_p^n(s) \end{pmatrix}$$

as in (5.6) with x replaced by x_n . With this choice it can easily be seen that if x_n converges to x , then $\omega^n(s)$ converges, and

$$\omega_i(s) := \lim_{n \rightarrow \infty} \omega_i^n(s) = 0 \quad \text{if } i > p, \quad \text{and} \quad \begin{pmatrix} \omega_1(s) \\ \vdots \\ \omega_p(s) \end{pmatrix}$$

is the right-hand side of (5.6). From (5.6) it is easily seen that $\omega(s)$ is meromorphic on a right half-plane and

$$\lim_{s \rightarrow \infty} \omega(s) = S(\infty)^{-1} \begin{pmatrix} \langle f_1, x \rangle \\ \vdots \\ \langle f_p, x \rangle \end{pmatrix} = 0 \quad \text{since } x \in \overline{\mathcal{V}_\Sigma(K)}.$$

Define $\xi(s) = (s - A)^{-1}x - (s - A)^{-1}B\omega(s)$; then $\xi(s)$ is meromorphic, $\xi(s)$ is strictly proper, $x = (s - A)\xi(s) - B\omega(s)$, and $\xi(s)$ is the limit of

$$\xi_n(s) = (s - A)^{-1}x_n - (s - A)^{-1}B\omega_n(s),$$

as $n \rightarrow \infty$, thus it is in K .

So x has a (ξ, ω) representation, with $\xi(\cdot)$ in $K_+(s)$; thus x is in $\mathcal{V}_\Sigma(K)$. \square

Remark. There exist innumerable examples such that $V \cap \text{Im } B \neq \bar{V} \cap \text{Im } B$, for V not closed.

From this lemma we can deduce two corollaries.

LEMMA 5.6. *Let V_1 and V_2 be closed, controlled invariant subspaces of X . If $(V_1 + V_2) \cap \text{Im } B = \overline{(V_1 + V_2)} \cap \text{Im } B$, then $\overline{V_1 + V_2}$ is controlled invariant.*

Proof. Consider $\mathcal{V}_\Sigma(\overline{V_1 + V_2})$; then by Lemma 5.3(b), $V_1 \subset \mathcal{V}_\Sigma(\overline{V_1 + V_2})$ and $V_2 \subset \mathcal{V}_\Sigma(\overline{V_1 + V_2})$. So by the linearity of $\mathcal{V}_\Sigma(\overline{V_1 + V_2})$, $V_1 + V_2$ is contained in $\mathcal{V}_\Sigma(\overline{V_1 + V_2})$. By Lemma 5.3(c), $\mathcal{V}_\Sigma(\overline{V_1 + V_2}) \subset \overline{V_1 + V_2}$; thus

$$\overline{\mathcal{V}_\Sigma(\overline{V_1 + V_2})} = \overline{V_1 + V_2}.$$

$$\overline{\mathcal{V}_\Sigma(\overline{V_1 + V_2})} \cap \text{Im } B = \overline{V_1 + V_2} \cap \text{Im } B = (V_1 + V_2) \cap \text{Im } B$$

$$\subset \mathcal{V}_\Sigma(\overline{V_1 + V_2}) \cap \text{Im } B \subset \overline{\mathcal{V}_\Sigma(\overline{V_1 + V_2})} \cap \text{Im } B.$$

So by Lemma 5.5, $\mathcal{V}_\Sigma(\overline{V_1 + V_2})$ is closed; thus it is equal to $\overline{V_1 + V_2}$. By Theorem 5.4 we have that $\overline{V_1 + V_2}$ is controlled invariant. \square

LEMMA 5.7. (a) *Let $V_n, n \geq 1$ be a nest of closed, linear, and controlled-invariant subspaces. Then $V := \bigcup_{n \geq 1} V_n$ is controlled invariant if $(\bigcup_{n \geq 1} V_n) \cap \text{Im } B = \overline{(\bigcup_{n \geq 1} V_n)} \cap \text{Im } B$.*

(b) *Consider the same nest V_n . Let B be one-dimensional and suppose there exists a closed subspace K such that $V_n \subset K$ and $\langle T(t)|\text{Im } B \rangle$ is not contained in K . Then $V := \bigcup_{n \geq 1} V_n$ is controlled invariant if and only if $\text{Im } B = b \notin V$.*

Proof. (a) Let V be $\bigcup_{n \geq 1} V_n$; then V_n is closed-loop invariant and contained in V for all n ; thus by Lemma 5.3(b), V_n is contained in $\mathcal{V}_\Sigma(V)$. Since $\{V_n\}$ is a nest, we have $\bigcup_{n \geq 1} V_n \subset \mathcal{V}_\Sigma(V)$. By definition of $\mathcal{V}_\Sigma(V)$ we have that $\mathcal{V}_\Sigma(V) \subset V = \bigcup_{n \geq 1} V_n$, $\bigcup_{n \geq 1} V_n \subset \mathcal{V}_\Sigma(V) \subset V$. Thus $\mathcal{V}_\Sigma(V) = V$.

$$V \cap \text{Im } B = \left(\bigcup_{n \geq 1} V_n \right) \cap \text{Im } B = \left(\bigcup_{n \geq 1} V_n \right)$$

$$\cdot \cap \text{Im } B \subset \mathcal{V}_\Sigma(V) \cap \text{Im } B \subset \overline{\mathcal{V}_\Sigma(V)} \cap \text{Im } B = V \cap \text{Im } B.$$

So $\mathcal{V}_\Sigma(V) \cap \text{Im } B = \overline{\mathcal{V}_\Sigma(V)} \cap \text{Im } B$, and by Lemma 5.6 and Theorem 5.4, $\mathcal{V}_\Sigma(V) = V$ and it is controlled-invariant. \square

(b) (If) See (a).

(Only if) Suppose V is controlled invariant and $b \in V$; then V is also $T(t)$ invariant. Furthermore since K is closed we have that $V \subset K$. $\langle T(t)|\text{Im } B \rangle$ is the smallest $T(t)$ invariant subspace containing $\text{Im } B = \text{span } \{b\}$. So $\langle T(t)|\text{Im } B \rangle \subset V \subset K \rightarrow \leftarrow$ \square

The condition in Lemma 5.5 is not easy to check, and it is not obvious that it is fulfilled even if A is a bounded operator. The next lemma will show that if A is a bounded operator, then $\mathcal{V}_\Sigma(K)$ is closed.

LEMMA 5.8. *If A is a bounded linear operator, then $\mathcal{V}^*(K)$ exists and it is equal to $\mathcal{V}_\Sigma(K)$.*

Proof. First, suppose that the following holds:

$$(5.7) \quad A(\overline{\mathcal{V}_\Sigma(K)}) \subset \overline{\mathcal{V}_\Sigma(K)} + \text{Im } B.$$

Then with Schmidt and Stern [16], $\overline{\mathcal{V}_\Sigma(K)}$ is closed-loop invariant. So with Lemma 5.3, $\overline{\mathcal{V}_\Sigma(K)} = \mathcal{V}^*(K)$; however, $\mathcal{V}^*(K) \subset \mathcal{V}_\Sigma(K)$, and so $\mathcal{V}_\Sigma(K) = \overline{\mathcal{V}_\Sigma(K)}$. Thus it remains to prove (5.7).

Let x be an element of $\mathcal{V}_\Sigma(K)$; then there exist $\xi(\cdot) \in K_+(s)$ and $\omega(\cdot) \in U_+(s)$ such that

$$(5.8) \quad x = (s - A)\xi(s) - B\omega(s).$$

With Lemma 3.3 we have that $x = \lim_{s \rightarrow \infty, s \in \mathbb{R}} s\xi(s)$. By rearranging (5.8), we have

$$As\xi(s) = s^2\xi(s) - sx - Bs\omega(s).$$

From Lemma 5.3(c) it follows that $As\xi(s) \in \mathcal{V}_\Sigma(K) + \text{Im } B \subset \overline{\mathcal{V}_\Sigma(K)} + \text{Im } B$. Since A is a bounded operator and $s\xi(s) \rightarrow x$ as $s \rightarrow \infty$, we have that $Ax \in \overline{\mathcal{V}_\Sigma(K)} + \text{Im } B$. Thus

$$A\mathcal{V}_\Sigma(K) \subset \overline{\mathcal{V}_\Sigma(K)} + \text{Im } B.$$

Since A is a bounded operator and $\overline{\mathcal{V}_\Sigma(K)} + \text{Im } B$ is a closed subspace we have proved (5.7). \square

The next lemma will give sufficient conditions for Theorem 5.4, which are easy to verify. These conditions first occur in a slightly stronger form in Curtain [2], where the aim was to give sufficient conditions for the existence of $\mathcal{V}^*(K)$ for bounded feedback laws.

LEMMA 5.9. *Let $\text{Ker } D$ be the kernel of a bounded linear operator D . Then $\mathcal{V}_\Sigma(\text{Ker } D)$ is closed if either of the following conditions holds.*

(a) *There exists a q in $\mathbb{N} \cup \{0\}$ such that DA^i has a bounded extension from X to Z (denoted by $\overline{DA^i}$) for $0 \leq i \leq q$, $\overline{DA^i}B = 0$ for $0 \leq i < q$, and $\overline{DA^q}Bu \neq 0$ for all $u \neq 0$ in U .*

(b) *$D(s - A)^{-1}B = 0$ for all s in $[\hat{s}, \infty)$ for some \hat{s} in \mathbb{R} .*

Proof. (a) Suppose that $x \in \mathcal{V}_\Sigma(\text{Ker } D)$ and condition (a) holds. From Lemma 3.3 we have that $x = \lim_{s \rightarrow \infty, s \in \mathbb{R}} s\xi(s)$ if

$$(5.9) \quad x = (s - A)\xi(s) - B\omega(s).$$

Using the fact that $\omega(s)$ is strictly proper, Lemma 3.3, and (5.9), we obtain

$$(5.10) \quad \lim_{\substack{s \rightarrow \infty \\ s \in \mathbb{R}}} A\xi(s) = 0.$$

If we premultiply (5.9) by D we get, since $\xi(s) \in \text{Ker } D$ and $DB = 0$,

$$(5.11) \quad Dx = D(s - A)\xi(s) - DB\omega(s) = sD\xi(s) - DA\xi(s) - DB\omega(s) = -DA\xi(s).$$

Using the fact that $A\xi(s)$ is strictly proper and D is a bounded operator, we get $Dx = 0$ and $DA\xi(s) = 0$. By induction it is easy to prove that $\overline{DA^i}x = 0$ and $\overline{DA^{i+1}}\xi(s) = 0$ for $0 \leq i < q$. Yet we shall consider $i = q$:

$$(5.12) \quad \begin{aligned} \overline{DA^q}x &= \overline{DA^q}(s - A)\xi(s) - \overline{DA^q}B\omega(s) \\ &= s\overline{DA^q}\xi(s) - \overline{DA^q}(A\xi(s)) - \overline{DA^q}B\omega(s) \\ &= 0 - \overline{DA^q}(A\xi(s)) - \overline{DA^q}B\omega(s). \end{aligned}$$

Using the properness of $A\xi(s)$ and $\omega(s)$, we get that $\overline{DA^q}x = 0$ for all x in $\mathcal{V}_\Sigma(\text{Ker } D)$. With the fact that $\overline{DA^q}$ is a bounded operator, we obtain that $\overline{DA^q}x = 0$ for all $x \in \overline{\mathcal{V}_\Sigma(\text{Ker } D)}$. From the last line of (a) we obtain that $\text{Im } B \cap \overline{\mathcal{V}_\Sigma(\text{Ker } D)} = \{0\}$. So $\text{Im } B \cap \overline{\mathcal{V}_\Sigma(\text{Ker } D)} = \text{Im } B \cap \mathcal{V}_\Sigma(\text{Ker } D)$, and by Lemma 5.5 $\mathcal{V}_\Sigma(\text{Ker } D)$ is closed.

(b) In the case that $D(s-A)^{-1}B=0$, $\mathcal{V}_{\Sigma}(\text{Ker } D)$ is equal to the subset in X of all x in $\text{Ker } D$ with $D(s-A)^{-1}x=0$. This follows since for x in $\mathcal{V}_{\Sigma}(\text{Ker } D)$, $(s-A)^{-1}x=\xi(s)-(s-A)^{-1}B\omega(s)$, holds and thus

$$D(s-A)^{-1}x=D\xi(s)-D(s-A)^{-1}B\omega(s)=0-0=0.$$

Suppose $D(s-A)^{-1}x=0$; then $x=(s-A)(s-A)^{-1}x-B0$ is a strictly proper (ξ, ω) representation with $\xi(s)=(s-A)^{-1}x$ in $\text{Ker } D$. So $\text{Im } B \subset \mathcal{V}_{\Sigma}(\text{Ker } D)$, which implies that $\mathcal{V}_{\Sigma}(\text{Ker } D)$ is closed by Lemma 5.5. \square

Let us remark that condition (a) in Lemma 5.9 is weaker, even if the input and output space are one-dimensional, than the condition that the transfer function $D(s-A)^{-1}B$ has only finitely many zeros in plus infinity, as can be seen from the next example.

Let X be a Hilbert space and A a generator of a C_0 -semigroup on X . Assume furthermore that $b, d \in X$ with $b \in D(A)$, $d \notin D(A^*)$ (the domain of the adjoint of A), $\langle d, b \rangle_x = 0$ and $\langle d, Ab \rangle_x \neq 0$. Then $\langle d, A \cdot \rangle_x$ does not have a bounded extension from X to \mathbb{C} , but the multiplicity of the zeros in plus infinity of the transfer function $\langle d, (s-A)^{-1}b \rangle$ is 1.

Using the characterization of $\mathcal{V}_{\Sigma}(\text{Ker } D)$ given in Lemma 5.3 and Theorem 5.4 we derive the following equivalent statements for the solvability of DDP.

THEOREM 5.10. *Assume that $\mathcal{V}_{\Sigma}(\text{Ker } D)$ is closed; then the following statements are equivalent:*

(a) DDP is solvable.

(b) $\text{Im } E \subset \mathcal{V}_{\Sigma}(\text{Ker } D)$.

(c) For every q in Q there exists an $\omega(\cdot)$ in $U_+(s)$ (see Definition 3.1) such that

$$(5.13) \quad D(s-A)^{-1}Eq = -D(s-A)^{-1}B\omega(s).$$

(d) There exists a $U(s)$ in $(L(Q, U))_+(s)$ such that

$$(5.14) \quad D(s-A)^{-1}BU(s) = D(s-A)^{-1}E.$$

(e) There exists a $U(s)$ in $(L(Q, U))_+(s)$ and $X(s)$ in $(L(Q, X))_+(s)$ such that

$$(5.15) \quad \begin{bmatrix} s-A & B \\ D & 0 \end{bmatrix} \begin{bmatrix} X(s) \\ -U(s) \end{bmatrix} = \begin{bmatrix} E \\ 0 \end{bmatrix}.$$

Proof. This follows trivially from Theorems 5.1 and 5.4 and Definitions 5.2 and 3.2. \square

Remark. Theorem 5.10 is the same as Theorems 3.4 and 3.6 in Hautus [9] for the finite-dimensional case.

We see from this theorem that, under an extra condition on the system, the solvability of DDP is equivalent to the solvability of a meromorphic matrix equation. We can easily show that the solvability of (5.14) is always a necessary condition for the solvability of DDP. However, it can be seen from Lemma 5.11(b) that it is not sufficient; of course in this case $\mathcal{V}_{\Sigma}(\text{Ker } D)$ will not be closed.

We shall conclude this section with some negative results on the existence of $\mathcal{V}^*(K)$ and closedness of $\mathcal{V}_{\Sigma}(K)$.

LEMMA 5.11. (a) $\mathcal{V}^*(K)$ need not exist.

(b) If $\mathcal{V}^*(K)$ exists, then $\mathcal{V}_{\Sigma}(K)$ need not be closed.

Proof. (a) See Pandolfi [14] for a counterexample for delay equations, or see Appendix B for a counterexample for partial differential equations. In the counterexample of Pandolfi the codimension of K is finite.

(b) Here we shall give only a sketch of the proof; for a precise proof we refer the reader to Appendix B.

Consider the system

$$\begin{aligned}
 \dot{x}_1(t) &= x_1(t-1) + x_2(t), \\
 \dot{x}_2(t) &= x_3(t) + x_2(t), \\
 \dot{x}_3(t) &= u(t) + q(t - \tfrac{1}{2}), \\
 y(t) &= x_1(t).
 \end{aligned}
 \tag{5.16}$$

The transfer function from input to output is

$$g_1(s) = \frac{1}{s(s - e^{-s})(s + 1)}, \tag{5.17}$$

and from disturbance to output it is

$$g_2(s) = \frac{e^{-s/2}}{s(s - e^{-s})(s + 1)}. \tag{5.18}$$

The usual realization of (5.16) over the state space \mathcal{M}^2 does not satisfy the conditions of (5.1). Therefore we shall construct another realization of (5.17) and (5.18) using the results of Curtain and Zwart [6].

From Curtain and Zwart [6] we have that (5.16) has a spectral realization (A, B, D, E) with state space l^2 , such that (A, B) is approximately controllable, (D, A) is initially observable, $\text{Im } B \subset D(A)$, $g_1(s) = D(s - A)^{-1}B$, and $g_2(s) = D(s - A)^{-1}E$. Since $g_1(s)$ has no zeros, we may conclude from Zwart [22] that $\mathcal{V}^*(\text{Ker } D) = \{0\}$. Thus DDP is not solvable; however, there exists a strictly proper function $U(s)$ such that $g_1(s)U(s) = g_2(s)$, $U(s) = e^{-s/2}$. So by Theorem 5.10, $\mathcal{V}_\Sigma(\text{Ker } D)$ cannot be closed in $X = l^2$. \square

Notice that we do not say anything about the more usual state space \mathcal{M}^2 , which might be different.

Remark. Even if we make $\mathcal{V}_\Sigma(\text{Ker } D)$ smaller by assuming the extra condition

$$\|\omega^{(n)}(s)\| \leq \frac{M(n-1)!}{(s-\alpha)^{n-1}}, \quad s > \alpha \quad \text{for some } M \text{ and } \alpha,$$

$\mathcal{V}_\Sigma(\text{Ker } D)$ still need not necessarily be closed, as can be seen from the same counterexample (see Appendix B).

6. Conclusions. In this paper we have shown that, for the system (1.1) under the extra conditions (C1) and (C2), there is equivalence between open- and closed-loop invariance and the concept of frequency invariance for closed subspaces. Using this last concept we define the largest frequency-invariant subspace contained in a given closed subspace, and as a corollary of the equivalence theorem we have that this largest frequency-invariant subspace is closed-loop invariant if and only if it is closed. Unfortunately this subspace need not necessarily be closed, even if the largest closed-loop invariant subspace exists. This motivates the following (still) open problem: under what conditions does the transfer function $f(s)$ have a realization (D, A, B) such that $f(s) = D(s - A)^{-1}B$, (C1) and (C2) hold, and the largest frequency-invariant subspace in $\text{Ker } D$ is closed?

Furthermore we have shown that, under the condition that $\mathcal{V}_\Sigma(\text{Ker } D)$ is a closed subspace, the solvability of the disturbance decoupling problem is equivalent to the solvability of a meromorphic matrix equation, and if $\mathcal{V}_\Sigma(\text{Ker } D)$ is not closed, then this equivalence does not necessarily hold.

The results as presented in this paper are of a more general nature than those in Zwart [22]; this was restricted to a subclass of spectral systems. In that paper necessary and sufficient conditions were given for the existence of the largest closed-loop (with bounded feedback law) invariant subspace in a closed subspace. These conditions were given in terms of the zeros of the transfer function; however, it was shown there that not only were the zeros in infinity important for the existence of $\mathcal{V}^*(\text{Ker } D)$, but the position of the finite zeros with respect to the poles of the system determined the existence of $\mathcal{V}^*(\text{Ker } D)$ (compare this with Lemma 5.9(a)).

We remark here without proof that theorems similar to those given in § 5 hold if we were to define $\mathcal{V}_\Sigma(K)$ as in Definition 5.2 with the extra condition that $\lim_{s \rightarrow \infty, s \in \mathbb{R}} s\omega(s)$ exists and $\mathcal{V}^*(K)$ is the largest closed-loop invariant subspace in K with a bounded feedback law (compare Theorems 4.4 and 4.6). Furthermore none of the results mentioned in this paper would change if we were to replace the definition of $Y_+(s)$ (Definition 3.1) by $Y_+(s) = \{f: \mathbb{R} \rightarrow Y \mid f \text{ is continuous on a interval } (r, \infty) \text{ and } \lim_{s \rightarrow \infty} f(s) = 0\}$ or if we were to replace the definition of open-loop invariance, Definition 2.5, by the existence of a distribution $u(\cdot)$ with values in U and its support contained in $[0, \infty)$ such that $\lim_{t \downarrow 0} \{\mathbb{1}_{[0, \infty)} * u\}(t) = 0$, for some M and α in \mathbb{R} , $|\{\mathbb{1}_{[0, \infty)} * u\}(t)| \leq Me^{\alpha t}$ almost everywhere on $[0, \infty)$ and the mild solution of (2.2) remains in V . Omitting the condition that $|\{\mathbb{1}_{[0, \infty)} * u\}(t)| \leq Me^{\alpha t}$ on $[0, \infty)$ for some M and α would give a more natural definition of open-loop invariance. However except for the case that the input space is one-dimensional we do not have a proof that this definition is equivalent to our present definition of open-loop invariance. Since the aim of this paper is to prove equivalence between open- and closed-loop invariance, we have not used this definition.

Theorem 5.4 concerning the existence of the largest closed-loop invariant subspace is very similar to Theorem 5.4 of Hautus [10], where controlled invariance is considered for systems over rings. As in systems over rings we have that the sum of two controlled-invariant subspaces need not necessarily be controlled-invariant.

Appendix A: A-bounded operators.

DEFINITION A1. Let A be a closed linear operator on a Banach space X with the resolvent set of A $\rho(A) \neq \emptyset$. A linear operator Q from X to a Banach space Z is A -bounded if $D(Q) \supset D(A)$ and there are constants a, b such that

$$\|Qx\|_Z \leq a\|x\|_X + b\|Ax\|_X \quad \text{for all } x \text{ in } D(A).$$

LEMMA A2. Let A be a closed linear operator with $\rho(A) \neq \emptyset$. For a linear operator Q from X to Z the following three properties are equivalent:

- (a) Q is A -bounded.
- (b) There exists a λ in $\rho(A)$ such that $Q(\lambda - A)^{-1}$ is in $L(X, Z)$.
- (c) The operator \mathbb{F} is in $L(W, Z)$, where $W = \{(x, Ax); x \in D(A)\}$, $\|(x, Ax)\|_W = \|x\|_X + \|Ax\|_X$ and $\mathbb{F}((x, Ax)) = Qx$.

Proof. (a) \Leftrightarrow (b). See Kato [11, p. 245] for the proof.

(c) \Leftrightarrow (a). See Kato [11, Remark 1.4, p. 191] for the proof. \square

DEFINITION A3. Let A be a closed linear operator with $\rho(A) \neq \emptyset$. A linear operator Q from X to Z is A -compact if $D(Q) \supset D(A)$ and if for any sequence $\{u_n\} \subset D(A)$ with both $\{u_n\}$ and $\{Au_n\}$ bounded, $\{Qu_n\}$ contains a convergent subsequence.

LEMMA A4. Suppose that A is a closed linear operator with $\rho(A) \neq \emptyset$. A linear operator Q from X to Z is A -compact if and only if $Q(\lambda - A)^{-1}$ is a compact operator from X to Z for some $\lambda \in \rho(A)$.

Proof. See Kato [11, Remark IV.1.12] for the proof. \square

COROLLARY A5. *Let A be the same as in Definition A1. An A -bounded operator Q with finite-dimensional range is A -compact.*

DEFINITION A6. Let A be a closed linear operator on a Banach space X with $\rho(A) \neq \emptyset$. The A -bound of an A -bounded operator Q is

$$\inf\{b \in [0, \infty) \mid \text{there exists an } a \in [0, \infty) \text{ such that}$$

$$\|Qx\| \leq a\|x\| + b\|Ax\| \text{ for all } x \text{ in } D(A)\}.$$

LEMMA A7. *Let A be the same as in Definition A1. If Q is an A -bounded operator with finite-dimensional range, then the A -bound of Q is zero.*

Proof. See Kato [11, p. 195] for the proof. \square

LEMMA A8. *Suppose A is a closed linear operator with $\rho(A) \neq \emptyset$, and let Q be an A -bounded operator. If $0 \in \rho(A)$ and $\|QA^{-1}\| < 1$, then 0 is an element of $\rho(A+Q)$ and $(A+Q)^{-1}$ is given by*

$$(A1) \quad (A+Q)^{-1} = A^{-1} \sum_{k=0}^{\infty} ((-Q)(A^{-1}))^k.$$

Remark. Equation (A1) is known as the second von Neuman series.

COROLLARY A9. *If A generates a C_0 -semigroup $T(t)$, then there are constants M and α such that $\|T(t)\| \leq Me^{\alpha t}$. Let Q be an A -bounded operator with A -bound b_0 , such that $b_0 + Mb_0 < 1$. Let V be a closed linear subspace such that $T(t)V \subset V$ and $Q(V \cap D(A)) \subset V$. Then there exists a λ_0 in \mathbb{R} such that for all λ , $\lambda \geq \lambda_0$, $(\lambda - A - Q)^{-1}$ exists and $(\lambda - A - Q)^{-1}V \subset V$.*

Proof. For all $\lambda > \alpha$ we have that $\lambda \in \rho(A)$ and $\|(\lambda - A)^{-1}\| \leq M/(\lambda - \alpha)$. Q has A -bound b_0 with $b_0 + Mb_0 < 1$. By the definition of the A -bound there exists a pair (a, b) such that $\|Qx\| \leq a\|x\| + b\|Ax\|$ and $b + Mb < 1$.

$$\begin{aligned} \|Q(\lambda - A)^{-1}x\| &\leq a\|(\lambda - A)^{-1}x\| + b\|(\lambda - A)^{-1}Ax\| \leq a\|(\lambda - A)^{-1}\|\|x\| + b\|x\| \\ &\quad + b\|\lambda\| \|(\lambda - A)^{-1}\|\|x\| \leq ((aM/(\lambda - \alpha)) + b + (b\|\lambda\|M/(\lambda - \alpha)))\|x\|. \end{aligned}$$

So for λ sufficiently large we have that $\|Q(-\lambda + A)^{-1}\| < 1$. With Lemma A8 this implies that $0 \in \rho((-\lambda + A) + Q)$ and $(\lambda - A - Q)^{-1} = (\lambda - A)^{-1} \sum_{k=0}^{\infty} ((Q)((\lambda - A)^{-1}))^k$. By the assumptions we have that $(Q)((\lambda - A)^{-1})V \subset V$, so $(\lambda - A - Q)^{-1}V \subset V$. \square

LEMMA A10. *If Q_1 and Q_2 are A -bounded operators with the A -bound of Q_1 , b_1 less than 1, then Q_2 is also $(A + Q_1)$ -bounded. $D(A + Q_1) := D(A)$.*

Proof. Since Q_1 is A -bounded with A -bound less than 1, there are constants (a, b) such that $\|Q_1x\| \leq a\|x\| + b\|Ax\|$ for all x in $D(A)$. If $x \in D(A)$, then $\|Ax\| = \|(A + Q_1)x - Q_1x\| \leq \|(A + Q_1)x\| + \|Q_1x\|$ or

$$(A2) \quad \|(A + Q_1)x\| \geq \|Ax\| - \|Q_1x\| \geq -a\|x\| + (1 - b)\|Ax\|.$$

Since b is less than 1 we obtain from (A2)

$$(A3) \quad \|Ax\| \leq \frac{1}{1-b} \{ \|(A + Q_1)x\| + a\|x\| \}.$$

Q_2 is A -bounded, so there are constants (a_2, b_2) such that

$$\begin{aligned} \|Q_2x\| &\leq a_2\|x\| + b_2\|Ax\| \stackrel{A.3}{\leq} a_2\|x\| + \frac{b_2}{1-b} \{ \|(A + Q_1)x\| + a\|x\| \} \\ &= \left(a_2 + \frac{ab_2}{1-b} \right) \|x\| + \frac{b_2}{1-b} \|(A + Q_1)x\|. \end{aligned}$$

Therefore Q_2 is $A + Q_1$ -bounded. \square

Appendix B: Proofs.

Proof of Lemma 2.2. (If) V is $T_{F_i}(t)$ invariant if and only if $(s - A - BF_i)^{-1}V \subset V$ (Theorem 4.3), $i = 1, 2$. By Lemma A8 there exists a real \hat{s} in the resolvent set of $(A + BF_1)$ and of $(A + BF_2)$ such that $[\hat{s}, \infty)$ is contained in the resolvent set of $(A + BF_1)$ and of $(A + BF_2)$, and

$$(B1) \quad (s - A - BF_2)^{-1} = (s - A - BF_1)^{-1} \sum_{k=0}^{\infty} ((BF_2 - BF_1)(s - A - BF_1)^{-1})^k$$

for all s in $[\hat{s}, \infty)$. From (B1) it is easy to see that $(s - A - BF_2)^{-1}V \subset V$ for all $s \geq \hat{s}$. So by Theorem 4.3, $T_{F_2}(t)V \subset V$.

Remark. (C2) was used in (B1), that is, $D(A) = D(A + BF_i)$.

(Only if) It is not hard to show that if a closed subspace V is $T_{F_i}(t)$ -invariant, then $(A + BF_i)(V \cap D(A + BF_i)) \subset V$, $i = 1, 2$. Let V be $T_{F_i}(t)$ -invariant; then

$$(B2a) \quad (A + BF_1)(V \cap D(A + BF_1)) \subset V,$$

$$(B2b) \quad (A + BF_2)(V \cap D(A + BF_2)) \subset V.$$

By simply subtracting (B2a) and (B2b) and using the fact that $D(A + BF_1) = D(A + BF_2) = D(A)$, we see that $\text{Im } B(F_1 - F_2)|_{V \cap D(A)} \subset V$. \square

Proof of Lemma 3.5. We shall give the proof of both assertions in one. Let τ be zero or 1, depending on the problem we are dealing with, and assume that every x in V has a (ξ, ω) representation with $\lim_{s \rightarrow \infty, s \in \mathbb{R}} s^\tau \omega(s)$ exists.

Let $b_1 \cdots b_p$ be a basis for $\mathcal{B}^0(V)$, the subspace of $\text{Im } B$ defined by (3.2), and let $b_{p+1} \cdots b_{p_1}$ be a basis for $\mathcal{B}^1(V) := \text{Im } B \cap V$. Then, since V is a subspace, $b_1 \cdots b_{p_1}$ is a basis for $\text{Im } B$.

Since V is frequency invariant and $\mathcal{B}^1(V) \subset V$, we have that there exist pairs $(\xi_i(s), \omega_i(s))$ such that $\xi_i(\cdot) \in V_+(s)$, $\omega_i(\cdot) \in U_+(s)$ and

$$(B3) \quad b_i = (s - A)\xi_i(s) - B\omega_i(s), \quad i = p + 1, \dots, p_1.$$

If we write

$$\omega_i(s) = \begin{pmatrix} \omega_{i1}(s) \\ \vdots \\ \omega_{ip_1}(s) \end{pmatrix},$$

then (B3) becomes

$$(B4) \quad b_i = (s - A)\xi_i(s) - \sum_{j=1}^{p_1} b_j \omega_{ij}(s), \quad i = p + 1, \dots, p_1.$$

We can write (B4) in matrix notation, and obtain

$$(B5) \quad \begin{pmatrix} b_{p+1} \\ \vdots \\ b_{p_1} \end{pmatrix} = Q(s) \begin{pmatrix} b_{p+1} \\ \vdots \\ b_{p_1} \end{pmatrix} + R(s) \quad \text{where } Q_{ij}(s) = -\omega_{ij}(s),$$

$$R_i(s) = (s - A)\xi_i(s) - \sum_{j=1}^p b_j \omega_{ij}(s).$$

Since $\omega_i(\cdot)$ is in $U_+(s)$, $Q(s)$ is a meromorphic function and $\lim_{s \rightarrow \infty, s \in \mathbb{R}} Q(s) = 0$. By rearranging (B5) we have

$$\begin{aligned}
 \begin{pmatrix} b_{p+1} \\ \vdots \\ b_{p_1} \end{pmatrix} &= (I - Q(s))^{-1} R(s) \\
 \text{(B6)} \quad &= (I - Q(s))^{-1} (s - A) \begin{pmatrix} \xi_{p+1}(s) \\ \vdots \\ \xi_{p_1}(s) \end{pmatrix} - (I - Q(s))^{-1} \begin{pmatrix} \sum_{j=1}^p b_j \omega_{p+1j}(s) \\ \vdots \\ \sum_{j=1}^p b_j \omega_{p_1j}(s) \end{pmatrix} \\
 &= (s - A)(I - Q(s))^{-1} \begin{pmatrix} \xi_{p+1}(s) \\ \vdots \\ \xi_{p_1}(s) \end{pmatrix} - \sum_{j=1}^p b_j (I - Q(s))^{-1} \begin{pmatrix} \omega_{p+1j}(s) \\ \vdots \\ \omega_{p_1j}(s) \end{pmatrix},
 \end{aligned}$$

by simple linear algebra.

This last formula implies that each b_i , $p+1 \leq i \leq p_1$, has a (ξ, ω) representation with $\hat{\xi}_i(\cdot)$ in $V_+(s)$ and $B\hat{\omega}_i(\cdot)$ in $\mathcal{B}_+^0(V)$, if we take $\hat{\xi}_i(s)$ to be the i th row of

$$(I - Q(s))^{-1} \begin{pmatrix} \xi_{p+1}(s) \\ \vdots \\ \xi_{p_1}(s) \end{pmatrix}$$

and $B\hat{\omega}_i(\cdot)$ to be the i th row of

$$\sum_{j=1}^p b_j (I - Q(s))^{-1} \begin{pmatrix} \omega_{p+1j}(s) \\ \vdots \\ \omega_{p_1j}(s) \end{pmatrix}.$$

Furthermore since

$$Q(\infty) = 0, \quad \text{then} \quad \lim_{\substack{s \rightarrow \infty \\ s \in \mathbb{R}}} s^\tau (\hat{\omega}_i(s)) = \lim_{\substack{s \rightarrow \infty \\ s \in \mathbb{R}}} s^\tau \begin{pmatrix} \omega_{i,1}(s) \\ \vdots \\ \omega_{i,p}(s) \end{pmatrix},$$

and by assumption this limit exists.

So we have proved the assertion of Lemma 3.5 for all elements in $\text{Im } B \cap V = \mathcal{B}^1(V)$. Let x be an arbitrary element of V ; then there exists a pair $(\xi(s), \omega(s))$, $\xi(\cdot) \in V_+(s)$ and $\omega(\cdot) \in U_+(s)$ such that

$$\begin{aligned}
 x &= (s - A)\xi(s) - B\omega(s) \\
 &= (s - A)\xi(s) - \sum_{i=1}^p b_i \omega_i(s) - \sum_{i=p+1}^{p_1} b_i \omega_i(s) \\
 \text{(B7)} \quad &= (s - A)\xi(s) - \sum_{i=p+1}^{p_1} ((s - A)\hat{\xi}_i(s) - B\hat{\omega}_i(s))\omega_i(s) - \sum_{i=1}^p b_i \omega_i(s) \\
 &= (s - A) \left(\xi(s) - \sum_{i=p+1}^{p_1} \hat{\xi}_i(s) \omega(s) \right) - \sum_{i=p+1}^{p_1} (B\hat{\omega}_i(s))\omega_i(s) - \sum_{i=1}^p b_i \omega_i(s);
 \end{aligned}$$

$\xi_i(s), \hat{\xi}_i(s) \in V_+(s)$ and $B\hat{\omega}_i(\cdot) \in \mathcal{B}_+^0(s)$; $\mathcal{B}^0 = \text{span}\{b_1, \dots, b_p\}$.

Thus x has a (ξ, ω) representation with $\tilde{\xi}(s) = \xi(s) - \sum_{i=p+1}^{p_1} \hat{\xi}_i(s) \omega(s) \in V_+(s)$, $B\tilde{\omega}(s) = \sum_{i=p+1}^{p_1} (B\hat{\omega}_i(s))\omega_i(s) - \sum_{i=1}^p b_i \omega_i(s) \in \mathcal{B}_+^0(s)$ and $\lim_{s \rightarrow \infty, s \in \mathbb{R}} s^\tau \tilde{\omega}(s)$ exists. \square

Proof of Lemma 4.1. Define X_A to be the graph of A , with the graph norm $\|(x, Ax)\|_A = \|x\| + \|Ax\|$, where $\|\cdot\|$ is the norm of X . A is a bounded operator from X_A to X , and V_1 is by definition closed in X_A .

If $v_1 \in V_1$, then there exist $v_2 \in V_2$ and $u \in U$ such that $Av_1 = v_2 + Bu$, and u and v_2 are uniquely determined if we assume that $u \in [\text{Ker } B]^\perp$ (the annihilator of $\text{Ker } B$) and $Bu \in \mathcal{B}^0(V_2)$.

Let F be defined by $Fv_1 = -u$, for all $v_1 \in V_1$. The operator F is linear, since u is uniquely determined. We shall show that F is a closed operator in X_A . Let us assume that (v_1^n, Fv_1^n) converges to $(v_1, -u)$; thus $v_1^n \rightarrow v_1$ and $Av_1^n \rightarrow Av_1$. We must prove that $Fv_1 = -u$; this is obvious since

$$w^n := Av_1^n - Bu^n = Av_1^n + BFv_1^n \text{ converges to } Av_1 - Bu =: w$$

and $v_1 \in V_1$ since V_1 is closed, and $Bu \in \mathcal{B}^0(V_2)$ since $\mathcal{B}^0(V_2)$ is closed.

Thus F is a closed operator from the whole V_1 , with induced norm of $\|\cdot\|_A$, to U . By the Closed Graph Theorem, F is a bounded operator on V_1 , norm $\|\cdot\|_A$. By the Hahn-Banach Theorem and the fact that, since $\text{Im } B$ is finite-dimensional, F has finite-dimensional range, F has a bounded extension on X_A . By Lemma A2, F is A -bounded on X . \square

Proof of Lemma 5.11. (a) The next example will show that $\mathcal{V}^*(K)$ need not exist.

Example. Let X be $L^2(0, 1)$ and A be the "heat operator," $A = d^2/dz^2$ with domain, $D(A) = \{x \in L^2(0, 1) | x'' \in L^2(0, 1); x(0) = x(1) = 0\}$,

$$B := b(z) = \begin{cases} 0, & z \in [0, \frac{1}{2}], \\ \sin 2\pi z, & z \in [\frac{1}{2}, 1], \end{cases}$$

$$K = \{x \in L^2(0, 1) | x(z) = 0; \text{ a.e. on } [0, \frac{1}{2}]\}.$$

Notice that $K \supset \text{Im } B$. So $\mathcal{B}^0(K) = \{0\}$, and with Lemma 3.6 we may conclude that, if K is frequency invariant, then every x in K has a unique (ξ, ω) representation with $B\omega(s) = 0$, which means that $(s - A)^{-1}K \subset K$. But this last inclusion would imply by Theorem 4.3 that $A^{-1}b = A^{-1}1_{[1/2, 1]}(z)\pi^2 \sin(\pi z)$ is an element of K . However,

$$A^{-1}(1_{[1/2, 1]}(z)\pi^2 \sin(\pi z)) = \begin{cases} -z, & z \in [0, \frac{1}{2}] \\ -z + 1 - \sin(\pi z), & z \in [\frac{1}{2}, 1] \end{cases} \text{ is not in } K.$$

So K cannot be frequency- (or equivalently closed-loop) invariant. Define for $n > 1$

$$e_n(z) = \begin{cases} 0 & z \in [0, \frac{1}{2}], \\ (1/4\pi^2 n(-1)^n(n^2 - 1))\{\sin 2\pi n z + n(-1)^n \sin 2\pi z\}, & z \in [\frac{1}{2}, 1], \end{cases}$$

and μ_n in \mathbb{C} by $\mu_n = -4\pi^2 n^2$.

Then (1) $\text{span}_{i=2, \dots, n} \{e_i\}$ is closed-loop invariant for all $n \in \mathbb{N}$;

(2) $\text{span}_{i \in \mathbb{N}/\{1\}} \{e_i\}$ is K .

Proof of (1). By a simple calculation we can show that $(A - \mu_n)e_n = b$. So $(s - A)e_n = -b + (s - \mu_n)e_n$ or $e_n = (s - A(e_n/s - \mu_n)) - b/(1/(\mu_n - s))$. Thus $\text{span}\{e_n\}$ is closed-loop invariant. By Lemma 5.6 we have that $\text{span}_{i=2, \dots, n} \{e_i\}$ is closed-loop invariant for all $n \in \mathbb{N}$.

Proof of (2). If $x \in K$, then

$$x = \begin{cases} 0 & \text{on } [0, \frac{1}{2}], \\ x' & \text{on } [\frac{1}{2}, 1] \end{cases}$$

with $x' \in L^2(\frac{1}{2}, 1)$. So $x'(z) = \sum_{n=1}^{\infty} \langle x'(z), 2 \sin(2\pi n z) \rangle_{L^2(1/2, 1)} 2 \sin(2\pi n z)$ and $\|x'\|^2 = \sum_{n=1}^{\infty} |\langle x'(z), 2 \sin(2\pi n z) \rangle_{L^2(1/2, 1)}|^2 < \infty$. Let $x \perp e_n$ for all $n > 1 \Rightarrow \langle x, e_n \rangle_{L^2(0, 1)} = 0 \Rightarrow \langle x', \sin 2\pi n z + n(-1)^n \sin 2\pi z \rangle_{L^2(1/2, 1)} = 0$. So

$$(B8) \quad \langle x', \sin 2\pi n z \rangle_{L^2(1/2, 1)} = -n(-1)^n \langle x' \sin 2\pi z \rangle_{L^2(1/2, 1)} \quad \forall n > 1,$$

$$(B9) \quad \infty > \|x'\|^2 \geq \sum_{n=2}^{\infty} |\langle x', \sin 2\pi n z \rangle_{L^2(1/2,1)}|^2 = \sum_{n=2}^{\infty} n^2 |\langle x', \sin 2\pi z \rangle_{L^2(1/2,1)}|^2.$$

Formula (B9) implies that $\langle x', \sin 2\pi z \rangle = 0$, and again with (B9) we have that $\langle x', \sin 2\pi n z \rangle_{L^2(1/2,1)} = 0$, for all n in $\mathbb{N}/\{0\}$, so $x' = 0$. Thus $x \equiv 0$ is the only vector in K perpendicular on all e_n , so $\text{span}_{i \in \mathbb{N}/\{1\}} \{e_i\}$ is K .

If $\mathcal{V}^*(K)$ were to exist, then it would necessarily be closed and contained in K , and by (1) it would have to contain $\text{span}_{i=2 \dots n} \{e_i\}$, for all $n \in \mathbb{N}$. This together with (2) would imply that $\mathcal{V}^*(K)$ equals K . This contradicts the fact that K is not closed-loop invariant. Thus $\mathcal{V}^*(K)$ cannot exist in this example. \square

Proof of Lemma 5.11(b). Consider the system (5.16) with transfer function $g_1(s)$ and $g_2(s)$. From Theorem 2.5 of Curtain and Zwart [6] we have that

$$(B10) \quad g(s) = \sum_{j=1}^{\infty} \frac{1}{z_j(1+z_j)^2(s-z_j)} + \frac{-1}{s} + \frac{1}{(1+e)(s+1)},$$

$$(B11) \quad g_2(s) = \sum_{j=1}^{\infty} \frac{1}{\sqrt{z_j}(1+z_j)^2(s-z_j)} + \frac{-1}{s} + \frac{\sqrt{e}}{(1+e)(s+1)},$$

where z_j is the j th zero of $s - e^{-s}$. With relations (B10) and (B11) we can make the following realization of (5.16): $X = l_2$, $A = \sum_{j=1}^{\infty} \lambda_j \langle \cdot, f_j \rangle f_j$, where f_j is j th orthogonal basis vector of l_2 , $\lambda_1 = 0$, $\lambda_2 = -1$ and $\lambda_{j+2} = z_j$, $j \geq 1$. For the domain of A we take all elements x of l_2 such that $\sum_{j=1}^{\infty} |\lambda_j \langle \cdot, f_j \rangle|^2 < \infty$. Furthermore we define

$$B = \sum_{j=1}^{\infty} b_j f_j, \quad b_j = 1, \quad j = 1, 2, \quad b_{j+2} = \frac{1}{(1+z_j)^{7/4}},$$

$$D = \langle \cdot, d \rangle, \quad d = \sum_{j=1}^{\infty} d_j f_j, \quad d_1 = -1, \quad d_2 = \frac{1}{(1+e)}, \quad d_{j+2} = \frac{1}{z_j(1+z_j)^{1/4}},$$

$$E = \sum_{j=1}^{\infty} \alpha_j f_j, \quad \alpha_1 = 1, \quad \alpha_2 = \sqrt{e}, \quad \alpha_{j+2} = \frac{\sqrt{z_j}}{(1+z_j)^{7/4}}.$$

Since $|z_j|$ is $O(j)$ we have that B , D , and E are bounded linear operators and $B \in D(A)$.

This realization satisfies (C1), and the conditions of Zwart [22]. We do not know if it satisfies (C2), therefore we cannot apply standard arguments. Of course we can define frequency invariance. Let V be a closed frequency invariant subspace in $\text{Ker } D$; we shall show that $V = \{0\}$. First we remark that $B \notin V$, since otherwise, by Lemma 3.5, every x in V can be written as $x = (s-A)\xi(s) - 0$. This in particular $(s-A)^{-1}B \in V \subset \text{Ker } D$, or $D(s-A)^{-1}B \equiv 0$, providing the contradiction.

So every x in V can be written as

$$(B12) \quad x = (s-A)\xi(s) + B\omega(s)$$

with $\omega(s) = \langle f, s(s-A)^{-1}x \rangle / \langle f, s(s-A)^{-1}B \rangle$, where $f \in X'$ such that $\langle f, B \rangle = 1$ and $f|_V = 0$ (as in the proof of Lemma 3.6). Let λ be an element of the resolvent set of A , $\rho(A)$. We shall prove that $(\lambda - A)(V \cap D(A))$ is closed-loop invariant with a bounded feedback law for the system $(A, (\lambda - A)B)$ and is contained in $\text{Ker } D(\lambda - A)^{-1}$. Then we are done, since the transfer function $D(\lambda - A)^{-1}(s-A)^{-1}(\lambda - A)B = D(s-A)^{-1}B = g_1(s)$ has no zeros, and so by Theorem 6.1 of Zwart [22], $(\lambda - A)(V \cap D(A)) = \{0\}$. Thus $V = \{0\}$, and since V is an arbitrary controlled invariant subspace in $\text{Ker } D$ we have that the largest controlled invariant subspace in $\text{Ker } D$ is $\{0\}$.

Thus it remains to show that $(\lambda - A)(V \cap D(A))$ is closed-loop invariant with a bounded feedback law. Let $(\lambda - A)x$ be an element of $(\lambda - A)(V \cap D(A))$, and since $B \in D(A)$ we have from (B12) that

$$(B13) \quad (\lambda - A)x = (s - A)\{(\lambda - A)\xi(s)\} + (\lambda - A)B\omega(s)$$

with $\xi(s)$ and $\omega(s)$ the same as in (B12). We shall show that $\lim_{s \rightarrow \infty} s\omega(s)$ exists. Let us first remark that since $x \in V \cap D(A)$ we have that

$$\langle f, s(s - A)^{-1}x \rangle = \langle f, (s - A + A)(s - A)^{-1}x \rangle = \langle f, A(s - A)^{-1}x \rangle = \langle f, (s - A)^{-1}Ax \rangle.$$

So

$$\begin{aligned} \lim_{\substack{s \rightarrow \infty \\ s \in \mathbb{R}}} s\omega(s) &= \lim_{\substack{s \rightarrow \infty \\ s \in \mathbb{R}}} s \frac{\langle f, (s - A)^{-1}x \rangle}{\langle f, s(s - A)^{-1}B \rangle} = \lim_{\substack{s \rightarrow \infty \\ s \in \mathbb{R}}} s \frac{\langle f, (s - A)^{-1}Ax \rangle}{\langle f, s(s - A)^{-1}B \rangle} \\ &= \lim_{\substack{s \rightarrow \infty \\ s \in \mathbb{R}}} \frac{\langle f, s(s - A)^{-1}Ax \rangle}{\langle f, s(s - A)^{-1}B \rangle} = \frac{\langle f, Ax \rangle}{\langle f, B \rangle} = \langle f, Ax \rangle. \end{aligned}$$

Since λ is an element of the resolvent set of A and V is a closed subspace we have that $(\lambda - A)(V \cap D(A))$ is a closed subspace too, and Theorem 4.6 gives the desired result. \square

Acknowledgment. I thank Ruth Curtain for her careful reading of this manuscript and for her valuable suggestions.

REFERENCES

- [1] G. BASILE AND G. MARRO, *Controlled and conditioned invariant subspaces in linear system theory*, J. Optim. Theory Appl., 3 (1969), pp. 306–315.
- [2] R. F. CURTAIN, *Invariance concepts in infinite dimensions*, SIAM J. Control Optim., 24 (1986), pp. 1009–1031.
- [3] ———, *(C, A, B)-pairs in infinite dimensions*, Systems Control Lett., 5 (1984), pp. 59–65.
- [4] ———, *Disturbance decoupling by measurement feedback with stability for infinite dimensional systems*, Internat. J. Control, 43 (1986), pp. 1723–1743.
- [5] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sci. 8, Springer-Verlag, Berlin, New York, 1978.
- [6] R. F. CURTAIN AND H. J. ZWART, *Spectral realisation for delay systems*, in Distributed Parameter Systems, Proc. Third International Conference on Control of Distributed Parameter Systems, Vorau, Austria, July 6–12, F. Kappel, K. Kunisch, and W. Schappacher, eds., Lecture Notes in Control and Information Sci. 102, Springer-Verlag, Berlin, New York, 1986, pp. 64–89.
- [7] G. DOETSCH, *Introduction to the Theory and Application of the Laplace Transformation*, Springer-Verlag, Berlin, New York, 1974.
- [8] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators—Part I: General Theory*, Interscience, New York, 1958.
- [9] M. L. J. HAUTUS, *(A, B)-invariant and stabilizability subspace, a frequency domain description*, Automatica, 16 (1980), pp. 703–707.
- [10] ———, *Controlled invariance in systems over rings*, Memorandum COSOR 82-01, Department of Mathematics and Computer Science, Eindhoven Univ. of Technology, Eindhoven, the Netherlands, 1982.
- [11] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, 1966.
- [12] T. G. KURTZ, *A general theorem on the convergence of operator semigroups*, Trans. Amer. Math. Soc., 148 (1980), pp. 23–32.
- [13] I. LASIECKA AND R. TRIGGIANI, *Finite rank, relatively bounded perturbations of semigroup generators, Part I*, Ann. Scuola Norm. Sup. Pisa, Cl. Sci. (4), 12 (1985), pp. 641–668.
- [14] L. PANDOLFI, *Disturbance decoupling and invariant subspaces for delay systems*, Appl. Math. Optim., 14 (1986), pp. 55–72.

- [15] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.
- [16] E. J. P. G. SCHMIDT AND R. J. STERN, *Invariance theory for infinite dimensional linear control systems*, Appl. Math. Optim., 6 (1980), pp. 113–122.
- [17] J. M. SCHUMACHER, *Algebraic characterizations of almost invariance*, Internat. J. Control, 38 (1983), pp. 107–124.
- [18] R. J. STERN, *Asymptotic holdability*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 1196–1198.
- [19] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, Berlin, New York, 1978.
- [20] J. ZABCZYK, *On decomposition of generators*, SIAM J. Control Optim., 16 (1978), pp. 523–534.
- [21] A. H. ZEMANIAN, *Distribution Theory and Transform Analysis, An Introduction to Generalized Functions, with Applications*, McGraw-Hill, New York, 1965.
- [22] H. J. ZWART, *Characterization of all controlled invariant subspaces for spectral systems*, SIAM J. Control Optim., 26 (1988), pp. 369–386.

EFFECTIVENESS AND ROBUSTNESS WITH RESPECT TO TIME DELAYS OF BOUNDARY FEEDBACK STABILIZATION IN ONE-DIMENSIONAL VISCOELASTICITY*

KENNETH B. HANNSGEN^{†‡}, YURIKO RENARDY^{†§}, AND ROBERT L. WHEELER^{†‡}

Abstract. The damping effect of a boundary feedback mechanism on torsional vibrations of a homogeneous viscoelastic rod is examined. A characteristic equation is derived for the oscillatory modes, and the solutions of this equation are studied by analytic and numerical methods, as functions of the feedback gain parameter and of a parameter for feedback delay. Results are compared to recent studies of elastic materials, where a feedback delay can cause exponential instability; here this phenomenon depends significantly on the short-time behavior of the viscoelastic memory kernel. Finally, an existence result is given, showing that the behavior of a weak solution corresponds in the expected way to the location of the characteristic roots.

Key words. boundary stabilization, viscoelasticity, time delays

AMS(MOS) subject classifications. primary 93D15; secondary 45K05

1. Introduction. We examine the damping effect of a boundary feedback mechanism on torsional vibrations of a homogeneous viscoelastic rod of uniform circular cross section. In the purely elastic case it is known [15] that this mechanism can force uniform exponential decay, but that time delays, which are inherent to physical feedback mechanisms, can introduce exponentially growing vibrations. When viscoelastic stress-strain laws are involved, exponential decay can occur even with fixed or free ends, but the decay can be slow in the low-frequency oscillating modes. We investigate whether boundary feedback can improve the rate of decay in these modes and under what conditions delays in the feedback can lead to unbounded vibrations. Analogous problems have been investigated for a variety of elastic bodies and structures [9]-[12], [14], [23]-[26], [28]; for the viscoelastic case, we have chosen a physical model that leads to the simplest equations of this type.

Specifically, we consider a homogeneous viscoelastic rod of length L and uniform circular cross section of radius R . For the constitutive law relating cross-sectional shear stress σ and shear strain γ at a point p in the rod, we use a linear Boltzmann model of the rate type [4] (see [29]):

$$(1.1) \quad \sigma(p, t) = G\gamma(p, t) + \int_0^\infty g(\tau) \frac{\partial}{\partial t} \gamma(p, t - \tau) d\tau.$$

The equilibrium stress modulus G is a positive constant (that is, the material is a solid), and $g(t)$ is completely monotonic and integrable on $(0, \infty)$ with $0 < g(0+) \leq \infty$.

We remark that complete monotonicity is a physically reasonable assumption that is satisfied by our examples. As noted below, some of our results (e.g., on the presence or absence of characteristic values in the right half-plane due to time delays) hold when g is merely positive, nonincreasing, and convex, with $-g'$ convex. By holding

* Received by the editors May 18, 1987; accepted for publication November 30, 1987. This research was partly supported by the Air Force Office of Scientific Research under grant AFOSR-86-0085.

[†] Department of Mathematics and Interdisciplinary Center for Applied Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-4097.

[‡] This research was partly supported by National Science Foundation grant DMS-8500947.

[§] This research was partly supported by National Science Foundation grant DMS-8615203 under the National Science Foundation Research Opportunities for Women Program and by the Defense Advanced Research Project Agency under grant F49620-87-C-0116.

throughout to the stronger hypothesis, we lose little generality overall and avoid repetitious statements of various lemmas for slightly different hypotheses on g .

Observe that if the strain history γ satisfies suitable decay conditions as $t \rightarrow -\infty$, then an integration by parts allows us to rewrite (1.1) as

$$(1.2) \quad \sigma(p, t) = G\gamma(p, t) + \frac{d}{dt} \int_0^t g(\tau) \gamma(p, t - \tau) d\tau + \int_t^\infty g'(\tau) \gamma(p, t - \tau) d\tau,$$

where the last integral on the right side of (1.2) is determined by the strain history for $t < 0$. Terms that arise from this contribution to the stress can be incorporated into a nonhomogeneous term. We do this throughout and make any required hypotheses directly of the resulting nonhomogeneous terms. As noted in § 5, our hypotheses on those nonhomogeneous terms have reasonable interpretations in terms of the original history value problem.

For $0 \leq y \leq L$, $\theta(y, t)$ denotes the angular displacement of the cross section at y from the equilibrium position. We assume that the angular displacement is known for $t \leq 0$, and we allow an initial jump in the angular velocity at $t = 0$, that is,

$$(1.3) \quad \theta(y, t) = \theta_0(y, t) \quad (t \leq 0), \quad \theta_t(y, 0^+) = \theta_1(y)$$

for $0 \leq y \leq L$. Then the constitutive law (1.2) together with a balance of angular momentum yields (compare [22, p. 42] for an elastic rod)

$$(1.4) \quad \mathcal{R} \rho \theta_{tt}(y, t) = \mathcal{R} \left\{ G \theta_{yy}(y, t) + \frac{\partial}{\partial t} \int_0^t g(\tau) \theta_{yy}(y, t - \tau) d\tau \right\} + B(y, t),$$

$0 < y < L$, $t > 0$, where ρ is the mass density of the rod, $\mathcal{R} = (\pi/2)R^4$, and $B(y, t)$ consists of a distributed applied torque together with the contribution from the initial history. We assume that the left end of the rod is fixed, i.e.,

$$(1.5) \quad \theta(0, t) = 0, \quad -\infty < t < \infty.$$

At the end $y = L$ we feed back a torque of the form

$$(1.6) \quad \tau(L, t) = -\kappa H(t - \varepsilon) \theta_t(L, t - \varepsilon), \quad t \geq 0,$$

where $\kappa \geq 0$ is the feedback gain, $\varepsilon \geq 0$ is a time delay in the feedback mechanism, and H denotes the Heaviside step function. We consider both the case where there is a concentrated tip mass at $y = L$ with moment of inertia $I_m > 0$ about the axis of the rod, and the case where no such tip mass is present, i.e., $I_m = 0$. This leads to the balance law

$$(1.7) \quad \begin{aligned} I_m \theta_{tt}(L, t) + \mathcal{R} \left\{ G \theta_y(L, t) + \frac{\partial}{\partial t} \int_0^t g(\tau) \theta_y(L, t - \tau) d\tau \right\} \\ = b(t) - \kappa H(t - \varepsilon) \theta_t(L, t - \varepsilon), \quad 0 \leq t < \infty, \end{aligned}$$

where $b(t)$ consists of any other applied torque at $y = L$ together with the contribution that arises from the initial history.

We normalize the length and radius of the rod by introducing the variables

$$(1.8) \quad \begin{aligned} E &= G\rho^{-1}L^{-2}, & a(t) &= g(t)\rho^{-1}L^{-2}, \\ I &= I_m\mathcal{R}^{-1}\rho^{-1}L^{-1}, & k &= \kappa\mathcal{R}^{-1}\rho^{-1}L^{-1}, \\ u(x, t) &= \theta(Lx, t), & u_0(x, t) &= \theta_0(Lx, t), & u_1(x) &= \theta_1(Lx), \\ F(x, t) &= B(Lx, t)\mathcal{R}^{-1}\rho^{-1} & (0 \leq x \leq 1), \\ f(t) &= b(t)\mathcal{R}^{-1}\rho^{-1}L^{-1}. \end{aligned}$$

Then (1.3)–(1.5) and (1.7) become

$$(1.9) \quad u_{tt}(x, t) = Eu_{xx}(x, t) + \frac{\partial}{\partial t} \int_0^t a(\tau) u_{xx}(x, t - \tau) d\tau + F(x, t), \quad 0 < x < 1, \quad t > 0,$$

$$(1.10) \quad u(x, 0) = u_0(x), \quad u_t(x, 0^+) = u_1(x), \quad 0 \leq x \leq 1,$$

$$(1.11) \quad u(0, t) = 0, \quad 0 \leq t < \infty,$$

$$(1.12) \quad \begin{aligned} Iu_{tt}(1, t) + Eu_x(1, t) + \frac{\partial}{\partial t} \int_0^t a(\tau) u_x(1, t - \tau) d\tau \\ = f(t) - kH(t - \varepsilon)u_t(1, t - \varepsilon), \quad t \geq 0. \end{aligned}$$

We remark that E and $a(t)$ have dimension t^{-2} , I is dimensionless, and k has dimension t^{-1} .

Let $A(t) = E + a(t)$ denote the (normalized) linear stress relaxation modulus. As mentioned above, we assume throughout that

$$(1.13) \quad A(t) = E + a(t) \text{ with } E > 0 \text{ and } a(t) \text{ completely monotonic, } a \in L^1(0, \infty), \text{ and } 0 < a(0^+) \leq \infty.$$

If we expect to achieve exponential decay, we must assume that $a(t)$ decays exponentially as $t \rightarrow \infty$, so we also require that

$$(1.14) \quad e^{\eta t} a(t) \in L^1(0, \infty) \quad \text{for some } \eta > 0.$$

(In the analysis, the lack of hypothesis (1.14) manifests itself in a branch cut along the negative real axis that extends all the way up to the origin.)

Some important special cases follow.

(i) $a(t) \equiv 0$ (excluded by (1.13)). This is the elastic case, and (1.9)–(1.12) (with $I = 0$) reduces to the feedback stabilization problem for the one-dimensional wave equation which was included in [15].

(ii) $A(0^+) < \infty$. Then

$$E\gamma(p, t) + \frac{\partial}{\partial t} \int_0^t a(\tau) \gamma(p, t - \tau) d\tau = A(0^+) \gamma(p, t) + \int_0^t a'(\tau) \gamma(p, t - \tau) d\tau.$$

This is linear viscoelasticity of the Boltzmann type [4]. In what follows the cases $A'(0^+) > -\infty$ and $A'(0^+) = -\infty$ will be distinguished.

$$(iii) \quad A(t) = E + \frac{\gamma}{\Gamma(1 - \alpha)} t^{-\alpha} e^{-\delta t}, \quad 0 < \alpha < 1, \quad \delta > 0$$

(Γ = the gamma function.) This is a fractional derivative model modified by an exponential decay factor. Fractional derivative models have been used successfully to fit experimental complex modulus data for some real materials [1]–[3], [30].

In § 2 we derive a characteristic equation for (1.9)–(1.12), solutions of which give the complex frequencies of asymptotic eigenvibrations. In § 2 we also examine the possible destabilizing effect of time delays in (1.12). The effectiveness of (1.12), when $\varepsilon = 0$, in stabilizing high-frequency vibrations (possibly a consideration when $A'(0^+) > -\infty$) is discussed in § 3. In the case where there is no concentrated tip mass, i.e., $I = 0$ in (1.12), we present analytic and numerical results showing the following.

(i) When $\varepsilon = 0$, all roots of the characteristic equation for (1.9)–(1.12) are in the left half-plane and bounded away from the imaginary axis so that solutions should decay exponentially. (This is justified in § 5.)

(ii) In the free-end and fixed-end cases ($k=0$, $k=\infty$, respectively) with $\varepsilon=0$, low-frequency modes decay slowly, but rates of decay can be improved by choosing k appropriately. Detailed numerical examples illustrating this are presented in § 4.

(iii) For sufficiently small positive ε , the absence or presence of resonance (with unbounded solutions) depends on the growth of $A(t)$ as $t \rightarrow 0^+$.

We also analyze the effect of a concentrated tip mass and show the following:

(iv) When $I > 0$ in (1.12), variation of k improves the decay rate for low-frequency modes, but does not affect the high-frequency modes. (In the case of regular kernels, i.e., $A'(0^+) > -\infty$, with $\varepsilon=0$, this result is in agreement with a recent result due to Desch and Miller [16] concerning exponential stabilization of abstract linear integrodifferential equations in Hilbert space. Namely, in this situation (1.9)–(1.12) can be formulated in a semigroup setting, and since $I > 0$, the operator (1.12) is a finite rank operator. Thus, by [16] the essential growth rate of the resolvent operator for the abstract integrodifferential equation cannot be changed by (1.12). See also Gibson [17].) Moreover, when $I > 0$, sufficiently small feedback delays ($\varepsilon > 0$) do not lead to unbounded solutions.

As mentioned above, we present detailed numerical examples illustrating the effectiveness of (1.12) (when $\varepsilon=0$) in stabilizing low-frequency modes, both when $I > 0$ and $I=0$. These numerical examples are discussed in § 4.

In § 5 we present theorems that guarantee the existence (in a weak sense) of the solutions to (1.9)–(1.12) discussed above. We also justify a Green's function representation which shows that the roots of the characteristic equation do determine the asymptotic decay of $u(x, t)$ as $t \rightarrow \infty$.

Finally, we postpone until § 6 the rather technical proofs of two lemmas.

2. The characteristic equation: robustness with respect to time delays. In this section we derive a Green's formula for the formal Laplace transform of the solution of (1.9)–(1.12), and we study how the location of the poles of this Green's formula is affected by time delays in (1.12). The interpretation of these results in terms of the existence of a solution and its asymptotic decay appears in § 5.

We consider an integrated version of (1.9)–(1.12), namely,

$$(2.1) \quad u_t(x, t) = \int_0^t A(t-\tau) u_{xx}(x, \tau) d\tau + F_0(x, t),$$

$$(2.2) \quad u(x, 0) = u_0(x),$$

$$(2.3) \quad u(0, t) = 0,$$

$$(2.4) \quad Iu_t(1, t) + \int_0^t A(t-\tau) u_x(1, \tau) d\tau = f_0(t) - kH(t-\varepsilon)[u(1, t-\varepsilon) - u_0(1)]$$

on $\{0 < x < 1, 0 \leq t < \infty\}$. Here A satisfies (1.13) and (1.14), I , k , and ε are nonnegative constants, and

$$F_0(x, t) = \int_0^t F(x, \tau) d\tau + u_1(x), \quad f_0(t) = \int_0^t f(\tau) d\tau + Iu_1(1).$$

Assume that $F_0(x, t)$ and $f_0(t)$ are exponentially bounded as $t \rightarrow \infty$. Then taking formal Laplace transforms in (2.1)–(2.4) we obtain (for $\text{Re } s$ sufficiently large)

$$(2.5) \quad \hat{A}(s) U_{xx}(x, s) - sU(x, s) = -[\hat{F}_0(x, s) + u_0(x)],$$

$$(2.6) \quad U(0, s) = 0,$$

$$(2.7) \quad I[sU(1, s) - u_0(1)] + \hat{A}(s) U_x(1, s) = -k e^{-\varepsilon s} \left[U(1, s) - \frac{u_0(1)}{s} \right] + \hat{f}_0(s),$$

where U stands for the Laplace transform \hat{u} . Proceeding formally, let

$$(2.8) \quad \alpha(s) = (s\hat{A}(s))^{1/2},$$

where we use the principal square root, i.e., $|\arg z^{1/2}| < \pi/2$ when $|\arg z| < \pi$ (see (2.15) below for justification), and define

$$(2.9) \quad \beta(s) = \alpha(s)/\hat{A}(s),$$

$$(2.10) \quad \Delta(s) = [Is + k e^{-\varepsilon s}] \sinh \beta(s) + \alpha(s) \cosh \beta(s).$$

Let s be such that $\Delta(s)$ is defined and $\Delta(s) \neq 0$. Then the general solution of (2.5)–(2.6) is given by

$$U(x, s) = C(s) \sinh \beta(s)x - \frac{1}{\alpha(s)} \int_0^x \sinh \beta(s)(x-y)[u_0(y) + \hat{F}_0(s, y)] dy,$$

and using the boundary condition (2.7) we see that the constant $C(s)$ is determined by

$$\begin{aligned} \Delta(s)C(s) &= \left(\frac{k e^{-\varepsilon s}}{s} + I \right) u_0(1) + \hat{f}_0(s) \\ &\quad + \int_0^1 \{ \alpha^{-1}(s)(k e^{-\varepsilon s} + Is) \sinh \beta(s)(1-y) + \cosh \beta(s)(1-y) \} \\ &\quad \cdot [u_0(y) + \hat{F}_0(y, s)] dy. \end{aligned}$$

Thus, for such s (2.5)–(2.7) has the unique solution

$$(2.11) \quad \begin{aligned} U(x, s) &= \frac{-1}{\Delta(s)} \int_0^1 G(x, y, s)[u_0(y) + \hat{F}_0(y, s)] dy \\ &\quad + \frac{\sinh \beta(s)x}{\Delta(s)} \left[\left(\frac{k e^{-\varepsilon s}}{s} + I \right) u_0(1) + \hat{f}_0(s) \right], \end{aligned}$$

where the Green's function is defined by

$$(2.12) \quad \begin{aligned} G(x, y, s) &= -\sinh \beta(s)x[\alpha^{-1}(s)(k e^{-\varepsilon s} + Is) \sinh \beta(s)(1-y) \\ &\quad + \cosh \beta(s)(1-y)] \quad \text{for } 0 \leq x < y \leq 1, \\ G(x, y, s) &= G(y, x, s) \quad \text{for } 0 \leq y < x \leq 1. \end{aligned}$$

We remark that $s\hat{A}(s) \rightarrow E$ as $s \rightarrow 0$ in $\operatorname{Re} s \geq 0$ since $a \in L^1(0, \infty)$. Thus, $\Delta(0) = E^{1/2} > 0$ and $(\sinh \beta(s))/s$ is continuous at $s = 0$; hence, the expression in (2.11) is continuous at $s = 0$ whenever $s\hat{F}_0(y, s)$ and $s\hat{f}_0(s)$ are.

We assume that $\hat{f}_0(s)$ and $\hat{F}_0(x, s)$ are analytic functions of s in an appropriate right half-plane that satisfy suitable decay conditions as $s \rightarrow \infty$ in this half-plane. Then $U(x, s)$ is meromorphic in a half-plane with poles at the zeros of $\Delta(s)$. In § 5, we use an argument based on Plancherel's theorem to obtain an exponentially weighted L^2 -estimate for the solution $u(x, t)$ of (2.1)–(2.4) that depends on the location of these poles. Conversely, if $\Delta(s_0) = 0$ with $\operatorname{Re} s_0 > \mu$, then we can pick F_0, f_0 , and u_0 so that (2.1)–(2.4) does not have a solution with $e^{-\mu t}u(\cdot, t) \in L^2(\mathbb{R}^+; L^2(0, 1))$. For the rest of this section as well as in §§ 3 and 4, we study the location of the zeros of the characteristic function $\Delta(s)$.

Recall that if (1.13) holds, then Bernstein's theorem yields a nondecreasing function μ on $[0, \infty)$ with $0 = \mu(0) < \mu(0^+) = E < \mu(\infty) = A(0^+) \leq \infty$ and $\mu(x) = \mu(x^-)$ for $0 < x < \infty$, such that

$$A(t) = \int_0^\infty e^{-xt} d\mu(x) \quad (t > 0).$$

Note that $\int_0^\infty d\mu(x)/x < \infty$ since $a \in L^1(0, \infty)$, and also observe that

$$(2.13) \quad \int_0^\infty d\mu(x) = A(0^+), \quad \int_0^\infty x d\mu(x) = -A'(0^+),$$

where $A'(0^+) = -\infty$ and $A(0^+) = \infty$ are both allowed in (2.13). Since $a \in L^1(0, \infty)$ the Laplace transform $\hat{A}(s)$ exists for $s = \sigma + i\tau$, $\sigma \geq 0$, $s \neq 0$, and it follows from the theory of Laplace and Stieltjes transforms [32, Chap. 8] that $\hat{A}(s)$ can be continued analytically to the slit plane $\mathbb{C} \setminus (-\infty, 0]$ by the formula

$$(2.14) \quad \hat{A}(s) = \int_0^\infty \frac{d\mu(x)}{s+x},$$

where the integral converges uniformly on compact subsets of $\mathbb{C} \setminus (-\infty, 0]$. (Since μ is real, $\hat{A}(\bar{s}) = \overline{\hat{A}(s)}$; this observation may be used without mention in the sequel.) Clearly $\hat{A}(\sigma) > 0$ when $\sigma > 0$, and by (2.14),

$$(2.15) \quad \operatorname{Im}(s\hat{A}(s)) = \int_0^\infty \frac{\tau x}{(\sigma+x)^2 + \tau^2} d\mu(x) \neq 0 \quad \text{when } \tau \neq 0.$$

If, in addition, (1.14) holds for some $\eta > 0$, then $\mu(x) \equiv E$ on $(0, \eta]$ and $\alpha(s)$ in (2.8) is defined and analytic in a slit plane $\mathbb{C} \setminus (-\infty, s^*]$, where $s^* < 0$ is defined by

$$(2.16) \quad s^* \hat{A}(s^*) = E + s^* \hat{a}(s^*) = 0.$$

The transform $\hat{A}(s)$ has a simple pole at $s = 0$, and the functions $\beta(s)$ and $\Delta(s)$ defined in (2.9) and (2.10) are analytic on $\mathbb{C} \setminus (-\infty, s^*]$ and $\beta(0) = 0$, $\Delta(0) = E^{1/2}$.

In order to study the question of when the equation

$$(2.17) \quad \Delta(s) = 0$$

has solutions in a right half-plane, it is necessary to have estimates that are uniform as $s \rightarrow \infty$ in a right half-plane. To obtain such estimates it is convenient to recall the following result due to Lindelöf (see [8, p. 2]). Fix σ_0 and suppose that there exists an $R > 0$ so that the function $h(s)$ is analytic in $G = \{\operatorname{Re} s > \sigma_0, |s| > R\}$, and so that h is bounded and continuous in the closure \bar{G} . If $h(\sigma_0 + i\tau) \rightarrow A$ as $|\tau| \rightarrow \infty$, then $h(s) \rightarrow A$ as $s \rightarrow \infty$ uniformly in $\{\operatorname{Re} s \geq \sigma_0\}$.

Our first lemma is an elementary result.

LEMMA 2.1. *Assume that (1.13) holds and let $0 \geq \sigma_0 > -\infty$ be given. Then*

(i)

$$(2.18) \quad s\hat{A}(s) \rightarrow A(0^+) \quad \text{as } s \rightarrow \infty,$$

$$(2.19) \quad \alpha(s)/s \rightarrow 0 \quad \text{as } s \rightarrow \infty,$$

and the convergence in (2.18) and (2.19) is uniform in $\{\operatorname{Re} s \geq \sigma_0\}$. (The right-hand side of (2.18) is infinity when $A(0^+) = \infty$.)

(ii) *If $A'(0^+) > -\infty$, then*

$$(2.20) \quad s^2 \hat{A}(s) - sA(0^+) - A'(0^+) \rightarrow 0 \quad \text{as } s \rightarrow \infty$$

uniformly in $\{\operatorname{Re} s \geq \sigma_0\}$.

Proof. Assume that $A(0^+) < \infty$. Then

$$(\sigma_0 + i\tau)\hat{A}(\sigma_0 + i\tau) = \int_0^\infty \frac{(\sigma_0 + i\tau)}{(\sigma_0 + i\tau) + x} d\mu(x) \rightarrow A(0^+)$$

as $|\tau| \rightarrow \infty$ by Lebesgue's dominated convergence theorem and (2.13). Since $s\hat{A}(s)$ is analytic and bounded in $\{\operatorname{Re} s \geq \sigma_0, |s| \geq 2|\sigma_0|\}$, Lindelöf's theorem yields (2.18) when $A(0^+) < \infty$. Similarly, if $A'(0^+) > -\infty$, then

$$s^2\hat{A}(s) - sA(0^+) = - \int_0^\infty \frac{sx}{s+x} d\mu(x)$$

is analytic and bounded in $\{\operatorname{Re} s \geq \sigma_0, |s| \geq 2|\sigma_0|\}$. Moreover, setting $s = \sigma_0 + i\tau$ in this expression and letting $|\tau| \rightarrow \infty$, we see by dominated convergence and (2.13) that $(\sigma_0 + i\tau)^2\hat{A}(\sigma_0 + i\tau) - (\sigma_0 + i\tau)A(0^+) \rightarrow A'(0^+)$ as $|\tau| \rightarrow \infty$; hence, (2.20) follows by Lindelöf's theorem.

To prove (2.18) when $A(0^+) = \infty$, let $s = \sigma_0 + i\tau$ and note that

$$\begin{aligned} |s\hat{A}(s)| &\geq |\tau| |\operatorname{Im} \hat{A}(s)| = \int_0^\infty \frac{\tau^2}{(\sigma_0 + x)^2 + \tau^2} d\mu(x) \\ &\geq \int_0^\infty \frac{\tau^2}{x^2 + \tau^2} d\mu(x) \geq \frac{1}{2} \int_0^{|\tau|} d\mu(x) \rightarrow \infty \end{aligned}$$

as $|\tau| \rightarrow \infty$. Thus, since $1/s\hat{A}(s)$ is analytic and bounded in $\{\operatorname{Re} s \geq \sigma_0, |s| \geq 2|\sigma_0|\}$, (2.18) is proved when $A(0^+) = \infty$.

Finally, rewriting (2.14) as

$$\hat{A}(s) = \int_0^\infty \frac{1+x}{s+x} \frac{d\mu(x)}{1+x},$$

and recalling that $\int_0^\infty (d\mu(x)/x) < \infty$, we see that $\hat{A}(s) \rightarrow 0$ as $s \rightarrow \infty$ uniformly in $\operatorname{Re} s \geq \sigma_0$ by the dominated convergence theorem, and, in particular, (2.19) holds by the definition of α . \square

As an easy consequence of Lemma 2.1 we get the following lemma.

LEMMA 2.2. Assume that (1.13) holds and that $A'(0^+) > -\infty$. Then for any $\sigma_0 > -\infty$,

$$(2.21) \quad \exp(-2\beta(s))/C \exp(-2sA(0^+)^{-1/2}) \rightarrow 1$$

as $s \rightarrow \infty$ uniformly in $\operatorname{Re} s \geq \sigma_0$, where

$$(2.22) \quad C = \exp(A'(0^+)/A(0^+)^{3/2}).$$

Proof. By (2.20) and a Taylor approximation to the square root,

$$s\alpha(s)A(0^+)^{-1} - sA(0^+)^{-1/2} - A'(0^+)(2A(0^+)^{3/2})^{-1} \rightarrow 0$$

as $s \rightarrow \infty$ uniformly in $\operatorname{Re} s \geq \sigma_0$. Also, using (2.18) and (2.20) we easily see that

$$\hat{A}(s)^{-1} - sA(0^+)^{-1} + A'(0^+)A(0^+)^{-2} \rightarrow 0$$

as $s \rightarrow \infty$ uniformly in $\operatorname{Re} s \geq \sigma_0$. Thus, since $\beta(s) = \alpha(s)/\hat{A}(s)$,

$$\beta(s) - sA(0^+)^{-1/2} + A'(0^+)(2A(0^+)^{3/2})^{-1} \rightarrow 0$$

as $s \rightarrow \infty$ uniformly in $\operatorname{Re} s \geq \sigma_0$, and (2.21) is proved. \square

The proof of the next lemma is somewhat technical and it is postponed until § 6.

LEMMA 2.3. Assume that (1.13) holds and that $A'(0^+) = -\infty$. Then for any $\sigma_0 > -\infty$,

$$(2.23) \quad \exp(-2\beta(s)) \rightarrow 0 \quad \text{as } s \rightarrow \infty$$

uniformly in $\operatorname{Re} s \geq \sigma_0$.

We remark that Lemmas 2.1–2.3 all hold when $\sigma_0 = 0$ with complete monotonicity in (1.13) weakened to a is positive, nonincreasing, and convex, with $-a'$ convex on $(0, \infty)$. (The proofs are sketched in [20].) In particular, as we noted in the Introduction, our results on the presence or absence of solutions of (2.17) in $\{\operatorname{Re} s \geq 0\}$ due to time delays in (2.4) all hold, with complete monotonicity replaced by this weaker assumption on a .

For the rest of this section we study robustness with respect to time delays, and investigate the question of when (2.17) has solutions in $\{\operatorname{Re} s \geq 0\}$. We will see that the behavior of the stress relaxation modulus $A(t)$ near $t = 0$, and the presence ($I > 0$) or absence ($I = 0$) of a concentrated tip mass, are of primary importance in this regard.

From (2.14) we write

$$(2.24) \quad \hat{A}(s) \equiv \varphi_\sigma(\tau) - i\psi_\sigma(\tau) = \int_0^\infty \frac{\sigma + x}{(\sigma + x)^2 + \tau^2} d\mu(x) - i \int_0^\infty \frac{\tau}{(\sigma + x)^2 + \tau^2} d\mu(x),$$

and, in particular, we obtain

$$\varphi_\sigma(\tau) > 0 (\sigma \geq 0, \tau \in \mathbb{R}, s \neq 0), \quad \tau\psi_\sigma(\tau) > 0 (\sigma \geq 0, \tau \neq 0).$$

By these two inequalities we see that

$$(2.25) \quad |\arg \alpha(s)| < \pi/4 \quad \text{when } \operatorname{Re} s \geq 0, \quad s \neq 0,$$

and that $\beta(s)$ defined by (2.9) is the same as $\beta(s) = (s/\hat{A}(s))^{1/2}$ (principal branch) when $\operatorname{Re} s \geq 0$. (We caution the reader that this formula for $\beta(s)$ need not be valid when $\operatorname{Re} s < 0$.) In particular,

$$(2.26) \quad |\arg \beta(s)| < \pi/2 \quad \text{when } \operatorname{Re} s \geq 0, \quad s \neq 0.$$

Finally, recall that $s = 0$ never satisfies (2.17) since $\Delta(0) = E^{1/2}$.

As an elementary consequence of (2.25) and (2.26) we obtain the following proposition.

PROPOSITION 2.1. *Let (1.13) hold.*

- (i) *If $k = 0$, then (2.17) has no solution in $\{\operatorname{Re} s \geq 0\}$.*
- (ii) *If $k > 0$ and I, ε are nonnegative, then (2.17) has no solution in $\{\sigma \geq 0, 0 \leq |\tau| < \pi/2\varepsilon\}$. In particular, (2.17) has no solution in $\{\operatorname{Re} s \geq 0\}$ when $\varepsilon = 0$.*

Proof. Rewrite (2.17) as

$$(2.17') \quad \exp(-2\beta(s)) = \frac{Is + k e^{-\varepsilon s} + \alpha(s)}{Is + k e^{-\varepsilon s} - \alpha(s)}.$$

By (2.26), $|\exp(-2\beta(s))| < 1$ when $\operatorname{Re} s \geq 0, s \neq 0$. If $k = 0$, then (2.25) implies that the modulus of the right-hand side of (2.17') is greater than or equal to 1, so (2.17') cannot hold when $\operatorname{Re} s \geq 0, s \neq 0$, and part (i) is proved. If $k > 0$ and $|\tau| < \pi/2\varepsilon$, then $|\arg k e^{-\varepsilon s}| < \pi/2$, so by (2.25) the modulus of the right-hand side of (2.17') is greater than 1, and (2.17') cannot hold. \square

Proposition 2.1 shows that for a given $\varepsilon > 0$, we can always be assured that (2.17) has no eigensolutions below the frequency $\pi/2\varepsilon$. On the other hand, for any $I \geq 0$, by rewriting (2.17) as

$$(2.17'') \quad k e^{-\varepsilon s} = -(\alpha(s) \coth \beta(s) + Is),$$

we see that any $s_0 \notin \mathbb{R}$ with $\operatorname{Re} s_0 > 0$ is a solution for some ε, k : choose s_0 , pick $\varepsilon > 0$ to make the arguments of (2.17'') agree, and then choose $k > 0$ so that the moduli match. Thus, any robustness result must impose restrictions on the sizes of k and ε .

The asymptotic behavior estimates in Lemmas 2.1–2.3, coupled with Proposition 2.1 for low frequencies, enables us to provide necessary restrictions on k and ε .

For the case where there is no tip mass we obtain the following theorem.

THEOREM 2.1. *Assume that (1.13) holds and that $I = 0$. Then*

(i) *If $A'(0^+) = -\infty$ and $0 < K < A(0^+)^{1/2}$, then there exists $\varepsilon_0 = \varepsilon_0(K) > 0$ such that (2.17) has no solution in $\{\operatorname{Re} s \geq 0\}$ whenever $0 \leq \varepsilon \leq \varepsilon_0$ and $0 \leq k \leq K$.*

(ii) *If $A'(0^+) = -\infty$ and $k > A(0^+)$, then for each $\varepsilon > 0$ (2.17) has infinitely many solutions in $\{\operatorname{Re} s > 0\}$.*

(iii) *Let $A'(0^+) > -\infty$ and define C by (2.22). If $K < A(0^+)^{1/2}(1 - C)/(1 + C)$, then there exists $\varepsilon_0 = \varepsilon_0(K) > 0$ so that the conclusion of part (i) holds.*

(iv) *Let $A'(0^+) > -\infty$ and define C by (2.22). If*

$$(2.27) \quad k > A(0^+)^{1/2}(1 - C)/(1 + C),$$

then there exists a dense open set $\mathcal{D} \subseteq (0, \infty)$ so that for each $\varepsilon \in \mathcal{D}$, (2.17) has infinitely many solutions in $\{\operatorname{Re} s > 0\}$.

We observe that $A(0^+)^{1/2}$ is the speed of propagation of shear disturbances in the scaled problem (1.9)–(1.12) [21], [27], while C defined by (2.22) is the height that a unit jump discontinuity in shear velocity has after it has propagated a distance 2 in a semi-infinite medium of the scaled material ($C = 0$ when $A'(0^+) = -\infty$) [21, p. 241], [27]. Thus, in the original unscaled problem (1.3)–(1.5) and (1.7), the critical feedback gain κ in (1.7) in the case when $g'(0^+) > -\infty$ is

$$\left(\frac{\pi}{2} R^4 \rho\right) \left(\frac{G + g(0^+)}{\rho}\right)^{1/2} \frac{1 - C}{1 + C},$$

where $(\pi/2)R^4\rho$ is the moment of inertia about its axis of a segment of the rod of length one, $((G + g(0^+))/\rho)^{1/2}$ is the speed of propagation of shear disturbances in the material, and $C = \exp(\rho^{1/2} L g'(0^+)/(G + g(0^+))^{3/2})$ is the height of a unit jump discontinuity in shear velocity after it has propagated a distance $2L$ in a semi-infinite medium of this material.

We also remark that parts (iii) and (iv) of Theorem 2.1 are exactly analogous to the results due to Datko, Lagnese, and Polis [15] for the damped wave equation

$$u_{tt} + 2bu_t + b^2u = Eu_{xx} \quad (0 \leq x \leq 1, b, E > 0)$$

with boundary conditions

$$u(0, t) = 0, Eu_x(1, t) = -ku_t(1, t - \varepsilon).$$

Proof of Theorem 2.1. (i) Fix $K < A(0^+)^{1/2}$. By (2.18) and (2.23) there exist $M = M(K) < 1$ and $\tau_0 = \tau_0(K) > 0$ such that for $0 \leq k \leq K$, $\varepsilon \geq 0$,

$$|e^{-2\beta(s)}| < M \quad \text{and} \quad \left| \frac{k e^{-\varepsilon s} + \alpha(s)}{k e^{-\varepsilon s} - \alpha(s)} \right| > M$$

when $s = \sigma + i\tau$ with $\sigma \geq 0$ and $|\tau| \geq \tau_0$, so (2.17') with $I = 0$ has no solution when $\sigma \geq 0, |\tau| \geq \tau_0$. Now let $\varepsilon_0 = \varepsilon_0(K) = \pi/2\tau_0(K)$ and combine the above with Proposition 2.1 to get that (2.17') ($I = 0$) has no solution in $\operatorname{Re} s \geq 0$ when $0 \leq k \leq K$, $0 \leq \varepsilon \leq \varepsilon_0$.

(ii) Fix $k > A(0^+)^{1/2}$ and $\varepsilon > 0$, and define $\sigma_0 > 0$ by $k \exp(-\varepsilon\sigma_0) = A(0^+)^{1/2}$. The numbers $z_n = \sigma_0 + (2n + 1)\pi i/\varepsilon$ ($n = 0, 1, 2, \dots$) are solutions of

$$h(s) = \frac{k e^{-\varepsilon s} + A(0^+)^{1/2}}{k e^{-\varepsilon s} - A(0^+)^{1/2}} = 0.$$

Now a Rouché's theorem argument using (2.18) and (2.23), and the fact that $h(s)$ is periodic with period $2\pi i/\varepsilon$, show that (2.17') with $I=0$ has a sequence of solutions s_n with $s_n - z_n \rightarrow 0$ ($n \rightarrow \infty$), and the proof of part (ii) is complete.

(iii) For $0 \leq k \leq K$, $\operatorname{Re} s \geq 0$,

$$\left| kA(0^+)^{-1/2} \frac{C \exp(-2sA(0^+)^{-1/2}) - 1}{C \exp(-2sA(0^+)^{-1/2}) + 1} \right| \leq KA(0^+)^{-1/2} \frac{1+C}{1-C} < 1$$

by hypothesis. Combining this with Lemma 2.2 and (2.18), we can find $\tau_0 = \tau_0(K) > 0$ so that

$$|k\alpha^{-1}(s) \tanh \beta(s)| < 1 \quad (\sigma \geq 0, |\tau| \geq \tau_0, k \leq K),$$

so (2.17'') (with $I=0$) has no solutions for $\sigma \geq 0$, $|\tau| \geq \tau_0$. Then, as in the proof of part (i), let $\varepsilon_0(K) = \pi/2\tau_0(K)$.

(iv) In this case the argument in [15] works virtually unchanged. Fix $\varepsilon > 0$. Note that s with $\operatorname{Re} s > 0$ is a solution of (2.17') (with $I=0$) if and only if it is a zero of

$$H(s, \varepsilon) = [\varphi(s) + k e^{-\varepsilon s}] + \exp(-2\beta(s))[\alpha(s) - k e^{-\varepsilon s}].$$

By Lemma 2.2 and (2.18),

$$H(s, \varepsilon) - A^{1/2}(0^+)G(s, \varepsilon) \rightarrow 0 \quad \text{as } s \rightarrow \infty$$

uniformly in $\operatorname{Re} s \geq 0$, where $G(s, \varepsilon)$ is defined by

$$G(s, \varepsilon) = 1 + C \exp(-2sA(0^+)^{-1/2}) + kA(0^+)^{-1/2} e^{-\varepsilon s} [1 - C \exp(-2sA(0^+)^{-1/2})].$$

Since $G(s, \varepsilon)$ is bounded and uniformly almost periodic in vertical strips, the argument used to prove Lemma 2.3 in [13] shows that if $G(s_0, \varepsilon) = 0$ for some $s_0 = \sigma_0 + i\tau_0$ with $\sigma_0 > 0$, then $H(s, \varepsilon)$ has an infinite number of zeros s_n in any vertical strip $\{|\operatorname{Re} s - \sigma_0| < \delta\}$, $\delta > 0$. Since (2.27) holds, Lemma 2 of [15] implies that there is a dense open set $\mathcal{D} \subseteq (0, \infty)$ so that when $\varepsilon \in \mathcal{D}$, $G(s_0(\varepsilon), \varepsilon) = 0$ for some $s_0 = s_0(\varepsilon)$ with $\sigma_0 > 0$, and the proof of part (iv) is complete. \square

We remark that for some values of $A(0^+)$ and $A'(0^+)$, and some k satisfying (2.27), there exists $\varepsilon > 0$ so that $G(s, \varepsilon)$ defined above has no zero in $\{\operatorname{Re} s > 0\}$. An example of this is sketched in [20]. In particular, when the hypotheses of part (iv) hold, we do not know if it is true that for each $\varepsilon > 0$, (2.17) has a solution in $\{\operatorname{Re} s > 0\}$, or if this is only true for ε in a dense open set $\mathcal{D} \subseteq (0, \infty)$. We suspect that the latter alternative holds.

When a tip mass is present ($I > 0$), the $I s$ terms dominate the right-hand side of (2.17') when $|s|$ is large, and in contrast to Theorem 2.1, we obtain the following theorem.

THEOREM 2.2. *Assume that (1.13) holds and that $I > 0$. Then for each $K > 0$, there exists $\varepsilon_0 = \varepsilon_0(K) > 0$ such that (2.17) has no solution in $\{\operatorname{Re} s \geq 0\}$ whenever $0 \leq \varepsilon \leq \varepsilon_0$ and $0 \leq k \leq K$.*

Proof. Using (2.19), and Lemma 2.2 when $A'(0^+) > -\infty$ or Lemma 2.3 when $A'(0^+) = -\infty$, we can find $M = M(K) < 1$ and $\tau_0 = \tau_0(K) > 0$ such that the left-hand side of (2.17') has modulus $< M$ and the right-hand side of (2.17') has modulus $> M$ whenever $\sigma \geq 0$ and $|\tau| \geq \tau_0$, $0 \leq k \leq K$, and $0 \leq \varepsilon$. Now let $\varepsilon_0 = \varepsilon_0(K) = \pi/2\tau_0(K)$ and use Proposition 2.1 as in the proof of Theorem 2.1(i) to complete the proof of Theorem 2.2. \square

In particular, we emphasize that even though the presence of a tip mass makes the feedback mechanism (2.4) with $\varepsilon = 0$ ineffective for exponential stabilization of high-frequency vibrations (possibly desirable when $A'(0^+) > -\infty$; see §§ 3 and 4), the presence of a tip mass precludes the extreme sensitivity to time delays exhibited in parts (ii) and (iv) of Theorem 2.1.

3. Effectiveness of boundary feedback: high frequencies. We now examine the influence of the boundary feedback (2.4) with $\varepsilon = 0$ on the eigensolutions of (2.17). We assume throughout that assumptions (1.13) and (1.14) hold. Three model examples satisfying these hypotheses are the following:

- (I) $A(t) = E + \gamma e^{-\delta t}$ ($E, \gamma, \delta > 0$); standard linear solid model.
- (II) $A(t) = E + (\gamma/\Gamma(1-\alpha))t^{-\alpha}e^{-\delta t}$ ($0 < \alpha < 1, E, \gamma, \delta > 0$), $\Gamma =$ gamma function; a modified "fractional derivative model" with exponential decay as $t \rightarrow \infty$.
- (III) $A(t) = E + (\gamma/\Gamma(1-\alpha)) \int_0^\infty \tau^{-\alpha} e^{-\delta \tau} d\tau$ ($0 < \alpha < 1, E, \gamma, \delta > 0$); an intermediate model with $A(0^+) < \infty$ and $A'(0^+) = -\infty$.

(For all three examples, (1.14) holds with any $\eta < \delta$.)

Throughout §§ 3 and 4 we assume that $\varepsilon = 0$. Recall (Proposition 2.1) that in this case (2.17) has no solution in $\{\operatorname{Re} s \geq 0\}$. Since $\Delta(s)$ is analytic in $\mathbb{C} \setminus (-\infty, s^*]$, where s^* is defined by (2.16), $\Delta(s)$ cannot have a sequence of zeros with a finite limit point in $\mathbb{C} \setminus (-\infty, s^*]$. It follows from the asymptotic estimates in Theorems 3.1 and 3.2 that for each $k \geq 0$, (2.17) has no solution in $\{\operatorname{Re} s \geq -d\}$ for some $d = d(k) > 0$.

In this section we examine the influence of the feedback (2.4) ($\varepsilon = 0$) on the high-frequency eigensolutions of (2.17). Detailed numerical calculations for specific stress relaxation moduli of the forms (I)–(III) that show how the low- to moderate-frequency solutions of (2.17) depend on the feedback gain k are presented in § 4.

THEOREM 3.1. *Assume that (1.13) holds and that $A'(0^+) = -\infty$. Let $k \geq 0$ and $\varepsilon = 0$. Then given $\sigma_0 > -\infty$, there exists $\tau_0 = \tau_0(\sigma_0) > 0$ so that (2.17) has no solutions in $\sigma \geq \sigma_0$, $|\tau| \geq \tau_0$. Moreover, τ_0 is independent of k when $I = 0$.*

Thus, when $A'(0^+) = -\infty$, solutions s_n of (2.17) ($\varepsilon = 0$) always satisfy $\operatorname{Re} s_n \rightarrow -\infty$ as $|\operatorname{Im} s_n| \rightarrow \infty$, and the corresponding vibrations decay at increasingly high exponential rates. The asymptotic location of these solutions for specific cases of our model examples (II) and (III) is discussed in § 4.

Proof. By Lemma 2.3, we can find $\tau_1 = \tau_1(\sigma_0) > 0$ so that $|\exp(-2\beta(s))| < \frac{1}{2}$ whenever $\sigma \geq \sigma_0$, $|\tau| \geq \tau_1$. By definition (2.10)

$$2\Delta(s) = (Is + k + \alpha(s)) e^{\beta(s)} - (Is + k - \alpha(s)) e^{-\beta(s)}.$$

If $I = 0$, we observe that $\operatorname{Re} \alpha(s) > 0$ for $s \in \mathbb{C} \setminus (-\infty, 0]$ by (2.8), and obtain

$$(3.1) \quad \left| \frac{\exp(\beta(s))}{\Delta(s)} \right| \leq 2 \left\{ |k + \alpha(s)| - \frac{1}{2} |k - \alpha(s)| \right\}^{-1} \\ \leq 4 |k + \alpha(s)|^{-1} \leq 4 |\alpha(s)|^{-1}$$

when $\sigma \geq \sigma_0$, $|\tau| \geq \tau_1$. In particular, $\Delta(s) \neq 0$ in $\sigma \geq \sigma_0$, $|\tau| \geq \tau_0 = \tau_1$.

When $I > 0$, Is is the dominant term of $Is + k \pm \alpha(s)$ by (2.19). Thus, by possibly increasing $\tau_0(\sigma_0)$,

$$(3.2) \quad \left| \frac{\exp(\beta(s))}{\Delta(s)} \right| \leq 4(I|s|)^{-1}$$

when $\sigma \geq \sigma_0$, $|\tau| \geq \tau_0$, and Theorem 3.1 is proved. \square

When $A'(0^+) > -\infty$ we proceed as follows. Define

$$N = \begin{cases} 1 & \text{when } I > 0, \\ (k - A(0^+)^{1/2})(k + A(0^+)^{1/2})^{-1} & \text{when } I = 0, \end{cases}$$

and define $G(s)$ by

$$(3.3) \quad G(s) = C^{-1} \exp(2sA(0^+)^{-1/2}) - N,$$

where C is given by (2.22). Let

$$F(s) = \exp(2\beta(s)) - \frac{Is + k - \alpha(s)}{Is + k + \alpha(s)}.$$

By (2.18), (2.19), and Lemma 2.2, for each $\sigma_0 > -\infty$

$$(3.4) \quad F(s) - G(s) \rightarrow 0 \quad \text{as } s \rightarrow \infty$$

uniformly in $\operatorname{Re} s \geq \sigma_0$. By considering moduli, we see that as $|\operatorname{Im} s| \rightarrow \infty$, zeros of $F(s)$ must approach the vertical line $\operatorname{Re} s = \sigma_*$, where

$$(3.5) \quad \sigma_* = \begin{cases} \frac{A'(0^+)}{2A(0^+)} & \text{when } I > 0, \\ \frac{1}{2} \left[\frac{A'(0^+)}{A(0^+)} + A(0^+)^{1/2} \log \left| \frac{k - A(0^+)^{1/2}}{k + A(0^+)^{1/2}} \right| \right] & \text{when } I = 0, \quad k \neq A(0^+)^{1/2}. \end{cases}$$

In the case where $I = 0$ and $k = A(0^+)^{1/2}$, then for each $\sigma_0 > -\infty$ we can find $\tau_0(\sigma_0) > 0$ such that $F(s)$ has no zeros in $\operatorname{Re} s \geq \sigma_0$, $|\tau| \geq \tau_0$. Moreover, except in the case where $I = 0$ and $k = A(0^+)^{1/2}$, an argument using Rouché's theorem shows that $F(s)$ has a sequence of zeros s_n satisfying $s_n - z_n \rightarrow 0$ as $|n| \rightarrow \infty$, where z_n are the zeros of the limit function $G(s)$ defined in (3.3) and ordered so that $|\operatorname{Im} z_n|$ increases as $|n|$ increases. Thus we have proved Theorem 3.2.

THEOREM 3.2. Assume that (1.13) holds and that $A'(0^+) > -\infty$. Let $k \geq 0$ and $\varepsilon = 0$.

(i) If $I + |k - A(0^+)^{1/2}| > 0$, define σ_* by (3.5). Then given $\sigma_0 > -\infty$ and $d > 0$, there exists $\tau_0 = \tau_0(\sigma_0, d) > 0$ such that all solutions $s = \sigma + i\tau$ of (2.17) in $\sigma \geq \sigma_0$, $|\tau| \geq \tau_0$ lie in the strip $|\sigma - \sigma_*| \leq d$. Moreover, (2.17) has a sequence of solutions s_n satisfying $s_n - z_n \rightarrow 0$ as $|n| \rightarrow \infty$, where z_n are the zeros of $G(s)$ defined in (3.3) and ordered so that $|\operatorname{Im} z_n|$ increases as $|n|$ increases.

(ii) If $I = 0$ and $k = A(0^+)^{1/2}$, then given $\sigma_0 > -\infty$, there exists $\tau_0 = \tau_0(\sigma_0) > 0$ so that (2.17) has no solutions in $\sigma \geq \sigma_0$, $|\tau| \geq \tau_0$.

In § 4 we discuss in more detail the asymptotic location of the high-frequency eigensolutions for the model example (I).

Note that when $I > 0$, the boundary feedback is ineffective at moving the high-frequency eigensolutions of (2.17). From a mechanical point of view, this is to be expected since rapidly oscillating torques have little effect on the concentrated moment of inertia at $x = 1$. Also, as we noted in the Introduction, Theorem 3.2 with $I > 0$ agrees with the recent abstract result of Desch and Miller [16] stating that the essential growth rate of the resolvent operator of certain linear integrodifferential equations in Hilbert space cannot be changed by compact perturbations. The related fact that an infinite-dimensional system of linear ordinary differential equations without damping cannot be exponentially stabilized by compact linear feedback is due to Gibson [17].

4. Numerical investigation of eigenvalues. In § 4.1, the numerical methods we use to locate eigenvalues are described. The general picture of the dependence of low modes on the parameters is rather similar for all three models mentioned in § 3: the standard linear solid (I), the fractional derivative model (II), and the intermediate model (III). We have chosen the first of these for a detailed discussion (§ 4.2). Our discussion of the latter two models is less exhaustive (§ 4.3).

In § 4.2.1, the eigenvalues of the standard linear solid model are classified. Their salient features are outlined and illustrated with specific numerical results obtained in a situation where the parameters are $O(1)$. Since we use a spectral method, such a choice of parameters has the best chance of catching all the different types of modes.

If parameters were of very different sizes, then our desired eigenvalues would reflect that, and more spectral modes would be necessary for their resolution.

In §§ 4.2.2 and 4.2.3, we highlight the effectiveness of the feedback parameter k in damping the system. This is best illustrated by numerical results that concern a nearly elastic situation rather than a situation with $O(1)$ parameters. In addition, the parameters are chosen so that the asymptotic formula for highly oscillatory modes places these well away from the imaginary axis. We note that in the elastic case, there are modes on the imaginary axis at $k=0$. In our nearly elastic situation, several of the least stable modes at $k=0$ are close to the imaginary axis. A small addition of k is shown to have a relatively large stabilizing effect on these modes. The effectiveness of small k decreases as the moment of inertia increases. This is consistent with the expectation that it takes more friction to stop a heavier object.

In § 4.2.3, the optimal value of k is discussed. This topic is complicated by the fact that, although the least stable mode at $k=0$ is the lowest complex conjugate pair, a higher mode may become the least stable one as k is increased. For large k , a mode on the negative real axis becomes the least stable one. In order to find the optimal value of k , we compute all relevant eigenvalues for the nearly elastic situation and note the least stable one for each k . This yields that for zero moment of inertia the optimal k is slightly less than the value at which low modes “loop up” (see, e.g., Fig. 1 in § 4.2.1). For nonzero moment of inertia, the optimal k is larger due to the fact that high modes lag behind low modes in their response to k . It is interesting that a large portion of the attainable stabilization is achieved while k is relatively small. The stabilization settles down somewhat for moderate k and then worsens at large k . Thus, the order of magnitude of stabilization achieved by the optimal k is also achieved in a wide interval of k .

4.1. Numerical schemes.

4.1.1. Chebyshev- τ method. We let $u(x, t) = e^{st}U(x)$ in the history value version of (1.9)–(1.12) (see (1.1)) with no external torques and $\varepsilon = 0$, and obtain

$$(4.1) \quad s^2 U(x) - \left[E + s \int_0^\infty a(\tau) e^{-s\tau} d\tau \right] U_{xx}(x) = 0, \quad 0 < x < 1,$$

$$(4.2) \quad s^2 IU(1) + skU(1) + \left[E + s \int_0^\infty a(\tau) e^{-s\tau} d\tau \right] U_x(1) = 0,$$

$$(4.3) \quad U(0) = 0,$$

$$(4.4) \quad a(t) = \gamma e^{-\delta t} \quad \text{for the standard linear solid,}$$

$$(4.5) \quad a(t) = (\gamma/\Gamma(1-\alpha)) t^{-\alpha} e^{-\delta t}$$

for the fractional derivative model modified by an exponential factor,

$$(4.6) \quad a(t) = \frac{\gamma}{\Gamma(1-\alpha)} \int_t^\infty \tau^{-\alpha} e^{-\delta\tau} d\tau \quad \text{for the intermediate model.}$$

The Chebyshev- τ method [18] is useful if (4.1)–(4.3) can be transformed into a matrix eigenvalue problem: determinant $(\underline{A} - \lambda(s)\underline{B}) = 0$, where s can be retrieved from $\lambda(s)$. This applies to the standard linear solid (4.4) and the two models (4.5) and (4.6), where the exponent α is a rational number. There is a limit to the size of an eigenvalue that can be approximated well with this method. This limit is set by the

storage capacity of the computer, since more Chebyshev modes are required to approximate an eigenfunction belonging to an eigenvalue of a larger magnitude, and by the accuracy available, since the matrix eigenvalue problem tends to become rather ill conditioned when a large number of modes is used. We use Newton's scheme for the highly oscillatory modes, as well as other modes for which we already have an idea of where to look, e.g., in the various limits of the parameters. On the other hand, it is virtually impossible to obtain a general picture of where all eigenvalues are with Newton's scheme, whereas the Chebyshev- τ method yields a complete picture of the location of low to moderate modes. The convergence rate of the latter for isolated eigenvalues of C^∞ -eigenfunctions is of infinite order [18]. Our computations were done on a VAX 11/785 computer.

In (4.1)–(4.3), we let $U(x) = \sum_{j=0}^N u_j T_j(z)$, where $z = 2x - 1$ and $T_j(z)$ is the Chebyshev polynomial of degree j . For the standard linear solid, the equations transform to $A_0^* + A_1^* s + A_2^* s^2 + A_3^* s^3 = 0$, where

$$A_0^* = \begin{bmatrix} -E\delta U_{xx}(x) \\ E\delta U_x(1) \\ U(0) \end{bmatrix}, \quad A_1^* = \begin{bmatrix} -(\gamma + E)U_{xx}(x) \\ (\gamma + E)U_x(1) + k\delta U(1) \\ 0 \end{bmatrix},$$

$$A_2^* = \begin{bmatrix} \delta U(x) \\ (I\delta + k)U(1) \\ 0 \end{bmatrix} \quad \text{and} \quad A_3^* = \begin{bmatrix} U(x) \\ IU(1) \\ 0 \end{bmatrix}.$$

When we use the orthogonality of $T_j(z)$, this equation is transformed into $(A_0 + A_1 s + A_2 s^2 + A_3 s^3)\underline{u} = 0$, where \underline{u} is the vector of coefficients u_j and A_i is an $N+1$ square matrix with the first $N-1$ rows devoted to the differential equation. Finally, the eigenvalue problem reduces to determinant $(A - sB) = 0$, where

$$A = \begin{bmatrix} 0 & 0 & A_0 \\ 1 & 0 & -A_1 \\ 0 & 1 & A_2 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -A_3 \end{bmatrix},$$

where A and B are $3(N+1)$ square matrices.

For the fractional derivative model modified by an exponential factor (4.5), we investigate in detail the case $\alpha = \frac{1}{2}$, for which (4.1)–(4.3) transform to $A_0^* + A_1^* \lambda + A_2^* \lambda^2 + A_3^* \lambda^3 + A_5^* \lambda^5 = 0$, where

$$\lambda = \sqrt{s + \delta}, \quad A_0^* = \begin{bmatrix} \gamma\delta U_{xx}(x) \\ -\gamma\delta U_x(1) \\ U(0) \end{bmatrix},$$

$$A_1^* = \begin{bmatrix} \delta^2 U(x) - EU_{xx}(x) \\ (I\delta - k)\delta U(1) + EU_x(1) \\ 0 \end{bmatrix}, \quad A_2^* = \begin{bmatrix} -\gamma U_{xx}(x) \\ \gamma U_x(1) \\ 0 \end{bmatrix},$$

$$A_3^* = \begin{bmatrix} -2\delta U(x) \\ (k - 2I\delta)U(1) \\ 0 \end{bmatrix}, \quad \text{and} \quad A_5^* = \begin{bmatrix} U(x) \\ IU(1) \\ 0 \end{bmatrix}.$$

In a manner analogous to the case of the standard linear solid, we obtain $\det(\underline{A} - \lambda(s)\underline{B}) = 0$, where

$$\underline{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & A_0 \\ 1 & 0 & 0 & 0 & -A_1 \\ 0 & 1 & 0 & 0 & A_2 \\ 0 & 0 & 1 & 0 & -A_3 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad \underline{B} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -A_5 \end{bmatrix},$$

where \underline{A} and \underline{B} are $5(N+1)$ square matrices. After calculating $\lambda(s)$, we discard those whose real parts are negative since they are not on the branch of interest. We then retrieve the eigenvalues $s = \lambda^2 - \delta$. This type of procedure can in principle be used for both models (4.5) and (4.6), provided the exponent α is a rational number.

4.1.2. Newton scheme. A Newton scheme, based on the characteristic equation

$$(4.7) \quad e^{2\beta(s)}(Is + \alpha(s) + k) - Is + \alpha(s) - k = 0, \quad \beta(s) = s/\alpha(s),$$

$$\alpha(s) = \begin{cases} \left(E + \frac{\gamma s}{s + \delta}\right)^{1/2} & \text{for the standard linear solid,} \\ \left(E + \frac{\gamma s}{(s + \delta)^{1-\alpha}}\right)^{1/2} & \text{for the fractional derivative model,} \\ \left(E - \frac{\gamma}{(s + \delta)^{1-\alpha}} + \frac{\gamma}{\delta^{1-\alpha}}\right)^{1/2} & \text{for the intermediate model} \end{cases}$$

is a cost-effective means of tracking specific eigenvalues when an initial guess is available. Initial guesses are obtained by computing eigenvalues with the Chebyshev- τ method for a particular parameter set, or with various limiting cases such as the asymptotics for large imaginary part, the case $I = k = 0$, or the case $|Is + k|$ large compared with $|\alpha(s)|$. The latter two cases are governed by

$$(4.8) \quad s^2 + c_n \alpha^2(s) = 0, \quad n = 1, 2, \dots,$$

where $c_n = (2n-1)^2 \pi^2 / 4$ and $n^2 \pi^2$, respectively. These are cubic equations for the standard linear solid:

$$(4.9) \quad s^3 + s^2 \delta + c_n (E + \gamma)s + c_n E \delta = 0$$

and quintic equations in $\lambda = \sqrt{s + \delta}$ for the other two models with exponent $\alpha = \frac{1}{2}$.

4.2. Standard linear solid model.

4.2.1. Location of eigenvalues. The general picture of the location of eigenvalues must first be obtained in order to keep a tab on the candidates for the least stable mode. At first sight, the picture for the case of zero moment of inertia appears to be quite different from the case of nonzero moment of inertia. However, the picture changes continuously from one case to the other, and it is of interest to present that transformation. We illustrate the main features with numerical results for the situation $E = \gamma = \delta = 1$.

For *zero moment of inertia*, we separate the eigenvalues into four classes below. Candidates for the least stable mode belong to Class 1 for low to moderate k and Class 4 for moderate to large k . Classes 2 and 3 are subject to bounds that prevent them from approaching the origin.

Class 1. These are the complex conjugate modes. Those with positive imaginary parts are included in Fig. 1 ($I = 0$, $E = \gamma = \delta = 1$). At $k = 0$ and at $k = \infty$, these satisfy (4.9). The third root of this equation is discussed under Class 2. The index n in (4.9) is used to index these modes. The modes at $k = 0$ and ∞ are interlaced in the complex

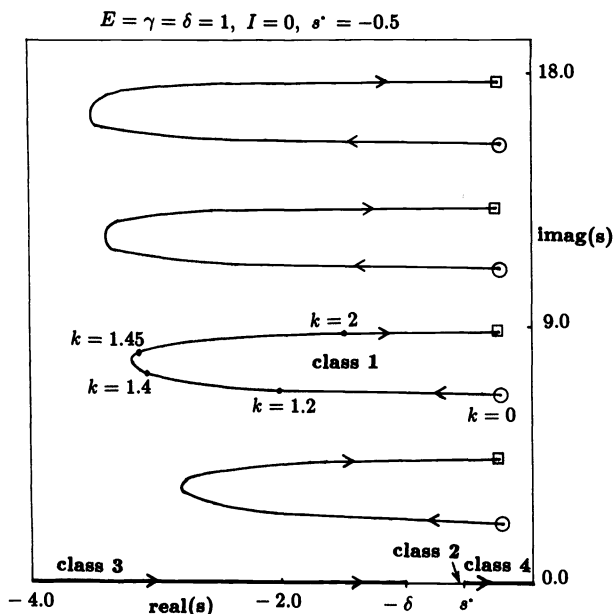


FIG. 1. $I=0$, $E=\gamma=\delta=1$, $k=1$ to ∞ , $s^*=-0.5$. All classes of eigenvalues are displayed. Complex conjugates at $k=0$ are circled and at $k=\infty$ are denoted by squares. Arrows indicate variation with k .

plane as shown in Fig. 1: the circles denote $k=0$ and the squares $k=\infty$. We do not use this notation for modes on the real axis because some lie close together and therefore the figure would lose clarity. The arrows show the trends as k increases. Modes travel away from the imaginary axis as k increases, loop up (around $k=\sqrt{E+\gamma}$ in the nearly elastic case), and then travel the other way as $k\rightarrow\infty$. Highly oscillatory modes have the asymptotic behavior

$$(4.10) \quad s \sim \frac{-\gamma\delta}{2(E+\gamma)} + \frac{\sqrt{E+\gamma}}{2} \log \left| \frac{k-\sqrt{E+\gamma}}{k+\sqrt{E+\gamma}} \right| \pm n\pi i\sqrt{E+\gamma}, \quad n \rightarrow \infty$$

for $k \neq \sqrt{E+\gamma}$, and if $k = \sqrt{E+\gamma}$

$$(4.11) \quad s \sim \frac{-\gamma\delta}{2(E+\gamma)} + \frac{\sqrt{E+\gamma}}{2} \log \left| \frac{\gamma\delta}{4\pi n(E+\gamma)^{3/2}} \right| \pm n\pi i\sqrt{E+\gamma}.$$

Here, n is the index used in (4.9).

The low eigenvalues at $k=0$ and ∞ in Fig. 1 are lined up almost vertically: this need not be the case if the parameters are chosen differently. In fact, when parameters are close to the elastic case, the lowest mode is very close to the imaginary axis and higher modes swing gradually toward the asymptote. The asymptotic formulas apply when the imaginary part of s is much larger than δ . In Fig. 1, δ is not large. Thus, the approach to the asymptote is fast. Even the low modes behave according to (4.10). On the other hand, a nearly elastic situation such as the one in § 4.2.2 has a large δ , so the asymptote is approached extremely slowly with n .

At $k=0$, the first mode in Fig. 1 is the least stable by a small margin. In the nearly elastic case, this margin is large. When k is increased, a higher mode can become the least stable one. That is, a picture of these modes at a fixed nonzero k can show a gradual swing of real parts toward zero as the index n increases, followed by a gradual swing of higher modes in the opposite direction toward the asymptote. We did not

encounter any situation numerically where there was more than one such swing as n increased. Reasons for having one swing are discussed in § 4.2.2.

Class 2. The characteristic equation (4.7) has an essential singularity at the value s^* defined in (2.16), where $\alpha(s^*) = 0$. In this example, $s^* = -E\delta/(E + \gamma)$. We comment that only those modes with $\operatorname{Re} s > s^*$ affect the exponential decay rate of the solution $u(x, t)$ of (2.1)–(2.4) as described in § 5 below, since u is recovered from its Laplace transform by an inversion integral along a vertical line $\operatorname{Re} s = \mu$, where $\mu > s^*$. The location of roots to the characteristic equation in the half-plane $\{\operatorname{Re} s \leq s^*\}$ is discussed here only for the sake of completeness.

In the interval $(-\delta, s^*)$, there are eigenvalues clustering toward s^* . At $k = 0$ and ∞ , these are the countable number of real roots of the cubic equation (4.9). These modes move to the right as k increases. Newton's scheme does not track these modes well because (obviously) it can converge to any member of the cluster. The Chebyshev- τ method resolves the modes furthest away from s^* well, and it takes more and more Chebyshev modes to converge to members closer to s^* . In Fig. 1, $s^* = -0.5$ and the cluster stays very close to this.

Class 3. On the negative real axis, an eigenvalue travels in from $-\infty$ for $k > \sqrt{E + \gamma}$ and approaches $-\delta$ as $k \rightarrow \infty$. This mode is absent when $k \leq \sqrt{E + \gamma}$. This mode enters the picture in Fig. 1 when k is approximately 1.6 and is already close to $-\delta$ when $k = 100$.

Class 4. An eigenvalue pops out on the right of the cluster point s^* when $k > 0$ and travels toward the origin as $k \rightarrow \infty$. This mode is absent at $k = 0$. This mode is inversely proportional to k as $k \rightarrow \infty$. Thus, there exists a value of k beyond which the use of k does more harm than good. For example, when k is larger than about 3 in Fig. 1, the Class 4 mode becomes less stable than the system is at $k = 0$.

As noted earlier, only those modes belonging to Classes 1 or 4 affect the exponential decay rate of the solution to (2.1)–(2.4) as described in § 5.

For *nonzero moment of inertia*, the only modes present at $k = 0$ are those analogous to Classes 1 (complex conjugates) and 2 (cluster). A mode corresponding to Class 4 pops out of the essential singularity s^* for $k > -Is^*$ and travels toward the origin. The dependence of these modes on k is quite different from the case of zero moment of inertia. As $k \rightarrow \infty$, there are modes corresponding to all four classes. Candidates for the least stable mode are again the complex conjugates for low k and the mode that adopts the Class 4 behavior for large k . The moment of inertia destabilizes the complex conjugates and their asymptote loses its dependence on k at the leading order:

$$s \sim \frac{-\gamma\delta}{2(E + \gamma)} \pm \pi \quad \text{in } \sqrt{E + \gamma} \quad \text{as } n \rightarrow \infty,$$

where the moment of inertia appears at $O(n^{-1})$ and the feedback parameter k at $O(n^{-2})$.

We present the overall picture of eigenvalues at large moment of inertia I and show how this changes as $I \rightarrow 0$. We again illustrate the main features with numerical results for the case $E = \gamma = \delta = 1$. For sufficiently large I , the complex conjugates loop down rather than up as k increases from zero. The first mode loops down to the negative real axis as shown in Fig. 2 ($I = 10$, $E = \gamma = \delta = 1$). This pair (the “dropper”) then becomes two real eigenvalues. One travels toward the origin as $k \rightarrow \infty$, just like a Class 4 eigenvalue. The other travels toward the essential singularity s^* . Meanwhile, when $k > -Is^*$, one eigenvalue (the “popper”) corresponding to Class 4 pops out of s^* and travels right. The upper diagram in Fig. 2 displays the details around s^* . The popper emerges for $k > 5$ and meets one of the droppers at $k = 5.11$. The two modes then become complex conjugates to hop over s^* and all members of the cluster (Class 2) except the one closest to $-\delta$. Note that there is a relatively large gap between

$$I = 10, E = \gamma = \delta = 1$$

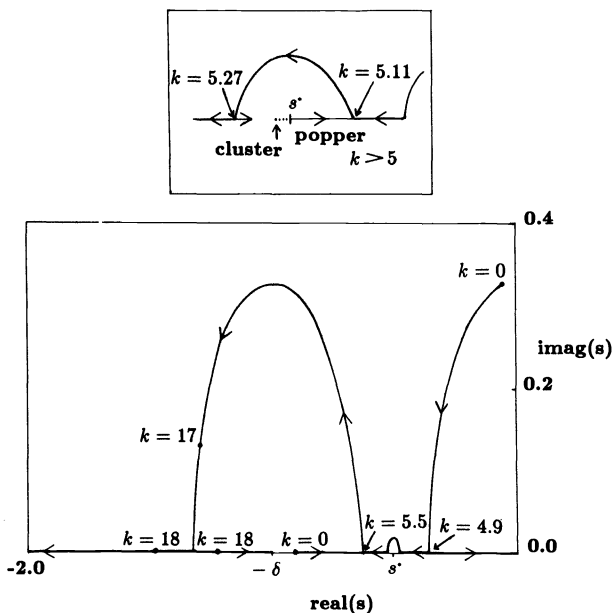


FIG. 2. $I = 10, E = \gamma = \delta = 1, k = 0$ to $20, s^* = -0.5$. The journey of the lowest complex conjugate mode and modes on the real axis.

this member and the rest of the cluster at $k = 0$. They land on the axis at $k = 5.27$ and become two reals. One travels right as the last member of the cluster. As $k \rightarrow \infty$, this cluster behaves just like Class 2. The other eigenvalue heads in the direction of $-\delta$ and at $k = 5.5$ and meets the remaining member of the original cluster, and both become complex conjugates to jump over $-\delta$. It takes a considerable amount of k to complete this jump. They land on the axis at k between 17 and 18, and then become two reals. One travels toward $-\delta$ as $k \rightarrow \infty$ as a Class 3 mode. The other goes out to $-\infty$ as $k \rightarrow \infty$: this mode goes out of the picture altogether as $I \rightarrow 0$, and the manner in which it does so will be explained.

The picture of the higher modes at large moment of inertia is exemplified in Fig. 3 ($I = 0.4, E = \gamma = \delta = 1$). As $k \rightarrow \infty$, the n th mode, $n = 2, 3, 4, \dots$, approaches the limit of the $(n-1)$ th mode of the case $I = 0$. As is evident from the characteristic equation (4.7), higher modes vary less with k and they lag behind low modes in their response to k . This is reflected in Fig. 3. At any one mode, the variation with k is less the larger the moment of inertia.

When I is decreased from the value in Fig. 2, the trajectory of the lowest complex conjugate pair over s^* gets closer to the two adjacent trajectories over the axis. Eventually, these merge. For example, at $I = 8, E = \gamma = \delta = 1$, the dropper lands on the axis to the left of $-\delta$ without landing on the right. There are two valleys on its trajectory before it lands: one to the right of s^* and a shallower valley to the left. As I decreases further, the valleys become less pronounced. Thus, in Fig. 3 ($I = 0.4, E = \gamma = \delta = 1$) no valleys are visible on the trajectory of the dropper. The dropper lands on the axis to the left of $-\delta$ at about $k = 2.8$, and becomes two real eigenvalues. One behaves like Class 3 as $k \rightarrow \infty$. The other travels to $-\infty$. The popper emerges from s^* for $k > -Is^* = 0.2$. As $k \rightarrow \infty$, this eigenvalue behaves like the Class 4 mode. The looping down of the higher modes remains qualitatively the same as at larger I .

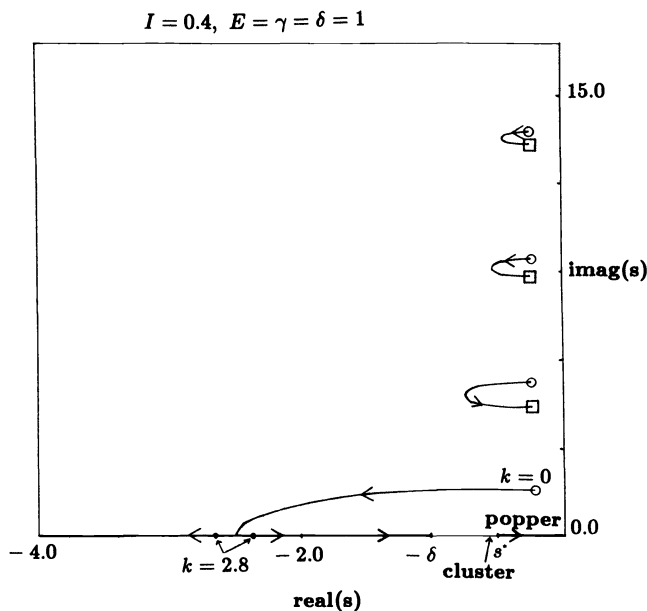


FIG. 3. $I = 0.4$, $E = \gamma = \delta = 1$, $k = 0$ to ∞ , $s^* = -0.5$. The complex conjugates at $k = 0$ are circled and at $k = \infty$ are denoted by squares.

When the moment of inertia is decreased further, the dropper's trajectory rises slightly and kisses the trajectory of the mode above it. At that point, the multiplicity of the eigenvalue is two. This is shown in the schematic drawings of Fig. 4, where I decreases from I^{*+} to I^{*-} . A specific example is the transition from $I = I^{*+} = 0.4$ in Fig. 3 to $I = I^{*-} = 0.01$ in Fig. 5. With successive exchanges as in Fig. 4, the dropping branch propagates upward and the sense of the loops changes from downward to upward. Figure 6 shows the fifth mode dropping at $I = 0.001$. The value of k at which the dropping branch reaches the real axis appears to decrease to $\sqrt{E + \gamma}$ and the junction on the real axis moves out to $-\infty$ as $I \rightarrow 0$.

4.2.2. Effectiveness of small k . In this section, we focus on a situation where, at $k = 0$, the least stable mode is extremely lightly damped compared with the asymptotic

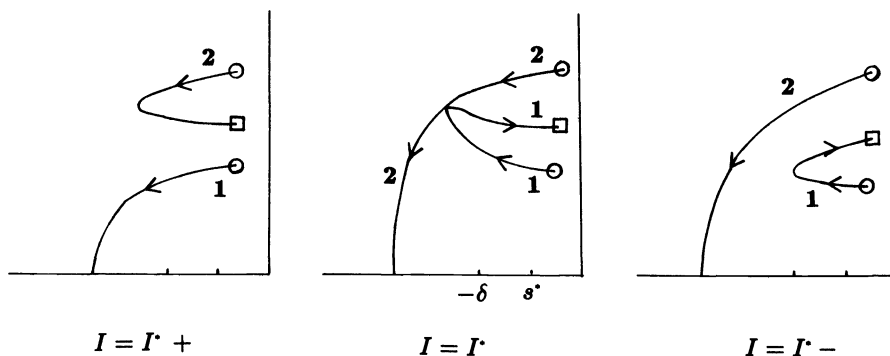


FIG. 4. Branches 1 and 2 represent the first two complex conjugate modes. Arrows on the branches indicate the direction of variation as k increases. Circles denote positions at $k = 0$ and the square at $k = \infty$. At $I = I^*$, branch 1 meets branch 2.

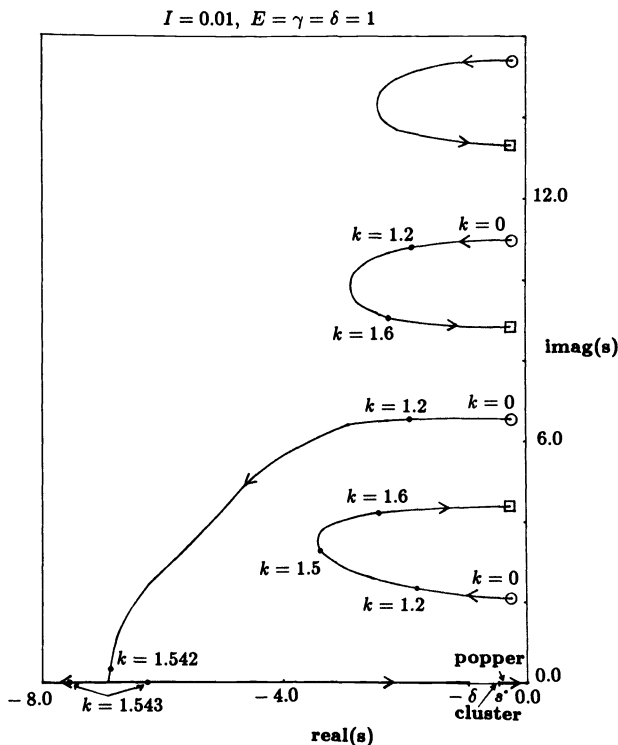


FIG. 5. $I = 0.01$, $E = \gamma = \delta = 1$, $s^* = -0.5$. The second branch drops down. The first branch loops up with k . Higher branches loop down with k . Complex conjugates at $k = 0$ are denoted by circles, and at $k = \infty$ by squares.

behavior of high modes. This occurs, e.g., when the parameters are close to the elastic case. We present numerical evidence to show that a small k can be very effective in damping the system when the moment of inertia is zero, and this effectiveness decreases as the moment of inertia increases. Moreover, since the least stable mode at $k = 0$ is less stable the larger the I , the end product after the addition of a small k is that the situation at a higher I is still less stable than at a lower I . A better decay rate for the larger I may be achieved by using a larger k , as we show in § 4.2.3. This is in fact consistent with the expectation that it takes more frictional force to stop a heavier object.

As an example of a nearly elastic situation, we take $E = 1$, $\gamma = 0.01$, $\delta = 1000$. Several of the Class 1 modes for $I = k = 0$ are listed in Table 1. These are the complex conjugate roots of (4.9). Here, many low modes have real parts that are much smaller than the asymptotic limit -4.95 of highly oscillatory modes. Due to the size of δ , the asymptote is achieved extremely slowly: the thousandth mode is still ten percent away. The approach to the asymptote at $I = k = 0$ is monotonic with n . For our parameters, the low complex conjugates are the least stable modes for small k and no other mode enters into our discussion.

For zero moment of inertia, Fig. 7 shows the dramatic shift in the first several modes as a response to a small value of the feedback gain parameter k . At $k = 0.01$ and 0.1 , their real parts become $O(k)$. For example, at $k = 0.01$, low modes are lined up at approximately $\text{Re } s = -0.01$ and higher modes curve back toward the asymptote. When $k = 0.1$, the low modes have shifted to $\text{Re } s = -0.1$. The modes still approach the asymptote $\text{Re } s = -4.96$ in a monotonic way. When $k = 0.1$, there is an $O(10^4)$ stabilization for the worst mode of the system.

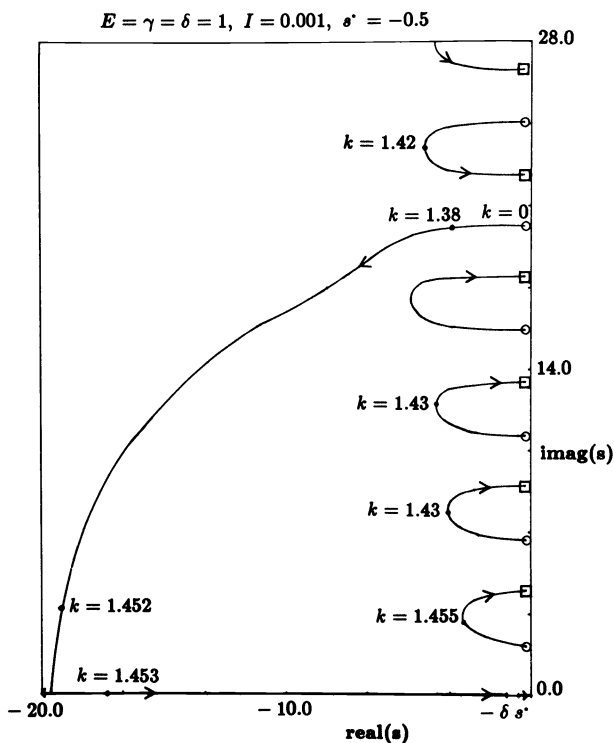


FIG. 6. $I = 0.001$, $E = \gamma = \delta = 1$, $s^* = -0.5$, $k = 0$ to ∞ . The first four branches loop up, the fifth branch drops down and higher branches loop down with k . Complex conjugates at $k = 0$ are denoted by circles and at $k = \infty$ by squares.

TABLE 1
Standard linear solid model.
 $k = 0$, $E = 1$, $\gamma = .01$, $\delta = 1000$, $I = 0$.

Index n	Complex conjugates
1	$-.125E - 4 \pm 1.57i$
2	$-.111E - 3 \pm 4.71i$
3	$-.309E - 3 \pm 7.85i$
10	$-.445E - 2 \pm 29.9i$
50	$-.118 \pm 156i$
100	$-.445 \pm 313i$
200	$-1.41 \pm 628i$
300	$-2.35 \pm 943i$
600	$-3.88 \pm 1891i$
900	$-4.41 \pm 2838i$

The *inclusion of moment of inertia* is displayed in Figs. 8-10 for $I = 0.1$, 1.0, and 10.0, respectively. It is evident from these figures that, for our values of k , the addition of the moment of inertia shifts these eigenvalues down and to the right toward the origin, and decreases the effect of k on the modes.

Figure 8 ($I = 0.1$) is drawn to the same scale as Fig. 7 ($I = 0$) and illustrates the addition of a small amount of I . In Fig. 8, the situation at $k = 0$ is, of course, similar

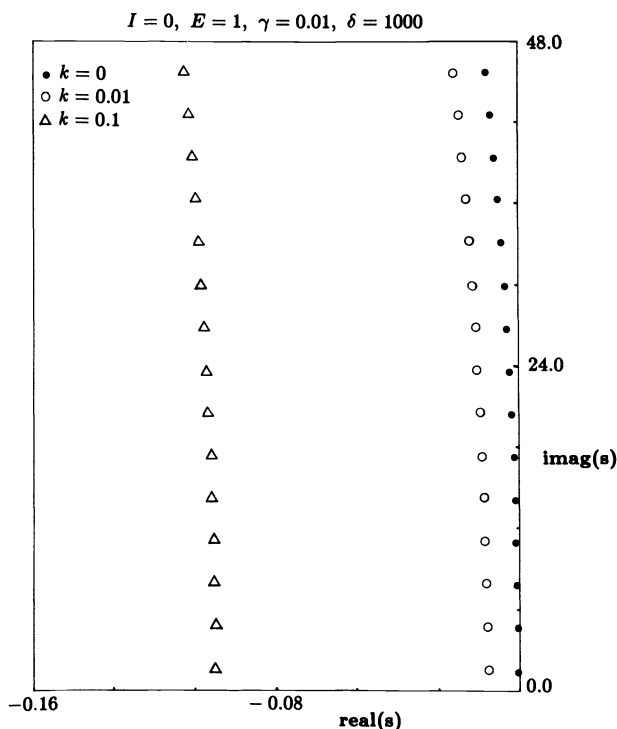


FIG. 7. $I = 0$, $E = 1$, $\gamma = 0.01$, $\delta = 1000$. The complex conjugate modes are stabilized much as k increases from 0 to 0.1. Other modes are out of this picture for this interval of k .

to that in Fig. 7. The worst mode at $k = 0$ is again the lowest with real part $O(10^{-5})$. However, the picture for $k \neq 0$ is obviously different from that for $I = 0$. Higher modes change little as k varies from 0 to 0.1. On the other hand, low modes are affected as dramatically as the case $I = 0$. We remark that at $k = 0.1$, intermediate modes *swing back* toward the imaginary axis and higher ones swing away toward the asymptote. Thus, the location of the worst mode moves around as k varies. After an extensive search, we find that the eleventh mode is the least stable at $k = 0.1$ with $s = -.014 \pm 34.8i$, and a few neighboring modes have comparable real parts. Therefore, at $I = 0.1$, the feedback parameter k has stabilized the worst decay rate by $O(10^3)$.

At first glance, we may become concerned that there may be other swing-backs of higher modes toward the imaginary axis. This does not occur for our parameters. This is because $|\alpha(s)|$ is almost a constant if $|s|$ is sufficiently large, and for our parameters, this holds even at moderate $|s|$. In addition, when $|Is|$ is much larger than the constant $|\alpha(s)|$, the characteristic equation becomes $e^{2\beta(s)} = 1$, regardless of k . The real parts of the roots of this cubic equation ((4.9) with $c_n = n^2\pi^2$) approach the asymptote as a monotonic function of n . Thus, the swing-back in Fig. 8 occurs at modes where $|s|$ is only moderately large; then the eigenvalues swing back to approach the solutions of the cubic equation.

In Fig. 9, the low modes are less stable than those of Figs. 7 ($I = 0$) and 8 ($I = 0.1$). At $I = 1.0$, the worst mode at $k = 0$ is the first at $s = -.37E - 5 \pm 0.86i$. At $k = 0.01$, it is the third at $s = -.44E - 3 \pm 6.4i$. At $k = 0.1$, it is the fifth at $s = -.14E - 2 \pm 12.6i$. Thus, there is an $O(10^2)$ improvement. In the scale of Fig. 10, only the first mode appears to vary but there is also much stabilization of the second mode. Here, the worst mode at $k = 0$ is the first at $s = -.5E - 6 \pm 0.31i$. At $k = 0.01$, it is the second at

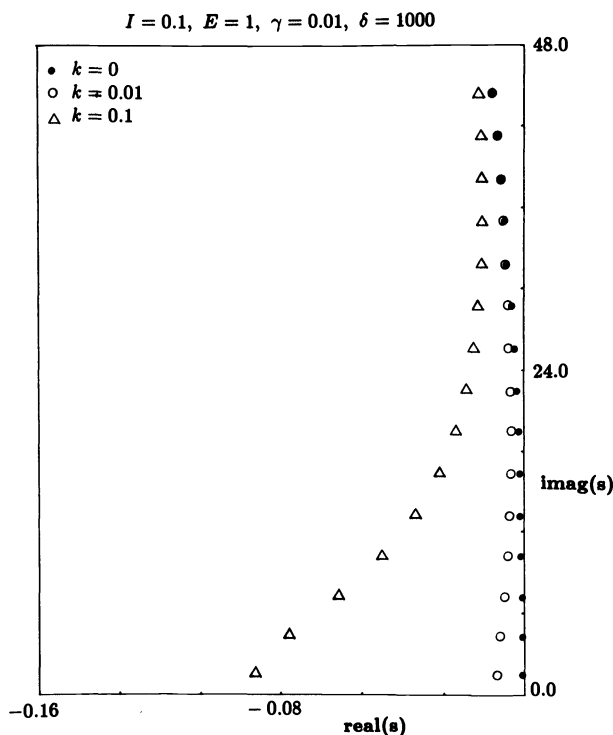


FIG. 8. $I = 0.1, E = 1, \gamma = 0.01, \delta = 1000$. The figure shows the stabilization of the complex conjugate modes as k is increased from 0 to 0.1. Other modes are out of this picture for this interval of k .

$s = -6E - 4 \pm 3.2i$. At $k = 0.1$, it is again the second at $s = -15E - 3 \pm 3.2i$. The overall stabilization is slightly worse than at $I = 1.0$ but is still $O(10^2)$.

4.2.3. The optimal choice for k . We search for an optimal choice for k in the nearly elastic situation discussed in the previous section. A number of questions arise, such as: by what order of magnitude does an optimal k stabilize the system? Is this magnitude sensitive to small changes in k ? Is it possible to choose k so that all modes are stabilized to the same level as the highly oscillatory modes? We investigate these questions with reference to the parameters $E = 1, \gamma = .01, \delta = 1000$. We show that if the moment of inertia is zero, it is possible to choose k so that the system is almost as stabilized as the high modes are at $k = 0$. For nonzero moment of inertia, the maximum amount of stabilization is not as great and the optimal k may be larger. For both $I = 0$ and $I \neq 0$, the magnitude of improvement achieved with the optimal k is not very sensitive to changes in k . The complex conjugate modes of Class 1 and the popper of Class 4 enter into our discussion.

At $I = 0, E = 1, \gamma = 0.01, \delta = 1000$, the dependence of low modes on k is qualitatively similar to Fig. 1, but with the lowest Class 1 modes being markedly closer to the imaginary axis at $k = 0$ and ∞ . For $k > 0$, the Class 4 mode pops out of the essential singularity $s^* = -990.099 \dots$, which is quite a distance away from the Class 1 modes. The low Class 1 modes loop up at about $\text{Re } s = -6$. The popper travels toward the origin as k increases, first slowly and then speeding up when k is about 1. When k is slightly larger than 1, the popper is close to the real parts of the low Class 1 modes; then, for large k , it is inversely proportional to k .

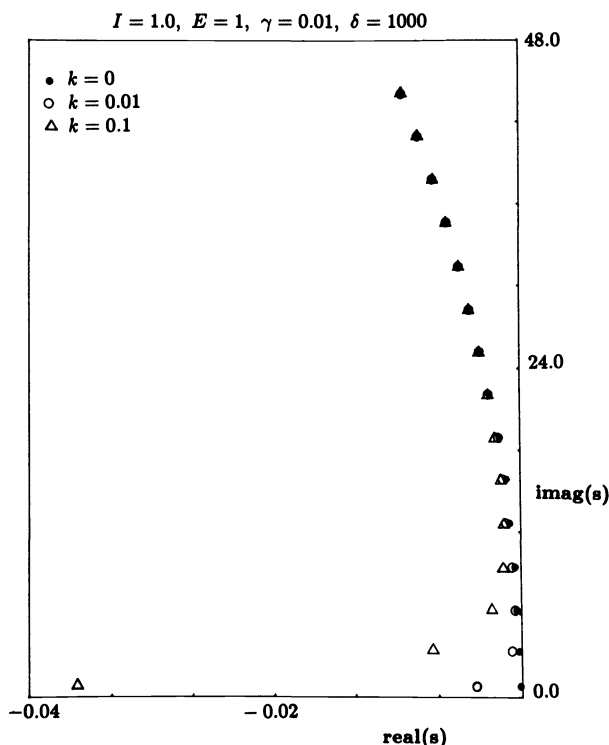


FIG. 9. $I = 1.0$, $E = 1$, $\gamma = 0.01$, $\delta = 1000$. The stabilization of the complex conjugate modes is displayed for k varying between 0 and 0.1. Other modes are out of this picture for this interval of k .

Since the asymptote for the high modes is maximally stabilized at $k = \sqrt{E + \gamma}$, we may guess that this value of k is optimal. However, when k has reached $\sqrt{E + \gamma} = 1.005$, the low modes have rounded their loops and are on their way back to the imaginary axis. (In situations that are not nearly elastic, it is possible that low modes round their loops at larger k and that $\sqrt{E + \gamma}$ is not a guideline.) When k is slightly less than $\sqrt{E + \gamma}$, the low modes appear to be as much damped as possible, and the high modes even more. There is a slight swing of intermediate modes toward the imaginary axis. The maximum amount of damping possible for each mode is different and accounts for the swing-back. Table 2 lists eigenvalues for $k = 1$, which is close to optimal. There is a swing-back around the 50th to 100th modes. The worst mode is around the 70th with real part around -4 , which is comparable to the asymptote for high modes at $k = 0$. Referring to Table 1 for $k = 0$, the optimal k improves the worst mode by $O(10^5)$. This order of magnitude in improvement can be achieved as long as the worst modes are pushed back to $\text{Re } s \approx -1.0$. In this sense, a wide interval of k , e.g., 0.7 to 1.3, achieves the same magnitude of improvement over the situation with zero k .

As an example of the situation with moment of inertia, we consider $I = 0.1$, $E = 1$, $\gamma = .01$, $\delta = 1000$. At $k = 0$, the worst mode is the first complex conjugate mode with $s = -1E - 4 \pm 1.4i$. The complex conjugates loop down as $k \rightarrow \infty$, as in Figs. 2 and 3. At $k = 1.23$, the lowest complex conjugate mode drops to the real axis at $\text{Re } s \approx -2$, which is quite far from the cluster point s^* . This behavior is reminiscent of Fig. 2. One eigenvalue travels toward s^* and will meet the popper. The other travels toward the origin and becomes the least stable mode for $k \geq 13$. For reasons that have been discussed in § 4.2.2, the lowest modes are affected first by the variation in k and higher

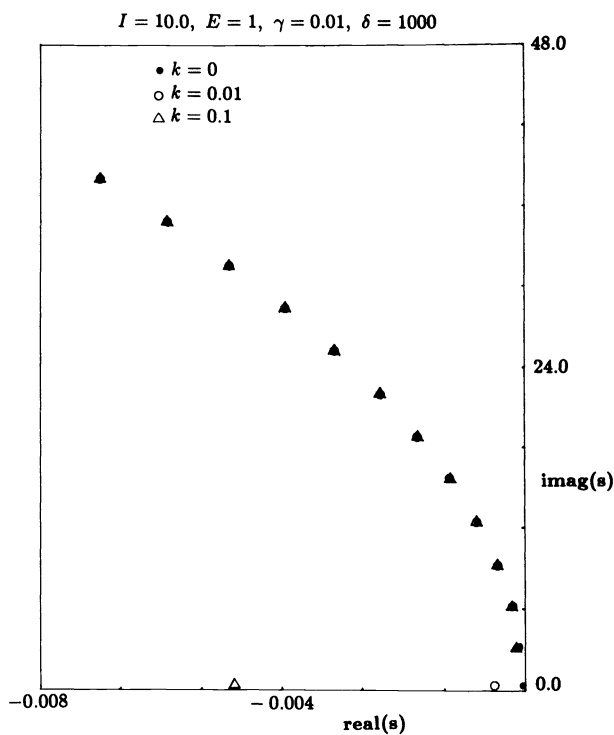


FIG. 10. $I = 10.0, E = 1, \gamma = 0.01, \delta = 1000$. The stabilization of the complex conjugate modes is displayed for k varying from 0 to 0.1. Other modes are out of this picture for this interval of k .

TABLE 2
Standard linear solid model.
 $k = 1, E = 1, \gamma = .01, \delta = 1000, I = 0$.

n	Complex conjugates
1	$-5.53 \pm 2.90i$
2	$-5.41 \pm 5.87i$
3	$-5.28 \pm 8.90i$
10	$-4.73 \pm 30.69i$
50	$-4.05 \pm 156i$
100	$-4.05 \pm 313i$
200	$-4.74 \pm 628i$
300	$-5.55 \pm 944i$
600	$-6.96 \pm 1891i$
900	$-7.46 \pm 2839i$

modes require larger k in order to move. Due to this lag, one of the higher modes is the least stable for most of $k \in [0, 13]$.

At $k = 1.2$, low complex conjugates start looping back toward the imaginary axis, but unlike the case in which $I = 0$, this does not indicate the optimal k . A search shows that k of about 8 is optimal. At $k = 8$, the worst mode is the 27th complex conjugate pair with $s = -.94E - 1 \pm 78.6i$. The improvement over the case of zero k is then $O(10^4)$. However, an improvement of $O(10^3)$ is gained for any k between 0.1 and 100, i.e., once the worst mode has been pushed to $\text{Re } s = O(0.01)$, the magnitude of improvement

is rather insensitive to k . For larger I , we find that the modes move around less, so that we expect the worst modes at larger I to be worse than at lower I . Thus, we expect the overall improvement achieved by an optimal k to diminish as I increases.

4.3. The fractional derivative model and the intermediate model. We begin with the fractional derivative model modified by an exponential factor (4.5), and end this section with a brief discussion of the intermediate model (4.6). Our results are based on computations for $\alpha = \frac{1}{2}$ in (4.5) and (4.6). For both models, the types of eigenvalues that arise and their journeys through the complex plane as parameters are varied are found to be reminiscent of the standard linear solid model.

For the fractional derivative model, our results concern the case $\alpha = \frac{1}{2}$, $E = 1$, $\gamma = .01$, $\delta = 5$. These parameters were chosen for the same reasons as those stated at the beginning of § 4 to justify the choice used in §§ 4.2.2 and 4.2.3: at $k = 0$, the system is close to criticality. For *zero moment of inertia*, the general picture of the eigenvalues as k varies from 0 to ∞ is analogous to Fig. 1, but the Class 3 mode is absent. At $k = 0$, there are modes corresponding to Classes 1, 2, and 4. The difference from the standard linear solid, apart from the asymptotics, is that there is now a branch cut along $(-\infty, -\delta)$ and no eigenvalues stay on that portion of the axis. The Class 2 cluster lies in $(-\delta, s^*)$, where s^* is the essential singularity $(1 - \sqrt{1 + 4\delta\gamma^2/E^2})E^2/2\gamma^2$. For $k > -Is^*$, a Class 4 mode pops out of s^* .

At small k , low complex conjugate modes are the least damped. The first several modes are listed in Table 3. The asymptotic formula for highly oscillatory modes is

$$s_n \sim [\pi^2(2n-1)^2\gamma/4]^{2/3} \left(-\frac{1}{2} \pm i \frac{\sqrt{3}}{2} \right) \quad \text{as } n \rightarrow \infty.$$

Observe that the highly oscillatory modes exhibit frequency proportional or “structural” damping. This behavior is approached slowly if parameters are nearly elastic, and fast if parameters are all moderate. Thus, for our parameters, even $s_{10,000} = -18300. \pm 51500i$ is not yet close to this formula. As k increases, a higher mode may be the least stable mode. For $k > 1$, the Class 4 popper becomes the least stable mode. Since the situation of interest is nearly elastic, and the modes are qualitatively similar to those of the standard linear solid, the search for an optimal k focuses on the low to moderate complex conjugates and the popper. We find that k close to one (i.e., slightly less than $\sqrt{E + \gamma}$) is again optimal and the worst mode recedes to real part -2.6 , an improvement of $O(10^3)$. The effect of k on the worst modes is shown in Table 3.

For *nonzero moment of inertia*, the overall journey of the modes is similar to the standard linear solid except that the branch cut $(-\infty, -\delta)$ is free of eigenvalues. For

TABLE 3
Fractional derivative model, $\alpha = \frac{1}{2}$, $E = 1$, $\gamma = .01$, $\delta = 5$, $I = 0$.
Table of real parts of low complex conjugate modes.
These are the least stable modes for $k \leq 1$.

n	$k = 0$	$k = 0.01$	$k = 0.1$	$k = 1.0$
1	$-.53E-2$	$-.15E-1$	$-.11$	-2.67
2	$-.4E-1$	$-.5E-1$	$-.14$	-2.62
3	$-.9E-1$	$-.1$	$-.19$	-2.58
4	$-.15$	$-.16$	$-.25$	-2.58
5	$-.21$	$-.22$	$-.31$	-2.59
6	$-.28$	$-.29$	$-.38$	-2.62
7	$-.36$	$-.37$	$-.46$	-2.66

example, computations at $I = 0.1$ reveal the overall picture to be analogous to Fig. 3 ($I = 0.4$, $E = \gamma = \delta = 1$ for the standard linear solid) for modes higher than the first. The first branch drops to the negative real axis to the right of s^* as in Fig. 2. The main difference is that after the eigenvalues jump over s^* and $-\delta$, they travel out, moving away from the real axis. For large negative real part, these complex conjugates have $\operatorname{Re} s \sim -k/I$ as $k \rightarrow \infty$ with nonzero imaginary parts. Figure 11 displays the effectiveness of k , both small and large, on the least stable modes. Here, $I = 0.1$ and $k = 0, 0.1$ and 1.0 . This figure is analogous to Fig. 8 ($I = 0.1$, $E = 1$, $\gamma = 0.01$, $\delta = 1000$, $k = 0, 0.01, 0.1$) for the standard linear solid. The modes on the real axis are out of this picture for our values of k . The worst mode at $k = 0$ is $-0.004 \pm 1.4i$ and this is stabilized 20-fold with $k = 0.1$. After the first several modes, however, the effectiveness of small k wears off and higher modes are close to the roots of $e^{2\beta(s)} = 1$, regardless of k . Just as in Figs. 8–10, increasing the moment of inertia results in a decrease in the effect of k , but the order of magnitude of improvement does not drop drastically. At $I = 10$, there is still a 20-fold improvement in the worst decay rate. In Fig. 11, the first two modes at $k = 1$ are out of the picture while intermediate modes display the same type of swing-back as in Fig. 8 for $k = 0.1$. After the modes have swung back, they approach the roots of $e^{2\beta(s)} = 1$.

We search for an optimal k for the case $I = 0.1$. At $k = 0$, the least stable mode is the lowest, but just as with the standard linear solid, when k is nonzero, the lowest is not always the least stable. Just as in Fig. 2, the lowest conjugate pair drops to the axis to the right of s^* . The junction occurs at $k = 1.23$. One mode travels toward s^*

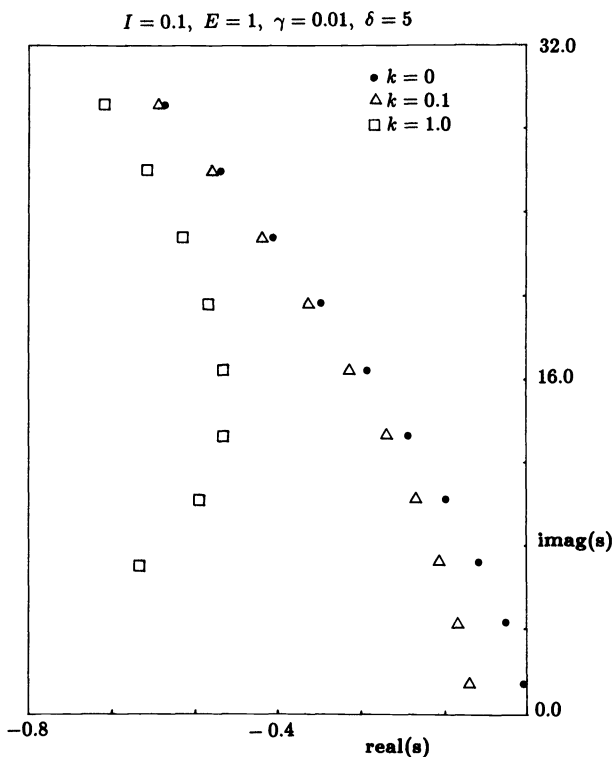


FIG. 11. Fractional derivative model with $\alpha = \frac{1}{2}$, $I = 0.1$, $E = 1$, $\gamma = 0.01$, $\delta = 5$. The stabilization of the complex conjugate modes is displayed for k varying from 0 to 1. Other modes are out of this picture for this interval of k .

and meets the popper. The other travels toward the origin. Meanwhile, one of the higher modes is the worst mode. The optimal k is close to $k = 1.8$, where the worst mode is $-.55 \pm 13i$. For k larger than about 2, one of the complex conjugates that dropped to the real axis becomes the least stable mode, approaching the origin as $k \rightarrow \infty$. The optimal k yields an $O(10^2)$ improvement in stabilization. This is an order of magnitude less than the improvement at $I = 0$. These magnitudes of improvement are attainable in a wide interval of k around the optimal value.

For the *intermediate model*, the main features are similar to the fractional derivative model already presented. For example, at zero moment of inertia, there are modes corresponding to Classes 1, 2, and 4 of the standard linear solid, and Class 3 is absent because there is a branch cut at $(-\infty, -\delta)$. The asymptotic behavior of high modes is different from the other models:

$$\operatorname{Im} s_n \sim \pi n \sqrt{E + \gamma/\sqrt{\delta}} \quad \text{as } n \rightarrow \infty \quad \text{and} \quad \operatorname{Re} s_n \sim -\gamma \pi^2 n^2 \sin \frac{\pi}{4} / (2 \operatorname{Im} s_n \sqrt{\operatorname{Im} s_n}).$$

Thus,

$$\operatorname{Re} s_n / \sqrt{\operatorname{Im} s_n} \rightarrow -\gamma \sin \frac{\pi}{4} / 2 \left(E + \frac{\gamma}{\sqrt{\delta}} \right).$$

The Class 4 mode pops out of the essential singularity $s^* = -\delta + \gamma^2 \delta / (E\sqrt{\delta} + \gamma)^2$ for $k > -Is^*$.

5. Existence and exponential decay. In this section we verify that our boundary stabilization problem has a solution (in a weak sense), whose rate of decay as $t \rightarrow \infty$ corresponds in the expected way to the location of the solutions of the characteristic equation (2.17) studied in §§ 3 and 4. We consider the problem (2.1)–(2.4), and we assume $k, \varepsilon, I \geq 0$, and

$$(5.1) \quad u_0 \in AC[0, 1] \quad \text{with } u_0(0) = 0.$$

We assume that the stress relaxation modulus $A(t)$ satisfies (1.13) and (1.14).

In this section, H^0 denotes the Hilbert space $L^2(0, 1)$ with norm $\|\cdot\|$; boldface denotes an element or operator in H^0 . We assume that $t \rightarrow \mathbf{F}_0(t) \equiv \mathbf{F}_0(\cdot, t) \in L^1_{\text{loc}}(\mathbb{R}^+, H^0)$, $f_0 \in L^1_{\text{loc}}(\mathbb{R}^+)$, and the Laplace transforms

$$\hat{f}_0(s) = \int_0^\infty e^{-st} f_0(t) dt \quad \text{and} \quad \hat{\mathbf{F}}_0(s) = \int_0^\infty e^{-st} \mathbf{F}_0(t) dt$$

converge for $\operatorname{Re} s > 0$. Moreover,

$$(5.2) \quad s\hat{f}_0(s) \text{ and } s\hat{\mathbf{F}}_0(s) \text{ have analytic extensions to } \{\operatorname{Re} s > -\eta\} \text{ that are bounded in each half-plane } \{\operatorname{Re} s > -\eta'\}, \eta' < \eta.$$

If, for example, $F_0(x, t) = u_1(x) + \int_0^t F(x, \tau) d\tau$, where u_1 and F arise from an initial jump in u_t and a memory term, respectively, as in § 1, then $u_1 \in H^0$ and

$$|u_x(1, t)| + \int_0^1 u_{xx}^2(x, t) dx \leq M < \infty \quad (t < 0),$$

together with (1.13), (1.14), will ensure that F_0 satisfies our requirements.

Let $\mathcal{D} = H^2 \cap H^1_0 \subset H^0$, and let $\mathbf{L}: \mathcal{D} \rightarrow H^0$ be the operator d^2/dx^2 . The adjoint of \mathbf{L} is $\mathbf{L}^*: H^0 \rightarrow \mathcal{D}^* = H^{-2}$. We let \mathbf{x} denote the identity function ($\mathbf{x}(x) = x$) in H^0 .

Define $\mathbf{R}(t): H^0 \rightarrow H^0$ by letting $\mathbf{w}(t) \equiv \mathbf{R}(t)\mathbf{w}_0$ be the solution of

$$\mathbf{w}'(t) = \int_0^t A(t-\tau) \mathbf{L} \mathbf{w}(\tau) d\tau, \quad \mathbf{w}(0) = \mathbf{w}_0$$

for $\mathbf{w}_0 \in \mathcal{D}$. It is known [6] that, under our hypotheses on $A(t)$, $\mathbf{R}(t)$ extends to a bounded operator on H^0 , and its operator norm satisfies

$$(5.3) \quad \|\mathbf{R}(t)\| \leq 1 \quad (0 \leq t < \infty), \quad \int_0^\infty \|\mathbf{R}(t)\| dt < \infty.$$

Furthermore,

$$(5.4) \quad \hat{\mathbf{R}}(s) = \hat{A}(s)^{-1}(\beta^2(s)\mathbf{I} - \mathbf{L})^{-1}.$$

Define $U(x, s)$ by (2.11), (2.12), for each s such that $\Delta(s) \neq 0$ and $0 \leq x \leq 1$. Let $\Phi(s) = U(1, s)$. By the elementary theory of boundary value problems, U is uniquely determined as the solution of

$$(5.5) \quad \begin{aligned} U_{xx}(x, s) - \beta^2 U(x, s) &= -[u_0(x) + \hat{F}_0(x, s)]/\hat{A}(s) \quad (\text{a.e.}), \\ \mathbf{U}(\cdot, s) - \mathbf{x}\Phi(s) &\in \mathcal{D}. \end{aligned}$$

To show that U is indeed the transform of a weak solution of (2.1)–(2.4), we shall show that $\mathbf{U}(\cdot, s)$ belongs to a Hardy space $\mathcal{H}^2(\{\operatorname{Re} s > \mu\}, H^0)$ with values in H^0 . The next result gives the required estimates.

LEMMA 5.1. (i) *Under the general assumptions of this section, suppose $\Delta(s) \neq 0$ ($\operatorname{Re} s \geq \mu \geq 0$). When $I = 0$, $\varepsilon > 0$, and $A(0^+) < \infty$, assume in addition that*

$$(5.6) \quad k < A(0^+)^{1/2} \frac{1-C}{1+C}$$

($C = 0$ if $A'(0^+) = -\infty$). Then

$$(5.7) \quad |U(x, s)| \leq \frac{Q}{1+|s|} \quad (\operatorname{Re} s > \mu, 0 \leq x \leq 1),$$

where Q is a constant that depends on I , k , μ , and ε , as well as on u_0 , f_0 , and F_0 .

(ii) *If $\varepsilon = 0$, then (5.7) holds for $\mu < 0$, provided that $\Delta(s) \neq 0$ in some larger half-plane $\{\operatorname{Re} s \geq \mu_1\}$, $\mu_1 < \mu$.*

Clearly, (5.7) implies $\mathbf{U}(\cdot, s) \in \mathcal{H}^2(\{\operatorname{Re} s > \mu\}, H^0)$, and $\Phi \in \mathcal{H}^2(\{\operatorname{Re} s > \mu\})$.

The proof of Lemma 5.1 involves estimates such as those employed in locating the zeros of Δ ; this proof appears in § 6. From Lemma 5.1 we can deduce our result on existence and asymptotic decay.

THEOREM 5.1. *Under the conditions of Lemma 5.1, there exist functions \mathbf{u} and φ , with $t \rightarrow e^{-\mu t} \mathbf{u}(t) \in L^2(\mathbb{R}^+, H^0)$, $t \rightarrow e^{-\mu t} \varphi(t) \in L^2(\mathbb{R}^+)$, such that $\hat{\mathbf{u}}(s) = \mathbf{U}(\cdot, s)$ and $\hat{\varphi} = \Phi$. Moreover, \mathbf{u} is the unique solution in $L^1_{\text{loc}}(\mathbb{R}^+, H^0)$ of*

$$(5.8) \quad \mathbf{u}(t) = \int_0^t B(t-\tau) \mathbf{L}^*[\mathbf{u}(\tau) - \mathbf{x}\varphi(\tau)] d\tau + \int_0^t \mathbf{F}_0(\tau) d\tau + \mathbf{u}_0$$

($B(t) = \int_0^t A(\tau) d\tau$), and we have the representation

$$(5.9) \quad \mathbf{u}(t) = \mathbf{R}(t)\mathbf{u}_0 + \int_0^t \mathbf{R}(t-\tau)\mathbf{F}_0(\tau) d\tau + \mathbf{x}\varphi(t) - \mathbf{z}(t),$$

where \mathbf{z} is defined by its transform

$$(5.10) \quad \hat{\mathbf{z}}(s) = s\Phi(s)[\mathbf{R}(\cdot)\mathbf{x}]^\wedge(s).$$

Remarks. Equation (5.8) is a weak, integrated form of (2.1)–(2.3), with the right boundary condition $\mathbf{u}(1, t) = \varphi(t)$ in place of (2.4).

Representation (5.9) can be used to define the weak solution \mathbf{u} when, instead of (5.7), we have only $\Phi \in \mathcal{H}^2$. This would happen in certain cases where (5.1) is weakened to $u_0 \in L^2(0, 1) \cap AC[x_0, 1]$, with $0 < x_0 < 1$, so that the integration by parts estimate of (6.15) in § 6 works only near $x = 1$.

Stronger hypotheses yield other representations and additional smoothness for u and φ . If, for example, $A(t) = E + \gamma t^{-1/2} e^{-\delta t}$, then $\beta(s) \sim s^{3/4} (s \rightarrow \infty)$, and $\mathbf{R}'(t)\mathbf{L}^{-\nu} \in L^1(\mathbb{R}^+, H^0)$ ($\nu > 0$) [7], and \mathbf{u} can be represented as

$$\mathbf{u}(t) = \mathbf{R}(t)\mathbf{u}_0 + \int_0^t \mathbf{R}(t-\tau)\mathbf{F}_0(\tau) d\tau - \int_0^t \mathbf{R}'(t-\tau)\mathbf{x}\varphi(\tau) d\tau.$$

If, in addition, $\mathbf{F}_0 = \mathbf{0}$, $f_0 = 0$, and u_0 is in $C^j[0, 1]$ with vanishing derivatives of order $0, 1, \dots, j-1$ at $x = 0, 1$, then we can improve the estimate of Lemma 5.1 by continuing to integrate by parts as in (6.15) and get

$$|\Phi(s)| \leq Q \left| \frac{e^\beta}{\beta^j \Delta} \right| \\ \sim \frac{Q}{|s|^{3j/4}} (I|s| + |s|^{1/4})^{-1} (s \rightarrow \infty).$$

By \mathcal{H}^2 theory, φ can then have derivatives in $L^2(\mathbb{R}^+)$. See [21], [27] for a systematic examination of regularity in related problems.

When $\mu \geq 0$, complete monotonicity can be replaced in Theorem 5.1 by the weaker conditions mentioned (for g) in § 1. When $k = 0$ and in the fixed-end case ($k = \infty$), the problem is self-adjoint, and stronger results can be obtained directly through separation of variables, as in [19].

Finally, if $\Delta(s_0) = 0$ with $\operatorname{Re} s_0 > 0$, and if we take $u_0(1) = 0$, $F_0 = 0$, $f_0 = 0$,

$$\int_0^1 G(1, y, s_0) u_0(y) dy = - \int_0^1 \sinh \beta(s_0)y u_0(y) dy \neq 0,$$

we see from (2.11) that Φ is unbounded at s_0 , and a solution with $u(1, \cdot) \in L^2(\mathbb{R}^+)$ cannot exist.

Proof of Theorem 5.1. By elementary \mathcal{H}^2 theory, the existence of \mathbf{u} and φ in the appropriate L^2 spaces, with transforms \mathbf{U} and Φ , respectively, is an immediate consequence of (5.7). Let $\mathbf{v}(t)$ denote the right-hand side of (5.9). By (5.3) and (5.7), relation (5.10) defines $\mathbf{z} \in L_{\text{loc}}^2(\mathbb{R}^+, H^0)$. Moreover, the second term of \mathbf{v} has the transform $\hat{\mathbf{R}}(s)\hat{\mathbf{F}}_0(s)$ in $\mathcal{H}^2(\{\operatorname{Re} s > 0\}, H^0)$, so it belongs to $L_{\text{loc}}^2(\mathbb{R}^+, H^0)$; the same is clearly true for the other terms of \mathbf{v} , and, indeed, $t \rightarrow e^{-\mu_0 t} \mathbf{v}(t) \in L^2(\mathbb{R}^+, H^0)$, with $\mu_0 = \max\{\mu, 0\}$. By evaluating Fourier coefficients with respect to the basis $\{\sin n\pi x\}_{n=1}^\infty$ in H^0 and taking transforms, we see that \mathbf{v} is the unique solution of (5.8). Finally, taking transforms in (5.9) and using (5.4), we see that

$$\hat{\mathbf{v}}(s) - \mathbf{x}\Phi(s) = (\mathbf{L} - \beta^2 \mathbf{I})^{-1} [\mathbf{x}s\Phi(s) - \hat{\mathbf{F}}_0(s) - \mathbf{u}_0] / \hat{A}(s).$$

Thus $\hat{\mathbf{v}}(s) - \mathbf{x}\Phi(s) \in \mathcal{D}$ and

$$\left(\frac{\partial^2}{\partial x^2} - \beta^2 \right) \hat{\mathbf{v}}(s)(x) = -(\hat{F}_0(x, s) + u_0(x)) / \hat{A}(s) \quad \text{a.e.,}$$

and we are back to (5.5), which determines U uniquely. This proves that $\hat{\mathbf{v}}(s) = \mathbf{U}(\cdot, s) = \hat{\mathbf{u}}(s)$ ($\operatorname{Re} s > \mu_0$), and the proof is complete. \square

6. Proofs of Lemmas 2.3 and 5.1.

Proof of Lemma 2.3. Fix $\sigma_0 < 0$. By (2.8), (2.9), and (2.24),

$$(6.1) \quad \beta^2(s) = s/\hat{A}(s) = \frac{\sigma\varphi - \tau\psi + i(\tau\varphi + \sigma\psi)}{\phi^2 + \psi^2} \quad (s = \sigma + i\tau).$$

Here to simplify notation we have written $\hat{A}(s) = \varphi_\sigma(\tau) - i\psi_\sigma(\tau) = \varphi - i\psi$. We begin by showing that there exists $\tau_0 = \tau_0(\sigma_0) > 0$ such that

$$(6.2) \quad \operatorname{Im} \beta^2(s) > 0 \quad \text{for } \tau \geq \tau_0, \quad 0 > \sigma \geq \sigma_0.$$

By (6.1) and (2.24), (6.2) also holds whenever $\tau > 0$ and $\sigma \geq 0$. Once (6.2) is proved, we note that, since $\beta(\bar{s}) = \overline{\beta(s)}$, the formula $\beta(s) = (s/\hat{A}(s))^{1/2}$ (principal square root) is valid whenever $\operatorname{Re} s \geq \sigma_0$ and $|s| \geq R$ for some sufficiently large R . In particular, $\exp(-2\beta(s))$ is bounded and analytic for $\operatorname{Re} s \geq \sigma_0$, $|s| \geq R$, so by Lindelöf's theorem and $\beta(\bar{s}) = \overline{\beta(s)}$, it suffices to prove that

$$(6.3) \quad \operatorname{Re} \beta(\sigma_0 + i\tau) \rightarrow \infty \quad \text{as } \tau \rightarrow \infty.$$

Returning to the proof of (6.2), note that by (2.24)

$$\begin{aligned} \tau(\tau\varphi + \sigma\psi) &= \int_0^\infty \frac{\tau^2(x+2\sigma)}{(\sigma+x)^2 + \tau^2} d\mu(x) \\ &\cong \int_0^{-2\sigma} \frac{2\tau^2\sigma}{(\sigma+x)^2 + \tau^2} d\mu(x) + \int_{-2\sigma}^{\tau-2\sigma} \frac{\tau^2(x+2\sigma)}{(\sigma+x)^2 + \tau^2} d\mu(x) \end{aligned}$$

when $\sigma_0 \leq \sigma \leq 0$, $\tau \geq -2\sigma_0$. By the dominated convergence theorem the first integral on the right-hand side of this inequality tends to $2\sigma \int_0^{-2\sigma} d\mu(x)$, as $\tau \rightarrow \infty$. The second integral is bounded from below by

$$2^{-1} \int_{-2\sigma}^{\tau-2\sigma} (x+2\sigma) d\mu(x) \geq 4^{-1} \int_{-4\sigma_0}^{\tau} x d\mu(x) \rightarrow \infty$$

as $\tau \rightarrow \infty$ by (2.13). Thus,

$$(6.4) \quad \tau(\tau\varphi + \sigma\psi) \rightarrow \infty \quad \text{as } \tau \rightarrow \infty \quad \text{uniformly for } \sigma_0 \leq \sigma \leq 0.$$

In particular, (6.2) is proved.

We now turn to the proof of (6.3). In order to simplify notation, we drop the subscript and show that (6.3) holds for a fixed $\sigma = \sigma_0 < 0$. We begin by showing that

$$(6.5) \quad \sigma\varphi_\sigma(\tau) \rightarrow 0 \quad \text{and} \quad \tau\psi_\sigma(\tau) \rightarrow A(0^+) \quad \text{as } \tau \rightarrow \infty.$$

To verify the first limit in (6.5), write

$$\sigma\varphi_\sigma(\tau) = \int_0^\infty \frac{\sigma(\sigma+x)(1+x)}{(\sigma+x)^2 + \tau^2} \frac{d\mu(x)}{1+x}.$$

Since

$$(6.6) \quad \hat{a}(0) = \int_{0^+}^\infty \frac{d\mu(x)}{x} < \infty,$$

the dominated convergence theorem shows that $\sigma\varphi_\sigma(\tau) \rightarrow 0$ as $\tau \rightarrow \infty$ and the first part of (6.5) is proved.

If $A(0^+) < \infty$, it follows from (2.24), (2.13), and the dominated convergence theorem that $\tau\psi_\sigma(\tau) \rightarrow A(0^+)$ as $\tau \rightarrow \infty$. If $A(0^+) = \infty$, then by (2.24) and (2.13)

$$\tau\psi_\sigma(\tau) \geq \frac{1}{2} \int_0^{\tau-\sigma} d\mu(x) \rightarrow \infty \quad \text{as } \tau \rightarrow \infty,$$

and the second part of (6.5) is proved.

Writing $\beta^2(s) = \rho_\sigma(\tau) e^{i\theta_\sigma(\tau)} \equiv \rho(\tau) e^{i\theta(\tau)} (s = \sigma + i\tau)$, we have by (6.1), (6.2), and (6.5) that

$$\rho(\tau) = |s/\hat{A}(s)|, \quad \tan \theta(\tau) = \frac{\tau\phi + \sigma\psi}{\sigma\phi - \tau\psi},$$

with $\theta(\tau) \in (\pi/2, \pi)$ for all large positive τ .

For τ such that $\tau\phi + \sigma\psi \geq \tau\psi - \sigma\phi > 0$, $\pi/2 < \theta(\tau) \leq 3\pi/4$, and $\operatorname{Re} \beta(\sigma + i\tau) \geq |s/\hat{A}(s)|^{1/2} \cos 3\pi/8 \rightarrow \infty (\tau \rightarrow \infty)$.

For the more difficult case where τ is such that

$$(6.7) \quad 0 < \tau\phi + \sigma\psi < \tau\psi - \sigma\phi,$$

we use the identity $2\cos^2 \theta/2 = 1 + \cos \theta$ and the estimate

$$(1+x^2)^{-1/2} \leq 1 - (x^2/4) (|x| < 1)$$

to get

$$(6.8) \quad \operatorname{Re} \beta(\sigma + i\tau) \geq 8^{-1/2} |s/\hat{A}(s)|^{1/2} \left[\frac{\tau\phi + \sigma\psi}{\tau\psi - \sigma\phi} \right].$$

If $A(0^+) < \infty$, note that (6.8) implies

$$\operatorname{Re} \beta(\sigma + i\tau) \geq 8^{-1/2} |s\hat{A}(s)|^{-1/2} \frac{\tau(\tau\phi + \sigma\psi)}{\tau\psi - \sigma\phi},$$

then use (6.4), (6.5), and $s\hat{A}(s) \rightarrow A(0^+)$ ($\tau \rightarrow \infty$) to conclude that (6.3) ($\sigma_0 = \sigma$) holds in this case.

Finally, consider the case where $A(0^+) = \infty$. Since (6.7) is assumed to hold, $\phi_\sigma(\tau) \leq 2\psi_\sigma(\tau)$ for all sufficiently large τ . Combining this with (6.5) and (6.8), we can find a constant $M > 0$ so that

$$(6.9) \quad \operatorname{Re} \beta(\sigma + i\tau) \geq M \frac{(\tau\psi)^{1/2}}{\tau\psi^2} (\tau\phi + \sigma\psi)$$

for all large τ . Note that

$$\int_{0^+}^{\infty} \frac{\tau}{(\sigma+x)^2 + \tau^2} d\mu(x) \geq (2\tau)^{-1} \int_{0^+}^{\tau-\sigma} d\mu(x),$$

and since $A(0^+) = \infty$, it follows from (2.13) that

$$(6.10) \quad \psi_\sigma(\tau) \leq 2 \int_{0^+}^{\infty} \frac{\tau}{(\sigma+x)^2 + \tau^2} d\mu(x)$$

for all large τ . Also, by (6.6) and the inequality due to Schwarz,

$$\begin{aligned}
 & \left\{ \int_{0^+}^{\infty} \frac{\tau}{(\sigma+x)^2 + \tau^2} d\mu(x) \right\}^2 \\
 (6.11) \quad & \leq \left\{ \int_{0^+}^{\infty} \frac{\tau^2}{x[(\sigma+x)^2 + \tau^2]} d\mu(x) \right\} \left\{ \int_0^{\infty} \frac{x}{(\sigma+x)^2 + \tau^2} d\mu(x) \right\} \\
 & \leq \hat{a}(0) \int_0^{\infty} \frac{x}{(\sigma+x)^2 + \tau^2} d\mu(x).
 \end{aligned}$$

Finally, the technique used to prove (6.4) can also be used to show that

$$\tau\phi + \sigma\psi \geq 2^{-1} \int_0^{\infty} \frac{\tau x}{(\sigma+x)^2 + \tau^2} d\mu(x)$$

for all large τ . Now, combining this inequality with (6.9)–(6.11) and using (6.5), we obtain

$$\operatorname{Re} \beta(\sigma + i\tau) \geq \frac{M}{8\hat{a}(0)} (\tau\psi_{\sigma}(\tau))^{1/2} \rightarrow \infty \quad \text{as } \tau \rightarrow \infty,$$

and the proof of (6.3) is complete. As we noted earlier, Lemma 2.3 is proved. \square

Proof of Lemma 5.1. We claim first that

$$(6.12) \quad \frac{1}{|\Delta(s)|} \leq \frac{Q|e^{-\beta}|}{I|s| + |\alpha|} \quad (\operatorname{Re} s > \mu).$$

(Throughout this proof, Q denotes some constant as in the statement of Lemma 5.1.) Since $\Delta(s) \neq 0$ for $\operatorname{Re} s \geq \mu$, we are concerned only with $\operatorname{Re} s \geq \mu$ and $|s|$ large. Write

$$(6.13) \quad 2\Delta(s) = e^{\beta} [Is + k e^{-\varepsilon s} + \alpha] \left(1 - e^{-2\beta} \frac{Is + k e^{-\varepsilon s} - \alpha}{Is + k e^{-\varepsilon s} + \alpha} \right).$$

First suppose that $\varepsilon = 0$, so that μ can be negative. By (2.15) and (2.16), $\mu > s^*$ and $\operatorname{Re} \alpha > 0$ ($\operatorname{Re} s \geq \mu$). When $A'(0^+) = -\infty$, (6.12) is immediate from (3.1), (3.2), and (2.19). When $A'(0^+) > -\infty$, an infinite sequence of zeros of Δ is asymptotic to the line $\operatorname{Re} s = \sigma_*$ of (3.5), so $\mu > \sigma_*$, except in the special case $I = 0$, $k = A(0^+)^{1/2}$, where σ_* is undefined. Hence, by (3.4) and (2.21), the last factor on the right in (6.13) is bounded away from zero for $\operatorname{Re} s \geq \mu$ and $|s|$ large. Using (2.18) and $\operatorname{Re} \alpha > 0$, we get (6.12).

Now suppose $\varepsilon > 0$, so that $\mu \geq 0$. When $I > 0$ or $A(0^+) = \infty$, (6.12) is evident from (6.13), together with (2.18), (2.19), and Lemma 2.3.

If, on the other hand, $A(0^+) < \infty$ and $I = 0$, we have

$$(6.14) \quad 2\Delta(s) = e^{\beta} \alpha (1 + e^{-2\beta}) \left[1 + \frac{k e^{-\varepsilon s}}{\alpha} \left(\frac{1 - e^{-2\beta}}{1 + e^{-2\beta}} \right) \right].$$

When $A'(0^+) = -\infty$, we have $e^{-2\beta} \rightarrow 0$ and $\alpha \rightarrow A(0^+)^{1/2}$ as $|s| \rightarrow \infty$ (Lemmas 2.1 and 2.3), and (5.6) yields (6.12). When $A'(0^+) > -\infty$, Lemmas 2.1 and 2.2 and (5.6) give us

$$\limsup_{\substack{|s| \rightarrow \infty \\ \operatorname{Re} s \geq \mu}} \left| \frac{k e^{-\varepsilon s}}{\alpha} \left(\frac{1 - e^{-2\beta}}{1 + e^{-2\beta}} \right) \right| \leq \frac{k}{A(0^+)^{1/2}} \left(\frac{1 - C}{1 + C} \right) < 1,$$

and (6.12) follows from (6.14). This establishes (6.12) in all cases.

Now consider $U(x, s)$ in (2.11). Since (5.2) holds and $\beta(s) = E^{-1/2}s + o(s)$ ($s \rightarrow 0$), U is analytic and locally bounded in $\{\operatorname{Re} s \geq \mu\}$, uniformly in $\{0 \leq x \leq 1\}$. Observe next that

$$(6.15) \quad \int_0^1 G(x, y, s) u_0(y) dy = G_1(x, 1, s) u_0(1) - \int_0^1 G_1(x, y, s) u_0'(y) dy,$$

where $G_1(x, y, s) = \int_x^y G(x, \tau, s) d\tau = O[(e^\beta/\beta)(Is/\alpha + 1)](|s| \rightarrow \infty, \operatorname{Re} s \geq \mu)$, uniformly in $0 \leq x, y \leq 1$. By (6.12) and (5.2), and since $\alpha\beta = s$, we get (5.7). This proves Lemma 5.1. \square

Acknowledgment. We thank Professor John A. Burns for helpful suggestions concerning the subject of this paper.

REFERENCES

- [1] R. L. BAGLEY, *Applications of generalized derivatives to viscoelasticity*, Ph.D. thesis, Air Force Inst. of Technology; Report TR-79-4103, Air Force Materials Laboratory, 1979.
- [2] R. L. BAGLEY AND P. J. TORVIK, *Fractional calculus—a different approach to the analysis of viscoelastically damped structures*, AIAA J., 21 (1983), pp. 741–748.
- [3] ———, *Fractional calculus in the transient analysis of viscoelastically damped structures*, AIAA J., 23 (1985), pp. 918–925.
- [4] L. BOLTZMANN, *Zur Theorie der elastischen Nachwirkung*, Ann. Physik (7) (1876), Ergänzungsband, pp. 624–625.
- [5] J. A. BURNS, E. M. CLIFF, AND R. E. MILLER, *Modeling and advanced control concepts*, Report, Mathematics Dept., Virginia Polytechnic Inst. and State Univ., Blacksburg, VA, 1986.
- [6] R. W. CARR AND K. B. HANNSEN, *A nonhomogeneous integrodifferential equation in Hilbert space*, SIAM J. Math. Anal., 10 (1979), pp. 961–984.
- [7] ———, *Resolvent formulas for a Volterra equation in Hilbert space*, SIAM J. Math. Anal., 13 (1982), pp. 459–483.
- [8] M. L. CARTWRIGHT, *Integral Functions*, in Cambridge Tracts in Mathematics and Mathematical Physics, 44, Cambridge University Press, Cambridge, 1962.
- [9] G. CHEN, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–273.
- [10] ———, *A note on the boundary stabilization of the wave equation*, SIAM J. Control Optim., 19 (1981), pp. 106–113.
- [11] G. CHEN, M. C. DELFOUR, A. M. KRALL, AND G. PAYRE, *Modeling, stabilization and control of serially connected beams*, SIAM J. Control Optim., 25 (1987), pp. 526–546.
- [12] G. CHEN, S. G. KRANTZ, D. W. MA, C. E. WAYNE, AND H. H. WEST, *The Euler–Bernoulli beam equation with boundary energy dissipation*, in Operator Methods for Optimal Control Problems, Marcel Dekker, New York, 1988, pp. 67–96.
- [13] R. DATKO, *A procedure for determination of the exponential stability of certain differential-difference equations*, Quart. Appl. Math., 36 (1978), pp. 279–292.
- [14] ———, *Not all feedback stabilized hyperbolic systems are robust with respect to small time delays in their feedbacks*, SIAM J. Control Optim., 26 (1988), pp. 697–713.
- [15] R. DATKO, J. LAGNESE, AND M. P. POLIS, *An example of the effect of time delays in boundary feedback stabilization of wave equations*, SIAM J. Control Optim., 24 (1986), pp. 152–156.
- [16] W. DESCH AND R. K. MILLER, *Exponential stabilization of Volterra integrodifferential equations in Hilbert space*, J. Differential Equations, 70 (1987), pp. 366–389.
- [17] J. S. GIBSON, *A note on stabilization of infinite dimensional linear oscillators by compact feedback*, SIAM J. Control Optim., 18 (1980), pp. 311–316.
- [18] D. GOTTLIEB AND S. A. ORSZAG, *Numerical Analysis of Spectral Methods: Theory and Applications*, CBMS-NSF Regional Conference Series in Applied Mathematics, 26, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1983.
- [19] K. B. HANNSEN AND R. L. WHEELER, *Uniform L^1 behavior in classes of integrodifferential equations with completely monotonic kernels*, SIAM J. Math. Anal., 15 (1984), pp. 579–594.
- [20] ———, *Time delays and boundary feedback stabilization in one-dimensional viscoelasticity*, in Proc. Third Internat. Conference on Distributed Parameter Systems, F. Kappel, K. Kunish, and W. Schappacher, eds., Springer-Verlag, Berlin, New York, 1987, pp. 136–152.

- [21] W. J. HRUSA AND M. RENARDY, *On wave propagation in linear viscoelasticity*, Quart. Appl. Math., 43 (1985), pp. 237–254.
- [22] H. KOLSKY, *Stress Waves in Solids*, Dover, New York, 1963.
- [23] J. U. KIM AND Y. RENARDY, *Boundary control of the Timoshenko beam*, SIAM J. Control Optim., 25 (1987), pp. 1417–1429.
- [24] J. LAGNESE, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163–182.
- [25] ———, *Boundary stabilization of linear elastodynamic systems*, SIAM J. Control Optim., 21 (1983), pp. 968–984.
- [26] I. LASIECKA AND R. TRIGGIANI, *Exponential uniform stabilization of the wave equation with $L_2(0, \infty; L_2(\Gamma))$ boundary feedback acting in the Dirichlet boundary conditions*, in Proc. 24th IEEE Conference on Decision and Control, Fort Lauderdale, FL, 1985, pp. 123–125.
- [27] J. PRUESS, *Positivity and regularity of hyperbolic Volterra equations in Banach space*, Math. Ann., 279 (1987), pp. 317–344.
- [28] J. P. QUINN AND D. L. RUSSELL, *Asymptotic stability and energy decay rates for solutions of hyperbolic equations with boundary damping*, Proc. Roy. Soc. Edinburgh, 77A (1977), pp. 97–127.
- [29] M. RENARDY, W. J. HRUSA, AND J. A. NOHEL, *Mathematical Problems in Viscoelasticity*, Longman, Essex, U.K., 1987.
- [30] L. ROGERS, *Operators and fractional derivatives for viscoelastic constitutive equations*, J. Rheology, 27 (1983), pp. 351–372.
- [31] D. L. RUSSELL, *Mathematical models for the elastic beam and their control-theoretic implications*, in Semigroups, Theory and Applications, Vol. 2, H. Brezis, M. G. Crandall, and F. Kappel, eds., Longman, New York, 1986, pp. 177–216.
- [32] D. V. WIDDER, *The Laplace Transform*, Princeton Univ. Press, Princeton, NJ, 1946.

THE STANDARD DECOMPOSED SYSTEM AND NONINTERACTING FEEDBACK CONTROL OF NONLINEAR SYSTEMS*

IN JOONG HA†

Abstract. Conditions for achieving noninteraction in nonlinear multivariable systems via the decomposition of state space are well established. The main contribution of this paper is to fully characterize the class of decomposing control laws. The characterization corresponds to a family of simple control laws which are applied to a standard decomposed system (SDS). The SDS is similar to the decomposed system of Isidori, Krener, Gori-Giorgi, and Monaco, but has a finer structure. The finer structure parallels the one used by Gilbert for linear systems. A weaker form of noninteraction, based on input-output behavior, is decoupling. Some connections between decomposition and decoupling are also established. An example illustrating the importance of the results is given.

Key words. nonlinear systems, standard decomposed system, noninteracting control, decomposition, decomposability, decomposing control law, decoupling, decoupling control law

AMS(MOS) subject classification. 93B10

1. Introduction. Consider a nonlinear system of the following form:

$$(1.1) \quad \dot{x}(t) = F(x, u) \triangleq X_0(x) + \sum_{i=1}^m X_i(x)u_i, \quad y(t) = H(x),$$

where X_i , $i = 1, \dots, m$ are vector fields on an n -dimensional manifold χ ; $H: \chi \rightarrow R^m$ is the output map of the system; and $u_i(t) \in R$, $y_i(t) \in R$ are the i th components of $u(t) \in R^m$, $y(t) \in R^m$, respectively. Decomposition (which is called noninteracting control in [8]) concerns the dynamic structure of systems in state space. Roughly speaking, the system (1.1) is decomposed if in an appropriate system of coordinates, it appears as a system having m independent subsystems such that for the i th subsystem, the input and output are u_i , y_i , respectively. Consider the following class of control laws:

$$(1.2) \quad u = K(x, \hat{u}) \triangleq \alpha(x) + \beta(x)\hat{u},$$

where $\alpha: \chi \rightarrow R^m$ and $\beta: \chi \rightarrow R^{m \times m}$. The system (1.1) is decomposable if there is a control law (1.2) such that the feedback system:

$$(1.3) \quad \dot{x} = F(x, \alpha(x) + \beta(x)\hat{u}), \quad y = H(x)$$

is decomposed. The corresponding control law is called a decomposing control law.

Conditions for (feedback) decomposition of nonlinear systems have been studied by many authors. See, e.g., [8]–[11], [13], [17]. The most common theoretical framework is some generalization of the geometric approach introduced by Wonham and Morse [12], [22] for linear systems. Here instead we emphasize decomposition structure and attack the previously neglected problem of fully characterizing the class of decomposing control laws. Characterizing the whole class of decomposing control laws has important practical implications because the use of subclasses may lead to undesirable closed-loop behavior such as instability. As will be seen, the class of decomposing control laws can be characterized by simple control laws applied to a standard decomposed system (SDS). The SDS is similar to the decomposed system of Isidori et al. [8] but has a finer structure. The structure parallels the one used by Gilbert [4] for linear systems.

* Received by the editors March 12, 1986; accepted for publication October 28, 1986. This research was supported by Air Force Office of Scientific Research grant F 49620-82-C-0089.

† Department of Control and Instrumentation Engineering, Seoul National University, Seoul, Korea.

A form of noninteracting weaker than decomposition is decoupling [2], [3], [5], [9], [14]–[16], [18]. Essentially, the system (1.1) is decoupled if for each $i = 1, \dots, m$, u_i affects only y_i . Note that decoupling concerns only the input–output map of systems, whereas decomposition concerns both the input–output map and the dynamic structure of systems in state space. Thus, decomposition may be more interesting from an engineering viewpoint. As part of our presentation we will establish some connections between the two concepts of noninteraction.

The paper is organized as follows. In this section, we introduce notation and some basic differential geometric tools used in later sections. Precise definitions of decomposition and decomposability are given. In § 2, necessary and sufficient conditions for a system to be decomposed and for a system to be decomposable are discussed. Section 3 contains the definition of the SDS and our main result: a characterization of the whole class of decomposing control laws. We also examine the relationship between the class of decomposing control laws and the class of the closed-loop decomposed systems. In § 4, an example is presented, which illustrates the application and significance of our results. Section 5 contains concluding remarks.

Let $M_{i,j}$ denote the set of integers $\{i, i+1, \dots, j\}$. We denote by $\{F, H, \chi\}$ the abstract system (1.1) defined on a manifold χ . At each $p \in \chi$, there exists a chart or coordinate neighborhood (U, ϕ) such that in the coordinates $z = \phi(x)$, the system (1.1) is described by

$$(1.4) \quad \dot{z}(t) = f(z(t)), \quad u(t) \triangleq f_0(z(t)) + \sum_{i=1}^m f_i(z(t))u_i(t), \quad y(t) = h(z(t)),$$

where the functions $f_i: \phi(U) \rightarrow R^n$, $i \in M_{0,m}$ are determined by, respectively, the local representation of the vector fields X_i in the chart. We denote this local representation by $\{f, h, U\}$. We denote by h_i, H_i the i th components of h, H , respectively. To simplify our definitions and proofs, all systems and control laws considered in this paper are assumed to be at least smooth (C^∞). (See [5] for the precise definitions of smoothness and real analyticity of systems and control laws.) All control laws considered in this paper are assumed to be nonsingular; $\beta(x)$ is nonsingular, $x \in \chi$.

Let T be a C^∞ -mapping from an n -dimensional smooth manifold χ into an m -dimensional smooth manifold $\hat{\chi}$. A smooth function $\hat{\phi}$ from $\hat{\chi}$ into R is T -related on χ to a smooth function ϕ from χ into R if $\phi(p) = \hat{\phi} \circ T(p)$, $p \in \chi$, where \circ denotes the function composition. For a smooth vector field Y on χ , Y_p denotes the tangent vector at $p \in \chi$ assigned by Y . A smooth vector field \hat{Y} on $\hat{\chi}$ is T -related on χ to a smooth vector field Y on χ if $\hat{Y}_{T(p)}\hat{\phi} = Y_p(\hat{\phi} \circ T)$, $p \in \chi$ for all C^∞ -functions $\hat{\phi}$ from $\hat{\chi}$ into R . The Lie bracket of two vector fields Y, Z on χ is denoted by $[Y, Z] \triangleq YZ - ZY$.

The following definition concerns state transformation between systems.

DEFINITION 1.1. Suppose for two systems $\{F, H, \chi\}$, $\{\hat{F}, \hat{H}, \hat{\chi}\}$, there exists a C^∞ -diffeomorphism $T: \chi \rightarrow \hat{\chi}$ such that (i) $\hat{\chi}_i$ is T -related on χ to X_i , $i \in M_{0,m}$, and (ii) \hat{H}_i is T -related on χ to H_i , $i \in M_{1,m}$. Then, $\{\hat{F}, \hat{H}, \hat{\chi}\}$ is T -related on χ to $\{F, H, \chi\}$.

Definitions similar to Definition 1.1 are found in the prior literature including [20].

Next, we introduce a general relation between systems, which takes into account both state and input-feedback transformations. Let T, α, β be mappings from χ into $\hat{\chi}$, R^m , and $R^{m \times m}$, respectively, such that $\beta(x)$ is nonsingular, $x \in \chi$. Define a mapping $J: \chi \times R^m \rightarrow \chi \times R^m$ by

$$(1.5) \quad J(x, u) \triangleq \begin{bmatrix} T(x) \\ -[\beta(x)]^{-1}\alpha(x) + [\beta(x)]^{-1}u \end{bmatrix}, \quad (x, u) \in \chi \times R^m.$$

We often write $J = \{T, \alpha, \beta\}$. Let $\{F, H, \chi\}^{a,\beta}$ denote the feedback system of $\{F, H, \chi\}$

corresponding to a control law $u = \alpha(x) + \beta(x)\hat{u}$. In other words, $\{F, H, \chi\}^{a,\beta}$ stands for the feedback system $\{\hat{F}, H, \chi\}$, where $\hat{F}(x, u) = \hat{F}(x, \alpha(x) + \beta(x)\hat{u})$.

DEFINITION 1.2. Suppose there exists a C^∞ -diffeomorphism $J: \chi \times R^m \rightarrow \hat{\chi} \times R^m$ defined by (1.5) such that $\{\hat{F}, \hat{H}, \hat{\chi}\}$ is T -related on χ to the system $\{F, H, \chi\}^{a,\beta}$. Then, $\{\hat{F}, \hat{H}, \hat{\chi}\}$ is J -feedback related on χ to $\{F, H, \chi\}$.

Similar definitions appear in the prior literature, including [7], [10].

Now, we define decomposition and decomposability. The definitions are similar to those in [8].

DEFINITION 1.3. $\{F, H, \chi\}$ is decomposed at $x_0 \in \chi$ if there exist: (a) an open neighborhood E of x_0 ; (b) an open subset $\bar{\chi}$ of R^n ; (c) a C^∞ -diffeomorphism $T: E \rightarrow \bar{\chi}$; (d) integers $\bar{s}_i \geq 1$, $i \in M_{1,m}$ and $\bar{s}_{m+1} \geq 0$ satisfying $n = \sum_{i=1}^{m+1} \bar{s}_i$; and (e) a system $\{\bar{F}, \bar{H}, \bar{\chi}\}$ which is T -related on E to $\{F, H, E\}$ such that its coordinate representation $\{\bar{f}, \bar{h}, \bar{\chi}\}$ has the form

$$(1.6) \quad \begin{aligned} \dot{\bar{x}}_i &= \bar{f}_i(\bar{x}_i) + \bar{g}_i(\bar{x}_i)\bar{u}_i, \quad \bar{y}_i = \bar{h}_i(\bar{x}_i), \quad i \in M_{1,m}, \\ \dot{\bar{x}}_{m+1} &= \bar{f}_{m+1}(\bar{x}) + \sum_{j=1}^m \bar{b}_j(\bar{x})\bar{u}_j, \end{aligned}$$

where $\bar{x}_i(t) \in R^{\bar{s}_i}$, $i \in M_{1,m+1}$, and $\bar{x} \triangleq (\bar{x}_1, \dots, \bar{x}_m, \bar{x}_{m+1})$. If $E = \chi$ in the above statement, $\{F, H, \chi\}$ is decomposed on χ .

DEFINITION 1.4. $\{F, H, \chi\}$ is decomposable at $x_0 \in \chi$ if there exists a control law $u = \alpha(x) + \beta(x)\hat{u}$ such that $\{F, H, \chi\}^{a,\beta}$ is decomposed at x_0 . $\{F, H, \chi\}$ is decomposable on χ if there exists a control law $u = \alpha(x) + \beta(x)\hat{u}$ such that $\{F, H, \chi\}^{a,\beta}$ is decomposed on χ .

Similar terminology arises in the precise definition of decoupling [5], [9]. It should be clear from definitions of decomposition and decoupling that a decomposed system is always decoupled. As will be discussed later in § 2, the converse statement is not necessarily true.

Let Z be a smooth vector field on χ . The codistribution Δ^\perp of a distribution Δ on χ is Z -invariant on χ if for any smooth function ϕ from χ into R , $d\phi \in \Delta^\perp$ on χ always implies $dZ\phi \in \Delta^\perp$ on χ . Let ψ_i , $i \in M_{1,k}$ be smooth functions from χ into R . The differentials $d\psi_i$, $i \in M_{1,k}$ are linearly independent on χ if $d\psi_i(p)$, $i \in M_{1,k}$ are linearly independent, $p \in \chi$. The first derivative of a smooth function $\psi: R^n \rightarrow R$ at x is denoted by $D\psi(x) \in R^{1 \times n}$. For each $i \in M_{1,m}$, M_i denotes the set $\{j: j \in M_{1,m}, j \neq i\}$. For each $i \in M_{1,m}$, let

$$(1.7) \quad \begin{aligned} \Delta_i^0(\{F, H, \chi\}) &\triangleq \{[X_{i_1}, [X_{i_2}, [\dots [X_{i_k}, X_j] \dots]]]: i \in \{0, i\}, r \in M_{1,k}, \\ &\quad k \in M_{0,\infty}, \text{ and } j \in M_i\}, \end{aligned}$$

where $[X_{i_1}, [X_{i_2}, [\dots [X_{i_k}, X_j] \dots]] \triangleq X_j$ if $k = 0$. Define $\Delta_i(\{F, H, \chi\})$ as the smallest subalgebra containing $\Delta_i^0(\{F, H, \chi\})$. Note that Δ_i^\perp is X_0 -invariant and X_i -invariant on χ . The identity map from χ onto χ is denoted by I .

2. Conditions for decomposition and decomposability. To state our results a variety of assumptions beyond smoothness and the nonsingularity of control laws are needed. To simplify the presentation we list them together here.

(A.1) The system $\{F, H, \chi\}$ satisfies the controllability rank condition on χ ([19]);

(A.2) $\Delta_i^\perp(\{F, H, \chi\})$ has a dimension $p_i \geq 1$ on χ , $i \in M_{1,m}$;

(A.3) There exist nonnegative integers d_i , $i \in M_{1,m}$ such that the following m -row

vector conditions are satisfied:

$$(2.1) \quad [X_1 X_0^k H_i(x) \cdots X_m X_0^k H_i(x)] = 0, \quad x \in \chi, \quad k \in M_{0, (d_i-1)} \text{ applies when } d_i > 0,$$

$$(2.2) \quad D_i^*(x) \triangleq [X_1 X_0^{d_i} H_i(x) \cdots X_m X_0^{d_i} H_i(x)] \neq 0, \quad x \in \chi.$$

The following theorem concerns necessary and sufficient conditions for local decomposition.

THEOREM 2.1. *The system $\{F, H, \chi\}$ is decomposed at $x_0 \in \chi$ if and only if there exist an open neighborhood E of x_0 and m involutive distributions Δ_i^* on E which has dimension $r_i < n$ such that on E ,*

- (i) $dH_i \in (\Delta_i^*)^\perp \subset \Delta_i^\perp, i \in M_{1,m};$
- (ii) $(\Delta_i^*)^\perp$ is X_0 -invariant and X_i -invariant, $i \in M_{1,m};$
- (iii) $(\Delta_i^*)^\perp, i \in M_{1,m}$ are mutually disjoint.

Theorem 2.1 is implied by Theorem 5.1 in [8]. Note that it is not easy to check for the existence of $\Delta_i^*, i \in M_{1,m}$ satisfying conditions specified in Theorem 2.1. This motivates the following result.

THEOREM 2.2. *Suppose that $\{F, H, \chi\}$ satisfies (A.1) and (A.2). Then $\{F, H, \chi\}$ is decomposed at each $x_0 \in \chi$ if and only if*

$$(2.3) \quad dH_i \in \Delta_i^\perp(\{F, H, \chi\}) \quad \text{on } \chi, \quad i \in M_{1,m}.$$

Apart from giving an easily verified condition for local decomposition, this result has other important implications. In [2], [5], [9], it was shown that (2.3) is a necessary and sufficient condition for decoupling of real analytic systems. From this and Theorem 2.1, we see that the conditions for decomposition are more complex than those for decoupling. Moreover, for real analytic systems satisfying the hypotheses of Theorem 2.2, the concepts of decomposition and decoupling are (at least locally) equivalent. It appears that this observation has not been made before. Proofs of Theorem 2.2 were obtained independently by Ha [6] and Nijmeijer [15].

We now turn to the question of when a system is decomposable by a control law. When (A.3) is satisfied, let $D^*(x)$ and $A^*(x)$ denote, respectively, the $(m \times m)$ and $(m \times 1)$ matrices of functions defined by

$$(2.4) \quad D^*(x) \triangleq \begin{bmatrix} D_1^*(x) \\ \vdots \\ D_m^*(x) \end{bmatrix}, \quad A^*(x) \triangleq \begin{bmatrix} X_0^{(d_1+1)} H_1(x) \\ \vdots \\ X_0^{(d_m+1)} H_m(x) \end{bmatrix}.$$

THEOREM 2.3. *Suppose $\{F, H, \chi\}$ satisfies (A.3). Then $\{F, H, \chi\}$ is decomposable at each $x_0 \in \chi$ if and only if*

$$(2.5) \quad D^*(x) \text{ is nonsingular at each } x \in \chi.$$

Furthermore, $u = [D^*(x)]^{-1}(\hat{u} - A^*(x))$ decomposes $\{F, H, \chi\}$ at each $x \in \chi$. That is, for $\alpha(x) \triangleq -[D^*(x)]^{-1}A^*(x)$ and $\beta(x) \triangleq [D^*(x)]^{-1}$, the system $\{F, H, \chi\}^{\alpha, \beta}$ is decomposed at each $x \in \chi$.

In [5], [6], [9], [18], it is shown that (2.5) is a necessary and sufficient condition for a smooth system to be decouplable on χ . Hence, Theorem 2.3 has the important implication that under assumption (A.3), decouplability and decomposability are locally equivalent.

The sufficiency of Theorem 2.3 was shown in [2], [6], [9]. We believe that the necessity of Theorem 2.3 is new. Since some details of the proofs are essential in the development of § 3, we give the brief proofs.

LEMMA 2.1. Suppose that $\{F, H, \chi\}$ satisfies (A.3). Let $\{\hat{F}, \hat{H}, \hat{\chi}\}$ be J -feedback related on χ to $\{F, H, \chi\}$ by $J = \{T, \alpha, \beta\}$. Let $\hat{X}_i, i \in M_{0,m}$ be vector fields corresponding to $\{\hat{F}, \hat{H}, \hat{\chi}\}$. Similarly, let $\hat{d}_i, \hat{D}^*, \hat{A}^*$ be determined by (2.1), (2.2), and (2.4) by replacing the X_i by the \hat{X}_i . Then:

- (i) $\{\hat{F}, \hat{H}, \hat{\chi}\}$ satisfies (A.3) with $\hat{d}_i = d_i, i \in M_{1,m}$;
- (ii) $\hat{D}^*(T(x)) = D^*(x)\beta(x), \hat{A}^*(T(x)) = A^*(x) + D^*(x)\alpha(x), x \in \chi$;
- (iii) $\hat{X}_0^k \hat{H}_i(T(x)) = X_0^k H_i(x), x \in \chi, k \in M_{0,d_i}$.

LEMMA 2.2. Suppose that a system $\{F, H, \chi\}$ satisfies (A.3) and (2.5). Then, $dX_0^k H_i, k \in M_{0,d_i}, i \in M_{1,m}$ are linearly independent on χ .

In [16], Lemma 2.1 is shown for the case when T is the identity map and $\chi = R^n$. Lemma 2.2 appears in [6] and [9].

PROOF OF THEOREM 2.3. Suppose there exists α, β such that $\{F, H, \chi\}^{\alpha, \beta}$ is decomposed at each $x_0 \in \chi$. Let $\{\hat{F}, \hat{H}, \hat{\chi}\} \triangleq \{F, H, \chi\}^{\alpha, \beta}$. Then, by Theorem 2.1,

$$(2.6) \quad dH_i \in \Delta_i^\perp(\{\hat{F}, H, \chi\}) \quad \text{on } \chi, \quad i \in M_{1,m}.$$

By the definition of the Lie bracket and Δ_i , this implies

$$(2.7) \quad \hat{X}_j \hat{X}_0^{d_i} H_i(x) = 0 \quad \text{on } \chi, \quad j \in M_i, \quad i \in M_{1,m}.$$

By Lemma 2.1(i) and the definition of $\{d_i, i \in M_{1,m}\}$, (2.7) implies

$$(2.8) \quad \lambda_i(x) \triangleq \hat{X}_i \hat{X}_0^{d_i} H_i(x) \neq 0, \quad x \in \chi, \quad i \in M_{1,m}.$$

On the other hand, by Lemma 2.1(ii), (2.7), and (2.8),

$$(2.9) \quad D^*(x)\beta(x) = \text{diag } \lambda_i(x), \quad x \in \chi.$$

Then, (2.5) is a direct consequence of (2.8), (2.9), and the nonsingularity assumption of control laws.

Next, assume (2.5). Let $\{\hat{F}, \hat{H}, \hat{\chi}\} \triangleq \{F, H, \chi\}^{\alpha, \beta}$ with $\alpha(x) \triangleq -[D^*(x)]^{-1}A^*(x)$ and $\beta(x) \triangleq [D^*(x)]^{-1}$. By Lemma 2.1, direct computation shows

$$(2.10) \quad \begin{aligned} \hat{X}_0^k \hat{H}_i(x) &= X_0^k H_i(x), \quad k \in M_{0,d_i}, \\ \hat{X}_0^{(d_i+1)} H_i(x) &= 0, \end{aligned}$$

$$(2.11) \quad \hat{X}_j \hat{X}_0^k H_i(x) = \begin{cases} 1 & \text{if } j = i \text{ and } k = d_i, \\ 0 & \text{otherwise.} \end{cases}$$

On the other hand, by Lemma 2.1(iii) and Lemma 2.2, $d\hat{X}_0^k H_i, k \in M_{0,d_i}, i \in M_{1,m}$ are linearly independent on χ . Let $T_{i,j} \triangleq \hat{X}_0^{(j-1)} H_i, i \in M_{1,(d_i+1)}, i \in M_{1,m}$. Let $T_i \triangleq (T_{i,1}, \dots, T_{i,(d_i+1)}), i \in M_{1,m}$. Let $p \triangleq \sum_{i=1}^m (d_i + 1)$ and $p_{m+1} \triangleq n - p$. Fix $x_0 \in \chi$. Because $d\hat{X}_0^k H_i, k \in M_{0,d_i}, i \in M_{1,m}$ are linearly independent on χ , it is possible to choose a C^∞ -mapping $T_{m+1}: \chi \rightarrow R^{p_{m+1}}$ such that $T = (T_1, \dots, T_m, T_{m+1})$ has rank n at x_0 . Then, it can be shown that there exist an open neighborhood E of x_0 and a system $\{\bar{F}, \bar{H}, \bar{\chi}\}$ with $\bar{\chi} = T(E), \bar{s}_i = d_i + 1, i \in M_{1,m}, \bar{s}_{m+1} = p_{m+1}$ meet the requirement of Definition 1.3. In particular, its coordinate representation $\{\bar{f}, \bar{h}, \bar{\chi}\}$ has the form (1.6) such that for each $i \in M_{1,m}$,

$$(2.12) \quad \bar{f}_i(\bar{x}_i) = \bar{A}_i \bar{x}_i, \quad \bar{g}_i(\bar{x}_i) = \bar{B}_i, \quad \bar{h}_i(\bar{x}_i) = \bar{C}_i \bar{x}_i,$$

where

$$(2.13) \quad \bar{A}_i \triangleq \begin{bmatrix} 0 \\ \vdots \\ I_{d_i} \\ 0 \quad \dots \quad 0 \end{bmatrix}, \quad \bar{B}_i \triangleq \begin{bmatrix} 0 \\ \vdots \\ 1 \end{bmatrix}, \quad \bar{C}_i \triangleq [1 \quad 0 \quad \dots \quad 0],$$

and I_{d_i} is the d_i by d_i identity matrix. Since $\{\hat{F}, H, \chi\}$ is decomposed at each $x_0 \in \chi$, $\{F, H, \chi\}$ is decomposable at each $x_0 \in \chi$ by the control law $u = [D^*(x)]^{-1}(\hat{u} - A^*(x))$.

Because of its importance in our subsequent developments, we henceforth reserve the notation $\{F^*, H, \chi\}$ for the system $\{F, H, \chi\}^{\alpha, \beta}$ with $\alpha(x) \triangleq -[D^*(x)]^{-1}A^*(x)$ and $\beta(x) \triangleq [D^*(x)]^{-1}$. It is well known [2], [3], [5], [9], [16], [18] that $\{F^*, H, \chi\}$ is decoupled on χ . In other words, the control law $u = [D^*(x)]^{-1}(\hat{u} - A^*(x))$ is a "global" decoupling control law. Note, however, that it is not necessarily a global decomposing control law.

3. Characterization of decomposing control laws. We begin this section by discussing an obvious class of decomposing control laws. Let $\{F, H, \chi\}$ be a system which satisfies (A.3) and (2.5). Let $\{\bar{F}, H, \chi\} \triangleq \{F^*, H, \chi\}$. In the proof of Theorem 2.3, we have shown that at each $x_0 \in \chi$, there exist: (a) an open neighborhood E of x_0 ; (b) a mapping T on E ; and (c) $\{\bar{F}, \bar{H}, T(E)\}$ which is T -related on E to $\{F, H, E\}$ such that the coordinated representation $\{\bar{f}, \bar{h}, T(E)\}$ has the form in (1.6) with the special structure (2.12). Now, suppose we choose the following control law for $\{\bar{F}, \bar{H}, T(E)\}$:

$$(3.1) \quad \bar{u}_i = \phi_i(\bar{y}_i^{(d)}, \dots, \bar{y}_i) + \psi_i(\bar{y}_i^{(d)}, \dots, \bar{y}_i)\bar{u}_i,$$

where ϕ_i, ψ_i are arbitrary C^∞ -functions of their arguments and $\bar{y}_i^{(d)}(t)$ is the d th derivative of $\bar{y}_i(t)$. Then the resulting closed-loop system retains the decomposed structure. Choosing the control law (3.1) for $\{\bar{F}, \bar{H}, T(E)\}$ corresponds to choosing for the original system $\{F, H, \chi\}$ a control law of the form $u = \alpha(x) + \beta(x)\bar{u}$, where

$$(3.2) \quad \alpha(x) = [D^*(x)]^{-1} \left\{ \begin{bmatrix} \eta_1(x) \\ \vdots \\ \eta_m(x) \end{bmatrix} - A^*(x) \right\},$$

$$(3.3) \quad \beta(x) = [D^*(x)]^{-1} \text{diag } \lambda_i(x),$$

where

$$(3.4) \quad \eta_i(x) \triangleq \phi_i(H_i(x), X_0 H_i(x), \dots, X_0^d H_i(x)),$$

$$(3.5) \quad \lambda_i(x) \triangleq \psi_i(H_i(x), X_0 H_i(x), \dots, X_0^d H_i(x)).$$

It is convenient to formalize the class (3.2)–(3.5).

DEFINITION 3.1. $S_0^\infty(\{F, H, \chi\})$ is the class of control laws $u = \alpha(x) + \beta(x)\bar{u}$ satisfying (3.2)–(3.5), where ϕ_i, ψ_i are arbitrary C^∞ -functions of their arguments.

From the above arguments, it is clear that a control law in $S_0^\infty(\{F, H, \chi\})$ decomposes $\{F, H, \chi\}$ at each $x_0 \in \chi$. However, we can find a more general class of decomposing control laws. It is necessary to introduce a more detailed structure for $\{\bar{f}, \bar{h}, \bar{\chi}\}$ than the one in (1.6) and (2.12).

DEFINITION 3.2. Let $\bar{\chi}$ be an open connected subset of R^n . A system $\{\bar{F}, \bar{H}, \bar{\chi}\}$ is a standard decomposed system (SDS) if its coordinate representation $\{\bar{f}, \bar{h}, \bar{\chi}\}$ has the following properties:

(1) There exist nonnegative integers $\bar{d}_i, i \in M_{1,m}$ and $\bar{p}_i, i \in M_{1,m+1}$, satisfying $n = \sum_{i=1}^{m+1} \bar{p}_i$, $\bar{p}_{m+1} \geq 0$, and $\bar{p}_i \geq \bar{d}_i + 1, i \in M_{1,m}$ so that $\{\bar{f}, \bar{h}, \bar{\chi}\}$ has the form

$$(3.6) \quad \dot{\bar{x}}_i = \bar{f}_i(\bar{x}_i, \bar{u}_i) \triangleq \begin{bmatrix} \bar{A}_i \bar{x}_i \\ \bar{\theta}_i(\bar{x}_i) \end{bmatrix} + \begin{bmatrix} \bar{B}_i \\ \bar{\gamma}_i(\bar{x}_i) \end{bmatrix} \bar{u}_i,$$

$$\bar{y}_i \triangleq \bar{h}_i(\bar{x}_i) = \bar{C}_i \bar{x}_i, \quad i \in M_{1,m},$$

$$(3.7) \quad \dot{\bar{x}}_{m+1} = \bar{f}_{m+1}(\bar{x}) + \sum_{i=1}^m \bar{b}_i(\bar{x}) \bar{u}_i,$$

where: $\bar{x}_i(t) \in R^{p_i}$, $i \in M_{1,m+1}$; $\bar{x} \triangleq (\bar{x}_1, \dots, \bar{x}_{m+1}) \in R^n$; \bar{A}_i , \bar{B}_i , \bar{C}_i are, respectively, $(\bar{d}_i + 1) \times \bar{p}_i$, $(\bar{d}_i + 1) \times 1$, $1 \times \bar{p}_i$ matrices such that

$$\bar{A}_i \triangleq \begin{bmatrix} 0 & & \vdots \\ \vdots & I_{d_i} & \vdots \\ 0 & \dots & 0 \end{bmatrix}, \quad \bar{B}_i \triangleq \begin{bmatrix} 0 \\ \vdots \\ 1 \end{bmatrix}, \quad \bar{C}_i \triangleq [1 \quad 0 \dots 0],$$

(2) Let $\bar{\chi}_i \triangleq [\bar{x}_i : \bar{x}] \triangleq (\bar{x}_i, \dots, \bar{x}_{m+1}) \in \bar{\chi}$. Each subsystem $\{\bar{f}_i, \bar{h}_i, \bar{\chi}_i\}$, $i \in M_{1,m}$, in (3.6) satisfies the controllability rank condition on $\bar{\chi}_i$.

(3) $\dim \Delta_i^\perp(\{\bar{F}, \bar{H}, \bar{\chi}\}) = \bar{p}_i$ on $\bar{\chi}$, $i \in M_{1,m}$.

Remark 3.1. The SDS in Definition 3.2 is a nonlinear version of the system introduced by Gilbert [4]. When $\{\bar{F}, \bar{H}, \bar{\chi}\}$ is a linear system, it can be shown that property (3) is equivalent to condition (iv) in Definition 6 of [4]. It is worth noting that properties (2) and (3) together imply that the SDS $\{\bar{F}, \bar{H}, \bar{\chi}\}$ satisfies (A.1).

Now, we are ready to state the following result.

THEOREM 3.1. Suppose that a system $\{F, H, \chi\}$ satisfies (A.1), (A.3), and (2.5). Further, assume that $\{F^*, H, \chi\}$ satisfies (A.2). Then, at each $x_0 \in \chi$, there exist: (a) an open neighborhood E of x_0 ; (b) an open connected subset $\bar{\chi}$ of R^n ; (c) a C^∞ -diffeomorphism $T: E \rightarrow \bar{\chi}$; and (d) a system $\{\bar{F}, \bar{H}, \bar{\chi}\}$, which is T -related on E to $\{F^*, H, E\}$ and is an SDS with $\bar{d}_i = d_i$, $\bar{p}_i = p_i$, $i \in M_{1,m}$, and $\bar{p}_{m+1} \triangleq p_{m+1} - \sum_{i=1}^m p_i$, where the p_i and d_i appear in (A.2) and (A.3).

For the proof of Theorem 3.1, we need the following lemmas.

LEMMA 3.1. Suppose that $\{\hat{F}, \hat{H}, \hat{\chi}\}$ is J -feedback related on χ to $\{F, H, \chi\}$ by $J = \{\alpha, \beta, T\}$. Then, if $\{F, H, \chi\}$ satisfies (A.1) on χ , $\{\hat{F}, \hat{H}, \hat{\chi}\}$ satisfies (A.1) on $T(\chi)$.

LEMMA 3.2. Suppose that $\{F, H, \chi\}$ satisfies (A.1) and (A.2). Then, at each point $x_0 \in \chi$, there exist $(\sum_{i=1}^m p_i)$ C^∞ -functions $\xi_{i,j}$, $j \in M_{1,p_i}$, $i \in M_{1,m}$, from an open neighborhood ν of x_0 into R such that

- (i) $d\xi_{i,j}$, $j \in M_{1,p_i}$, $i \in M_{1,m}$ are linearly independent on ν ;
- (ii) $d\xi_{i,j} \in \Delta_i^\perp(\{F, H, \chi\})$ on ν , $j \in M_{1,p_i}$, $i \in M_{1,m}$.

Lemma 3.1 seems to be well known. The proof of Lemma 3.2 is omitted because of limited space. It can be found in [6].

Proof of Theorem 3.1. Let $\{\hat{F}, H, \chi\} \triangleq \{F^*, H, \chi\}$. By given hypotheses and Lemma 3.1, $\{\hat{F}, H, \chi\}$ satisfies (A.1). Fix $x_0 \in \chi$. Then, by Lemma 3.2, there exist an open neighborhood ν of x_0 and $(\sum_{i=1}^m p_i)$ C^∞ -functions $\phi_{i,j}: \nu \rightarrow R$, $j \in M_{1,p_i}$, $i \in M_{1,m}$ such that on ν ,

$$(3.8) \quad d\phi_{i,j}, \quad j \in M_{1,p_i}, \quad i \in M_{1,m} \quad \text{are linearly independent,}$$

$$(3.9) \quad d\phi_{i,j} \in \Delta_i^\perp(\{\hat{F}, H, \chi\}), \quad j \in M_{1,p_i}, \quad i \in M_{1,m}.$$

As was shown in the proof of Theorem 2.3, $\{\hat{F}, H, \chi\}$ is decomposed at x_0 . Therefore, by (i) of Theorem 2.1, there exists an open neighborhood $\hat{\nu} \subset \nu$ of x_0 such that

$$(3.10) \quad dH_i \in \Delta_i^\perp(\{\hat{F}, H, \chi\}) \quad \text{on } \hat{\nu}, \quad i \in M_{1,m}.$$

Since Δ_i^\perp is \hat{X}_0 -invariant on χ , this implies

$$(3.11) \quad d\hat{X}_0^k H_i \in \Delta_i^\perp(\{\hat{F}, H, \chi\}) \quad \text{on } \hat{\nu}, \quad k \in M_{0,d_i}, \quad i \in M_{1,m}.$$

This, (3.11), and Lemma 2.2 show

$$(3.12) \quad p_i \geq d_i + 1, \quad i \in M_{1,m}.$$

Next, we show that there exist an open neighborhood $W \subset \hat{v}$ of x_0 and a basis of $\Delta_i^\perp(\{\hat{F}, H, \chi\})$ on W which contains $d\hat{X}_0^k H_i$, $k \in M_{0,d_i}$. By (3.8), (3.9), and (3.11), for each $i \in M_{1,m}$, there exist an open neighborhood $V_i \subset \hat{v}$ of x_0 and a C^∞ -function $\psi_{i,j}$ from an appropriate subset of R^{p_i} into R , $j \in M_{1,(d_i+1)}$ such that

$$(3.13) \quad T_{i,j}(x) \triangleq \hat{X}_0^{(j-1)} H_i(x) = \psi_{i,j}(\phi_{i,1}(x), \dots, \phi_{i,p_i}(x)), \quad x \in V_i, \quad j \in M_{1,(d_i+1)}.$$

By Lemma 2.1(iii), Lemma 2.2, and (3.8), $D\psi_{i,j}(\phi_{i,1}(x_0), \dots, \phi_{i,p_i}(x_0))$, $j \in M_{1,(d_i+1)}$, are linearly independent $(1 \times p_i)$ row vectors. Now, for each $i \in M_{1,m}$, let $r_i \triangleq p_i - d_i - 1$ and choose $r_i(1 \times p_i)$ row vectors $\eta_{i,j}$ such that $D\psi_{i,j}(\phi_{i,1}(x_0), \dots, \phi_{i,p_i}(x_0))$, $j \in M_{1,(d_i+1)}$ and $\eta_{i,j}$, $j \in M_{1,r_i}$ are linearly independent. Let

$$(3.14) \quad E_i \triangleq (\phi_{i,1}, \dots, \phi_{i,p_i}), \quad T_{i,(d_i+1+j)} \triangleq \eta_{i,j} E_i, \quad j \in M_{1,r_i}, \quad i \in M_{1,m}.$$

Then, by the construction of $T_{i,j}$, $j \in M_{1,p_i}$, $i \in M_{1,m}$,

$$(3.15) \quad dT_{i,j}(x_0), \quad j \in M_{1,p_i}, \quad i \in M_{1,m} \quad \text{are linearly independent,}$$

$$(3.16) \quad dT_{i,j} \in \Delta_i^\perp(\{\hat{F}, H, \chi\}) \quad \text{on } V_i, \quad j \in M_{1,p_i}, \quad i \in M_{1,m}.$$

Let $V \triangleq V_1 \cap \dots \cap V_m$ and $p_{m+1} \triangleq n - \sum_{i=1}^m p_i$. If $p_{m+1} \geq 1$, choose a C^∞ -mapping T_{m+1} from V into $R^{p_{m+1}}$ such that T has rank n at x_0 , where

$$(3.17) \quad T \triangleq (T_1, \dots, T_m, T_{m+1}), \quad T_i \triangleq (T_{i,1}, \dots, T_{i,p_i}), \quad i \in M_{1,m}.$$

Then, by the Local Inverse Function Theorem [1], there exists an open neighborhood $W \subset V$ of x_0 such that

$$(3.18) \quad T \text{ is a } C^\infty\text{-diffeomorphism on } W,$$

$$(3.19) \quad \{dT_{i,j}(p), j \in M_{1,p_i}\} \text{ is a basis of } (\Delta_i^\perp(\{\hat{F}, H, \chi\}))_p, \quad p \in W.$$

Now, using (3.18) and (3.19), we show property (1) of Definition 3.2. Since Δ_i^\perp is \hat{X}_0 -invariant and \hat{X}_i -invariant on χ , (3.19) implies that

$$(3.20) \quad d\hat{X}_0 T_{i,j}, d\hat{X}_i T_{i,j} \in \Delta_i^\perp(\{\hat{F}, H, \chi\}) \quad \text{on } W, \quad j \in M_{1,p_i}, \quad i \in M_{1,m}.$$

Then, (3.19) and (3.20) with the Constant Mapping Theorem [21, p. 18] imply that there exist an open connected neighborhood $E \subset W$ of x_0 and C^∞ -functions $\bar{\theta}_{i,j}$, $\bar{\gamma}_{i,j}$ from appropriate subsets of R^{p_i} into R , $j \in M_{1,r_i}$, $i \in M_{1,m}$ such that

$$(3.21) \quad \begin{aligned} \hat{X}_0 T_{i,(d_i+1+j)}(x) &= \bar{\theta}_{i,j}(T_i(x)), \\ \hat{X}_i T_{i,(d_i+1+j)}(x) &= \bar{\gamma}_{i,j}(T_i(x)), \quad x \in E. \end{aligned}$$

On the other hand, by (3.18), there exist C^∞ -functions $\bar{f}_{m+1,j}$, $\bar{b}_{i,j}$, $i \in M_{1,m}$, $j \in M_{1,p_{m+1}}$ defined on appropriate subsets of R^n such that

$$(3.22) \quad \hat{X}_0 T_{m+1,j}(x) = \bar{f}_{m+1,j}(T(x)), \quad \hat{X}_i T_{m+1,j}(x) = \bar{b}_{i,j}(T(x)), \quad x \in E.$$

Let $\bar{\chi} \triangleq T(E)$. Let $\bar{x} \triangleq (\bar{x}_1, \dots, \bar{x}_m, \bar{x}_{m+1}) \triangleq (T_1(x), \dots, T_{m+1}(x))$. Let $\bar{\theta}_i = (\bar{\theta}_{i,1}, \dots, \bar{\theta}_{i,r_i})$, $\bar{\gamma}_i = (\bar{\gamma}_{i,1}, \dots, \bar{\gamma}_{i,r_i})$, $i \in M_{1,m}$. Let $\bar{f}_{m+1} \triangleq (\bar{f}_{m+1,1}, \dots, \bar{f}_{m+1,p_{m+1}})$, $\bar{b}_i = (\bar{b}_{i,1}, \dots, \bar{b}_{i,p_{m+1}})$, $i \in M_{1,m}$. Define vector fields \bar{X}_i , $i \in M_{0,m}$ by

$$(3.23) \quad \bar{X}_0(\bar{x}) \triangleq \sum_{i=1}^m \left(\sum_{j=1}^{d_i} \bar{x}_{i,(j+1)} \frac{\partial}{\partial \bar{x}_{i,j}} + \sum_{j=d_i+2}^{p_i} \bar{\theta}_{i,(j-d_i-1)}(\bar{x}_i) \frac{\partial}{\partial \bar{x}_{i,j}} \right) + \sum_{j=1}^{p_{m+1}} \bar{f}_{m+1,j}(\bar{x}) \frac{\partial}{\partial \bar{x}_{m+1,j}},$$

$$(3.24) \quad \bar{X}_i(\bar{x}) \triangleq \frac{\partial}{\partial \bar{x}_{i,(d_i+1)}} + \sum_{j=d_i+2}^{p_i} \bar{\gamma}_{i,(j-d_i-1)}(\bar{x}_i) \frac{\partial}{\partial \bar{x}_{i,j}} + \sum_{j=1}^{p_{m+1}} \bar{b}_{i,j}(\bar{x}) \frac{\partial}{\partial \bar{x}_{m+1,j}}, \quad i \in M_{1,m},$$

$$(3.25) \quad \bar{H}_i(\bar{x}) \triangleq \bar{x}_{i,1}, \quad i \in M_{1,m}$$

where $\bar{x}_{i,j}$ is the j th component of \bar{x}_i . Let $\{\bar{F}, \bar{H}, \bar{\chi}\}$ be the system constructed as above. By (2.10), (2.11), and (3.21)–(3.25), the coordinate representation $\{\bar{f}, \bar{h}, \bar{\chi}\}$ of $\{\bar{F}, \bar{H}, \bar{\chi}\}$ has the form indicated in (1) of Definition 3.2, where $\bar{d}_i = d_i$, $i \in M_{1,m+1}$.

Let \bar{Y}_i be a C^∞ -vector field in $\Delta_i(\{\bar{F}, \bar{H}, \bar{\chi}\})$. Then, using (3.23) and (3.24), we can show that if $\bar{Y}_i = \sum_{j=1}^{m+1} \sum_{k=1}^{p_i} \gamma_{j,k}(\cdot) \partial / \partial \bar{x}_{j,k}$ is a local representation of \bar{Y}_i on $\bar{\chi}$,

$$(3.26) \quad \gamma_{i,k}(\bar{x}) = 0, \quad \bar{x} \in \bar{\chi}, \quad k \in M_{1,p_i}.$$

By Lemma 3.1, $\{\bar{F}, \bar{H}, \bar{\chi}\}$ must satisfy (A.1). Thus, (3.26) implies property (2) of Definition 3.2. Property (3) follows from the fact that by (3.18), $(\Delta_i)_p(\{\bar{F}, \bar{H}, \bar{\chi}\})$ and $(\Delta_i)_{T(p)}(\{\bar{F}, \bar{H}, \bar{\chi}\})$ are isomorphic at each $p \in E$.

Next, we state a converse result.

THEOREM 3.2. *Suppose that at each $x_0 \in \chi$, there exist: (a) an open neighborhood E of x_0 ; (b) an open connected subset $\bar{\chi}$ of R^n ; and (c) an SDS $\{\bar{F}, \bar{H}, \bar{\chi}\}$ which is J -feedback related on E to $\{F, H, E\}$ by $J = \{T, \alpha, \beta\}$. Then the following properties hold:*

- (i) $\{F, H, \chi\}$ satisfies (2.5), (A.1), and (A.3) with $d_i = \bar{d}_i$, $i \in M_{1,m}$;
- (ii) $\{F^*, H, \chi\}$ satisfies (A.2) with $p_i = \bar{p}_i$, $i \in M_{1,m}$;
- (iii) $\alpha(x) = -[D^*(x)]^{-1}A^*(x)$ and $\beta(x) = [D^*(x)]^{-1}$.

Proof. By Remark 3.1, $\{\bar{F}, \bar{H}, \bar{\chi}\}$ satisfies (A.1). By Lemma 3.1, this implies that $\{F, H, \chi\}$ satisfies (A.1). Direct computation shows that $\{\bar{F}, \bar{H}, \bar{\chi}\}$ satisfies (A.3) with $d_i = \bar{d}_i$, $i \in M_{1,m}$, $\bar{D}^*(x) = I_m$, and $\bar{A}^*(x) = 0$. By this, Lemma 2.1, and the nonsingularity assumption of control laws, we see that $\{F, H, \chi\}$ satisfies (2.5), (A.3) with $d_i = \bar{d}_i$, $i \in M_{1,m}$, and furthermore (iii). Since $\{\bar{F}, \bar{H}, \bar{\chi}\}$ is J -feedback related on E to $\{F, H, E\}$ by $J = \{T, \alpha, \beta\}$, (iii) implies that $\{\bar{F}, \bar{H}, \bar{\chi}\}$ is T -related on E to $\{F^*, H, \chi\}$. Consequently, $(\Delta_i)_q(\{F^*, H, \chi\})$ and $(\Delta_i)_{T(p)}(\{\bar{F}, \bar{H}, \bar{\chi}\})$ are isomorphic at each $q \in E$. This implies (ii).

Now, we introduce another class of control laws.

DEFINITION 3.3. $S^\infty(\{F, H, \chi\})$ is the class of control laws $u = \alpha(x) + \beta(x)\hat{u}$ satisfying (3.2), (3.3), and

$$(3.27) \quad d\eta_i, d\lambda_i \in \Delta_i^\perp(\{F^*, H, \chi\}) \quad \text{on } \chi.$$

By Lemma 2.1(iii) and (3.11), the smooth functions η_i, λ_i in (3.4) and (3.5) satisfy (3.27). Thus,

$$(3.28) \quad S_0^\infty \subset S^\infty.$$

In general, S_0^∞ is a proper subset of S^∞ . See the discussion which appears later in Remark 3.2. Knowledge of a more general class of decomposing control laws allows more flexibility in choosing a decomposing control law. This will be illustrated in § 4.

The following result shows that S^∞ is the whole class of local decomposing control laws for a special class of smooth systems.

THEOREM 3.3. *Suppose that $\{F, H, \chi\}$ satisfies the hypotheses for Theorem 3.1. Then a control law $u = \alpha(x) + \beta(x)\hat{u}$ decomposes $\{F, H, \chi\}$ at each $x_0 \in \chi$ if and only if it belongs to $S^\infty(\{F, H, \chi\})$.*

Proof. Suppose a control law $u = \alpha(x) + \beta(x)\hat{u}$ decomposes $\{F, H, \chi\}$ at each $x_0 \in \chi$. Let $\{\bar{F}, H, \chi\} \triangleq \{F, H, \chi\}^{\alpha, \beta}$. Let $\bar{X}_i, i \in M_{0,m}$ be vector fields corresponding to $\{\bar{F}, H, \chi\}$. Then, by (i) of Theorem 2.1 and the definition of the Lie bracket,

$$(3.29) \quad \bar{X}_j \bar{X}_{i_1} \bar{X}_{i_2} \cdots \bar{X}_{i_k} H_i = 0 \quad \text{on } \chi, \\ i_q \in \{0, i\}, \quad q \in M_{1,k}, \quad k \in M_{0,\infty}, \quad j \in M_i, \quad i \in M_{1,m}.$$

Let $\hat{X}_i, i \in M_{0,m}$ be vector fields corresponding to $\{F^*, H, \chi\}$. Define C^∞ -mappings $\eta: \chi \rightarrow R^m, \tau: \chi \rightarrow R^{m \times m}$ by

$$(3.30) \quad \eta(x) \triangleq D^*(x)\alpha(x) + A^*(x), \quad \tau(x) \triangleq D^*(x)\beta(x).$$

Then, through the same arguments used for the proof of the necessity of Theorem 3.1 in [5], it can be shown that (3.29) implies

$$(3.31) \quad \tau_{i,j} = 0 \quad \text{on } \chi, \quad i \neq j,$$

$$(3.32) \quad \hat{X}_j \hat{X}_{i_1} \hat{X}_{i_2} \cdots \hat{X}_{i_k} \eta_i = \hat{X}_j \hat{X}_{i_1} \hat{X}_{i_2} \cdots \hat{X}_{i_k} \tau_{i,i} = 0 \quad \text{on } \chi, \\ i_q \in \{0, i\}, \quad q \in M_{1,k}, \quad k \in M_{0,\infty}, \quad j \in M_i, \quad i \in M_{1,m},$$

where $\tau_{i,j}$ is the (i, j) th component of τ . By the definitions of Δ_i and the Lie bracket, this implies that the control law $u = \alpha(x) + \beta(x)\hat{u}$ must belong to $S^\infty(\{F, H, \chi\})$.

Next, suppose that a control law $u = \alpha(x) + \beta(x)\hat{u}$ belongs to $S^\infty(\{F, H, \chi\})$. Fix $x_0 \in \chi$. Then, by Theorem 3.1, there exist an open neighborhood E of x_0 and a mapping $T: E \rightarrow R^n$ such that $\{\bar{F}, \bar{H}, \bar{\chi}\}$, which is J -feedback related on E to $\{F, H, E\}$ by $J = \{T, -(D^*)^{-1}A^*, (D^*)^{-1}\}$, is a standard decomposed system with $\bar{\chi} = T(E), \bar{d}_i = d_i, i \in M_{1,m}$, and $\bar{p}_i = p_i, i \in M_{1,m}$. The mapping T constructed by (3.13), (3.14), and (3.17) satisfies (3.18) and (3.19) on E . This with (3.27) implies that there exist an open neighborhood $U \subset E$ of x_0 and C^∞ -functions $\bar{\eta}_i, \bar{\lambda}_i$, defined on appropriate subsets of $R^{p_i}, i \in M_{1,m}$ such that

$$(3.33) \quad \eta_i(x) = \bar{\eta}_i(T_i(x)), \quad \lambda_i(x) = \bar{\lambda}_i(T_i(x)), \quad x \in U.$$

Let $\bar{\eta} \triangleq (\bar{\eta}_1, \dots, \bar{\eta}_m)$ and $\bar{\tau} \triangleq \text{diag } \bar{\lambda}_i$. Let $J_1 \triangleq \{T, 0, I_m\}$ and $J_2 \triangleq \{I, \eta, \text{diag } \lambda_i\}$. Then, $\{F, H, U\}^{\alpha, \beta}$ is J_3 -feedback related on $T(U)$ to $\{\bar{F}, \bar{H}, T(U)\}$ by $J_3 \triangleq J_2 \circ J_1^{-1}$. Direct computation shows that $J_2 \circ J_1^{-1} = \{T^{-1}, \bar{\eta}, \bar{\tau}\}$. The form of the standard decomposed system, the form of $\bar{\eta}, \bar{\tau}$, and Definition 1.3 imply $\{F, H, \chi\}^{\alpha, \beta}$ is decomposed at x_0 .

If $\{F, H, \chi\}$ is an SDS, we might expect intuitively from its special structure that its decomposing control laws are of the form $\bar{u}_i = \bar{\eta}_i(\bar{x}_i) + \bar{\lambda}_i(\bar{x}_i)\bar{u}_i, i \in M_{1,m}$. The next theorem states that this is really the case.

DEFINITION 3.4. Let $\{\bar{F}, \bar{H}, \bar{\chi}\}$ be an SDS. $\bar{S}^\infty(\{\bar{F}, \bar{H}, \bar{\chi}\})$ is the class of control laws $\bar{u} = \bar{\alpha}(\bar{x}) + \bar{\beta}(\bar{x})\bar{u}$ satisfying

$$(3.34) \quad \bar{\alpha}(\bar{x}) = \begin{bmatrix} \bar{\eta}_1(\bar{x}_1) \\ \vdots \\ \bar{\eta}_m(\bar{x}_m) \end{bmatrix}, \quad \bar{\beta}(\bar{x}) = \text{diag } \bar{\lambda}_i(\bar{x}_i),$$

where $\bar{\eta}_i, \bar{\lambda}_i$ are C^∞ -functions from $\bar{\chi}_i$ into $R, i \in M_{1,m}$.

THEOREM 3.4. The control law $\bar{u} = \bar{\alpha}(\bar{x}) + \bar{\beta}(\bar{x})\bar{u}$ decomposes an SDS $\{\bar{F}, \bar{H}, \bar{\chi}\}$ on $\bar{\chi}$ if and only if it belongs to $\bar{S}^\infty(\{\bar{F}, \bar{H}, \bar{\chi}\})$.

The sufficiency of Theorem 3.4 is obvious. The proof for the necessity of Theorem 3.4 is omitted because of limited space. It can be found in [6]. Property (3) of the SDS is essential in obtaining Theorem 3.4. As might be expected, there is a one-to-one correspondence between control laws in $S^\infty(F, H, \chi)$ and control laws in $\bar{S}^\infty(\bar{F}, \bar{H}, \bar{\chi})$.

THEOREM 3.5. Suppose that $\{F, H, \chi\}$ satisfies the hypotheses of Theorem 3.1. Let $x_0 \in \chi$. Let $E, T, \{\bar{F}, \bar{H}, \bar{\chi}\}$ be the open neighborhood of x_0 , the mapping, and the SDS given by Theorem 3.1. Then, there exists an open neighborhood $U \subset E$ of x_0 such that:

(i) For every $u = \alpha(x) + \beta(x)\hat{u}$ in $S^\infty(\{F, H, U\})$, there exists a unique control law $\bar{u} = \bar{\alpha}(\bar{x}) + \bar{\beta}(\bar{x})\bar{\hat{u}}$ in $\bar{S}^\infty(\{\bar{F}, \bar{H}, T(U)\})$ such that $\{\bar{F}, \bar{H}, T(U)\}^{\bar{\alpha}, \bar{\beta}}$ is T -related on U to $\{F, H, U\}^{\alpha, \beta}$. Conversely, for every $\bar{u} = \bar{\alpha}(\bar{x}) + \bar{\beta}(\bar{x})\bar{\hat{u}}$ in $\bar{S}^\infty(\{\bar{F}, \bar{H}, T(U)\})$, there exists a unique control law $u = \alpha(x) + \beta(x)\hat{u}$ in $S^\infty(\{F, H, U\})$ such that $\{F, H, U\}^{\alpha, \beta}$ is T^{-1} -related on $T(U)$ to $\{\bar{F}, \bar{H}, T(U)\}^{\bar{\alpha}, \bar{\beta}}$.

(ii) Let $u = \alpha(x) + \beta(x)\hat{u}$, $\bar{u} = \bar{\alpha}(\bar{x}) + \bar{\beta}(\bar{x})\bar{\hat{u}}$ be control laws in $S^\infty(\{F, H, U\})$, $\bar{S}^\infty(\{\bar{F}, \bar{H}, T(U)\})$, respectively. Suppose they are in the one-to-one correspondence described in (i). Then,

$$(3.35) \quad \alpha(x) = [D^*(x)]^{-1}\{\bar{\alpha}(T(x)) - A^*(x)\}, \quad \beta(x) = [D^*(x)]^{-1}\bar{\beta}(T(x)).$$

(iii) In particular, when T is a C^∞ -diffeomorphism on χ and χ is connected, (i) and (ii) hold with $U = \chi$.

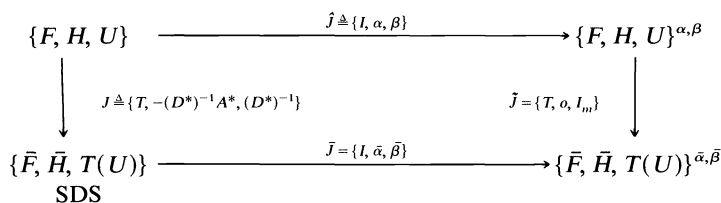


FIG. 3.1. A schematic description of Theorem 3.5, where $u = \alpha(x) + \beta(x)\hat{u}$, $\bar{u} = \bar{\alpha}(\bar{x}) + \bar{\beta}(\bar{x})\bar{\hat{u}}$ are control laws in $S^\infty(\{F, H, U\})$, $\bar{S}^\infty(\{\bar{F}, \bar{H}, T(U)\})$, respectively.

Proof. First consider (i). Suppose $u = \alpha(x) + \beta(x)\hat{u}$ belongs to $S^\infty(\{F, H, U\})$. Then, following the second part of the proof for Theorem 3.3 leads to the fact that there exist an open neighborhood $U \subset E$ of x_0 and C^∞ -functions $\bar{\eta}_i, \bar{\lambda}_i$, defined on appropriate subsets of R^{p_i} , $i \in M_{1,m}$ such that (3.33) holds. Note that given T and U , the $\bar{\eta}_i$ and $\bar{\lambda}_i$ are unique. Define $\bar{\alpha} \triangleq (\bar{\eta}_1, \dots, \bar{\eta}_m)$ and $\bar{\beta} \triangleq \text{diag } \bar{\lambda}_i$. Then $\bar{u} = \bar{\alpha}(\bar{x}) + \bar{\beta}(\bar{x})\bar{\hat{u}}$ belongs to $\bar{S}^\infty(\{\bar{F}, \bar{H}, T(U)\})$. Furthermore, $\{\bar{F}, \bar{H}, T(U)\}^{\bar{\alpha}, \bar{\beta}}$ is T -related on U to $\{F, H, U\}^{\alpha, \beta}$. Next, consider the converse statement. Suppose that $\bar{u} = \bar{\alpha}(\bar{x}) + \bar{\beta}(\bar{x})\bar{\hat{u}}$ belongs to $\bar{S}^\infty(\{\bar{F}, \bar{H}, T(U)\})$. Define α, β by (3.35). Then, by (3.19), it follows that $u = \alpha(x) + \beta(x)\hat{u}$ belongs to $S^\infty(\{F, H, U\})$. Clearly, $\{F, H, U\}^{\alpha, \beta}$ is T^{-1} -related on U to $\{\bar{F}, \bar{H}, T(u)\}^{\bar{\alpha}, \bar{\beta}}$.

Part (ii) has been shown implicitly above. Part (iii) follows, since the given hypotheses imply that (3.18) and (3.19) hold on χ and $T(\chi)$ is connected.

Remark 3.2. Suppose that (A.3), (2.5), and $n = \sum_{i=1}^m (d_i + 1)$ are satisfied. Then, it can be shown that the hypotheses of Theorem 3.5 are satisfied with $p_i = d_i + 1$, $i \in M_{1,m}$ and $T = G$, where $G \triangleq (G_1, \dots, G_m)$, $G_1 \triangleq (G_{1,1}, \dots, G_{1,d_1+1})$, and $G_{i,j} \triangleq X_0^{(j-1)} H_i$. It follows from part (ii) of Theorem 3.5 that, at least locally, $S^\infty(\{F, H, \chi\}) = S_0^\infty(\{F, H, \chi\})$. When G is a C^∞ -diffeomorphism on χ and χ is connected, (iii) confirms that $S^\infty(\{F, H, \chi\}) = S_0^\infty(\{F, H, \chi\})$. Note that for this case, we do not need to solve the partial differential equations given by (3.27) to characterize $S^\infty(\{F, H, \chi\})$. If G is not a C^∞ -diffeomorphism, $S^\infty(\{F, H, \chi\}) = S_0^\infty(\{F, H, \chi\})$ is not necessarily true. This is shown through Example 4.2 in [5].

Remark 3.3. If, in Definition 3.3, C^∞ is replaced by C^ω (real analytic), the definition of S^ω is obtained. In [5], we showed that S^ω is the whole class of real analytic control laws which decouple a real analytic system $\{F, H, \chi\}$ on χ . Therefore,

under real analyticity and the hypotheses of Theorem 3.1, the whole class of decomposing control laws and the whole class of decoupling control laws are locally identical.

4. An example. In this section, we present an example which illustrates the significance of the results developed in the previous sections. The example is also used in [5] for the discussion of decoupling. For this example, $S_c^\infty(\{F, H, \chi\})$ is a proper subset of $S^\infty(\{F, H, \chi\})$. While there is no control law in $S_0^\infty(\{F, H, \chi\})$ which decomposes $\{F, H, \chi\}$ on χ with Bounded Input-Bounded State (BIBS) stability, there are many control laws in $S^\infty(\{F, H, \chi\})$ which decomposes $\{F, H, \chi\}$ on χ with BIBS stability.

Let us consider a smooth system $\{F, H, R^3\}$ with $m = 2$ and

$$(4.1) \quad X_0(x) \triangleq (x_2 + x_1 x_3) \frac{\partial}{\partial x_2},$$

$$(4.2) \quad X_1(x) \triangleq \frac{\partial}{\partial x_1} + (1 + x_1 - x_3) \frac{\partial}{\partial x_2} - \frac{\partial}{\partial x_3},$$

$$(4.3) \quad X_2(x) \triangleq \frac{\partial}{\partial x_1} + (1 - x_3) \frac{\partial}{\partial x_2},$$

$$(4.4) \quad H_1(x) \triangleq x_1, \quad H_2(x) \triangleq x_1 + x_3.$$

Direct computation shows that (A.3) is satisfied with

$$(4.5) \quad d_1 = d_2 = 0, \quad D^*(x) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad A^*(x) = 0.$$

Thus, by Theorem 2.3, $\{F, H, R^3\}$ is decomposable at each $x \in R^3$. Actually, $\{F, H, R^3\}$ is decomposable on R^3 since the open set E defined in the proof for the sufficiency of Theorem 2.3 can be chosen to be R^3 .

Let \hat{X}_i , $i \in M_{0,2}$ be the vector fields corresponding to the decomposed system $\{F^*, H, R^3\}$. Then, by (4.5), we have

$$(4.6) \quad \hat{X}_0(x) = (x_2 + x_1 x_3) \frac{\partial}{\partial x_2},$$

$$(4.7) \quad \hat{X}_1(x) = \frac{\partial}{\partial x_1} + (1 + x_1 - x_3) \frac{\partial}{\partial x_2} - \frac{\partial}{\partial x_3},$$

$$(4.8) \quad \hat{X}_2(x) = -x_1 \frac{\partial}{\partial x_2} + \frac{\partial}{\partial x_3}.$$

From (4.6)–(4.8), it can be easily shown that

$$(4.9) \quad \Delta_1(\{F^*, H, R^3\}) = \text{span} \{\hat{X}_2\},$$

$$(4.10) \quad \Delta_2(\{F^*, H, R^3\}) = \text{span} \{\hat{X}_1, [\hat{X}_0, \hat{X}_1]\}$$

and that (A.2) is satisfied with $p_1 = 2$, $p_2 = 1$. Consequently, all hypotheses of Theorem 3.5 are satisfied.

Define C^∞ -functions $T_{i,j}$, $j \in M_{1,p_i}$, $i \in M_{1,2}$ by

$$(4.11) \quad T_{1,1}(x) \triangleq H_1(x), \quad T_{1,2}(x) \triangleq x_2 + x_1 x_3, \quad T_{2,1}(x) \triangleq H_2(x).$$

Let $T \triangleq (T_{1,1}, T_{1,2}, T_{2,1})$. Then, we can easily show that T is a C^∞ -diffeomorphism from R^3 onto R^3 and $\{dT_{i,j}(q), j \in M_{1,p_i}\}$ is a basis of $(\Delta_i)_q(\{F^*, H, R^3\})$, $q \in R$, $i \in M_{1,2}$. Let $\{\bar{F}, \bar{H}, R^3\}$ be an SDS whose coordinate representation is

$$(4.12) \quad \begin{bmatrix} \dot{\bar{x}}_{1,1} \\ \dot{\bar{x}}_{1,2} \end{bmatrix} = \begin{bmatrix} 0 \\ \bar{x}_{1,2} \end{bmatrix} + \bar{u}_i \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \bar{y}_i = \bar{x}_{1,1},$$

$$\dot{\bar{x}}_2 = \bar{u}_2, \quad \bar{y}_2 = \bar{x}_2.$$

Then we can check that the above T and $\{\bar{F}, \bar{H}, R^3\}$ with $E = R^3$ are those described in Theorems 3.1 and 3.5.

By (3.35) and (4.5), $S^\infty(\{\bar{F}, \bar{H}, R^3\})$ is given by

$$(4.13) \quad \alpha(x) = \begin{bmatrix} \phi_1(x_1, x_2 + x_1 x_3) - \phi_2(x_1 + x_3) \\ \phi_2(x_1 + x_3) \end{bmatrix},$$

$$(4.14) \quad \beta(x) = \begin{bmatrix} \psi_1(x_1, x_2 + x_1 x_3) & \vdots & -\psi_2(x_1 + x_3) \\ 0 & \vdots & \psi_2(x_1 + x_3) \end{bmatrix},$$

where ϕ_i, ψ_i , $i \in M_{1,2}$ are arbitrary C^∞ -functions of their arguments such that $\Psi_1(z_1, z_2) \neq 0$, $(z_1, z_2) \in R^2$ and $\psi_2(z_3) \neq 0$, $z_3 \in R$. On the other hand, by Definition 3.1 and (4.5), $S_0^\infty(\{F, H, R^3\})$ is given by

$$(4.15) \quad \alpha(x) = \begin{bmatrix} \hat{\phi}_1(x_1) - \hat{\phi}_2(x_1 + x_3) \\ \hat{\phi}_2(x_1 + x_3) \end{bmatrix},$$

$$(4.16) \quad \beta(x) = \begin{bmatrix} \hat{\psi}_1(x_1) & \vdots & -\hat{\psi}_2(x_1 + x_3) \\ 0 & \vdots & \hat{\psi}_2(x_1 + x_3) \end{bmatrix},$$

where $\hat{\phi}_i, \hat{\psi}_i$, $i \in M_{1,2}$ are arbitrary C^∞ -functions of their arguments such that $\hat{\psi}_i(z) \neq 0$, $z \in R$, $i \in M_{1,2}$. From (4.13)–(4.16), we see that $S^\infty(\{F, H, R^3\}) \neq S_0^\infty(\{F, H, R^3\})$ but $S_0^\infty(\{F, H, R^3\}) \subset S^\infty(\{F, H, R^3\})$.

By Theorem 3.4, $\bar{S}^\infty(\{\bar{F}, \bar{H}, R^3\})$ is given by

$$(4.13)' \quad \bar{\alpha}(\bar{x}) = \begin{bmatrix} \bar{\phi}_1(\bar{x}_{1,1}, \bar{x}_{1,2}) \\ \bar{\phi}_2(\bar{x}_2) \end{bmatrix},$$

$$(4.14)' \quad \bar{\beta}(\bar{x}) = \begin{bmatrix} \bar{\psi}_1(\bar{x}_{1,1}, \bar{x}_{1,2}) & 0 \\ 0 & \bar{\psi}_2(\bar{x}_2) \end{bmatrix},$$

where $\bar{\phi}_i, \bar{\psi}_i$, $i \in M_{1,2}$ are arbitrary C^∞ -functions of their arguments such that $\bar{\psi}_1(\bar{x}_1, \bar{x}_2) \neq 0$, $(\bar{x}_1, \bar{x}_2) \in R^2$ and $\bar{\psi}_2(\bar{x}_3) \neq 0$, $\bar{x}_3 \in R$. As is indicated by Theorem 3.5, there is a one-to-one correspondence between the control laws of $S^\infty(\{F, H, R^3\})$ in (4.13) and (4.14) and those of $\bar{S}^\infty(\{\bar{F}, \bar{H}, R^3\})$ in (4.13)' and (4.14)'.

Using the SDS, it is easy to see how to choose control laws which decompose $\{F, H, R^3\}$ in a stable way. Suppose we want to decompose $\{F, H, R^3\}$ on R^3 with BIBS stability. First, consider $\bar{u} = \bar{\alpha}(\bar{x}) + \bar{\beta}(\bar{x})\bar{u}$ where $\bar{\alpha}, \bar{\beta}$ satisfy (4.13)' and (4.14)'. Let $\{\tilde{f}, \tilde{h}, R^3\}$ be the coordinate representation of $\{\bar{F}, \bar{H}, R^3\}^{\bar{\alpha}, \bar{\beta}}$. Then, $\{\tilde{f}, \tilde{h}, R^3\}$ is described by

$$(4.17) \quad \begin{bmatrix} \dot{\tilde{x}}_{1,1} \\ \dot{\tilde{x}}_{1,2} \end{bmatrix} = \begin{bmatrix} \bar{\phi}_1(\tilde{x}_{1,1}, \tilde{x}_{1,2}) \\ \tilde{x}_{1,2} + \bar{\phi}_1(\tilde{x}_{1,1}, \tilde{x}_{1,2}) \end{bmatrix} + \begin{bmatrix} \bar{\psi}_1(\tilde{x}_{1,1}, \tilde{x}_{1,2}) \\ \bar{\psi}_1(\tilde{x}_{1,1}, \tilde{x}_{1,2}) \end{bmatrix} \tilde{u}_1, \quad \tilde{y}_1 = \tilde{x}_{1,1},$$

$$\dot{\tilde{x}}_2 = \bar{\phi}_2(\tilde{x}_2) + \bar{\psi}_2(\tilde{x}_2)\tilde{u}_2, \quad \tilde{y}_2 = \tilde{x}_2.$$

Note that $\{\bar{f}_1, \bar{h}_1, R^2\}$, $\{\bar{f}_2, \bar{h}_2, R\}$ in (4.12) are controllable linear systems. Therefore, there are many choices of $\bar{\phi}_1, \bar{\phi}_2$ so that $\{\bar{F}, \bar{H}, R^3\}^{\bar{\alpha}, \bar{\beta}}$ is decomposed on R^3 with BIBS stability. For such a control law $\bar{u} = \bar{\alpha}(\bar{x}) + \bar{\beta}(\bar{x})\bar{u}$, choose a control law $u = \alpha(x) + \beta(x)\hat{u}$ by (3.35). Then, $\{\bar{F}, \bar{H}, R^3\}^{\bar{\alpha}, \bar{\beta}}$ is T -related on R^3 to $\{F, H, R^3\}^{\alpha, \beta}$. Recall that T is a C^∞ -diffeomorphism on R^3 . Furthermore, by a special form of T in (4.11), it follows that for any constant b , $\{x \in R^3: |T(x)| \leq b\}$ is bounded. These observations imply that $\{F, H, R^3\}^{\alpha, \beta}$ is decomposed on R^3 with BIBS stability. Thus, we have shown that there are many control laws $u = \alpha(x) + \beta(x)\hat{u}$ in $S^\infty(\{F, H, \chi\})$ which decompose $\{F, H, R^3\}$ on R^3 in a stable way.

Similarly, using a control law in $S_0^\infty(\{F, H, R^3\})$ causes (4.17) to be replaced by

$$(4.17)' \quad \begin{aligned} \begin{bmatrix} \dot{\tilde{x}}_{1,1} \\ \dot{\tilde{x}}_{1,2} \end{bmatrix} &= \begin{bmatrix} \hat{\phi}_1(\tilde{x}_{1,1}) \\ \tilde{x}_{1,2} + \hat{\phi}_1(\tilde{x}_{1,1}) \end{bmatrix} + \begin{bmatrix} \hat{\psi}_1(\tilde{x}_{1,1}) \\ \hat{\psi}_1(\tilde{x}_{1,1}) \end{bmatrix} \tilde{u}_1, & \tilde{y}_1 &= \tilde{x}_{1,1}, \\ \dot{\tilde{x}}_2 &= \hat{\phi}_2(\tilde{x}_2) + \hat{\psi}_2(\tilde{x}_2)\tilde{u}_2, & \tilde{y}_2 &= \tilde{x}_2. \end{aligned}$$

From (4.17)', we see that there is no $\hat{\phi}_1$ and $\hat{\psi}_1$ such that for every bounded \tilde{u}_1 , $\tilde{x}_{1,2}$ is bounded. By the special structure of T in (4.11), this implies that there is no control law in $S_0^\infty(\{F, H, R^3\})$, which decomposes $\{F, H, R^3\}$ on R^3 in a stable way.

5. Conclusion. In this paper, we have presented a complete characterization of decomposing control laws for a general class of smooth nonlinear systems. The two concepts, decoupling and decomposition are not globally equivalent for nonlinear systems. Some conditions under which the two concepts are at least locally equivalent are found.

We would like to emphasize the practical importance of a standard decomposed system. For the design of decomposed control systems, it may be more convenient to deal with the standard decomposed system instead of the original system. This advantage comes from the simplicity of the results for standard decomposed systems. Specifically, the class of decomposing control laws for the standard decomposed system is given by (3.34) (see Theorem 3.4) and for each decomposing control law in this class the decomposing control law for the original system can be obtained through the J -feedback relation (see Theorem 3.5). In general, the J -feedback relation which transforms the original system into the standard decomposed system requires the solutions of a set of first-order partial differential equations except for the case described in Remark 3.2. However, in some applications the J -feedback relation may be found by inspection or rather simple manipulation of the dynamic equations for the original system. This is the case for the robotic manipulators discussed in [6].

The results in this paper can be extended to more general class of nonlinear systems: (i) systems in which $F(x, u)$ is not affine in u ; (ii) systems in which the $y_i(t)$ are vectors instead of scalars. Decoupling conditions for the above more general cases, obtained by Fliess [3] and Nijmeijer [15], are useful for the extension because under their assumptions, the two concepts of decoupling and decomposition are locally equivalent. However, characterization of the whole class of decomposing control laws for the general cases becomes less explicit and more complex.

Acknowledgment. This paper contains some of the results in the author's Ph.D. dissertation. The author thanks Professor E. G. Gilbert for his sincere and excellent guide during the author's doctoral study and preparation of this paper.

REFERENCES

- [1] W. M. BOOTHBY, *An Introduction to Differential Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.
- [2] D. CLAUDE, *Découplage des systèmes nonlineaires, séries génératrices noncommutatives et algèbres de Lie*, SIAM J. Control Optim., 24 (1986), pp. 562–578.
- [3] M. FLIESS, *A new approach to the noninteracting control problem in nonlinear systems theory*, in Proc. 23rd Allerton Conference on Communication, Control, and Computing, Univ. of Illinois Press, 1985, pp. 123–129.
- [4] E. G. GILBERT, *The decoupling of multivariable systems by state variable feedback*, SIAM J. Control Optim., 7 (1969), pp. 50–64.
- [5] I. J. HA AND E. G. GILBERT, *A complete characterization of decoupling control laws for a general class of nonlinear systems*, IEEE Trans. Automat. Control., 31 (1986), pp. 823–830.
- [6] I. J. HA, *Nonlinear decoupling theory with applications to robotics*, Ph.D. thesis, Center for Robotics and Integrated Manufacturing Report RSD-TR-8-85, Univ. of Michigan, Ann Arbor, MI, 1985.
- [7] L. R. HUNT, H. LUKSIC, AND R. SU, *Exact linearization of input–output system*, Internat. J. Control, 43 (1986), pp. 247–255.
- [8] A. ISIDORI, A. J. KRENER, C. GORI-GIORGI, AND S. MONACO, *Nonlinear decoupling via feedback: a differential geometric approach*, IEEE Trans. Automat. Control, 26 (1981), pp. 331–345.
- [9] A. ISIDORI, *Nonlinear Control Systems: An Introduction*, Lecture Notes in Control and Information Sci. 72, Springer-Verlag, Berlin, New York, 1985.
- [10] A. ISIDORI AND A. J. KRENER, *On feedback equivalence of nonlinear systems*, Systems Control Lett., 2 (1982), pp. 118–121.
- [11] S. H. MIKHAIL AND W. H. WONHAM, *Local decomposability and the disturbance decoupling in nonlinear autonomous systems*, in Proc. Allerton Conf. Commun. Contr. Comput., 1978, pp. 664–669.
- [12] A. S. MORSE AND W. M. WONHAM, *Status of noninteracting control*, IEEE Trans. Automat. Control, 16 (1971), pp. 568–581.
- [13] H. NIJMEIJER, *Feedback decomposition of nonlinear control systems*, IEEE Trans. Automat. Control, 28 (1983), pp. 861–862.
- [14] ———, *Zeros at infinity for affine nonlinear control systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 566–573.
- [15] ———, *The regular local noninteracting control problem for nonlinear systems*, SIAM J. Control Optim., 24 (1986), pp. 1232–1245.
- [16] W. A. PORTER, *Diagonalization and inverses for nonlinear systems*, Internat. J. Control, 10 (1970), pp. 252–264.
- [17] W. RESPONDEK, *On decomposition of nonlinear control systems*, Systems Control Lett., 1 (1982), pp. 301–308.
- [18] S. N. SINGH AND W. J. RUGH, *Decoupling in a class of nonlinear systems by state variable feedback*, ASME J. Dynamic Systems Measurement Control, 94 (1972), pp. 323–329.
- [19] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
- [20] H. J. SUSSMANN, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Systems Theory, 10 (1977), pp. 263–284.
- [21] F. W. WARNER, *Foundations of Differentiable Manifolds and Lie Groups*, Scott–Foresman, Glenview, IL, 1971.
- [22] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, 2nd edition, Springer-Verlag, New York, 1979.

NOTE ON BOUNDARY STABILIZATION OF WAVE EQUATIONS*

JOHN E. LAGNESE†

Abstract. An energy decay rate is obtained for solutions of wave type equations in a bounded region in \mathbb{R}^n whose boundary consists partly of a nontrapping reflecting surface and partly of an energy absorbing surface. Unlike most previous results on this subject, the results presented here are potentially valid for regions having connected boundaries.

Key words. wave equations, boundary stabilization, exponential stability

AMS(MOS) subject classifications. 93D15, 35L05

Let Ω be a bounded, open, connected set in \mathbb{R}^n ($n \geq 2$) and let Γ denote its boundary. Assume that Γ is piecewise smooth and consists of two parts, Γ_0 and Γ_1 , with $\Gamma_1 \neq \emptyset$ and relatively open in Γ , and Γ_0 either empty or having a nonempty interior. We set $\Sigma_0 = \Gamma_0 \times (0, \infty)$, $\Sigma_1 = \Gamma_1 \times (0, \infty)$. Let k be an $L^\infty(\Gamma_1)$ function satisfying $k(x) \geq 0$ almost everywhere on Γ_1 . Consider the problem

$$(1) \quad w'' - \Delta w = 0 \quad \text{in } \Omega \times (0, \infty),$$

$$(2a) \quad \frac{\partial w}{\partial \nu} = -kw' \quad \text{on } \Sigma_1,$$

$$(2b) \quad w = 0 \quad \text{on } \Sigma_0,$$

$$(3) \quad w(0) = w^0, \quad w'(0) = w^1 \quad \text{in } \Omega$$

where $' = d/dt$ and ν is the unit normal of Γ pointing towards the exterior of Ω .

Associated with each solution of (1.1) is its total *energy* at time t :

$$E(t) = \frac{1}{2} \int_{\Omega} (w'^2 + |\nabla w|^2) dx.$$

A simple calculation shows that

$$E'(t) = - \int_{\Gamma_1} kw'^2 d\Gamma \leq 0;$$

hence $E(t)$ is nonincreasing. The question of interest is the following. Under what conditions is it true that there is an *exponential decay rate* for $E(t)$, i.e.,

$$(4) \quad E(t) \leq C e^{-\omega t} E(0), \quad t \geq 0$$

for some *positive* ω ?

The first person to establish (4) for solutions of (1)–(3) was Chen [1], under the following assumptions: $k(x) \geq k_0 > 0$ on Γ_1 , and there is a point $x_0 \in \mathbb{R}^n$ such that

$$(5) \quad (x - x_0) \cdot \nu \leq 0, \quad x \in \Gamma_0,$$

$$(6) \quad (x - x_0) \cdot \nu \geq \gamma > 0, \quad x \in \Gamma_1.$$

Chen slightly relaxed (5) and (6) in a later paper [2]. The most general result to date in terms of the assumed geometrical conditions on Γ appears in [6]. There it is proved

* Received by the editors April 13, 1987; accepted for publication (in revised form) November 25, 1987. This research was supported by the Air Force Office of Scientific Research under grant AFOSR-86-0162.

† Department of Mathematics, Georgetown University, Washington, D.C. 20057.

that (4) is valid provided there exists a vector field $h(x) = [h_1(x), \dots, h_n(x)] \in C^2(\bar{\Omega})$ such that

$$(7) \quad h \cdot \nu \leq 0 \quad \text{on } \Gamma_0,$$

$$(8) \quad h \cdot \nu \geq \gamma > 0 \quad \text{on } \Gamma_1,$$

$$(9) \quad \text{The matrix } (\partial h_i / \partial x_j + \partial h_j / \partial x_i) \text{ is positive definite on } \bar{\Omega}.$$

This last result has subsequently been re-proved by Lasiecka and Triggiani [8] and Triggiani [10] using methods different from those in [6]. In all of the papers cited, the estimate (4) was obtained from estimates on $\int_0^\infty E(t) dt$ by employing a result of Datko [3] (later extended by Pazy [9]).

An important observation is that when Γ is smooth, conditions (5) and (6) (respectively, (7) and (8)) together force $\bar{\Gamma}_0 \cap \bar{\Gamma}_1 = \emptyset$. Thus if $\Gamma_0 \neq \emptyset$, the above results cannot apply to regions Ω having a connected boundary. However, in a recent paper [5], Kormornik and Zuazua succeeded in relaxing Chen's condition (6) to

$$(10) \quad (x - x_0) \cdot \nu \geq 0 \quad \text{on } \Gamma_1,$$

thus *in principle* allowing for regions with smooth connected boundaries (see Remark 3), but at the expense of replacing the boundary condition (2a) by

$$(11) \quad \frac{\partial w}{\partial \nu} = -((x - x_0) \cdot \nu) w' \quad \text{on } \Sigma_1.$$

In addition, the proof in [5] gives, in a simple and natural way, *explicit* estimates of the constants C and ω in (4) in terms of the geometry of Ω , more specifically, in terms of the constants μ_0 and μ_1 which appear in (16), (17) below.

Remark 1. The proofs in [1], [2], [6], [8], [10] *implicitly* contain estimates of ω since we may estimate ω in terms of the constant K satisfying

$$\int_0^\infty E(t) dt \leq KE(0).$$

However, except in the case of (5), (6), K contains other constants whose explicit values (in terms of Ω and the vector field h) seem difficult to determine.

Remark 2. If $\bar{\Gamma}_0 \cap \bar{\Gamma}_1 \neq \emptyset$, the proof of Komornik and Zuazua is restricted to dimension $n = 2$. The reason is that their proof requires more regularity than solutions of (1), (2b), (3), (11) will generally possess, regardless of the smoothness of the initial data. (Such regularity is available if $\bar{\Gamma}_0 \cap \bar{\Gamma}_1 = \emptyset$.) However, when $n = 2$, the regularity requirement may be weakened by using a certain a priori estimate recently obtained by Grisvard [4]. Grisvard's proof is very technical, and the presence of the factor $(x - x_0) \cdot \nu(x)$ in the boundary condition (11) is crucial to the proof.

The purpose of this paper is to extend the result of [5] in two ways: first, by replacing the specific vector field $x - x_0$ in (5) and (10) by a general vector field $h(x)$ satisfying (7), (9), and

$$(12) \quad h \cdot \nu \geq 0 \quad \text{on } \Gamma_1;$$

second, by replacing the boundary condition (11) by

$$(13) \quad \frac{\partial w}{\partial \nu} = -k^*(h \cdot \nu) w' \quad \text{on } \Sigma_1$$

where $k^* \in L^\infty(\Gamma_1)$ satisfies $k^* \geq k_0 > 0$ on Γ_1 . Note that if $h \cdot \nu \geq \gamma > 0$ on Γ_1 , boundary condition (2a) may be written as (13) with $k^* = k/(h \cdot \nu)$. Hence, in this situation, we

recover (a sharpened form of) the main result of [6] (see the theorem below). The decay rate ω that will be obtained depends on (among other parameters of the problem) the gain k^* . If k^* is the constant function k_0 and h is a radial field $x - x_0$, it will be seen that there is a unique value of k_0 which *maximizes* ω . Furthermore, the optimal k_0 can be expressed explicitly in terms of computable parameters.

The formal statements of the two results to be proved are as follows.

THEOREM. *Let w be a regular solution to (1), (2b), and (13). Then there is a positive constant ω such that*

$$\int_0^\infty E(s) ds \leq \frac{1}{\omega} E(0),$$

$$\int_t^\infty E(s) ds \leq e^{-\omega t} \int_0^\infty E(s) ds, \quad t \geq 0.$$

COROLLARY. *Under the hypotheses of the theorem,*

$$E(t) \leq e \cdot e^{-\omega t} E(0), \quad t \geq \frac{1}{\omega}.$$

Remark 3. The geometric conditions (7), (9), and (12) admit a much larger class of configurations $(\Omega, \Gamma_0, \Gamma_1)$ than do conditions (5) and (10) considered in [5]. But as in [5], we are faced with the problem of lack of (proven) regularity of solutions when $\bar{\Gamma}_0 \cap \bar{\Gamma}_1 \neq \emptyset$. It may be the case that by extending Grisvard's estimate to the present situation, the regularity assumption can be weakened, at least if $n = 2$. But that has not been proved at this time.

Remark 4. As in [5], we will obtain an "explicit" estimate of ω . However, as in [6], [8], [10], this estimate will contain a constant which cannot be easily determined in terms of Ω and the vector field h , unless h is a radial field.

Remark 5. The theorem and corollary may be extended to generalized wave equations with time-independent coefficients as in [6] but under the weaker condition (12) and also to linear elastodynamic systems (cf. [6, p. 167], [7]). We omit the details.

Proof of the corollary. Since $E(t)$ is nonincreasing, for every $\tau > 0$

$$\tau E(t + \tau) \leq \int_t^\infty E(s) ds \leq \frac{1}{\omega} e^{-\omega t} E(0)$$

or

$$(14) \quad E(t + \tau) \leq \frac{e^{\omega\tau}}{\omega\tau} e^{-\omega(t+\tau)} E(0), \quad \tau > 0.$$

The first factor on the right has its minimum at $\tau = 1/\omega$ and for this value of τ (14) becomes

$$E\left(t + \frac{1}{\omega}\right) \leq e \cdot e^{-\omega(t+1/\omega)} E(0), \quad t \geq 0.$$

Proof of the theorem. We assume that $\Gamma_0 \neq \emptyset$, since there are simple examples which show that the theorem is *false* if $\Gamma_0 = \emptyset$.

Define the matrix $H = (\partial h_i / \partial x_j + \partial h_j / \partial x_i)$. By assumption we have

$$(15) \quad H\xi \cdot \xi \geq h_0 |\xi|^2, \quad \xi \in \mathbb{R}^n, \quad x \in \Omega, \quad h_0 > 0.$$

Since multiplication of h by a positive constant leaves Γ_0 and Γ_1 invariant, we may (and do) assume that $h_0 = 1$ in (15).

Define constants μ_0 and μ_1 by

$$(16) \quad \int_{\Gamma_1} v^2 dx \leq \mu_0 \int_{\Omega} |\nabla v|^2 dx,$$

$$(17) \quad \int_{\Omega} v^2 dx \leq \mu_1 \int_{\Omega} |\nabla v|^2 dx$$

for all $v \in H^1(\Omega)$ such that $v = 0$ on Γ_0 . For $\varepsilon > 0$ and fixed, define

$$F_\varepsilon(t) = E(t) + \varepsilon \rho(t)$$

where

$$\rho(t) = 2(w', h \cdot \nabla w) + ((h_{j,j} - 1)w, w').$$

We note that

$$|\rho(t)| \leq C_0 E(t);$$

hence

$$(18) \quad (1 - \varepsilon C_0)E(t) \leq F_\varepsilon(t) \leq (1 + \varepsilon C_0)E(t)$$

where C_0 depends on h and μ_1 . We will show that for ε sufficiently small,

$$(19) \quad F'_\varepsilon(t) \leq -\varepsilon E(t) + C\varepsilon \int_{\Omega} w^2 dx$$

where C depends on h , μ_0 , and μ_1 .

Remark 6. In what follows we will use the notation $\varphi_{,j} = \partial \varphi / \partial x_j$. The standard summation convention for repeated indices will also be used throughout. Since some of the calculations which follow are similar to those in [6], not all details will be provided. The interested reader is referred to [6] for a more leisurely derivation of some of the estimates below.

We have

$$(20) \quad \rho'(t) = 2(w'', h \cdot \nabla w) + 2(w', h \cdot \nabla w') + ((h_{j,j} - 1)w', w') + ((h_{j,j} - 1)w, w'').$$

From (1), (2) we have

$$(21) \quad (w'', v) + (\nabla w, \nabla v) + b(w', v) - \int_{\Gamma_0} \frac{\partial w}{\partial \nu} v d\Gamma = 0 \quad \forall v \in H^1(\Omega)$$

where

$$b(w', v) = \int_{\Gamma_1} k^*(h \cdot \nu) w' v d\Gamma.$$

We use (21) to calculate $(w'', h \cdot \nabla w)$ and $((h_{j,j} - 1)w, w'')$ in (20). We have

$$(22) \quad (w'', h \cdot \nabla w) = -(\nabla w, \nabla(h \cdot \nabla w)) - b(w', h \cdot \nabla w) + \int_{\Gamma_0} \frac{\partial w}{\partial \nu} h \cdot \nabla w d\Gamma.$$

A direct calculation gives

$$(23) \quad (\nabla w, \nabla(h \cdot \nabla w)) = \int_{\Omega} h_{i,j} w_{,i} w_{,j} dx - \frac{1}{2} \int_{\Omega} h_{j,j} |\nabla w|^2 dx + \frac{1}{2} \int_{\Gamma} h \cdot \nu |\nabla w|^2 d\Gamma.$$

Similarly,

$$(24) \quad ((h_{j,j} - 1)w, w'') = - \int_{\Omega} (h_{j,j} - 1)|\nabla w|^2 dx - \int_{\Omega} h_{j,ij} w w_{,i} dx - b(w', (h_{j,j} - 1)w).$$

We also have

$$(25) \quad (w', h \cdot \nabla w') = \frac{1}{2} \int_{\Gamma_1} (h \cdot \nu) w'^2 d\Gamma - \frac{1}{2} \int_{\Omega} h_{j,j} w'^2 dx.$$

Use of (22)–(25) in (20) gives

$$(26) \quad \begin{aligned} \rho'(t) = & -2 \int_{\Omega} h_{i,j} w_{,i} w_{,j} dx + \int_{\Omega} |\nabla w|^2 dx - \int_{\Omega} w'^2 dx - \int_{\Omega} h_{j,ij} w w_{,i} dx \\ & - \int_{\Gamma} (h \cdot \nu) |\nabla w|^2 d\Gamma + 2 \int_{\Gamma_0} \frac{\partial w}{\partial \nu} h \cdot \nabla w d\Gamma + \int_{\Gamma_1} (h \cdot \nu) w'^2 d\Gamma \\ & - 2b(w', h \cdot \nabla w) - b(w', (h_{j,j} - 1)w). \end{aligned}$$

The integrals over Γ_0 , viz.,

$$(27) \quad 2 \int_{\Gamma_0} \frac{\partial w}{\partial \nu} h \cdot \nabla w d\Gamma - \int_{\Gamma_0} h \cdot \nu |\nabla w|^2 d\Gamma = \int_{\Gamma_0} h \cdot \nu \left(\frac{\partial w}{\partial \nu} \right)^2 d\Gamma \leq 0.$$

We also have the estimates

$$(28) \quad \begin{aligned} |b(w', h \cdot \nabla w)| &= \left| \int_{\Gamma_1} k^*(h \cdot \nu) w' (h \cdot \nabla w) d\Gamma \right| \\ &\leq \frac{\alpha}{2} \int_{\Gamma_1} h \cdot \nu |h \cdot \nabla w|^2 d\Gamma + \frac{1}{2\alpha} \int_{\Gamma_1} k^{*2} (h \cdot \nu) w'^2 d\Gamma \\ &\leq \frac{1}{2} \int_{\Gamma_1} h \cdot \nu |\nabla w|^2 d\Gamma + \frac{1}{2} k_M^2 C_1 \int_{\Gamma_1} (h \cdot \nu) w'^2 d\Gamma \end{aligned}$$

for an appropriate α , where $k_M = \sup \{k^*(x) \mid x \in \Gamma_1\}$,

$$(29) \quad |b(w', (h_{j,j} - 1)w)| \leq k_M^2 \frac{C_2}{2\delta} \int_{\Gamma_1} (h \cdot \nu) w'^2 d\Gamma + \frac{\delta}{2} \mu_0 \int_{\Omega} |\nabla w|^2 dx,$$

$$(30) \quad \left| \int_{\Omega} h_{j,ij} w w_{,i} dx \right| \leq \frac{C_3}{2\delta} \int_{\Omega} w^2 dx + \frac{\delta}{2} \int_{\Omega} |\nabla w|^2 dx,$$

where C_1 , C_2 , and C_3 depend on h only and where $\delta > 0$ will be chosen below. Use of (27)–(30) and (15) (recall that $h_0 = 1$) in (26) yields

$$\begin{aligned} \rho'(t) \leq & - \int_{\Omega} (w'^2 + |\nabla w|^2) dx + \frac{\delta}{2} (\mu_0 + 1) \int_{\Omega} |\nabla w|^2 dx \\ & + \left[k_M^2 \left(\frac{C_1 + C_2}{2\delta} \right) + 1 \right] \int_{\Gamma_1} (h \cdot \nu) w'^2 d\Gamma + \frac{C_3}{2\delta} \int_{\Omega} w^2 dx. \end{aligned}$$

Choosing $\delta = 1/(\mu_0 + 1)$ we obtain

$$(31) \quad \rho'(t) \leq -E(t) + (k_M^2 C_4 + 1) \int_{\Gamma_1} (h \cdot \nu) w'^2 d\Gamma + C_5 \int_{\Omega} w^2 dx$$

where $C_4 = C_1 + C_2/(2\delta)$, $C_5 = C_3/(2\delta)$. Since $k^* \geq k_0 > 0$ on Γ_1 , we obtain from (31)

$$\begin{aligned}
 F'_\varepsilon(t) &= E'(t) + \varepsilon \rho'(t) = - \int_{\Gamma_1} k^*(h \cdot \nu) w'^2 d\Gamma + \varepsilon \rho'(t) \\
 (32) \quad &\leq -\varepsilon E(t) + \varepsilon C_5 \int_{\Omega} w^2 dx + [\varepsilon(k_M^2 C_4 + 1) - k_0] \int_{\Gamma_1} (h \cdot \nu) w'^2 d\Gamma \\
 &\leq -\varepsilon E(t) + \varepsilon C_5 \int_{\Omega} w^2 dx
 \end{aligned}$$

provided

$$\varepsilon(k_M^2 C_4 + 1) \leq k_0.$$

This establishes (19).

Let $\beta > 0$ and consider

$$\begin{aligned}
 (33) \quad \int_t^\infty e^{-\beta(s-t)} F'_\varepsilon(s) ds &= -F_\varepsilon(t) + \beta \int_t^\infty e^{-\beta(s-t)} F_\varepsilon(s) ds \\
 &\leq -\varepsilon \int_t^\infty e^{-\beta(s-t)} E(s) ds + \varepsilon C_5 \int_t^\infty e^{-\beta(s-t)} |w(\cdot, s)|^2 ds.
 \end{aligned}$$

From (18), $F_\varepsilon(s) \geq 0$ provided

$$\varepsilon C_0 \leq 1.$$

From Theorem 2 of [6], we have the estimate

$$(34) \quad \int_t^\infty e^{-\beta(s-t)} |w(\cdot, s)|^2 ds \leq C_\eta^* E(t) + \eta \int_t^\infty e^{-\beta(t-s)} E(s) ds$$

where $\eta > 0$ is arbitrary and C_η^* is a constant independent of β . Therefore (33), (34) imply

$$(35) \quad \varepsilon \int_t^\infty e^{-\beta(s-t)} E(s) ds \leq F_\varepsilon(t) + \varepsilon C_5 \left[C_\eta^* E(t) + \eta \int_t^\infty e^{-\beta(s-t)} E(s) ds \right]$$

where $\varepsilon = \min(1/C_0, k_0/C_4)$. Choosing $\eta = 1/qC_5$ ($q > 1$) in (35) gives the estimate

$$(36) \quad \frac{(q-1)}{q} \varepsilon \int_t^\infty e^{-\beta(s-t)} E(s) ds \leq F_\varepsilon(t) + \varepsilon C_5 C_{1/q}^* E(t) \leq (1 + \varepsilon K_q) E(t)$$

where $K_q = C_0 + C_5 C_{1/q}^*$ does not depend on β . Define $\omega_q = (q-1)\varepsilon/(q(1 + \varepsilon K_q))$ and let $\beta \rightarrow 0$ in (36) to obtain

$$(37) \quad \int_t^\infty E(s) ds \leq \frac{1}{\omega_q} E(t), \quad t \geq 0, \quad q > 1.$$

The conclusions of the theorem with $\omega = \omega_2 = \varepsilon/2(1 + \varepsilon K_2)$ (for example) follow easily from (37).

Remark 7. Since $\omega = \varepsilon/2(1 + \varepsilon K_2)$ increases monotonically to $1/2K_2$ as $\varepsilon \rightarrow \infty$, the maximum decay rate will be achieved when the largest admissible value of ε is selected, that is,

$$\varepsilon = \min \left[\frac{k_0}{k_M^2 C_4 + 1}, \frac{1}{C_0} \right]$$

for a given gain $k^*(x)$. Let us assume that $k^*(x) \equiv k_0$, and consider the problem

$$\max \{ \omega(k_0) \mid k_0 > 0 \}.$$

In general, it will be very difficult to determine that value of k_0 which maximizes $\omega(k_0)$ (assuming that such a value exists), since the constants C_η^* (and therefore K_2) depend on k_0 in an unknown way. However, suppose that $h(x) = x - x_0$. Then the constant C_3 in (30) is zero; hence $C_5 = 0$, and by retracing the steps following (32) ((34) is no longer needed) we find that

$$\omega = \frac{\varepsilon}{1 + \varepsilon C_0}.$$

Therefore,

$$\begin{aligned} \max \{ \omega(k_0) \mid k_0 > 0 \} &= \max_{k_0 > 0} \min \left[\frac{k_0}{k_0^2 C_4 + 1}, \frac{1}{C_0} \right] \\ &= \min \left\{ \max_{k_0 > 0} \left[\frac{k_0}{k_0^2 C_4 + 1} \right], \frac{1}{C_0} \right\}. \end{aligned}$$

It is easy to see that the function $k \rightarrow k/(k^2 C_4 + 1)$ has a unique maximum which occurs at the value $k = 1/\sqrt{C_4}$. Therefore

$$\max \{ \omega(k_0) \mid k_0 > 0 \} = \min \left[\frac{1}{2\sqrt{C_4}}, \frac{1}{C_0} \right],$$

this value being achieved when

$$\begin{aligned} k_0 &= \frac{1}{\sqrt{C_4}} \quad \text{if } \frac{1}{2\sqrt{C_4}} \leq \frac{1}{C_0}, \\ k_0 &= \frac{1}{2C_4} [C_0 \pm (C_0^2 - 4C_4)^{1/2}] \quad \text{if } \frac{1}{2\sqrt{C_4}} \geq \frac{1}{C_0}. \end{aligned}$$

REFERENCES

- [1] G. CHEN, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl. (9), 58 (1979), pp. 249-274.
- [2] ———, *A note on boundary stabilization of the wave equation*, SIAM J. Control Optim., 19 (1981), pp. 106-113.
- [3] R. DATKO, *Extending a theorem of Liapunov to Hilbert spaces*, J. Math. Anal. Appl., 32 (1970), pp. 610-616.
- [4] P. GRISVARD, *Contrôlabilité exacte des solutions de certains problèmes mixtes pour l'équation des ondes dans un polygone ou un polyèdre*, J. Math. Pures Appl., to appear.
- [5] V. KORMORNIK AND E. ZUAZUA, *Stabilisation frontière de l'équation des ondes: Une méthode directe*, C. R. Acad. Sci. Paris, Ser. I Math., 305 (1987), pp. 605-608.
- [6] J. LAGNESE, *Decay of solutions of the wave equation in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163-182.
- [7] ———, *Boundary stabilization of linear elastodynamic systems*, SIAM J. Control Optim., 21 (1983), pp. 968-984.
- [8] I. LASIECKA AND R. TRIGGIANI, *Uniform exponential energy decay of the wave equation in a bounded region with $L_2(0, \infty; L_2(\Gamma))$ -feedback control in the Dirichlet boundary condition*, J. Differential Equations, 66 (1987), pp. 340-390.
- [9] A. PAZY, *On the applicability of Lyapunov's theorem in Hilbert space*, SIAM J. Math. Anal., 3 (1972), pp. 291-294.
- [10] R. TRIGGIANI, *Wave equation on a bounded domain with boundary dissipation: An operator approach*, J. Math. Anal. Appl., to appear.

STABILIZATION OF AN UNCERTAIN LINEAR SYSTEM IN WHICH UNCERTAIN PARAMETERS ENTER INTO THE INPUT MATRIX*

IAN R. PETERSEN†

Abstract. This paper presents a procedure for stabilizing a class of uncertain linear systems. The uncertain systems under consideration are described by state equations in which the input matrix depends on a matrix of uncertain parameters. This matrix of uncertain parameters may be time-varying; however, it is constrained by a bound on its induced norm.

The stabilization procedure presented involves the repeated solution of an algebraic Riccati equation. When a positive definite solution to this Riccati equation is obtained, this solution is used to construct a stabilizing linear control law. Another important result contained in this paper can be stated roughly as follows: For the class of uncertain systems under consideration, if a system can be stabilized via nonlinear control, then it is also possible to stabilize the system via linear control.

Key words. uncertain systems, Riccati equations, stabilization, state feedback

AMS(MOS) subject classification. 93015

1. Introduction. A problem of recurring interest in recent years concerns the stabilization via state feedback of uncertain linear systems containing unknown but bounded, time-varying, uncertain parameters (see, e.g., [1]–[6]). This paper considers uncertain linear systems in which the uncertain parameters enter only into the “input matrix” (see also [7]). The uncertain parameters are in the form of an uncertain matrix $F(t)$ contained in a set of the form $\{F: F'F \leq I\}$. As in [1]–[6], the approach taken in this paper relies on the use of quadratic Lyapunov functions. This leads us to consider a notion of quadratic stabilizability. The main result of this paper is a necessary and sufficient condition for a given uncertain linear system to be quadratically stabilizable.

The procedure presented for constructing the stabilizing control law involves the repeated solution of an algebraic Riccati equation. When a positive-definite solution to the Riccati equation is obtained, this solution is then used to construct a stabilizing linear feedback control law. The fact that the stabilization procedure generates a linear control law leads to an important corollary. This corollary can be stated roughly as follows: If an uncertain system (in the class under consideration) is quadratically stabilizable via nonlinear control, then it is also stabilizable via linear control.

2. Systems and definitions. The uncertain linear systems considered in this paper are described by state equations of the form

$$(\Sigma) \quad \begin{aligned} \dot{x}(t) &= Ax(t) + (B + DF(t)E)u(t), \\ F(t)'F(t) &\leq I \end{aligned}$$

where $x \in \mathbb{R}^n$ is the *state*, $u(t) \in \mathbb{R}^m$ is the *control input*, and $F(t) \in \mathbb{R}^{p \times q}$ is a *matrix of uncertain parameters*. It is assumed that the elements of $F(\cdot)$ are Lebesgue measurable functions such that $I - F(t)'F(t)$ is positive semidefinite for all $t \geq 0$. Furthermore, it is assumed that the $q \times m$ matrix E is of full rank.

* Received by the editors May 4, 1987; accepted for publication (in revised form) October 23, 1987. This work was supported by the Australian Research Grants Scheme.

† Department of Electrical Engineering, Australian Defence Force Academy, Canberra ACT 2600, Australia.

We will be concerned with stabilizing uncertain systems of the form (Σ) using state feedback control. In particular, the stability of the resulting closed-loop system will be established using a quadratic Lyapunov function. Hence, we introduce the following definition (see also [3]).

DEFINITION 2.1. The uncertain linear system (Σ) is said to be *quadratically stabilizable* if there exists a continuous feedback control $p(\cdot): \mathbb{R}^n \rightarrow \mathbb{R}^m$ with $p(0) = 0$, an $n \times n$ positive-definite symmetric matrix P , and a constant $\alpha > 0$ such that the following condition holds: Given any admissible uncertainty $F(\cdot)$, the Lyapunov derivative corresponding to the closed-loop system

$$(\Sigma_{cl}) \quad \dot{x}(t) = Ax(t) + (B + DF(t)E)p(x(t))$$

satisfies the inequality

$$(2.1) \quad \begin{aligned} L(x, t) &:= x'[A'P + PA]x + 2x'P(B + DF(t)E)p(x) \\ &\leq -\alpha \|x\|^2 \end{aligned}$$

for all nonzero $x \in \mathbb{R}^n$ and $t \geq 0$. In this inequality, $\|\cdot\|$ denotes the standard Euclidean norm.

If (2.1) is satisfied, it is straightforward to verify that the corresponding closed-loop system is uniformly and asymptotically stable (see, e.g., [8]).

Remark. In the definition above, we have allowed for the use of the nonlinear control $u(t) = p(x(t))$. However, it will be shown in the sequel that for any system of the form (Σ) , attention can be limited to the use of linear control.

Notation. The following notation will be used in the sequel. Given any symmetric $k \times k$ matrices M and N , the notation $M > N$ denotes the fact that $M - N$ is positive definite. Similarly, $M \geq N$ denotes the fact that $M - N$ is positive semidefinite. Also, $\lambda_{\min}[M]$ denotes the minimum eigenvalue of the matrix M . Throughout the paper, $\|x\|$ will refer to the Euclidean norm of a vector x .

We now present a condition which is necessary and sufficient for the quadratic stabilizability of the system (Σ) . In order to state this condition, we first define an $(m - q) \times m$ matrix Φ as follows: Let Φ be any full rank $(m - q) \times m$ matrix such that¹

$$(2.2) \quad \{x \in \mathbb{R}^m: \Phi x = 0\} = \{x \in \mathbb{R}^m: x = E'\eta \text{ for some } \eta \in \mathbb{R}^q\}.$$

Condition 1. Let $Q \in \mathbb{R}^{n \times n}$ be a given positive definite symmetric weighting matrix and let $\Phi \in \mathbb{R}^{(m-q) \times m}$ be a full rank matrix satisfying (2.2). Then the system (Σ) is said to satisfy *Condition 1* if there exists an $\varepsilon > 0$ such that the Riccati equation

$$(2.3) \quad A'P + PA - PBE'(EE')^{-2}EB'P - \frac{1}{\varepsilon}PB\Phi'\Phi B'P + PDD'P + \varepsilon Q = 0$$

has a positive definite solution P .

We will now show that the success or failure of this condition is independent of the choice of the matrices Q and Φ .

THEOREM 2.1. Suppose there exist a positive definite matrix $Q \in \mathbb{R}^{n \times n}$, a full rank matrix $\Phi \in \mathbb{R}^{(m-q) \times m}$ satisfying (2.2), and a constant $\varepsilon > 0$ such that the Riccati equation (2.3) has a positive definite solution. Then given any positive definite weighting matrix $\tilde{Q} \in \mathbb{R}^{n \times n}$ and any full-rank matrix $\tilde{\Phi} \in \mathbb{R}^{(m-q) \times m}$ satisfying (2.2), there exists a constant

¹ If $m = q$, we let $\Phi = 0$.

$\varepsilon^* > 0$ such that the Riccati equation

$$(2.4) \quad A'P + PA - PBE'(EE')^{-2}EB'P - \frac{1}{\tilde{\varepsilon}}PB\tilde{\Phi}'\tilde{\Phi}B'P + PDD'P + \tilde{\varepsilon}\tilde{Q} = 0$$

has a positive definite solution for all $\tilde{\varepsilon} \in (0, \varepsilon^*]$.

Proof. If (2.3) has a positive definite solution P , then the positive definite matrix $S = P^{-1}$ satisfies the Riccati equation

$$(2.5) \quad -AS - SA' - \varepsilon SQS + BE'(EE')^{-2}EB' + \frac{1}{\varepsilon}B\Phi'\Phi B' - DD' = 0.$$

We now choose the constant $\varepsilon^* > 0$ such that $\varepsilon^*\tilde{Q} < \varepsilon Q$ and $\Phi'\Phi/\varepsilon^* \geq \tilde{\Phi}'\tilde{\Phi}/\varepsilon$. Hence, by Lemma 1 of [9] and the results of [10], it follows that the Riccati equation

$$-AS - SA' - \tilde{\varepsilon}S\tilde{Q}S + BE'(EE')^{-2}EB' + \frac{1}{\tilde{\varepsilon}}B\tilde{\Phi}'\tilde{\Phi}B' - DD' = 0$$

has a maximal symmetric solution S^+ for all $\tilde{\varepsilon} \in (0, \varepsilon^*]$. Furthermore, from the main result of [11], it follows that $S^+ \geq S > 0$. However, the positive definite matrix $P^+ = (S^+)^{-1}$ must satisfy the Riccati equation (2.4). This is the required result. \square

3. The main result.

THEOREM 3.1. *Condition 1 is a necessary and sufficient condition for the uncertain system (Σ) to be quadratically stabilizable. Furthermore, if Condition 1 is satisfied, then the linear control law*

$$(3.1) \quad u(t) = -\left\{\frac{1}{2\varepsilon}\Phi'\Phi + E'(EE')^{-2}E\right\}B'Px(t)$$

is a suitable stabilizing control law.

In proving this theorem, the following lemma will be used. This lemma is a corollary of a theorem first proved by P. Finsler in 1937 (see [16]).

LEMMA 3.1 (see Appendix A for proof). *Let M and V be given symmetric matrices and let W be a matrix such that*

$$(3.2) \quad x'Mx < 0$$

for all nonzero vectors x such that $x'Vx \leq 0$ and $Wx = 0$. Then there exist constants $\sigma > 0$ and $\lambda > 0$ such that

$$M - \sigma V - \lambda W'W < 0.$$

Proof of Theorem 3.1. (Sufficiency.) Suppose that the system (Σ) satisfies Condition 1 and let P be the positive-definite solution to the Riccati equation (2.3). The required state feedback control law is then constructed according to (3.1). This leads to the closed-loop uncertain system

$$\dot{x}(t) = Ax(t) - (B + DF(t)E)\left\{\frac{1}{2\varepsilon}\Phi'\Phi + E'(EE')^{-2}E\right\}B'Px(t),$$

$$F(t)'F(t) \leq I.$$

Let $F(\cdot)$ be any admissible uncertainty. It follows that the Lyapunov derivative corresponding to this closed-loop system and the Lyapunov function $V(x) = x'Px$ is given by

$$L(x, t) = 2x'P\left\{A - \frac{1}{2\varepsilon}B\Phi'\Phi B'P - BE'(EE')^{-2}EB'P - DF(t)(EE')^{-1}EB'P\right\}X(t).$$

In order to establish an upper bound for this Lyapunov derivative, we first establish the following claim.

CLAIM 1. Given any matrix $F \in \mathbb{R}^{p \times q}$ such that $F'F \leq I$, then

$$-2x'PDF(EE')^{-1}EB'Px \leq x'PDD'Px + x'PBE'(EE')^{-2}EB'Px.$$

To establish this claim, we observe that for any $x \in \mathbb{R}^n$,

$$\begin{aligned} 0 &\leq \|D'Px + F(EE')^{-1}EB'Px\|^2 \\ &= x'PDD'Px + 2x'PDF(EE')^{-1}EB'Px + x'PBE'(EE')^{-1}F'F(EE')^{-1}EB'Px \\ &\leq x'PDD'Px + 2x'DF(EE')^{-1}EB'Px + x'PBE'(EE')^{-2}EB'Px, \end{aligned}$$

and hence the claim follows.

Applying this claim to the above expression for the closed-loop Lyapunov derivative, we obtain the following upper bound:

$$\begin{aligned} L(x, t) &\leq x' \left[A'P + PA - \frac{1}{\varepsilon} PB\Phi'\Phi B'P \right. \\ &\quad \left. - 2PBE'(EE')^{-2}EB'P + PDD'P + PBE'(EE')^{-2}EB'P \right] x \\ &= x' \left[A'P + PA - \frac{1}{\varepsilon} PB\Phi'\Phi B'P - PBE'(EE')^{-2}EB'P + PDD'P \right] x \end{aligned}$$

for all $x \in \mathbb{R}^n$ and $t \geq 0$. Furthermore, by the Riccati equation (2.3), it follows that

$$\begin{aligned} L(x, t) &\leq -\varepsilon x'Qx \\ &\leq -\varepsilon \lambda_{\min}[Q] \|x\|^2 \end{aligned}$$

for all $x \in \mathbb{R}^n$ and $t \geq 0$. That is, (2.1) is satisfied with $\alpha = \varepsilon \lambda_{\min}[Q] > 0$. Hence the system (Σ) is quadratically stabilizable.

(Necessity.) Suppose the system (Σ) is quadratically stabilizable. It follows that there exist a control law $p(\cdot): \mathbb{R}^n \rightarrow \mathbb{R}^m$, a positive definite matrix P , and a constant $\alpha > 0$ such that the following condition holds:

$$(3.3) \quad x'[A'P + PA]x + 2x'P(B + DFE)p(x) + \alpha \|x\|^2 \leq 0$$

for all $x \in \mathbb{R}^n$ and all $F \in \mathbb{R}^{p \times q}$ such that $F'F \leq I$. We now define a set $N \subset \mathbb{R}^n$ as follows (see also [3]):

$$N := \{x \in \mathbb{R}^n: (B + DFE)'Px = 0 \text{ for some matrix } F: F'F \leq I\}.$$

It follows from (3.3) that

$$x'[A'P + PA]x < 0$$

for all nonzero $x \in N$. We now derive an alternate expression for the set N .

CLAIM 2. $N = \{x \in \mathbb{R}^n: x'PBE'(EE')^{-2}EB'Px - x'PDD'Px \leq 0; \Phi B'Px = 0\}$.

In order to establish this claim, we first let $x \in N$ be given. It follows that there exists a matrix F such that $F'F \leq I$ and

$$(3.4) \quad B'Px = -E'F'D'Px.$$

Hence, $EB'Px = -EE'F'D'Px$ and therefore

$$\begin{aligned} x'PBE'(EE')^{-2}EB'Px &= x'PDFF'D'Px \\ &\leq x'PDD'Px. \end{aligned}$$

Furthermore, it follows from (3.4) that

$$\Phi B'Px = -\Phi E'F'Dx = 0.$$

Thus $x \in \{x \in \mathbb{R}^n : x'PBE'(EE')^{-2}EB'Px - x'PDD'Px \leq 0; \Phi B'Px = 0\}$.

Conversely, let $x \in \mathbb{R}^n$ be given such that

$$(3.5) \quad \Phi B'Px = 0$$

and

$$(3.6) \quad x'PBE'(EE')^{-2}EB'Px \leq x'PDD'Px.$$

It follows from (3.5) and the definition of Φ that $B'Px \in \{x \in \mathbb{R}^n; x = E'\eta \text{ for some } \eta \in \mathbb{R}^q\}$. Thus, there exists an $\eta \in \mathbb{R}^q$ such that $B'Px = E'\eta$. Hence, using (3.6), we conclude that

$$\begin{aligned} x'PDD'Px &\geq \eta'EE'(EE')^{-2}EE'\eta \\ &= \eta'\eta. \end{aligned}$$

That is,

$$(3.7) \quad \|\eta\| \leq \|D'Px\|.$$

We now consider two cases.

Case 1. $D'Px = 0$. In this case, it follows from (3.7) that $\eta = 0$, and hence $B'Px = 0$. Thus, given $F = 0$, it follows that $x \in N$.

Case 2. $D'Px \neq 0$. In this case, let

$$F := \frac{D'Px\eta'}{x'PDD'Px}.$$

Since this matrix is of rank 1, it follows that the induced norm of F is

$$\|F\| = \frac{\|D'Px\| \|\eta\|}{\|D'Px\|^2} = \frac{\|\eta\|}{\|D'Px\|} \leq 1,$$

that is, $F'F \leq I$. Furthermore

$$\begin{aligned} (B + DFE)'Px &= B'Px - \frac{E'\eta x'PDD'Px}{x'PDD'Px} \\ &= B'Px - E'\eta \\ &= 0. \end{aligned}$$

Hence, $x \in N$. This completes the proof of the claim.

Using the above claim, we now conclude that

$$x'[A'P + PA]x < 0$$

for all nonzero $x \in \mathbb{R}^n$ such that $x'PBE'(EE')^{-2}EB'Px - x'PDD'Px \leq 0$ and $\Phi B'Px = 0$. It now follows from Lemma 3.1 that there exist constants $\sigma > 0$ and $\lambda > 0$ such that

$$A'P + PA - \sigma PBE'(EE')^{-2}EB'P + \sigma PDD'P - \lambda PB\Phi'\Phi B'P < 0.$$

We now define \tilde{Q} to be the positive definite matrix

$$\tilde{Q} := \lambda[-A'P - PA + \sigma PBE'(EE')^{-2}EB'P - \sigma PDD'P + \lambda PB\Phi'\Phi B'P].$$

Letting $\varepsilon = \sigma/\lambda$, it follows that the positive-definite matrix $\tilde{P} := \sigma P$ satisfies the Riccati equation

$$A'\tilde{P} + \tilde{P}A + \tilde{P}B E'(EE')^{-2} E B'\tilde{P} + \tilde{P}D D'\tilde{P} - \frac{1}{\varepsilon} \tilde{P}B \Phi' \Phi B'\tilde{P} + \varepsilon \tilde{Q} = 0.$$

Thus, Condition 1 is satisfied. This completes the proof of the theorem. \square

Nonlinear versus linear control. An interesting corollary to the theorem above concerns the utility of nonlinear control in the stabilization of uncertain systems of the form (Σ) . A quadratically stabilizable system (Σ) is said to be *quadratically stabilizable via linear control* if the control law $p(\cdot)$ in Definition 2.1 is linear. By this definition, Corollary 3.1 follows directly from Theorem 3.1.

COROLLARY 3.1. *If an uncertain system of the form (Σ) is quadratically stabilizable (via nonlinear control) then it will also be quadratically stabilizable via linear control.*

Remark. This corollary generalizes the main result of [7] to the case of multi-input uncertain systems. As in [7], the proof of our main result relies on the form of the uncertainty bounding set $\{F: F'F \leq 1\}$ and on the fact that uncertainty enters only into the B matrix. For general uncertain systems in which these conditions do not hold, this corollary is not true (see [14]).

Observation. In light of Theorems 2.1 and 3.1, we can now state a procedure for testing if the system (Σ) is quadratically stabilizable. First, the positive definite weighting matrix Q and the full rank matrix Φ (satisfying (2.2)) are chosen. It follows from Theorem 2.1 that these choices can be arbitrary. For successively smaller values of $\varepsilon > 0$, we then attempt to find a positive definite solution to the Riccati equation (2.3). In practice, the existence of such a solution would be determined by applying a standard algorithm such as that given in [15].

If a positive definite solution to (2.3) is found for some $\varepsilon > 0$, then the required stabilizing state feedback can be constructed as in (3.1). If no positive definite solution to (2.3) is found, it then follows that the system (Σ) is not quadratically stabilizable.

4. Numerical example. Consider the uncertain system

$$\dot{x}(t) = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} x(t) + \begin{bmatrix} \alpha f(t) \\ 1 \end{bmatrix} u(t), \quad |f(t)| \leq 1.$$

For any $\alpha > 0$, this system is an uncertain system of form (Σ) where the matrices A , B , D , and E are defined as follows:

$$A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad D = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}, \quad E = 1.$$

We will apply the method described above to determine the range of values of α for which this system is quadratically stabilizable.

In order to apply our method, we first choose the matrices Q and Φ . For this example, we can let $Q = I$ and $\Phi = 0$. Using these matrices, we now determine if the corresponding Riccati equation (2.3) has a positive definite solution. Indeed, for any $\alpha > 0$, we obtain the following positive definite solution to (2.3):

$$P = \begin{bmatrix} (1 - \sqrt{1 - \varepsilon \alpha^2})/\alpha^2 & 0 \\ 0 & 1 + \sqrt{1 + \varepsilon} \end{bmatrix}$$

provided $\varepsilon < 1/\alpha^2$. Thus, we conclude that this system is quadratically stabilizable for all $\alpha > 0$. Furthermore, the required feedback control law is

$$u(t) = -[0 \quad 1 + \sqrt{1 + \varepsilon}]x(t).$$

5. Conclusions. The main result of this paper is a necessary and sufficient condition for the quadratic stabilizability of uncertain systems of form (Σ) . For this class of uncertain systems, the uncertain parameters enter only into the input matrix. Thus, the results of this paper complement the results of [5] in which the uncertain parameters enter only into the system matrix.

Appendix A: Proof of Lemma 3.1.

Proof of Lemma 3.1. Let Θ be a matrix whose columns form a set of basis vectors for $\{x \in \mathbb{R}^n: Wx = 0\}$ and suppose that condition (3.2) holds. It follows that

$$(A1) \quad \xi' \Theta' M \Theta \xi < 0$$

for all nonzero ξ such that $\xi' \Theta' V \Theta \xi \leq 0$. We now consider a number of cases.

Case 1. There exists an $\alpha \geq 0$ such that $\Theta' M \Theta = \alpha \Theta' V \Theta$. In this case, suppose ξ is given such that $\xi' \Theta' V \Theta \xi \leq 0$. If ξ is nonzero, it follows from (A1) that

$$0 > \xi' \Theta' M \Theta \xi = \alpha \xi' \Theta' V \Theta \xi,$$

which leads to a contradiction. Thus, $\xi = 0$. Hence, $\xi' \Theta' V \Theta \xi > 0$ for all nonzero ξ . It now follows that there exists a $\sigma > 0$ such that $\Theta' [M - \sigma V] \Theta < 0$, i.e.,

$$x' [M - \sigma V] x < 0$$

for all nonzero x such that $Wx = 0$. We now use Finsler's Theorem (in this instance, a special case of Finsler's Theorem will suffice; see [12] and [13]) to conclude that there exists a $\lambda > 0$ such that $M - \sigma V - \lambda W' W < 0$.

Case 2. There exists an $\alpha < 0$ such that $\Theta' M \Theta = \alpha \Theta' V \Theta$. In this case, suppose ξ is given such that $\xi' \Theta' V \Theta \xi = 0$. If ξ is nonzero, it follows from (A1) that

$$0 > \xi' \Theta' M \Theta \xi = \alpha \xi' \Theta' V \Theta \xi = 0,$$

which leads to a contradiction. Thus, $\xi = 0$. Hence

$$(A2) \quad \xi' \Theta' V \Theta \xi \neq 0$$

for all nonzero ξ .

CLAIM. $\Theta' V \Theta$ is sign definite.

In order to establish this claim, we observe that the function

$$f(\xi) := \xi' \Theta' V \Theta \xi$$

is continuous. Thus, if $f(\xi)$ changes sign on the set $\{\xi: \|\xi\| = 1\}$, there exists a point ξ^* such that $\|\xi^*\| = 1$ and $f(\xi^*) = 0$. This contradicts (A2). Hence $\Theta' V \Theta$ must be sign definite.

We now consider two subcases.

Case 2(a). $\Theta' V \Theta$ is positive definite. In this case, the proof of the theorem proceeds as in Case 1 above.

Case 2(b). $\Theta' V \Theta$ is negative definite. In this case, it follows from (A1) that

$$x' M x < 0$$

for all nonzero x such that $Wx = 0$. As in Case 1, we now use Finsler's Theorem to conclude that there exists a $\lambda > 0$ such that $M - \lambda W' W < 0$. Furthermore, using a continuity argument, it follows that there exists a $\sigma > 0$ such that $M - \sigma V - \lambda W' W < 0$.

Case 3. $\Theta' M \Theta$ is not a scalar multiple of $\Theta' V \Theta$. In this case, it follows from (A1) and Finsler's Theorem (see [12]) that there exists a constant σ such that

$$(A3) \quad \Theta' M \Theta - \sigma \Theta' V \Theta < 0.$$

We now consider two subcases.

Case 3(a). $\sigma < 0$. In this case, it follows from (A3) that, given any nonzero ξ such that $\xi' \Theta' V \Theta \xi > 0$, then $\xi' \Theta' M \Theta \xi < 0$. This result, combined with (A1), shows that the matrix $\Theta' M \Theta$ is negative definite. The proof now proceeds as in Case 2(b).

Case 3(b). $\sigma > 0$. In this case, we have

$$x'(M - \sigma V)x < 0$$

for all nonzero x such that $Wx = 0$. Using Finsler's Theorem as in Case 1, it now follows that there exists a constant $\lambda > 0$ such that $M - \sigma V - \lambda W' W < 0$. This completes the proof of the lemma. \square

REFERENCES

- [1] G. LEITMANN, *Guaranteed asymptotic stability for some linear systems with bounded uncertainties*, J. Dynamical Systems Measurement Control, 101 (1979), pp. 212-216.
- [2] E. NOLDUS, *Design of robust state feedback laws*, Internat. J. Control, 35 (1982), pp. 935-944.
- [3] B. R. BARMISH, *Necessary and sufficient conditions for quadratic stabilizability of an uncertain linear system*, J. Optim. Theory Appl., 46 (1985), pp. 399-408.
- [4] I. R. PETERSEN AND C. V. HOLLOT, *A Riccati equation approach to the stabilization of uncertain linear systems*, Automatica, 22 (1986), pp. 397-411.
- [5] I. R. PETERSEN, *A stabilization algorithm for a class of uncertain linear systems*, Systems Control Lett., 8 (1987), pp. 351-357.
- [6] ———, *Quadratic stabilizability of uncertain linear systems containing both constant and time-varying uncertain parameters*, J. Optim. Theory Appl., to appear.
- [7] I. R. PETERSEN AND B. R. BARMISH, *The stabilization of uncertain single input linear systems via linear control*, in Proc. 6th International Conference on the Analysis and Optimization of Systems, Nice, 1984.
- [8] M. VIDYSAGAR, *Nonlinear Systems Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1978.
- [9] I. DERESE AND E. NOLDUS, *Design of linear feedback laws for bilinear systems*, Internat. J. Control, 31 (1980), pp. 219-237.
- [10] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 621-634.
- [11] H. K. WIMMER, *Monotonicity of maximal solutions of algebraic Riccati equations*, Systems Control Lett., 5 (1985), pp. 317-319.
- [12] F. UHLIG, *A recurring theorem about pairs of quadratic forms and extensions: a survey*, Linear Algebra Appl., 25 (1979), pp. 219-237.
- [13] D. H. JACOBSON, *Extensions of Linear-Quadratic Control, Optimization and Matrix Theory*, Academic Press, New York, 1977.
- [14] I. R. PETERSEN, *Quadratic stabilizability of uncertain linear systems: existence of a nonlinear stabilizing control does not imply existence of a linear stabilizing control*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 291-293.
- [15] P. VAN DOOREN, *A generalized eigenvalue approach for solving Riccati equations*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 121-125.
- [16] P. FINSLER, *Über das vorkommen definiten und semidefiniten Formen in scharen quadratischen Formen*, Comment. Math. Helv., 9 (1936/37), pp. 188-192.

ON THE STABILIZATION OF UNCERTAIN LINEAR SYSTEMS VIA BOUND INVARIANT LYAPUNOV FUNCTIONS*

KEMIN ZHOU† AND PRAMOD P. KHARGONEKAR‡

Abstract. This paper considers the problem of stabilizing a class of systems with an arbitrarily large uncertainty bounding set. In particular, the stabilization of linear uncertain systems via a bound-invariant Lyapunov function is examined. Necessary and sufficient conditions for the existence of a bound invariant Lyapunov function are derived for systems with block structured uncertainty and a class of systems with scalar linear uncertainty.

Key words. quadratic stabilizability, bound invariant Lyapunov function, block structured uncertainty, linear uncertainty, constrained Lyapunov problem, robust control, Lyapunov stability, feedback stabilization, state-feedback

AMS(MOS) subject classifications. 93C15, 93C60, 93D05, 93D15

1. Introduction. In this paper, we are concerned with the problem of robust stabilization of uncertain systems described by differential equations which contain time-varying uncertain parameters. One popular approach to this problem today is the so-called “quadratic stabilization” approach. It involves constructing a suitable quadratic Lyapunov function for the closed-loop system. In Barmish [2], necessary and sufficient conditions for quadratic stabilizability are given. However, these conditions are rather difficult to check in general. Some sufficient conditions which are easy to check are given by various authors, see, for example, Hollot and Barmish [6], Leitmann [7], [8], Noldus [10], Petersen [11], Petersen and Hollot [14], Thorp and Barmish [15], Willems and Willems [16], and Zhou and Khargonekar [17]. Necessary and sufficient conditions are obtained for systems with unstructured uncertainties by Petersen [12], [13] and Zhou and Khargonekar [18].

In a recent paper of Hollot [5], an interesting concept called a bound-invariant Lyapunov function (BILF) is introduced to study the class of stabilizable uncertain systems (see also Meilakhs [9]). Roughly, a system is said to be stabilizable via a BILF if the system is stabilizable via a quadratic Lyapunov function which is independent of the uncertainty bounding set. Indeed, systems satisfying the so-called “matching condition” (in, e.g., Petersen [11]) admit BILFs. Necessary and sufficient conditions are given for the stabilizability via BILF (see § 2 for more details) for systems with *rank one* linear uncertainties and these conditions are very easy to check. It is our desire to generalize this technique to more general classes of uncertain systems. We first derive necessary and sufficient conditions for the existence of BILFs for the class of systems with block structured uncertainties. We indicate that the problem can be reduced to a constrained Lyapunov problem (CLP) introduced by Galimidi and Barmish [3]. Thus, our results yield constructive techniques for obtaining stabilizing controllers. Moreover, we give a procedure to deal with systems with scalar linear uncertainties. Necessary and sufficient conditions are also obtained for stabilizability via BILFs for a subclass of systems with scalar linear uncertainties.

* Received by the editors August 31, 1987; accepted for publication (in revised form) December 17, 1987. This research was supported in part by the National Science Foundation under grant ECS-8451519, and in part by grants from Honeywell and GE Corporations.

† Center for Control Science and Dynamical Systems, University of Minnesota, Minneapolis, Minnesota 55455.

‡ Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota 55455.

2. Dynamical system, assumptions, and definitions. Consider the uncertain dynamical system described by the differential equation

$$(2.1) \quad \frac{dx}{dt} = [A + \Delta A(q(t))]x + Bu, \quad t \geq 0$$

where $x(t) \in \mathbb{R}^n$ is the state, $u(t) \in \mathbb{R}^m$ is the control, and $q(t)$ is the model uncertainty which is restricted to the prescribed bounding set \mathbb{Q} in \mathbb{R}^r . The following standard assumptions are understood:

- (A1) The matrix ΔA is a continuous matrix function of q .
- (A2) The uncertainty $q(\cdot): [0, \infty) \rightarrow \mathbb{Q}$ is Lebesgue measurable.
- (A3) The bounding set \mathbb{Q} is a compact set in \mathbb{R}^r .
- (A4) The nominal system (A, B) is stabilizable.
- (A5) $m < n$ and B has full column rank.

We will consider the stabilization of such systems by state feedback using Lyapunov stability theory. In particular, we consider the case where the Lyapunov function is a quadratic Lyapunov function. This leads to the notion of quadratic stabilizability (see Barmish [2]).

DEFINITION 2.2. System (2.1) is said to be *quadratically stabilizable (via linear control)* if there exists an $m \times n$ real constant matrix K , an $n \times n$ positive definite symmetric matrix P , and a constant $\alpha > 0$ such that, for any admissible uncertainties $q(t)$, the closed-loop system with state feedback $u(t) = Kx(t)$ and Lyapunov function $V(x) = x^T P x$ has the following property:

$$(2.3) \quad \mathcal{L}(x, t) := \frac{dV}{dt} = x^T [(A + \Delta A(q(t)))^T P + P(A + \Delta A(q(t)))]x + 2x^T P B K x \leq \alpha \|x\|^2$$

for all $x \in \mathbb{R}^n$ and $t \geq 0$.

As is well known, if the inequality above holds, it is standard that the closed-loop system is uniformly asymptotically stable at the equilibrium point $x = 0$, for any given admissible uncertainties $q(t)$.

DEFINITION 2.4. Let Θ be any $n \times (n - m)$ left unitary matrix whose columns form a set of basis vectors for the linear space

$$\mathcal{N}(B^T) := \{x \in \mathbb{R}^n: B^T x = 0\}.$$

Thus, $B^T \Theta = 0$ and $\Theta^T \Theta = I$.

The following result is from Barmish [2].

LEMMA 2.5. *The system (2.1) is quadratically stabilizable if and only if there exists a positive definite symmetric matrix S ($S := P^{-1}$) such that*

$$(2.6) \quad \Theta^T [(A + \Delta A(q))S + S(A + \Delta A(q))^T] \Theta < 0$$

for all $q \in \mathbb{Q}$. Furthermore, if the inequality above holds, then $V(x) = x^T S^{-1} x$ is a quadratic Lyapunov function for the closed-loop system and the control law can be taken as

$$K := -\gamma B^T S^{-1}$$

where $\gamma > 0$ is some sufficiently large scalar.

Most of the previous research is concerned with the construction of a positive definite symmetric matrix $S > 0$ such that (2.6) holds for a given uncertainty bounding set \mathbb{Q} . However, for those systems with uncertainties satisfying the so-called “matching

condition," we can construct a positive definite symmetric matrix S independent of the size of the bounding set \mathbb{Q} . To see that, suppose $\Delta A(q) = BM(q)$ for some continuous matrix function $M(q)$; then $\Theta^T \Delta A(q) = \Theta^T BM(q) = 0$ and (2.6) is reduced to $\Theta^T (AS + SA^T) \Theta < 0$, which is independent of the bounding set \mathbb{Q} . Hence S can be chosen independent of the uncertainty bounding set. A particular choice of S can be easily obtained using standard linear system theory. This has motivated us to ask what class of systems admits a quadratic Lyapunov function which is independent of the uncertainty bounding set. We now introduce a notion which is central to this paper.

DEFINITION 2.7. System (2.1) is said to be stabilizable via a (quadratic) bound-invariant Lyapunov function (BILF) if there exists a positive definite symmetric matrix P such that the following holds: Given any (arbitrarily large) bounding set, there exists a state feedback law $u(t) = Kx(t)$ such that

$$x^T [(A + \Delta A(q) + BK)^T P + P(A + \Delta A(q) + BK)] x < 0$$

for all nonzero $x \in \mathbb{R}^n$ and all $q \in \mathbb{Q}$.

Remark 2.8. From the definition of BILF, it is clear that, for the given system uncertainty structure, (2.1) is quadratically stabilizable via a BILF if and only if there exists a positive definite symmetric matrix S which is independent of the uncertainty bounding set such that (2.6) holds. This observation will be used in the following sections to characterize the uncertain systems which admit BILFs.

Remark 2.9. We would like to remind the readers that although there is a fairly large class of systems with only state matrix uncertainties which admit BILFs, their corresponding stabilizing feedback gains typically depend on the bounding set \mathbb{Q} (i.e., the scalar γ in Lemma 2.5 depends on the bounding set). (See Hollot [4], [5] for examples.) With this in mind, we should note that if we use the technique in Barmish [1] we may run into some difficulties in this context. Recall that Theorem 3.1 in Barmish [1] states that a system with uncertain input matrix (and possibly with state matrix uncertainties as well) is quadratically stabilizable via linear control if and only if a so-called "augmented system" is quadratically stabilizable. This new augmented system is of higher dimension; however, its input matrix is constant independent of uncertainties. Now we might consider applying the BILF technique in this paper and in Hollot [5] to this new system. If this new system admits a BILF, then it follows that the original system with input matrix uncertainties (and possibly with state matrix uncertainties) can be stabilized by a constant state feedback *independent* of the uncertainty bounding set! Clearly, this would be possible only for a limited class of uncertainties. Thus, we can conclude that in a large number of cases, the augmented system will not be stabilizable via BILF. However, it is possible that the original system with input (and state) matrix uncertainties is not only quadratically stabilizable but also is quadratically stabilizable via BILF.

3. Bound invariant Lyapunov function for block structured uncertainties. In this section, we consider BILF for block structured uncertain systems, i.e., the uncertainty ΔA can be described as

$$(3.1) \quad \Delta A = \sum_{i=1}^p D_i F_i(t) E_i^T$$

where $D_i \in \mathbb{R}^{n \times l_i}$ and $E_i \in \mathbb{R}^{n \times k_i}$ are constant known matrices, and $F_i(t) \in \mathbb{R}^{l_i \times k_i}$ are any norm-bounded time-varying Lebesgue measurable matrix functions.

The problem of finding BILF for system (2.1) with uncertainty structure (3.1) can be stated as follows: find a positive definite symmetric matrix S such that (2.6) holds for any norm-bounded $F_i(t)$.

Remark 3.2. The uncertainty structure (3.1) obviously includes rank 1 uncertainties ($k_i = l_i = 1$) as a special case; hence the result presented here can be thought of as a generalization of Holot's result for rank one uncertainties [5].

Before we present our result, let us first note the following simple fact.

LEMMA 3.3. *Let $x \in \mathbb{R}^n$, $y \in \mathbb{R}^n$ if*

$$xy^T + yx^T = 0 \quad \text{and} \quad x \neq 0;$$

then $y = 0$.

We will not present the proof of this fact here, since it is very easy. This fact will play a key role in the proof of the results in this section and in the next section.

LEMMA 3.4. *Consider system (2.1) satisfying assumptions (A1)–(A5) with the block structured uncertainty described in (3.1). Then the system is stabilizable via a BILF if and only if there exists a positive definite symmetric matrix S such that*

- (i) $\Theta^T(AS + SA^T)\Theta < 0$;
- (ii) $\Theta^T D_i = 0$, or $E_i^T S \Theta = 0$, $i = 1, 2, \dots, p$.

Proof. (Sufficiency.) Suppose there exists a positive definite symmetric matrix S such that (i) and (ii) hold for $i = 1, 2, \dots, p$. Then it is clear that

$$\Theta^T \left\{ \left[A + \sum_{i=1}^p D_i F_i(t) E_i^T \right] S + S \left[A + \sum_{i=1}^p D_i F_i(t) E_i^T \right]^T \right\} \Theta = \Theta^T (AS + SA^T) \Theta < 0$$

where $F_i(t)$ are arbitrarily bounded matrix functions. Hence using Lemma 2.5 we conclude that the system is stabilizable via a BILF.

(Necessity.) Now suppose system (2.1) with (3.1) is stabilizable via a BILF. It follows from Lemma 2.5 and the definition of BILF that there exists a positive definite symmetric matrix S such that the following holds:

$$(3.5) \quad \Theta^T \left\{ \left[A + \sum_{i=1}^p D_i F_i(t) E_i^T \right] S + S \left[A + \sum_{i=1}^p D_i F_i(t) E_i^T \right]^T \right\} \Theta < 0$$

for any arbitrarily bounded matrix functions $F_i(t)$. Let $F_i(t) = 0$, $i = 1, 2, \dots, p$; then (3.5) reduces to (i). Now we only need to show that (ii) must also be true. Suppose $\Theta^T D_i \neq 0$ for some i ; we need to show $E_i^T S \Theta = 0$. To do that, let us choose a special class of uncertainties. Let $F_i(t) = r_i(t) u_i v_i^T$ be such that $u_i \in \mathbb{R}^{l_i}$, $v_i \in \mathbb{R}^{k_i}$, and $\Theta^T D_i u_i \neq 0$, and let $r_i(t)$ be any arbitrarily bounded measurable scalar function. Then an argument such as that in Holot [5, p. 165] shows that

$$\Theta^T D_i u_i v_i^T E_i^T S \Theta + \Theta^T S E_i v_i u_i^T D_i^T \Theta = 0.$$

This implies $\Theta^T S E_i v_i = 0$ from Lemma 3.3. Since v_i is arbitrary, we get $\Theta^T S E_i = 0$. This proves the necessity. \square

Condition (ii) can be written in a more compact form. To do that, let Z_0 denote the index set

$$Z_0 = \{i: \Theta^T D_i \neq 0\}.$$

DEFINITION 3.6. Let E be any $n \times k$ matrix whose columns form a set of basis vectors for the linear space

$$\mathbb{H}_0 = \text{span} \{\text{columns of } E_i, i \in Z_0\}, \quad k = \dim \mathbb{H}_0.$$

Now condition (ii) can be written as

$$(ii)' \quad E^T S \Theta = 0.$$

Hence system (2.1) with uncertainties (3.1) is stabilizable via a BILF if and only if there exists a positive definite symmetric matrix S such that (i) and (ii)' hold. The necessary and sufficient conditions for the existence of such a matrix are first given by Hollot in his thesis [4] and a recent paper [5], but in a slightly different formulation.

We would like to point out here that the existence of a solution $S > 0$ to (i) and (ii) is equivalent to the existence of a solution to the so-called Constrained Lyapunov Problem (CLP) formulated by Galimidi and Barmish [3]. This can be seen by noticing that the columns of Θ form a set of basis vectors of $\mathbb{N}(B^T)$. Hence (ii)' can be rewritten as

$$(ii)'' \quad E^T S = \Omega B^T$$

for some matrix $\Omega \in \mathbb{R}^{k \times m}$.

The issue of the existence of an $S > 0$ and Ω such that (i) and (ii)'' hold is exactly the CLP.

Remark 3.7. It is easy to see from condition (ii)'' that it is necessary to have $\text{rank}(E) \leq \text{rank}(B)$ (i.e., $k \leq m$) for the stabilizability of the system via BILF. It is also easy to see that the existence of a BILF does *not* depend on the choice of the matrix E .

Now we can state one of the main results of this paper.

THEOREM 3.8. Consider system (2.1) with the block structured uncertainties described in (3.1). Let E be an $n \times k$ matrix given by the Definition 3.6 and if $m > k$ then let Ψ be any $m \times (m - k)$ orthonormal matrix satisfying

$$E^T B \Psi = 0, \quad \Psi^T \Psi = I$$

and define

$$A_* := \Theta^T A \Theta - \Theta^T A B B^T E (E^T B B^T E)^{-1} E^T \Theta, \quad B_* := \Theta^T A B \Psi.$$

Then the system is stabilizable via a BILF if and only if $E^T B$ is right-invertible and either of the following is true:

- (a) $k = m$ and A_* is a stability matrix;
- (b) $k < m$ and the pair (A_*, B_*) is stabilizable.

Proof. Using Lemma 3.4 we can obtain conditions (i) and (ii) (or (i) and (ii)'); then the result follows from the result of Hollot [4], [5] or Galimidi and Barmish [3]. \square

If the system is stabilizable via a BILF, then the following algorithm can be used to construct a positive definite symmetric matrix S satisfying conditions (i) and (ii) in Lemma 3.4.

ALGORITHM FOR COMPUTING S .

- (1) Choose K_* such that

$$\hat{A} = A_* + B_* K_*$$

is asymptotically stable. (Take $K_* = 0$ if A_* is stable.)

- (2) Choose any positive definite symmetric matrix F such that

$$\hat{A}F + F\hat{A}^T < 0.$$

- (3) Define G as

$$G := [\Psi K_* - B^T E (E^T B B^T E)^{-1} E^T \Theta] F.$$

- (4) Choose any positive definite symmetric matrix H such that

$$\Sigma := \begin{bmatrix} F & G^T \\ G & H \end{bmatrix}$$

is positive definite. For example, $H = \delta I$ such that

$$\delta > \lambda_{\max}(G^T G) / \lambda_{\min}(F).$$

(5) A solution of S is now given by

$$S = [\Theta \quad B] \Sigma \begin{bmatrix} \Theta^T \\ B^T \end{bmatrix}.$$

4. Systems with scalar linear uncertainties. We consider system (2.1) with scalar linear uncertainties in this section. We show that some conditions similar to the ones in the previous section give the necessary and sufficient conditions for quadratic stabilizability via BILF for a class of high-rank scalar linear uncertainties.

To be specific, let us assume that the system uncertainty ΔA is described as follows:

$$(4.1) \quad \Delta A = \sum_{i=1}^p q_i(t) A_i,$$

where A_i are constant matrices (not necessarily rank one), and $q_i(t)$ are in a compact set.

Define the following index sets:

$$\mathbb{Z}_1 = \{i: \text{rank}(\Theta^T A_i) = 1\}, \quad \mathbb{Z}_2 = \{i: \text{rank}(\Theta^T A_i) > 1\}.$$

Clearly, $\text{rank}(\Theta^T A_i) = 1$ does not mean $\text{rank}(A_i) = 1$, for example,

$$A_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Theta = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Let d_i and e_i be any vectors such that

$$\Theta^T A_i = d_i e_i^T \quad \text{for } i \in \mathbb{Z}_1.$$

Let D_i and E_i be any matrices such that

$$\text{rank}(D_i) = \text{rank}(E_i) = \text{rank}(\Theta^T A_i) \quad \text{and} \quad \Theta^T A_i = D_i E_i^T \quad \text{for } i \in \mathbb{Z}_2.$$

The following two linear vector spaces are subspaces of \mathbb{R}^n :

$$\mathbb{H}_1 = \text{span}\{e_i, i \in \mathbb{Z}_1\}, \quad \mathbb{H}_2 = \text{span}\{\text{columns of } E_i, i \in \mathbb{Z}_2\}.$$

Remark 4.2. The decomposition of $\Theta^T A_i$ as $d_i e_i^T$ or $D_i E_i^T$ is not unique. However, it is clear that the linear spaces \mathbb{H}_1 and \mathbb{H}_2 do not depend on the specific decompositions.

Let \mathbb{H} denote the linear space spanned by \mathbb{H}_1 and \mathbb{H}_2 , i.e., \mathbb{H} is the smallest linear vector space containing \mathbb{H}_1 and \mathbb{H}_2 . Note that a set of basis of \mathbb{H} can be chosen from the vectors e_i , $i \in \mathbb{Z}_1$, and the columns of E_i , $i \in \mathbb{Z}_2$.

DEFINITION 4.3. Let E be any $n \times k$ matrix whose columns form a set of basis vectors for the linear space \mathbb{H} and $k = \dim(\mathbb{H})$.

LEMMA 4.4. System (2.1) with scalar linear uncertainties (4.1) is stabilizable via a BILF if there exists a positive definite symmetric matrix S such that

- (i) $\Theta^T (AS + SA^T) \Theta < 0$;
- (ii) $E^T S \Theta = 0$.

Furthermore if $\mathbb{H}_1 = \mathbb{H}$ (i.e., $\mathbb{H}_1 \supseteq \mathbb{H}_2$), then (i) and (ii) are also necessary for the system to be stabilizable via BILF.

Remark 4.5. The second part of this lemma can also be stated as follows: if all the columns of E can be chosen from e_i , $i \in \mathbb{Z}_1$, then (i) and (ii) are necessary and sufficient for the system to be stabilizable via a BILF.

Proof of Lemma 4.4. Recall (see Hollot [4], [5]) that system (2.1) with scalar linear uncertainties (4.1) is stabilizable via a BILF if and only if there exists a positive definite symmetric matrix S such that

$$\Theta^T(AS + SA^T)\Theta < 0 \quad \text{and} \quad \Theta^T(A_iS + SA_i^T)\Theta = 0, \quad i = 1, 2, \dots, p.$$

(Sufficiency.) We consider three different cases:

(a) $i \notin \mathbb{Z}_1 \cup \mathbb{Z}_2$: from the definitions of \mathbb{Z}_1 and \mathbb{Z}_2 , we know $\Theta^T A_i = 0$.

(b) $i \in \mathbb{Z}_1$: from the definition of E , there exists a $\xi_i \in \mathbb{R}^k$ such that $e_i = E\xi_i$. Hence

$$\Theta^T(A_iS + SA_i^T)\Theta = d_i e_i^T S \Theta + \Theta^T S e_i d_i^T = d_i \xi_i^T E^T S \Theta + \Theta^T S E \xi_i d_i^T = 0.$$

(c) $i \in \mathbb{Z}_2$: similar to (b), there exists a $\Pi_i \in \mathbb{R}^{k \times \text{rank}(E_i)}$ such that $E_i = E\Pi_i$. Hence

$$\Theta^T(A_iS + SA_i^T)\Theta = D_i E_i^T S \Theta + \Theta^T S E_i D_i^T = D_i \Pi_i^T E^T S \Theta + \Theta^T S E \Pi_i D_i^T = 0.$$

Cases (a), (b), and (c) together show that $\Theta^T(A_iS + SA_i^T)\Theta = 0$ for all $i = 1, 2, \dots, p$. This proves the sufficiency.

Now assume $\mathbb{H}_1 = \mathbb{H}$; we show that (i) and (ii) are also necessary. Since $\Theta^T(A_iS + SA_i^T)\Theta = 0$ for all $i = 1, 2, \dots, p$, it follows that

$$\Theta^T(A_iS + SA_i^T)\Theta = 0, \quad i \in \mathbb{Z}_1$$

or

$$d_i e_i^T S \Theta + \Theta^T S e_i d_i^T = 0, \quad i \in \mathbb{Z}_1.$$

Since $d_i \neq 0$, $i \in \mathbb{Z}_1$, using Lemma 3.3 we conclude that

$$e_i^T S \Theta = 0, \quad i \in \mathbb{Z}_1.$$

Clearly $E^T S \Theta = 0$, because the columns of E are a set of basis vectors of $\mathbb{H}_1 (= \mathbb{H})$. \square

Now we summarize our results in the following theorem. (The proof is exactly the same as for Theorem 3.8.)

THEOREM 4.6. *Consider system (2.1) with the scalar linear uncertainties described in (4.1). Then the system is stabilizable via a BILF if (and only if, in case $\mathbb{H}_1 = \mathbb{H}$, i.e., $\mathbb{H}_1 \supseteq \mathbb{H}_2$), $E^T B$ is right invertible, and either of the following is true:*

(a) $k = m$ and A_* is a stability matrix;

(b) $k < m$ and the pair (A_*, B_*) is stabilizable,

where A_* and B_* are defined exactly as in Theorem 3.8.

Example 4.7. Consider system (2.1) with uncertainty (4.1); here

$$A_1 = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then

$$\Theta = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \Theta^T A_1 = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\Theta^T A_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \Theta^T A_3 = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$\text{rank}(\Theta^T A_1) = \text{rank}(\Theta^T A_2) = 1$, and $\text{rank}(\Theta^T A_3) = 2$,

$$\mathbb{H}_1 = \text{span} \left\{ \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\} \supseteq \mathbb{H}_2 = \text{span} \left\{ \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\},$$

$$E = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad E^T B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Since $\mathbb{H}_1 \supseteq \mathbb{H}_2$, the conditions given in Theorem 4.6 are necessary and sufficient for such a system to be stabilizable via a BILF.

For illustration, let us take

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{bmatrix};$$

then we can compute

$$A_* = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix},$$

which is stable. Since $m = k = 2$, we claim that the system is stabilizable via a BILF.

To obtain a positive definite symmetric matrix S satisfying Lemma 4.4, we use the algorithm in § 3. Since A_* is stable, we can let $K_* = 0$. Hence $\hat{A} = A_*$. An F can be obtained by solving $\hat{A}F + F\hat{A}^T = -I$. Hence

$$F = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \quad \text{and} \quad G = \begin{bmatrix} \frac{1}{2} & -1 \\ 0 & 0 \end{bmatrix}.$$

Now choose $\delta = 2 > \lambda_{\max}(G^T G) / \lambda_{\min}(F) = 1.82$ and $H = \delta I$. Then

$$\Sigma := \begin{bmatrix} F & G^T \\ G & H \end{bmatrix}.$$

A solution of S can now be obtained:

$$S = [\Theta B] \Sigma \begin{bmatrix} \Theta^T \\ B^T \end{bmatrix} = \begin{bmatrix} \frac{3}{2} & -\frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{2} & 1 & -1 & 0 \\ \frac{1}{2} & -1 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

REFERENCES

- [1] B. R. BARMISH, *Stabilization of uncertain systems via linear control*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 848-850.
- [2] ———, *Necessary and sufficient conditions for quadratic stabilizability of an uncertain linear svstems*, J. Optim. Theory Appl., 46 (1985), pp. 399-408.
- [3] A. R. GALIMIDI AND B. R. BARMISH, *The constrained Lyapunov problem and its application to robust output feedback stabilization*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 410-419.
- [4] C. V. HOLLOT, *Construction of quadratic Lyapunov functions for a class of uncertain linear systems*, Ph.D. thesis, University of Rochester, Rochester, NY, 1984.

- [5] C. V. HOLLOT, *Bound invariant Lyapunov functions: a means for enlarging the class of stabilizable uncertain systems*, Internat. J. Control, 46 (1987), pp. 161-184.
- [6] C. V. HOLLOT AND B. R. BARMISH, *Optimal quadratic stabilizability of uncertain linear systems*, in Proc. 18th Allerton Conference on Communication, Control and Computation, University of Illinois, Monticello, IL, 1980.
- [7] G. LEITMANN, *On the efficacy of nonlinear control in uncertain linear systems*, J. Dynamic Systems Measurement Control, 102 (1981), pp. 95-102.
- [8] ———, *Guaranteed asymptotic stability for some linear systems with bounded uncertainties*, J. Dynamic Systems Measurement Control, 101 (1979), pp. 212-216.
- [9] A. M. MEILAKHS, *Design of stable control systems subject to parametric perturbations*, Automat. i Telemekh., 10 (1978), pp. 5-16.
- [10] E. NOLDUS, *Design of robust state feedback laws*, Internat. J. Control, 35 (1982), pp. 935-944.
- [11] I. R. PETERSEN, *Structural stabilization of uncertain systems: necessity of the matching conditions*, SIAM J. Control Optim., 23 (1985), pp. 286-296.
- [12] ———, *A stabilization algorithm for a class of uncertain linear systems*, Systems Control Lett., 8 (1987), pp. 351-357.
- [13] ———, *Stabilization of an uncertain linear system in which uncertain parameters enter into the input matrix*, 1987, to appear.
- [14] I. R. PETERSEN AND C. V. HOLLOT, *A Riccati equation approach to the stabilization of uncertain linear systems*, Automatica, 22 (1986), pp. 397-411.
- [15] J. S. THORP AND B. R. BARMISH, *On guaranteed stability of uncertain linear systems via linear control*, J. Optim. Theory Appl., 35 (1981), pp. 559-579.
- [16] J. L. WILLEMS AND J. C. WILLEMS, *Robust stabilization of uncertain systems*, SIAM J. Control Optim., 21 (1983), pp. 352-375.
- [17] K. ZHOU AND P. P. KHARGONEKAR, *Stabilization of uncertain systems via linear feedback control*, submitted.
- [18] ———, *Robust stabilization of linear systems with norm bounded time varying uncertainty*, Systems Control Lett., to appear.

SMOOTH OPTIMIZATION METHODS FOR MINIMAX PROBLEMS*

R. A. POLYAK†

Abstract. The classical discrete minimax problem is considered. It is transformed into an equivalent problem by a monotone transformation of the initial functions. It was found that the classical Lagrangian of the equivalent problem has a number of important properties both in primal and dual spaces in convex as well as in nonconvex cases.

In particular, the classical Lagrangian of the equivalent problem, being as smooth as the initial functions, has the main advantages of augmented Lagrangians. This makes it possible to construct a multiplier method for the minimax problem and a general method for the simultaneous solution of the primal and the dual problems.

These methods are based on the theory of methods of smooth optimization and preserve the main advantages of the latter for nonsmooth minimax problems.

Key words. minimax problem, monotone transformation, multiplier method, smooth optimization, dual problems

AMS(MOS) subject classification. 90C20

1. Introduction. A growing interest in problems of nonsmooth optimization and, in particular, in minimax problems is due to the special role these problems have in modern optimization theory (see, for example, [11], [24], [25]). Of the large number of papers that have appeared recently on this subject (see [10], [15], [16], [25], [26], [35]) the research summarized in [35] plays an important role. In it methods of generalized gradient and their variants are studied and developed. However, the rate of convergence of these methods is not high. Even in the case of ellipsoid methods ([18], [23], [35]) that have polynomial complexity applied to linear minimax problems (see [10] for a proof), the rate of convergence is estimated (see [35]) by the ratio $q_n = 1 - (2n^2)^{-1}$ (n is the space dimension), i.e., $q_n \rightarrow 1$ as $n \rightarrow \infty$. In these methods the properties of convexity and smoothness of functions that appear in the minimax problem cannot essentially be used for acceleration of convergence since the gradient of $F(x) = \max \{f_i(x) | i = \overline{1, m}\}$ is not smooth, not to mention the absence of higher-order smoothness even when $f_i(x)$, $i = \overline{1, m}$ are sufficiently smooth.

Therefore our goal is to construct methods for solving minimax problems that will preserve the convergence rate of smooth optimization methods (assuming some properties of smoothness and convexity of $f_i(x)$, $i = \overline{1, m}$) without substantially increasing the number of computations needed at each step.

This is achieved by application of a monotone transformation to initial functions and subsequent use of the classical Lagrangian of the equivalent problem.

In this paper it is established that the classical Lagrangian of the equivalent problem shares all the advantages of augmented Lagrangians (see [7], [19], [29], [34]) in both convex and nonconvex cases. Moreover, it has the same order of smoothness as $f_i(x)$.

It allows us to use all means of smooth optimization techniques for the solution of minimax problems, including methods of Newtonian and quasi-Newtonian type.

* Received by the editors August 7, 1985; accepted for publication (in revised form) December 22, 1987.

† c/o Dr. Claude Lemarechal, Institut Nationale de Recherche en Informatique et Automatique, Domaine de Voluceau, Rocquencourt, Le Chesnay, France.

In addition, smoothness properties of the Lagrangian on the initial space give a deeper insight into convexity and smoothness properties of the dual function that can be used for construction of a method for simultaneous solution of the primal and the dual problem. The rate of convergence of the dual problem is defined by the product of the primal and dual spaces.

The main results of the paper were obtained in 1980 and 1981 and announced in [30] and [31]. In 1983 we learned from [3] about the possibility of developing the multiplier method for the minimax problem using the function

$$Q_c(g(x), \mu) = c^{-1} \log (\sum \mu_i \exp (cg_i(x))).$$

However, our main results are not presented in [3], or in the subsequent papers [4]–[6].

It seems impossible to obtain these results by a direct reformulation of the minimax problem as a constrained optimization problem, and still preserve the initial condition.

Besides, the corresponding class of augmented Lagrangians \hat{P}_1 , as was mentioned in ([3, p. 309]), had been insufficiently investigated.

The class of functions \hat{P}_1 was first carefully studied in convex as well as nonconvex cases in [33]. In particular, using a monotone transformation and ordinary Lagrangian for the equivalent problem it was possible to essentially improve the barrier and the center methods.

Some other approaches to solving discrete minimax problems based on the replacement of $F(x)$ by sequences of smooth functions are considered in [2], [8], [16], [32].

2. Monotone transformation of the minimax problem. In this section we introduce a monotone transformation of $f_i(x)$, $i = \overline{1, m}$ and study properties of the classical Lagrangian of the equivalent problem.

We suppose that $f_i(x) \in C^1$, $i = \overline{1, m} : R^n \rightarrow R^1$ and consider the following problem:

$$(1) \quad x^* \in \operatorname{Argmin} \{F(x) | x \in R^n\}.$$

The existence of x^* is guaranteed by, for example, the following condition:

$$(2) \quad \text{There exists } c > 0 \text{ such that } \Omega = \{x : F(x) \leq c\} \text{ is compact.}$$

Let $\Psi(t) : R^1 \rightarrow R^1$ be a strictly convex and increasing function. Then functions $\tilde{f}_i(x, k) = k^{-1} \Psi(kf_i(x))$, where $k > 0$ define a minimax problem equivalent to the original one, i.e.,

$$x^* \in \operatorname{Argmin} \left\{ \max_{1 \leq i \leq m} \tilde{f}_i(x, k) | x \in R^n \right\}.$$

For $k > 0$, consider on $R \times S_m$, $S_m = \{u : \sum u_i = 1, u_i \geq 0, i = \overline{1, m}\}$ the ordinary Lagrangian function

$$A(x, u, k) = k^{-1} \sum_{i=1}^m u_i \Psi(kf_i(x)).$$

It appears that $A(x, u, k)$ has a number of advantages over the Lagrangian $L(x, u) = \sum_{i=1}^m u_i g_i(x)$ of the initial problem. For definiteness we shall take $\psi(t) = \exp t$, so that

$$A(x, u, k) = k^{-1} \sum_{i=1}^m u_i \exp (kf_i(x)).$$

We observe that $M(x, k) = \sum_{i=1}^m \exp (kf_i(x))$ was already considered by Motzkin in 1952 [22].

The results that follow will show that the relation between $A(x, u, k)$ and $M(x, k)$ is the same as between augmented Lagrangians and penalty functions, despite the fact that $A(x, u, k)$ is an ordinary Lagrangian for an equivalent problem.

In what follows we suppose that there exists a Kuhn-Tucker point $z^* = (x^*, u^*)$:

$$(3) \quad \begin{aligned} L'_x(z^*) &= \sum u_i^* f'_i(x^*) = 0, \quad u_i^*(F(x^*) - f_i(x^*)) = 0, \quad i = \overline{1, m}, \\ \sum_{i=1}^m u_i^* &= 1, \quad u_i^* \geq 0. \end{aligned}$$

Let $f(x) = (f_1(x), \dots, f_m(x))$, $I^* = \{i: F^* = F(x^*) = f_i(x^*)\} = \{1, \dots, r\}$, $\bar{f}(x) = (f_1(x), \dots, f_r(x))$ be a vector of active functions, let $f'(x) = J(f(x))$ be its Jacobi matrix and let $\bar{f}'(x) = J(\bar{f}(x))$. Set $e = (1, \dots, 1) \in R^n$, $\bar{e} = (1, \dots, 1) \in R^r$. Sometimes we shall use the condition

$$(4) \quad \text{Rank}(\bar{f}'(x^*), -\bar{e}^T) = r, \quad u_i^* > 0, \quad i = \overline{1, r},$$

which together with the condition

$$(5) \quad (L''_{xx}(z^*)y, y) \geq \lambda \|y\|^2, \quad \lambda > 0 \quad \forall y: \bar{f}'(x^*)y = 0$$

is a second-order sufficient condition in the minimax problems.

Set $S(x^*, \varepsilon) = \{x: \|x - x^*\| \leq \varepsilon\}$, $S(u^*, \varepsilon) = \{u \in S_m: \|u - u^*\| \leq \varepsilon\}$, and $S(z^*, \varepsilon) = S(x^*, \varepsilon) \times S(u^*, \varepsilon)$. Sometimes we shall also use the condition

$$(6) \quad \|f''_i(x) - f''_i(y)\| \leq L\|x - y\| \quad \forall (x, y) \in S(x^*, \varepsilon) \times S(x^*, \varepsilon).$$

We shall use the following version of the well-known theorem of Debreu [1].

PROPOSITION 1. Let $A = A^T: R^n \rightarrow R^n$; $B: R^n \rightarrow R^r$,

$$(7) \quad \begin{aligned} U &= \text{diag } u_i: R^r \rightarrow R^r, \quad u = (u_1, \dots, u_r) > 0. \\ Bx = 0 &\Rightarrow (Ax, x) \geq \lambda \|x\|^2, \quad \lambda > 0. \end{aligned}$$

Then for any $\gamma < \lambda$ there exists a $k_0 > 0$ such that for $k \geq k_0$

$$(8) \quad ((A + kB^T \cup B)x, x) \geq \gamma \|x\|^2 \quad \forall x \in R^n.$$

The proof is similar to that of Debreu's Theorem.

The following theorem describes the properties of $A(x, u, k)$.

THEOREM 1. Assume that conditions (3)–(5) are satisfied and $f_i \in C^2$. Then there exist $\varepsilon > 0$, $k_0 > 0$, and a convex neighborhood $V(k_0, \varepsilon)$ of x^* (dependent on k_0 and ε) such that for $k \geq k_0$ and for all $u \in S(u^*, \varepsilon)$ the following hold:

(1) There exists $\hat{x} \equiv x_k(u) \in \text{int } V(k_0, \varepsilon)$ such that

$$A'_x(\hat{x}, u, k) = 0.$$

(2) $A(\cdot, u, k)$ is strongly convex in $V(k_0, \varepsilon)$ so that

$$\hat{x} = \underset{x \in V(k_0, \varepsilon)}{\text{argmin}} A(x, u, k).$$

(3) For \hat{x} and $\hat{u} \equiv \hat{u}_k(u) = (\hat{u}_1, \dots, \hat{u}_m): \hat{u}_i = u_i \exp(kf_i(\hat{x})) \cdot (\sum u_j \exp(kf_j(\hat{x})))^{-1}$, $i = \overline{1, m}$ the following estimates hold:

$$(9) \quad \|\hat{x} - x^*\| \leq ck^{-1}\|u - u^*\|, \quad \|\hat{u} - u^*\| \leq ck^{-1}\|u - u^*\|,$$

where c is independent of k .

(4) If $f_i(x)$, $i = \overline{1, m}$ are convex there exists $\hat{x} = \underset{x \in R^n}{\text{argmin}} \{A(x, u, k)\}$ and (2)–(3) are true for \hat{x} and \hat{u} .

Proof. For $(x, u) \in S(z^*, \varepsilon)$ and $\kappa \in R^1$ we consider the functions

$$\hat{U}_i(x, u, \kappa) = \begin{cases} u_i \exp(\cdot |\kappa|^{-1} f_i(x)) (\sum u_j \exp(|\kappa|^{-1} f_j(x)))^{-1}, & \kappa \neq 0, \\ 0, & \kappa = 0. \end{cases}$$

If $\varepsilon > 0$ is small enough, then the definition is correct, because $\sum u_j \exp(|\kappa|^{-1} f_j(x)) > 0$. The functions $\hat{U}_i(x, u, \kappa) \equiv \hat{U}_i(\cdot)$, $i > r$ are smooth on $S(z^*, \varepsilon) \times R^1$ and $\hat{U}'_{ixux}(x, u, 0) = 0$, $i > r$. In fact, there is a $\sigma > 0$: $F^* - f_i(x^*) > \sigma$, $i > r$, $u_i^* \geq \sigma$, $i \leq r$. Choose an $\varepsilon > 0$ such that $|f_i(x) - f_i(x^*)| \leq \delta$, $|u_i - u_i^*| \leq \delta$, for all $(x, u) \in S(z^*, \varepsilon)$, $\delta < \sigma(2m)^{-1}$. Then for $\kappa \neq 0$ we have

$$\begin{aligned} & \sum u_j \exp(|\kappa|^{-1} f_j(x)) \\ &= \sum_{j=1}^r u_j \exp(|\kappa|^{-1} f_j(x)) + \sum_{j=r+1}^m u_j \exp(|\kappa|^{-1} f_j(x)) \\ &\geq (\sigma - \delta)r \exp(|\kappa|^{-1}(F^* - \delta)) - \delta(m - r) \exp(|\kappa|^{-1}(F^* - \sigma + \delta)) \\ &\geq 2^{-1}(\sigma - \delta)r \exp(|\kappa|^{-1}(F^* - \delta)). \end{aligned}$$

For $i > r$ we have

$$\begin{aligned} & 0 < \exp(|\kappa|^{-1} f_i(x)) (\sum u_j \exp(|\kappa|^{-1} f_j(x)))^{-1} \\ &\leq \exp(|\kappa|^{-1}(F^* - \sigma + \delta)) [2^{-1}(\sigma - \delta)r \exp(|\kappa|^{-1}(F^* - \delta))]^{-1} \\ &= 2(\sigma - \delta)^{-1} r^{-1} \exp(-|\kappa|^{-1}(\sigma - 2\delta)). \end{aligned}$$

Therefore there exists $c > 0$ independent of κ and such that $\hat{U}_i(x, u, \kappa) \leq c \exp(-c|\kappa|^{-1})$, $\kappa \neq 0$, so $\hat{U}_i(\cdot)$ is continuous in $S(z^*, \varepsilon) \times (-\varepsilon, \varepsilon)$, $i > r$. Further,

$$\begin{aligned} \hat{U}'_{ix}(\cdot) &= |\kappa|^{-1} \hat{U}_i(\cdot) (f'_i(x) - \sum \hat{U}_j(\cdot) f'_j(x)), \\ \hat{U}'_{iu_i}(\cdot) &= (1 - \hat{U}_i(\cdot)) \exp(|\kappa|^{-1} f_i(x)) (\sum u_j \exp(|\kappa|^{-1} f_j(x)))^{-1}, \\ \hat{U}'_{iu_j}(\cdot) &= -\hat{U}_i(\cdot) \exp(|\kappa|^{-1} f_j(x)) (\sum u_j \exp(|\kappa|^{-1} f_j(x)))^{-1}, \quad j \neq i, \\ \hat{U}'_{ix}(\cdot) &= -(\kappa|\kappa|)^{-1} \hat{U}_i(\cdot) (f_i(x) - \sum \hat{U}_j(\cdot) f_j(x)). \end{aligned}$$

Therefore, taking into account the estimate for $\hat{U}_i(\cdot)$, we obtain for the full derivative $\hat{U}'_{ixux} = (\hat{U}'_{ix}(\cdot); \hat{U}'_{iu}(\cdot); \hat{U}'_{ix}(\cdot))$ that there is a $c > 0$ independent of κ such that $\|\hat{U}'_i(\cdot)\| \leq c|\kappa|^{-2} \exp(-c|\kappa|)$. Therefore $\hat{U}'_i(x, u, 0)$ exists and $\hat{U}'_i(x, u, 0) = 0$, so $U'_i(x, u, \kappa)$ is continuous. Set $\rho(x, u, \kappa) = \sum_{j=r+1}^m \hat{U}_j(\cdot) f'_j(x)$; then $\rho(x, u, \kappa)$ is smooth on $S(z^*, \varepsilon) \times R$ and $\rho(x, u, 0) = 0$, $\rho'(x, u, 0) = 0$. Let $\bar{u}^* = (u_1^*, \dots, u_r^*)$, and $S(\bar{u}^*, \varepsilon) = \{u = (u_1, \dots, u_r): \|u - \bar{u}^*\| \leq \varepsilon\}$. On $S(x^*, \varepsilon) \times S(\bar{u}^*, \varepsilon) \times (-\varepsilon, \varepsilon) \times S(u^*, \varepsilon) \times (-\varepsilon, \varepsilon)$ we consider the map $\Phi(x, \hat{u}, t, u, \kappa): R^{n+r+m+2} \rightarrow R^{n+r+1}$ defined by $\Phi(x, \hat{u}, t, u, \kappa) = (\sum_{i=1}^r \hat{u}_i f'_i(x) + \rho(x, u, \kappa); f_i(x) - F^* + t - \kappa \ln \hat{u}_i u_i^{-1}, i = \overline{1, r}; \sum_{i=1}^r \hat{u}_i + \sum_{i=r+1}^m \hat{U}_j(x, u, \kappa) - 1)$. Since $f_i(x^*) = F^*$, $i = \overline{1, r}$, $\sum u_i^* = 1$, $\sum_{i=1}^r u_i^* f'_i(x^*) = 0$, $\rho(x^*, u^*, 0) = 0$, $\sum_{j=r+1}^m \hat{U}_j(x^*, u^*, 0) = 0$ we have $\Phi(x^*, \bar{u}^*, 0, u^*, 0) = 0$. Then setting $\bar{f}'(x^*) = \bar{f}'$, $L''_{xx}(z^*) = L''_{xx}$ we get

$$\Phi'_{x\hat{u}t} = \Phi'_{x\hat{u}t}(x^*, \bar{u}^*, 0, u^*, 0) = \begin{pmatrix} L''_{xx} & \bar{f}'^T & 0 \\ \bar{f}' & 0 & \bar{e}^T \\ 0 & \bar{e} & 0 \end{pmatrix}.$$

The matrix $\Phi'_{x\hat{u}t}$ is nonsingular. Indeed, set $w = (y, v, \tau)$, $y \in R^n$, $v \in R^r$, $\tau \in R$. Then $\Phi'_{x\hat{u}t} w = 0$ implies $L''_{xx} y + \bar{f}'^T v = 0$, $\bar{f}' y + \tau \bar{e} = 0$, $(\bar{e}, v) = 0$. Taking the inner product of the second equality with \bar{u}^* and taking into account the Kuhn-Tucker relations, we obtain $(\sum_{i=1}^r u_i^* f'_i(x^*), y) + \tau = 0$, so $\tau = 0$. It implies $\bar{f}' y = 0$. Taking the inner product

at the first equality and y we obtain $(L''_{xx}y, y) + (\bar{f}'y, v) = 0$, so $y = 0$ by (5) and $\bar{f}'^T v = 0$, $(v, \bar{e}) = 0$. The last equalities together with (4) give $v = 0$. Since $\Phi'_{x\hat{u}t} w = 0$ implies $w = 0$ the matrix $\Phi'_{x\hat{u}t}$ is nonsingular in a neighborhood of $(x^*, \bar{u}^*, 0, u^*, 0)$. Since $f_i(x) \in C^2$, the implicit function theorem (see [20]) suggests that in this neighborhood there exists a unique smooth vector function $(x(u, \kappa), \hat{U}(u, \kappa), t(u, \kappa))$:

$$\Phi(x(u, \kappa), \hat{U}(u, \kappa), t(u, \kappa), u, \kappa) \equiv \Phi(\cdot) \equiv 0$$

such that

$$\sum_{i=1}^m \hat{u}_j(x(u, \kappa), u, \kappa) f'_j(x(u, \kappa)) = 0, \\ \hat{u}_i(x(u, \kappa), u, \kappa) = u_i \exp(|\kappa|^{-1} f_i(x(u, \kappa))) (\sum u_j \exp(|\kappa|^{-1} f_j(x(u, \kappa))))^{-1}, \quad i = \overline{1, m}.$$

For $\kappa \neq 0$ equality $\Phi(\cdot) = 0$ is equivalent to $A'_x(x, u, k) = 0$, $k = \kappa^{-1}$. Differentiating the map $\Phi(\cdot)$ with respect to u , we obtain $\Phi'_{x\hat{u}t}(\cdot) \times w(\cdot) + \Phi'_u(\cdot) = 0$, where $w(\cdot) = (x'_u(u, \kappa), u'_u(u, \kappa), t'_u(u, \kappa))$, i.e., $w(\cdot) = -\Phi'^{-1}_{x\hat{u}t}(\cdot) \Phi'_u(\cdot)$. Since $\Phi'_{x\hat{u}t}$ is a nonsingular matrix and $f_i(x) \in C^2$, there are $\varepsilon > 0$ and $c_1 > 0$, $c_2 > 0$ that are independent of κ and such that the matrix $\Phi'_{x\hat{u}t}(\cdot)$ is nonsingular in $S(u^*, \varepsilon) \times (-\varepsilon, \varepsilon)$ and $\|\Phi'^{-1}_{x\hat{u}t}(\cdot)\| \leq c_1$, $\|\Phi'_u(\cdot)\| \leq c_2 \kappa$. Therefore there exists $c > 0$ not dependent on κ and such that $\|w(\cdot)\| \leq c\kappa$. Since $\Phi(x^*, \bar{u}^*, 0, u^*, \kappa) = 0$ for all κ ($|\kappa| < \varepsilon$) we have $x(u^*, \kappa) = x^*$, $\hat{u}(u^*, \kappa) = u^*$, so

$$\|x(u, \kappa) - x^*\| \leq c\kappa \|u - u^*\|, \quad \|\hat{u}(u, \kappa) - u^*\| \leq c\kappa \|u - u^*\|.$$

Setting $k = \kappa^{-1} > 0$; $\hat{x} \equiv x_k(u) \equiv x(u, \kappa)$; and $\hat{u} \equiv \hat{u}_k(u) = (\hat{u}(u, \kappa), \hat{u}_i(x, u, \kappa), i = \overline{r+1, m})$, we obtain the estimates (9).

Finally, the strong convexity of $A(x, u, k)$ in κ in the neighborhood of $\hat{x} = x(u, \kappa)$ follows from $f_i(x) \in C^2$ and the relation

$$A''_{xx}(\hat{x}, u, k) = \sum u_i \exp(k f_i(\hat{x})) f''_i(\hat{x}) + k \sum u_i \exp(k f_i(\hat{x})) f'_i{}^T(\hat{x}) f'_i(\hat{x}) \\ = (\sum u_i \exp(k f_i(\hat{x})) (L''_{xx}(\hat{z}) + k \sum \hat{u}_i f'_i{}^T(\hat{x}) f'_i(\hat{x})))$$

if we take into account Proposition 1 and estimate (9) for $k > 0$ sufficiently large. Due to strong convexity $A(x, u, k)$ in x , the necessary condition $A'_x(\hat{x}, u, k) = 0$ is sufficient for \hat{x} to be a minimizer. Observe that we have not assumed that $f_i(x)$ are convex. If so, the condition $A'_x(\hat{x}, u, k) = 0$, $u \geq 0$, $k > 0$, together with the positive definiteness of matrix $A''_{xx}(\hat{x}, u, k)$, gives (4), and the proof of Theorem 1 is complete. \square

The local results (1)–(3) of Theorem 1 are valid under weaker conditions than (4)–(5) despite the fact that these are the standard second-order sufficient condition for the minimax problem. Consider an example.

Put $I^* = \{1, 2, 3, 4\}$; $f_1(x_1, x_2) = f_1(\cdot) = (x_1 - 1)^2 + x_2^2$; $f_2(\cdot) = (x_1 + 1)^2 + x_2^2$; $f_3(\cdot) = x_1^2 + (x_2 - 1)^2$; $f_4(\cdot) = x_1^2 + (x_2 + 1)^2$. In this case condition (4) is not satisfied and the set $Y: f'(x^*)y = 0$ consists of a single point $y = 0$, which makes (5) meaningless. However, (1)–(3) of Theorem 1 remain true if we take $I_1 = \{1, 2\}$ or $I_2 = \{3, 4\}$ as I^* .

In the first case, conditions similar to (4), (5) are satisfied for $z^* = (x^*; u^*) = (0; 0; \frac{1}{2}; \frac{1}{2}; 0; 0)$, and in the second case they are satisfied for $z^* = (x^*; u^*) = (0; 0; 0; 0; \frac{1}{2}; \frac{1}{2})$.

Moreover, the results of Theorem 1 remain true if we replace the convex functions $f_1(\cdot)$ and $f_2(\cdot)$ by the nonconvex functions $\tilde{f}_1(\cdot) = -(x_1 - 1)^2 + x_2^2$; $\tilde{f}_2(\cdot) = -(x_1 + 1)^2 + x_2^2$. This shows that (1)–(3) of Theorem 1 hold not only without convexity of $f_i(x)$, $i = \overline{1, r}$, but with conditions similar to (4)–(5) satisfied for any minimal set $I \subset I^*$.

The set $I \subset I^*$ is called minimal if

$$\min \left\{ \left\| \sum_{i \in I} u_i f'_i(x^*) \right\| \mid \sum_{i \in I} u_i = 1, u_i \geq 0, i \in I \right\} = \left\| \sum_{i \in I} u_i^* f'_i(x^*) \right\| = 0$$

and

$$\min \left\{ \left\| \sum_{i \in I \setminus j} u_i f'_i(x^*) \right\| \left\| \sum_{i \in I \setminus j} u_i = 1, u_i \geq 0, i \in I \setminus j \right\} \right\} > 0$$

for all $j \in I$.

It is easy to show that there is a one-to-one correspondence between minimal sets and vertices of the Kuhn-Tucker polyhedron $Q(x^*) = \{u \in S_r: \sum_{i=1}^r u_i f'_i(x^*) = 0\}$. Moreover, for any minimal set I , the vectors $(f'_i(x^*), -1)$ are linearly independent and $u_i^* > 0, i \in I$. Thus, condition (4) is always satisfied for the minimal set. Therefore (1)–(3) of Theorem 1 remain true if, instead of (4), (5), we assume that

$$(5') \quad (L''_{xx}(z^*)y, y) \geq \lambda \|y\|^2, \quad \lambda > 0 \quad \forall y: f'_i(x^*)y = 0, \quad i \in I,$$

where u^* are vertices of the Kuhn-Tucker polyhedron, which correspond to I .

3. Multiplier method. Theorem 1 allows us to realize the method of multipliers:

$$(10) \quad \begin{aligned} x^{S+1} &= \operatorname{argmin}_x \exp(-kF(x^S))A(x, u^S, k), \\ u^{S+1} &= (u_i^{S+1} = u_i^S \exp(kf_i(x^{S+1}))(\sum u_j^S \exp(kf_j(x^{S+1}))^{-1}, i = \overline{1, m}). \end{aligned}$$

Under the conditions of Theorem 1 we obtain the estimate

$$(11) \quad \|x^S - x^*\| \leq ck^{-S}, \quad \|u^S - u^*\| \leq ck^{-S}$$

with $c > 0$ independent of k .

In order to realize the method (10), if $f_i(x)$ are convex it is enough to know $u^0 \in S(u^*, \varepsilon)$; if $f_i(x)$ are nonconvex we must know $z^0 \in S(z^*, \varepsilon)$.

The next lemmas allow us to obtain $z^0 \in S(z^*, \varepsilon)$. The second difficulty, which we must overcome in order to realize (10), is to change the infinite procedure of smooth optimization to find x^S to a finite one and preserve the estimates above.

Let $\mu \geq 0, \quad U(x, u, k) = (u_i(x, u, k) = u_i \exp(kf_i(x)) \cdot (\sum u_j \exp(kf_j(x)))^{-1}, i = \overline{1, m})$.

PROPOSITION 2. *If the conditions of Theorem 1 are satisfied, there exist $k_0 > 0$ and $c > 0$ independent of k such that for all $k \geq k_0$ and $\tilde{z} = (\tilde{x}, \tilde{u})$*

$$\|\exp(-kF(x))A'_x(\tilde{x}, u, k)\| \leq \mu k^{-1} \|U(\tilde{x}, u, k) - u\|, \quad \tilde{u} = U(\tilde{x}, u, k)$$

the following estimate holds:

$$\begin{aligned} \|\tilde{x} - x^*\| &\leq c(1 + \mu)k^{-1} \|u - u^*\|, \\ \|\tilde{u} - u^*\| &\leq c(1 + \mu)k^{-1} \|u - u^*\| \quad \forall u \in S(u^*, \varepsilon). \end{aligned}$$

The proof is as in Theorem 5 of [29].

Proposition 2 allows us, in principle, to overcome the second difficulty. Indeed, if $k > 0$ is large enough and $f_i(x) \in C^2$ then it is sufficient, beginning from some step S_0 , to make only one step of the Newton or quasi-Newton method of smooth optimization (see [27], [28]) to obtain $\tilde{z}^s = (\tilde{x}^s, \tilde{u}^s)$ which satisfies (11). The next lemmas allow us to obtain $z \in S(z^*, \varepsilon)$.

4. Lemmas. Let $f_i(x) \in C, X^* = \{x: F(x) = F^*\}, d(x, x^*) = \min \{\|x - y\| | y \in X^*\}$.

If (2) is satisfied then $F^* > -\infty$. Set $N(x, k) = (M(x, k))^{1/k}; F_k^* = \ln \inf_{x \in R^n} N(x, k); X_k^* = \{x | \ln N(x, k) = F_k^*\}$.

LEMMA 1. *If (2) is satisfied then*

$$(12) \quad F^* \leq F_k^* \leq k^{-1} \ln m + F^*,$$

$$(13) \quad X_k^* \neq \emptyset, \quad \text{if } k > (C - F^*)^{-1} \ln m,$$

$$(14) \quad \lim_{k \rightarrow \infty} \lim \{d(x, X^*) | x \in X_k^*\} = 0.$$

Proof. We have

$$(15) \quad \exp(F(x)) \leq N(x, k) \leq m^{1/k} \exp(F(x)) \quad \forall x.$$

Let $x^* \in X^*$. Then

$$\exp(F_k^*) = \inf_{x \in R^n} N(x, k) \leq N(x^*, k) \leq m^{1/k} \exp(F(x^*)) = \exp(k^{-1} \ln m + F^*).$$

The latter implies $F_k^* \leq k^{-1} \ln m + F^*$. For $\varepsilon > 0$ and $x \in \{x: N(x, k) \leq \inf_x N(x, k) + \varepsilon\}$ we have $\exp(F^*) \leq \exp(F(x)) \leq N(x, k) \leq \inf_x N(x, k) + \varepsilon \leq \exp(F_k^*) + \varepsilon$. Since $\varepsilon > 0$ is arbitrary, $F^* \leq F_k^*$. Therefore (12) is proved.

Set

$$\Omega_k = \{x: F(x) \leq k^{-1} \ln m + F^*\},$$

$$\bar{\Omega}_k = \{x: N(x, k) \leq m^{1/k} \exp(F^*)\}.$$

The following inclusions hold: $X^* \subset \bar{\Omega}_k \subset \Omega_k \subset \Omega$. Indeed, consider $x^* \in X^*$. Then $N(x^*, k) \leq m^{1/k} \exp(F^*)$, i.e., $x^* \in \bar{\Omega}_k$. If $x \in \bar{\Omega}_k$ then $\exp(F(x)) \leq N(x, k) \leq m^{1/k} \exp(F^*)$. So $F(x) \leq k^{-1} \ln m + F^*$, i.e., $x \in \Omega_k$. Finally, $k > (C - F^*)^{-1} \ln m$ implies that $F(x) \leq k^{-1} \ln m + F^* \leq C$ for $x \in \Omega_k$. So $\Omega_k \subset \Omega$ in view of (2).

Let $d(k) = \max \{d(x, X^*) | x \in \Omega_k\}$. It is obvious that $d(k) \rightarrow 0$ as $k \rightarrow \infty$. Therefore $\Omega_k \rightarrow X^*$ in the Hausdorff metric. The inclusion $X^* \subset \bar{\Omega}_k \subset \Omega_k \subset \Omega$ implies that $\bar{\Omega}_k \neq \emptyset$ and is bounded. Therefore the continuity of $N(x, k)$ implies (13). Also $X_k^* \subset \bar{\Omega}_k$ implies $X_k^* \subset \Omega_k$. Therefore, taking into account that $d(k) \rightarrow 0$, we obtain (14).

COROLLARY. *Let*

$$x(k) = \operatorname{argmin} \{M(x, k) | x \in R^n\},$$

$$u(k) = (u_i(k) = \exp(kf_i(x(k))(\sum \exp(kf_i(x(k))))^{-1}, i = \overline{1, m}), \quad z(k) = (x(k), u(k)).$$

If $X^ = x^*$ and condition (4) holds, i.e., $z^* = (x^*, u^*)$ is unique, then $\lim_{k \rightarrow \infty} z(k) = z^*$. Since $M'_x(x(k), k) = \sum \exp(kf_i(x(k)))f'_i(x(k)) = 0$, we have $\sum u_i(k)f'_i(x(k)) = 0$, $\{x(k)\} \subset \Omega$, $\{u(k)\} \subset S_m$. Therefore, for all $\{z(k_i)\}$: $\lim_{k_i \rightarrow \infty} z(k_i) = \tilde{z}$, we have $\sum \tilde{u}_i f'_i(\tilde{x}) = 0$, $\tilde{u}_i(F(\tilde{x}) - f_i(\tilde{x})) = 0$; i.e., $\tilde{z} = z^*$. Taking into account that z^* is unique we get $\lim_{k \rightarrow \infty} z(k) = z^*$.*

Remark. Lemma 1 and some facts stated in [32] allow us to attach global character to the local results of Theorem 1 if we can obtain $x(k)$ for sufficiently large $k > 0$.

The estimate of distance between $x(k) \in X_k^*$ and x^* is related to uniqueness conditions for $z^* = (x^*, u^*)$.

LEMMA 2. *Suppose $f_i(x) \in C^2$, $i = \overline{1, m}$, and conditions (2)–(5) are satisfied. Then there exist $k_0 > 0$ and $c > 0$ which do not depend on $k \geq k_0$ and are such that for all $k \geq k_0$ we have the following:*

(1) *The estimates*

$$(16) \quad \|x(k) - x^*\| \leq ck^{-1}, \quad \|u(k) - u^*\| \leq ck^{-1}$$

are true.

(2) The function $M(x, k)$ is strongly convex in x in a neighborhood of $x(k)$.

Proof. It follows from Lemma 1 that $x(k)$ exists if $k_0 > (c - F^*)^{-1} \ln m$. Consider the following functions in a neighborhood of x^* and for $\kappa \in R^1$:

$$V_i(x, \kappa) = \begin{cases} \exp(|\kappa|^{-1} f_i(x)) \left(\sum_{j=1}^m \exp(|\kappa|^{-1} f_j(x)) \right)^{-1}, & \kappa \neq 0, \\ 0, & \kappa = 0. \end{cases}$$

If $i > r$, then there is a $c > 0$ independent of κ such that $V_i(x, \kappa) \leq c \exp(-c|\kappa|^{-1})$ for all $(x, \kappa) \in S(x^*, \varepsilon) \times (-\varepsilon, \varepsilon)$. Therefore $V_i(x, \kappa)$ is continuous in this neighborhood $(x^*, 0)$. We have $\|V'_i(x, \kappa)\| = \|(V'_{ix}(\cdot), V'_{i\kappa}(\cdot))\| = \|\kappa|^{-1} V_i(\cdot)(f'_i(x) - \sum_{j=1}^m V_j(\cdot)f'_j(x))\| - (\kappa|\kappa|)^{-1} V_i(\cdot)(f_i(x) - \sum_{j=1}^m V_j(\cdot)f_j(x))\| \leq c|\kappa|^{-2} \exp(-c|\kappa|^{-1})$. Setting $V'_i(x, 0) = 0$, $i > r$; $\rho(x, \kappa) = \sum_{i=r+1}^m V_i(\cdot)f'_i(x)$, we see that $V_i(x, \kappa)$ is smooth as well as $\rho(x, \kappa)$, and $\rho(x, 0) = 0$, $\rho'(x, 0) = 0$, for all $(x, \kappa) \in S(x^*, \varepsilon) \times (-\varepsilon, \varepsilon)$. Now consider the map $\bar{\Phi}(x, v, t, \kappa) = (\sum_{i=1}^r v_i f'_i(x) + \rho(x, \kappa); f_i(x) - F^* + t - \kappa \ln v_i, i = \overline{1, r}; \sum_{i=1}^r v_i + \sum_{i=r+1}^m V_i(x, \kappa) - 1): R^{n+r+2} \rightarrow R^{n+r+1}$. Taking into account the Kuhn-Tucker relations and the equalities $f_i(x^*) = F^*$, $i = \overline{1, r}$, $\rho(x^*, 0) = 0$, $\rho'(x^*, 0) = 0$, we conclude that $\bar{\Phi}(x^*, \bar{u}^*, 0, 0) = 0$ and $\bar{\Phi}'_{xvt} = \Phi'_{x\bar{u}t}$, so $\bar{\Phi}'_{xvt}$ is a nonsingular matrix. Therefore the implicit function theorem gives us the following: if $\varepsilon > 0$ is small enough, then there is a unique smooth vector-function $y(\kappa) = (x(\kappa), v(\kappa), t(\kappa)) \equiv y(\cdot): \bar{\Phi}(x(\cdot), v(\cdot), t(\cdot), \cdot) \equiv \bar{\Phi}(\cdot) \equiv 0$; that is, $\sum_{i=1}^r V_i(\cdot)f'_i(x(\cdot)) + \rho(x(\cdot), \cdot) = 0$, $f_i(x(\cdot)) - F^* + t(\cdot) - \kappa \ln V_i(\cdot) = 0$, $i = \overline{1, r}$ and $\sum_{i=1}^r V_i(\cdot) + \sum_{j=r+1}^m V_j(x(\cdot), \cdot) = 1$, $v(0) = \bar{u}^*$, $t(0) = 0$. Furthermore, from $\bar{\Phi}(\cdot) = 0$ we get $\bar{\Phi}'_{xvt}(\cdot)y'(\cdot) + \Phi_x^{-1}(\cdot) = 0$. Taking into account that $\bar{\Phi}_{xvt} = \Phi'_{x\bar{u}t}$, $f_i(x) \in C^2$, we obtain $\|\bar{\Phi}'_{xvt}(\cdot)\| \leq c_1$; then $\|\bar{\Phi}'_x(\cdot)\| \leq c_2$, and moreover c_1 and c_2 are independent of κ . Therefore $\|y'(\cdot)\| \leq c$ for $c = c_1 \cdot c_2$; that is, $\|x'(\kappa)\| \leq c$, $\|v'(\kappa)\| \leq c$. Using the inequalities and setting $k = \kappa^{-1}$, we obtain (16).

Now we prove the strong convexity $M(x, k)$ in a neighborhood of $x(k)$.

Let $U^* = \text{diag } u_i^*: R^r \rightarrow R^r$, $U(k) = \text{diag } u_i(k): R^m \rightarrow R^m$. Then $M''_{xx}(x, k) = k(\sum \exp(kf_i(x))f''_i(x) + \sum \exp(kf_i(x)f'_i{}^T(x)f''_i(x)))$ so that

$$\begin{aligned} M''_{xx}(x(k), k) &= M''_{xx}(\cdot, k) \\ &= (k \sum \exp(kf_i(\cdot))) \cdot (\sum u_i(k)f''_i(\cdot) + k \sum u_i(k)f'_i{}^T(\cdot)f'_i(\cdot)) \\ &= k \sum \exp(kf_i(\cdot))(L''_{xx}(\cdot) + kf'^T(\cdot)U(k)f'(\cdot)). \end{aligned}$$

Using estimate (16), Proposition 1, and the fact that $f_i \in C^2$, we have, for $k > 0$ large enough, $(M''_{xx}(\cdot, k)\xi, \xi) \approx kM(x^*, k)((L''_{xx}(z^*) + k\bar{f}'^T(x^*)U^*\bar{f}'(x^*))\xi, \xi) \geq \gamma\|\xi\|^2$, $\gamma > 0$, for all $\xi \in R^n$; therefore $M(x, k)$ is strongly convex in a neighborhood of $x(k)$.

Lemma 1 and 2 show that (10) can be realized with the estimate $\|z^s - z^*\| \leq ck^{-s}$, starting, for example, from $u^0 = (m^{-1}, \dots, m^{-1}) \in S_m$ if obtaining $x(k)$, $u(k)$ is possible for sufficiently large $k > 0$. Convexity $f_i(x)$, $i = \overline{1, m}$ is sufficient for this.

The rate of convergence can be improved by increasing k . But when k increases, the function $A(x, u, k)$ becomes ill-conditioned in x , which makes it more difficult to search for the minimum of $A(x, u, k)$. Therefore we cannot succeed in obtaining rapidly convergent processes using only $A(x, u, k)$. It appears that such processes can be obtained by applying important properties of the problem dual to (1). We shall now study these properties.

5. Dual problem. When proving Theorem 1 we found that there exists

$$x_k(u) = \underset{x \in V^*}{\operatorname{argmin}} A(x, u, k)$$

where

$$V^* = \begin{cases} R^n & \text{when all the functions } f_i \text{ are convex} \\ V(k_0, \varepsilon) & \text{if not} \end{cases}$$

if k is sufficiently large and (3)–(5) are satisfied. Therefore we have a function $\varphi_k(u) = A(x_k(u), u, k)$ defined on $S(u^*, \varepsilon)$. For $u \in S_m \setminus S(u^*, \varepsilon)$ we set $\varphi_k(u) = \inf_x A(x, u, k)$. In the neighborhood $S(u^*, \varepsilon)$, smoothness properties of $\varphi_k(u)$ are determined by the corresponding properties of $f_i(x)$, $i = \overline{1, m}$. In particular,

$$\begin{aligned} \varphi'_{ku}(u) &= A'_x(x_k(u), u, k)x'_k(u) + A'_u(x_k(u), u, k) \\ &= A'_u(x_k(u), u, k) \\ &= k^{-1}(\exp(kf_1(x_k(u))), \dots, \exp(kf_m(x_k(u)))) \end{aligned}$$

since $A'_x(x_k(u), u, k) = 0$. Furthermore, $\varphi''_{kuu}(u) = A''_{ux}(x_k(u), u, k)x'_k(u) = A''_{ux}(\cdot)x'_k(\cdot)$.

In order to determine $x'_k(u)$ we shall differentiate $A'_x(x_k(u), u, k) \equiv 0$ with respect to u . We obtain $A''_{xx}(\cdot)x'_k(u) + A''_{xu}(\cdot) = 0$. Therefore $x'_k(u) = (A''_{xx}(\cdot))^{-1}A''_{xu}(\cdot)$ and $\varphi''_{kuu}(u) = -A''_{ux}(\cdot)(A''_{xx}(\cdot))^{-1}A''_{xu}(\cdot)$. These formulas give the relation between smoothness properties of $\varphi_k(u)$ and corresponding properties of $f_i(x)$, $i = \overline{1, m}$. In particular, the continuity of $f'_i(x)$ implies the continuity of $\varphi''_{kuu}(u)$ and (6) implies that $\varphi''_{kuu}(u)$ satisfies a Lipschitz condition. Consider the problem dual to (1):

$$(17) \quad \tilde{u} = \operatorname{argmax} \{ \varphi_k(u) | u \in S_m \}.$$

The following theorem holds.

THEOREM 2. *Let the conditions (3)–(5) be satisfied and $f_i(x) \in C^2$, $i = \overline{1, m}$:*

(1) *Then there exists $k_0 > 0$ such that for $k \geq k_0$ the solution of the dual problem (17) exists and the strict form of sufficient optimality condition is satisfied for (17).*

(2) *There exists $\varepsilon > 0$ such that for $k \geq k_0$ the function $A(x, u, k)$ is strongly convex in $x \in S(x^*, \varepsilon)$, concave in u , and has a unique saddle point on $S(x^*, \varepsilon) \times S_m$, that is,*

$$(18) \quad A(x, u^*, k) \geq A(x^*, u^*, k) = \varphi_k(u^*) \geq A(x^*, u, k).$$

(3) *If $f_i(x)$, $i = \overline{1, m}$ are convex, then (18) holds on $R^n \times S_m$, and if instead of (5) we assume strong convexity one of $f_i(x)$, $i \in I^*$ then conditions (1)–(3) hold for all $k > 0$.*

Proof. (1) Since $A(x, u, k)$ is strongly convex in x , there exists $x_k(u) = \operatorname{argmin}_{x \in V^*} A(x, u, k)$ and $\varphi'_k(u) = k^{-1}(\exp(kf_1(x_k(u))), \dots, \exp(kf_m(x_k(u))))$ for all $u \in S(u^*, \varepsilon)$. Moreover, $x_k(u^*) = x^*$, $\varphi'_k(u^*) = k^{-1}(\exp kF^*, \dots, \exp kF^*; \exp kf_{r+1}(x^*), \dots, \exp kf_m(x^*))$. Consider the Lagrangian of (17):

$$L(u, \lambda) = \varphi_k(u) + \sum \lambda_i u_i + \lambda_0 (\sum u_i - 1).$$

Setting $\lambda_0^* = -k^{-1} \exp kF^*$ and $\lambda_i^* = k^{-1}(\exp kF^* - \exp kf_i(x^*))$, we obtain that the Kuhn–Tucker relations for the problem (17) are satisfied at (x^*, λ^*) :

$$k^{-1} \exp(kf_i(x^*)) + \lambda_i^* + \lambda_0^* = 0, \quad i = \overline{1, m},$$

i.e.,

$$\mathcal{L}'_u(u^*, \lambda^*) = 0, \quad \lambda_i^* \geq 0, \quad u_i^* \geq 0, \quad \lambda_i^* u_i^* = 0, \quad \sum u_i^* = 1.$$

The function $\varphi_k(u)$ is concave no matter whether $f_i(x)$ are convex or not, so u^* is a solution of (17).

Let $e_i = (0, 0, \dots, 0, \overline{1}, 0, \dots, 0)$, $e = (1, \dots, 1) \in R^m$. Since $F^* > f_i(x^*)$, $i = \overline{r+1, m}$, then $\lambda_i^* > 0$, $i = \overline{r+1, m}$. The gradients of active constraints in (17), e_{r+1}, \dots, e_m , e are linearly independent, i.e., for (17) the conditions of form (4) are satisfied. We shall now show that the conditions of form (5) are also satisfied for (17). They have the following form:

$$(19) \quad (e_i, v) = 0, \quad i = \overline{r+1, m}, \quad (e, v) = 0 \text{ implies } -(\varphi''_{kuu}(u^*)v, v) \geq \mu \|v\|^2, \quad \mu > 0.$$

Since $\varphi''_{kuu}(u^*) = -A''_{ux}(A''_{xx})A''_{xu}$ at $x = x^*$ and $u = u^*$ we have $(-\varphi''_{kuu}(u^*)v, v) = (A''_{xx}^{-1}A''_{xu}v, A''_{xu}v)$. There (19) is equivalent to the condition

$$(e_i, v) = 0, \quad i = \overline{r+1, m}, \quad (e, v) = 0 \text{ implies } \|A''_{xu}v\| \geq \mu \|v\|, \quad \mu > 0,$$

i.e.,

$$v_i = 0, \quad i = \overline{r+1, m}, \quad \sum_{i=1}^r v_i = 0 \text{ implies } \left\| \sum_{i=1}^r v_i \exp(kf_i(x^*))f'_i(x^*) \right\| \geq \mu \|v\|.$$

But the last condition is equivalent to the condition

$$\sum_{i=1}^r v_i = 0 \text{ implies } \left\| \sum_{i=1}^r v_i f'_i(x^*) \right\| \geq \mu \|v\|, \quad \mu > 0,$$

which is equivalent to (4), so the condition of type (5) for the dual problem (17) has been proved.

Therefore, for (17), the second-order sufficient conditions in a strict form (see, for example, [28, p. 47]) are satisfied if these conditions are satisfied for (1).

(2) The strong convexity of $A(x, u, k)$ in $x \in S(x^*, \varepsilon)$ for all $u \in S(u^*, \varepsilon)$ follows from Theorem 1 and the fact that $f_i(x) \in C^2$ if k is large enough.

In particular, $A''_{xx}(x^*, u^*, k)$ is a positive definite matrix, so together with $A'_x(x^*, u^*, k) = 0$ it gives the left inequality (18) and uniqueness of x^* . Since $A(x, u, k)$ is a linear function of u , the function $\varphi_k(u)$ is concave in u no matter whether or not $f_i(x)$, $i = \overline{1, m}$ are convex and the uniqueness of u^* is a consequence of (4). The right-hand side of (18) follows from $A(x^*, u, k) \leq k^{-1} \exp(kF^*)$, for all $u \in S_m$.

(3) The left inequality in (18) follows from the convexity $A(x, u^*, k)$ on x and (3); the right one follows from $A(x^*, u, k) \leq k^{-1} \exp(kF^*)$, for all $u \in S_m$. The uniqueness of x^* follows from the strong convexity of $A(x, u^*, k)$ on x , and the uniqueness of u^* follows from (4).

COROLLARY. *The restriction of the cost function $\varphi_k(u)$ to the manifold of active constraints of the dual problem (17) is a strongly concave function, i.e., the restrictions of $\varphi_k(u)$ to the manifold $\{u = (u_1, \dots, u_m) \geq 0: u_i = 0, i = \overline{r+1, m}, \sum_{i=1}^m u_i = 1\}$ are strongly concave.*

We shall consider this problem in more detail: Set

$$\bar{\varphi}_k(u) = \min_{x \in V^*} k^{-1} \sum_{i=1}^r u_i \exp(kf_i(x)) = \min_{x \in V^*} A_r(x, u, k) \quad \forall u \in \bar{S}(u^*, \varepsilon),$$

$$\bar{S}(u^*, \varepsilon) = \{u = (u_1, \dots, u_r) \geq 0: \|u - u^*\| \leq \varepsilon, u^* = (u_1^*, \dots, u_r^*)\},$$

$$\mathfrak{L} = \{u = (u_i, \dots, u_r): \sum u_i = 1\}.$$

Let the matrix $P: R^r \rightarrow R^r$ be the orthogonal projector into \mathfrak{L} of the vectors $u \in \bar{S}(u^*, \varepsilon)$ and let $\bar{\varphi}_{k\mathfrak{L}}(u) = \bar{\varphi}_k(Pu)$ be the restriction of the function $\bar{\varphi}_k(u)$ to \mathfrak{L} . Then $\bar{\varphi}'_{k\mathfrak{L}}(u) = PA'_{ru}(\cdot)$ is the gradient of the restriction $\bar{\varphi}_k(u)$ to \mathfrak{L} . The Hessian of the restriction $\bar{\varphi}_k(u)$ to \mathfrak{L} is defined by

$$\bar{H}_{k\mathfrak{L}}(u) = \bar{H}_{k\mathfrak{L}}(\cdot) = P\bar{\varphi}''_{kLU}(\cdot)P = -PA''_{ru}(\cdot)(A''_{rxx}(\cdot))^{-1}A''_{r \times u}(\cdot)P.$$

In order for $-\bar{H}_{k\mathfrak{L}}(\cdot)$ to be positive definite it is sufficient (see [21]) that the matrix $A''_{rx}(u)$ be positive definite and $A''_{rxu}(u)$ be nonsingular on \mathfrak{L} . The first property was proved in Theorem 1. Since $\bar{\varphi}_k(u)$ is considered in a small neighbourhood of u^* , we have $u_i > 0$, $i = \overline{1, r}$. Furthermore,

$$A''_{rxu}(u) = \left(\sum_{i=1}^r u_i \exp(kf_i(x_k(u))) \right) T(u, k),$$

$$T(u, k) = (\hat{u}_1 u_1^{-1} f'_1(x_k(u)), \dots, \hat{u}_r u_r^{-1} f'_r(x_k(u)))$$

so that for sufficiently large k we obtain $T(u^*, k) \approx (f'_1(x^*), \dots, f'_r(x^*)) = \bar{f}'^T(x^*)$. Therefore the nonsingularity of the operator $A''_{rxu}(u)$ on \mathfrak{L} is a consequence of condition (4).

Remark 1. Let I be any minimal set, and let u^* be a vertex of the Kuhn–Tucker polyhedron corresponding to I :

$$S_I(u^*, \varepsilon) = \left\{ u: \sum_{i \in I} u_i = 1, u_i \geq 0, i \in I, u_i = 0, i \in \bar{I}, \|u - u^*\| \leq \varepsilon \right\}.$$

Then the strong concavity property for $\varphi_{Ik}(u) = \min_{x \in V^*} k^{-1} \sum_{i \in I} u_i \exp(kf_i(x))$ on $S_I(u^*, \varepsilon)$ still holds if we replace (5) by (5').

Remark 2. The results of Theorems 1 and 2 are not true for the classical Lagrangian function $L(x, u)$ corresponding to the original problem (1), since (4)–(5) do not ensure in general that $L(x, u)$ is convex in x and $\operatorname{argmin}_{x \in V^*} L(x)$ may not exist or may not coincide with x^* .

Finally, note that $\varphi_k(u)$ is concave, and that if conditions (4)–(6) are satisfied, then $\varphi_k(u)$ is strongly concave on \mathfrak{L} in a neighbourhood of u^* with its Hessian satisfying a Lipschitz condition. These properties are used to find an approximation for the Lagrange multipliers. Consider this problem in more detail.

As a result of one step of (10) for sufficiently large k we can obtain $u \in S(u^*, \varepsilon)$ and isolate the set of active constraints I^* . After that, the search for u^* is reduced to the determination of $u^* = (u_1^*, \dots, u_\gamma^*)$

$$(20) \quad u^* = \operatorname{argmax} \{ \bar{\varphi}_k(u) | u \in \bar{S}(u^*, \varepsilon) \cap \mathfrak{L} \}.$$

Every method of smooth optimization of $\bar{\varphi}_k(u)$ on \mathfrak{L} determines some relaxation operator $R: \bar{S}(u^*, \varepsilon) \cap \mathfrak{L} \rightarrow \bar{S}(u^*, \varepsilon) \cap \mathfrak{L}$, i.e., $\bar{\varphi}_k(Ru) > \bar{\varphi}_k(u)$ and $R^S u \rightarrow u^*$, for all $u \in \bar{S}(u^*, \varepsilon) \cap \mathfrak{L}$. In particular, the gradient method has its relaxation operator defined by

$$(21) \quad Ru = u + tPA'_{ru}(\cdot), \quad t > 0,$$

and the relaxation operator corresponding to the Newton method is

$$(22) \quad Ru = u + \xi$$

where ξ is the normal solution of the system

$$\bar{H}_{k\mathfrak{L}}(u)\xi = -PA'_{ru}(\cdot).$$

The properties of these operators are determined by the properties of the corresponding methods of smooth optimization. The convergence rates of the approximation of u_i^* , $i \in I^*$ and, consequently, x^* are determined not only by the rate of convergence of the corresponding method of smooth optimization but also by the estimate from (9) because the optimization of $\bar{\varphi}_k(u)$ on \mathfrak{L} is always accompanied by a step of (10).

Under the conditions of Theorem 1 the matrix $\bar{H}_{k\lambda}(u)$ is continuous and has a negative spectrum. Therefore for the operator (22) we obtain a sequence $\{u^{S+1} = Ru^S\}_{S=0}^\infty$ such that $\|u^{S+1} - u^*\| \leq q_S \|u^S - u^*\|$, $q_S \rightarrow 0$. If condition (6) is satisfied then there exists $\lambda > 0$: $\|u^{S+1} - u^*\| \leq \lambda \|u^S - u^*\|^2$. The use of Newtonian and quasi-Newtonian methods allows us to obtain relaxation operators with corresponding properties of convergence. This makes it possible to formulate the following general method, which we shall consider for the convex $f_i(x)$, $i = \overline{1, m}$:

Let $u^0 = (m^{-1}, \dots, m^{-1}) \in R^m$ and let $k > 0$ be large enough. Assume that $z^S = (x^S, u^S)$ has already been found. Then define $z^{S+1} = (x^{S+1}; u^{S+1})$:

$$(23) \quad x^{S+1} = \operatorname{argmin} \{A_r(x, Ru^S, k) | x \in R^n\},$$

$$(24) \quad u^{S+1} \times \left(u_i^{S+1} = \bar{u}_i^S \exp(kf_i(x^{S+1})) \left(\sum_{i=1}^r \bar{u}_i^S \exp(kf_i(x^{S+1})) \right)^{-1}, i = \overline{1, r} \right)$$

where $\bar{u}^S = (\bar{u}_1^S, \dots, \bar{u}_r^S) = Ru^S$.

Let operator R have one of the following properties:

- (1°) $\|Ru - u^*\| \leq q \|u - u^*\|, \quad q < 1,$
- (2°) $\|Ru - u^*\| \leq q(u) \|u - u^*\|, \quad q(u) \rightarrow 0 \quad \text{as } u \rightarrow u^*,$
- (3°) $\|Ru - u^*\| \leq \lambda \|u - u^*\|^2.$

If $f_i(x)$, $i = \overline{1, m}$ is convex, (3)–(6) are satisfied, and k is large enough, then for the sequences $\{z^S\}_{S=0}^\infty$ generated by method (23), (24), using relaxation operators with properties (1°)–(3°) we obtain the following estimation:

- (1°°) $\|z^S - z^*\| \leq (cqk^{-1})^S;$
- (2°°) $\|z^S - z^*\| \leq (ck^{-1})^S \prod_{i=1}^S q_i, q_S \rightarrow 0 \quad \text{as } S \rightarrow \infty;$
- (3°°) $\|z^S - z^*\| \leq (c\lambda k^{-1})^{2^{S-1}} q_0^{2^S}, \quad q_0 < 1.$

The last estimates follow directly from the properties of operators R and the inequalities

$$\|x^{S+1} - x^*\| \leq \frac{c}{k} \|Ru^S - u^*\|, \quad \|u^{S+1} - u^*\| \leq \frac{c}{k} \|Ru^S - u^*\|$$

which follow from Theorem 1.

Acknowledgments. It is my pleasure to express my gratitude to Drs. G. D. Maistrovsky and M. E. Primak, and to the reviewers whose remarks contributes to the improvement of this work. I am grateful to Prof. D. Bertsekas, who kindly sent me many of his results.

REFERENCES

- [1] K. ARROW, L. HURWICZ, AND H. UZAVA, *Studies in Linear and Nonlinear Programming*, Stanford University Press, Stanford, CA, 1958.
- [2] D. BERTSEKAS, *Nondifferential optimization via approximation*, Math. Programming Stud., 3 (1975), pp. 1–25.
- [3] ———, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1981.
- [4] ———, *Approximation procedures based on the method of multipliers*, J. Optim. Theory Appl., 23 (1977), pp. 487–510.
- [5] ———, *A new algorithm for solution of resistive networks involving diodes*, IEEE Trans. Circuits and Systems, CS-23 (1976), pp. 599–608.

- [6] D. BERTSEKAS, G. LAUER, N. SANDELL, AND T. POSBERCH, *Optimal short-term scheduling of large scale power system*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 432-443.
- [7] D. BERTSEKAS, *A convergence analysis of the method of multipliers for nonconvex constrained optimization*, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 1978.
- [8] C. CHARALAMBOUS, *Nonlinear least Pth optimization and nonlinear programming*, Math. Programming, 12 (1977), pp. 195-225.
- [9] ———, *Acceleration of the least Pth algorithm for nonlinear programming*, Math. Programming, 17 (1979), pp. 270-297.
- [10] V. A. DAUGAVET AND V. N. MALOZYOMOV, *Quadratic convergence of some method of linearization for discrete minimax problems*, U.S.S.R. Comput. Math. and Math. Phys., 21 (1981), pp. 435-443.
- [11] V. DEMYANOV AND V. MALOZYOMOV, *Introduction to Minimax*, John Wiley, New York, 1974.
- [12] A. V. FIACCO AND G. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.
- [13] E. G. GOLSHTEIN AND N. V. TRETYAKOV, *On augmented Lagrangians for convex programming*, Econom. Math. Methods, 16 (1980), pp. 772-777.
- [14] K. GROSSMAN AND A. KAPLAN, *Nonlinear Programming Based on Unconstrained Optimization*, Nauka, Moscow, 1981. (In Russian.)
- [15] S. P. HAN, *Variable metric methods for minimizing a class of nondifferentiable functions*, Math. Programming, 20 (1981), pp. 1-13.
- [16] J. HALD AND K. MADSEN, *Combined LP and quasi Newton methods for minimax optimization*, Math. Programming, 20 (1981), pp. 49-62.
- [17] L. V. KANTOROVICH AND G. P. AKILOV, *Functional Analysis*, Nauka, Moscow, 1973.
- [18] L. G. KHACHIAN, *Polynomial algorithm in linear programming*, Soviet Math. Dokl., 244 (1979), pp. 1093-1096.
- [19] B. KORT AND D. BERTSEKAS, *A New Penalty Function for Constrained Minimization*, Proc. IEEE Conference on Decision and Control, New Orleans, LA, 1972.
- [20] ———, *Combined primal-dual penalty method for convex programming*, SIAM J. Control Optim., 14 (1976), pp. 268-294.
- [21] T. D. MAISTROVSKY AND U. V. OLKHOVSKY, *On the convergence of the method of the steepest descent in constrained optimization problems*, U.S.S.R. Comput. Math. and Math. Phys., 15 (1975), pp. 844-859.
- [22] T. S. MOTZKIN, *New technique for linear inequalities and optimization*, in Project SCOOP Symposium on Linear Inequalities and Programming, Planning Research Division, U.S. Air Force, Washington, D.C., 1952.
- [23] A. S. NEMIROVSKY AND D. B. YUDIN, *Complexity of Problems and Effectiveness of Optimization Methods*, Nauka, Moscow, 1979. (In Russian.)
- [24] ———, *Nonsmooth optimization*, in Proc. IIASA Workshop, C. Lemarechal and P. Mifflin, eds., Pergamon Press, Oxford, 1978.
- [25] ———, *Nondifferentiable optimization*, Math. Programming Stud., 3 (1975).
- [26] E. A. NURMINSKY, *Numerical Methods for Deterministic and Stochastic Minimax Problems*, Naukova Dumka, Kiev, 1979. (In Russian.)
- [27] J. M. ORTEGA AND W. C. RHEINOLDT, *Iterative solution of Nonlinear Equations of Several Variables*, Academic Press, New York, 1970.
- [28] V. N. PSHENICHNY AND U. M. DANILIN, *Numerical methods in extremal problems*, MIR, Moscow, 1978.
- [29] B. T. POLYAK AND N. N. TRETYAKOV, *Penalty method for conditional extremum problems*, U.S.S.R. Comput. Math. and Math. Phys., 13 (1973), pp. 34-46.
- [30] R. A. POLYAK, *Smooth optimization methods for solving nonlinear extremal and equilibrium problems with constraints*, Abstracts of the papers, Eleventh International Symposium on Mathematical Programming, Bonn, 1982.
- [31] ———, *Smooth optimization method in discrete minimax problems*, Ann. New York Acad. Sci. (1983), pp. 133-135.
- [32] ———, *On the best convex Chebyshev approximation*, Soviet Math. Dokl., 12 (1971), pp. 538-540.
- [33] ———, *Controlled processes in the extremal and equilibrium problems*, Dept. Manuscript, Viniti, Moscow, 1986. (In Russian.)
- [34] R. ROCKAFELLAR, *The multipliers method of Hestens and Powell applied to convex programming*, J. Optim. Theory Appl., 12 (1973), pp. 555-562.
- [35] N. Z. SHOR, *Minimization Methods for Nondifferentiable Functions*, Lecture Notes in Computat. Math., Springer-Verlag, Berlin, New York, 1985.

GALERKIN APPROXIMATION FOR OPTIMAL LINEAR FILTERING OF INFINITE-DIMENSIONAL LINEAR SYSTEMS*

A. GERMANI[†], L. JETTO[‡], AND M. PICCIONI[§]

Abstract. In this paper an implementable approximation of the infinite-dimensional Kalman filter is proposed for linear distributed systems corrupted by white noise. It relies on a Galerkin-type treatment of both the Riccati and the filter equation, and it is shown to converge to the best linear estimate of the state for each sample path of the noise. A basic tool is the study of the Riccati equation on the Hilbert space of Hilbert-Schmidt (HS) operators on the state space. Numerical results are given for a case concerning delay-differential equations.

Key words. infinite-dimensional systems, Galerkin approximation, linear filtering

AMS(MOS) subject classifications. primary 93E11; secondary 93E25

1. Introduction. The theory of filtering for linear systems evolving in Hilbert spaces is well established, both for systems described by cylinder finitely additive Gaussian measures [1] and for those described by Itô stochastic differential equations [5], [7].

We refer to the former approach because it also allows us to deal with disturbances on the output which are actually “white,” whereas a Hilbert-valued Wiener process with identity covariance does not correspond to a probability measure [1]. On the other hand, when the disturbance on the output has nuclear covariance, the two approaches can be shown to be equivalent by the same arguments as in the finite-dimensional case [8]. The evolution of the best linear estimate of the state is described by a linear system driven by the observation process through a gain operator. Such an operator is obtained by solving an infinite-dimensional Riccati equation [6], which as usual does not depend on the observations and yields the covariance of the estimation error. Formally, the situation is entirely analogous to the classical finite-dimensional Kalman filter, but the filter cannot be instrumented because of the infinite-dimensionality.

The idea can be conceptually divided into two steps. First, the original system is replaced by a finite-dimensional one whose state variable approximates the projection of the true state on a given finite-dimensional subspace. A fictitious finite-dimensional output is introduced and the corresponding Kalman filter is derived. Second, such fictitious output is replaced by a suitable finite-rank transformation of the actual output, yielding a finite-dimensional system whose solution is shown to converge to the best linear filter when the approximating subspace “converges” to the whole space.

The main tool for proving such a result consists of a smoothness property of the solution of the infinite-dimensional Riccati equation. In fact, the solution of such an equation is a continuous function with values in the Hilbert space of Hilbert-Schmidt (HS) operators on the state space. This is because the noise affecting the state equation is supposed to have a nuclear covariance operator, which is almost unavoidable even in the finitely additive approach [1]. For this reason the convergence of the solution of the approximate Riccati equation can be established with the unique assumption

* Received by the editors July 3, 1986; accepted for publication November 30, 1987.

[†] Istituto di Analisi dei Sistemi ed Informatica, Viale Manzoni 30, 00185, Rome, Italy and Dipartimento EE, Università de l'Aquila, 67100 Monteluco, Italy.

[‡] Dipartimento di Elettronica e Automatica, Università di Ancona, Via Brecce Bianche, 67100, Ancona, Italy.

[§] Dipartimento di Matematica, Seconda Università di Roma “Tor Vergata,” Via O. Raimondo, 00173, Rome, Italy.

that the approximate semigroups converge in the Trotter-Kato sense. We do not need any assumption of the behaviour of the adjoint semigroups. Lemmas 1 and 3 clarify such a situation. This makes such a class of approximation schemes much more tractable than for the dual problem of optimal control [3], [9], [11]–[13]. In the latter case the convergence of both the semigroups and their adjoints must be guaranteed. For this reason, the choice of approximating subspaces becomes quite delicate when the domain of the infinitesimal generator of the semigroup governing the system and the domain of its adjoint have a nondense intersection, which happens, for example, general hereditary systems. In this case the choice of approximating subspaces appears to be a nontrivial issue because the simplest way of projecting the evolution on a subspace does not work [12], [13]. We must mention that, in [12], such a problem is solved for some approximation scheme for hereditary systems, and the trace-norm convergence of the solutions of the resulting Riccati equations are obtained.

The paper is organized as follows. The discussion of the Riccati equation in the space of HS operators is given in § 2. In § 3 these results and the assumption about the convergence of the approximate semigroups allow us to prove the convergence of the filter for each sample path of the noise. We also outline how such a result can be extended to an infinite interval under a suitable stability condition. Section 4 is devoted to the discussion of an example concerning hereditary systems with approximating subspaces generated by first-order splines.

2. The Riccati equation in the space of Hilbert-Schmidt operators. Let H be a real separable Hilbert space. An HS operator S on H is a bounded linear operator such that

$$\sum_{i=1}^{\infty} \|Se_i\|^2 < +\infty$$

where $\{e_i\}$ is an orthonormal basis. The space $\mathcal{N}(H)$ of all HS operators on H is a Hilbert space with the inner product

$$[S, U]_{\text{HS}} = \sum_{i=1}^{\infty} (Se_i, Ue_i),$$

which is independent of the basis $\{e_i\}$ [1]. The space of self-adjoint HS operators is a subspace of $\mathcal{N}(H)$, which will be denoted by $\mathcal{N}_S(H)$. The cone of nonnegative definite operators in $\mathcal{N}_S(H)$ will be denoted by $\mathcal{N}_S^+(H)$. We will omit the reference to the space H when this is clear from the context.

We need to recall that \mathcal{N} is a (left and right) ideal of the space $\mathcal{L}(H)$ of linear bounded operators on H , such that

$$\|SL\|_{\text{HS}} \leq \|S\|_{\text{HS}}\|L\|, \quad S \in \mathcal{N}, \quad L \in \mathcal{L}(H);$$

furthermore,

$$(2.1) \quad \|S\|_{\text{HS}} = \|S^*\|_{\text{HS}}, \quad S \in \mathcal{N}.$$

We denote by $C(0, T; \mathcal{N}_S^+)$ the closed cone in $C(0, T; \mathcal{N}_S)$ of \mathcal{N}_S^+ -valued continuous functions.

In this section we will prove there exists a unique solution in $C(0, T; \mathcal{N}_S^+)$ for the following Riccati equation:

$$(2.2) \quad P(t) = T(t)P_0T^*(t) + \int_0^t T(t-s)[\Lambda - P(s)\Sigma p(s)]T^*(t-s) ds$$

where the following hold:

- (a) $\{T(t)\}$ is a strongly continuous semigroup of operators on H ;
 (b) Λ and P_0 are in \mathcal{N}_S^+ and Σ is nonnegative definite (not necessarily Hilbert-Schmidt).

First we need to prove a lemma that clarifies the strong smoothness action of HS operators. Such a lemma will be used repeatedly in the sequel.

LEMMA 1. *Let $\{G_m(t), t \in [0, T]\}$ be a sequence of strongly continuous $\mathcal{L}(H)$ -valued functions, strongly convergent to $\{G(t), t \in [0, T]\}$, uniformly on $[0, T]$. Let \mathcal{K} be a compact subset in \mathcal{N} . Then $\|G_m(t)N - G(t)N\|_{\text{HS}}$ converges to zero, uniformly with respect to $N \in \mathcal{K}$, $t \in [0, T]$.*

Proof. By definition

$$\|G_m(t)N - G(t)N\|_{\text{HS}}^2 = \sum_{i=1}^{\infty} \|(G_m(t) - G(t))Ne_i\|^2.$$

Since by hypothesis $\{G_m(t)\}$ is strongly bounded, by the uniform boundedness principle there exists a constant M such that

$$\|G_m(t)\|^2 + \|G(t)\|^2 \leq M^2/2, \quad t \in [0, T], \quad m = 1, 2, \dots.$$

On the other hand, for any $\varepsilon > 0$ there exists a finite covering of \mathcal{K} with open spheres S_j of radius $\sqrt{2}\varepsilon/4M$ centered in N_j , $j = 1, \dots, \nu$.

Let us define an integer \bar{n} such that

$$\sup_{j=1, \dots, \nu} \sum_{i=\bar{n}+1}^{\infty} \|N_j e_i\|^2 \leq \frac{\varepsilon^2}{8M^2}$$

so that

$$\begin{aligned} \sup_{N \in \mathcal{K}} \sum_{i=\bar{n}+1}^{\infty} \|Ne_i\|^2 &\leq \sup_{j=1, \dots, \nu} \sup_{N \in S_j} \sum_{i=\bar{n}+1}^{\infty} \|Ne_i\|^2 \\ &\leq 2 \sup_{j=1, \dots, \nu} \sup_{N \in S_j} \|N - N_j\|_{\text{HS}}^2 + 2 \sup_{j=1, \dots, \nu} \sum_{i=\bar{n}+1}^{\infty} \|N_j e_i\|^2 \\ &\leq \frac{\varepsilon^2}{4M^2} + 2 \sup_{j=1, \dots, \nu} \sum_{i=\bar{n}+1}^{\infty} \|N_j e_i\|^2 \leq \frac{\varepsilon^2}{2M^2}. \end{aligned}$$

Furthermore, note that for any e_i , the linear function ϕ_i from \mathcal{N} to H that maps N into Ne_i is continuous, so that $\phi_i(\mathcal{K})$ is compact in H . By consequence the set $\mathcal{K}^* = \bigcup_{i=1}^{\bar{n}} \phi_i(\mathcal{K})$ is compact and

$$\begin{aligned} \sup_{N \in \mathcal{K}} \|G_m(t)N - G(t)N\|_{\text{HS}}^2 &\leq \sup_{N \in \mathcal{K}} \sum_{i=1}^{\bar{n}} \|(G_m(t) - G(t))Ne_i\|^2 + M^2 \sup_{N \in \mathcal{K}} \sum_{i=\bar{n}+1}^{\infty} \|Ne_i\|^2 \\ &\leq \bar{n} \sup_{x \in \mathcal{K}^*} \|(G_m(t) - G(t))x\|^2 + \frac{\varepsilon^2}{2}. \end{aligned}$$

Now the linear operator from H to $C(0, T; H)$, which maps x into $[G_m(\cdot) - G(\cdot)]x$, converges pointwise to zero; therefore it converges uniformly on compact subsets of H [14] so that there exists \bar{m} such that for $m \geq \bar{m}$

$$\sup_{t \in [0, T]} \sup_{N \in \mathcal{K}} \|G_m(t)N - G(t)N\|_{\text{HS}} \leq \varepsilon. \quad \square$$

COROLLARY 1. *Under the hypotheses of Lemma 1, for each $N \in \mathcal{K}$ let $J_{N,m}$ be a sequence converging to N , uniformly in \mathcal{K} . Then $\|G_m(t)J_{N,m} - G(t)N\|_{\text{HS}}$ converges to zero uniformly with respect to $N \in \mathcal{K}$, $t \in [0, T]$.*

Proof. It suffices to note that

$$\|G_m(t)J_{N,m} - G(T)N\|_{\text{HS}} \leq \|G_m(t)\| \|J_{N,m} - N\|_{\text{HS}} + \|(G_m(t) - G(t))N\|_{\text{HS}}$$

which goes to zero by Lemma 1. \square

For each $t \geq 0$, let us define the linear bounded operator on \mathcal{N}_S by

$$(2.3) \quad \mathcal{R}_T(t)X = T(t)XT^*(t), \quad X \in \mathcal{N}_S.$$

It is immediately apparent that this constitutes an algebraic semigroup. We need to show that \mathcal{R} is also strongly continuous. Since we also deal with time-varying perturbations of $\{T(t)\}$ we state such a result for mild evolution operators [7].

THEOREM 1. *Let $\{U(t, s), 0 \leq s \leq t \leq T\}$ be a mild evolution operator on H and let*

$$(2.4) \quad S_U(t, s)X = U(t, s)XU^*(t, s), \quad X \in \mathcal{N}_S.$$

Then $\{S_U(t, s), 0 \leq s \leq t \leq T\}$ is a mild evolution operator on \mathcal{N}_S .

Proof. We need only show the following:

(a) $\mathcal{S}(\cdot, s)$ is strongly continuous in $[s, T]$;

(b) $\mathcal{S}(t, \cdot)$ is strongly continuous in $[0, t]$.

For (a), denote by C a bound for the norm of all operators $U(t, s)$

$$\begin{aligned} \|(\mathcal{S}_U(t+h, s) - \mathcal{S}_U(t, s))X\|_{\text{HS}} &= \|U(t+h, s)XU^*(t+h, s) - U(t, s)XU^*(t, s)\|_{\text{HS}} \\ &\leq \|U(t+h, s)XU^*(t+h, s) \\ &\quad - U(t+h, s)XU^*(t, s)\|_{\text{HS}} \\ &\quad + \|U(t+h, s)XU^*(t, s) - U(t, s)XU^*(t, s)\|_{\text{HS}} \\ &\leq 2C\|(U(t+h, s) - U(t, s))X\|_{\text{HS}} \end{aligned}$$

which goes to zero for $h \rightarrow 0$ in view of Lemma 1. In a similar way (b) is easily proved. \square

Note that the evolution operators $\mathcal{S}_U(t, s)$ leave the cone of nonnegative definite operators invariant.

When we use definition (2.3), we can rewrite (2.2) as a nonlinear mild equation on \mathcal{N}_S :

$$(2.5) \quad P(t) = \mathcal{R}_T(t)P_0 + \int_0^t \mathcal{R}_T(t-s)[\Lambda - P(s)\Sigma P(s)] ds.$$

For proving the existence of the solution we will use the Picard sequence, as suggested in [1]. First we need to prove some lemmas. For any $D \in C(0, T; \mathcal{N})$ and $x \in H$ the mild equation in H ,

$$z(t) = T(t-s)x - \int_s^t T(t-u)D(u)z(u) du,$$

has a unique continuous solution $U(t, s)x$ which defines a mild evolution operator on H [7].

Moreover, given D continuous in the HS norm, we can apply Lemma 1 to show that

$$(2.6) \quad U(t, s) = T(t-s) - \int_s^t T(t-u)D(u)U(u, s) du$$

where the integral is Riemann on the space of HS operators (note that the integrand is continuous in $0 \leq s \leq u \leq t \leq T$). Obviously, $U(t, s)$ is also the unique solution of such an equation.

A uniform bound on the norm of $U(t, s)$, which depends on D only through its sup-norm, can be found by using Gronwall's inequality.

LEMMA 2. *The equation in \mathcal{N}_S ,*

$$(2.7) \quad X(t) = \mathcal{R}_T(t)X_0 + \int_0^t \mathcal{R}_T(t-s)[F(s) - X(s)D^*(s) - D(s)X(s)] ds,$$

with $F(\cdot) \in C(0, T; \mathcal{N}_S)$, $D(\cdot) \in C(0, T; \mathcal{N})$, $X_0 \in \mathcal{N}_S$ has the unique continuous solution

$$(2.8) \quad X(t) = \mathcal{S}_U(t, 0)X_0 + \int_0^t \mathcal{S}_U(t, s)F(s) ds$$

where $U(t, s)$ is given by (6). Moreover such a solution is in $C(0, T; \mathcal{N}_S^+)$ provided $X_0 \in \mathcal{N}_S^+$ and $F \in C(0, T; \mathcal{N}_S^+)$.

Proof. The proof is accomplished by showing that (2.8) satisfies (2.7). By linearity it suffices to consider the case in which the forcing term $F(s)$ is identically zero. Then

$$\begin{aligned} X(t) &= \mathcal{S}_U(t, 0)X_0 = U(t, 0)X_0U^*(t, 0) \\ &= \left[T(t) - \int_0^t T(t-s)D(s)U(s, 0) ds \right] \\ &\quad \cdot X_0 \left[T^*(t) - \int_0^t U^*(u, 0)D^*(u)T^*(t-u) du \right] \\ &= T(t)X_0T^*(t) - \int_0^t T(t-s)D(s)U(s, 0)X_0T^*(s)T^*(t-s) ds \\ &\quad - \int_0^t T(t-u)T(u)X_0U^*(u, 0)D^*(u)T^*(t-u) du \\ &\quad + \int_0^t \int_0^s T(t-s)D(s)U(s, 0)X_0U^*(u, 0)D^*(u)T^*(t-u) du ds \\ &\quad + \int_0^t \int_s^t T(t-s)D(s)U(s, 0)X_0U^*(u, 0)D^*(u)T^*(t-u) du ds \\ &= T(t)X_0T^*(t) - \int_0^t T(t-s)D(s)U(s, 0)X_0 \\ &\quad \cdot \left[T^*(s) - \int_0^s U^*(u, 0)D^*(u)T^*(s-u) du \right] T^*(t-s) ds \\ &\quad - \int_0^t T(t-u) \left[T(u) - \int_0^t T(u-s)D(s)U(s, 0) ds \right] \\ &\quad \cdot X_0U^*(u, 0)D^*(u)T^*(t-u) du = T(t)X_0T^*(t) \\ &\quad - \int_0^t T(t-s)D(s)X(s)T^*(t-s) ds - \int_0^t T(t-s)X(s)D^*(s)T^*(t-s) ds \end{aligned}$$

where in the last line (2.6) is used so that the desired equation, (2.7), is obtained with $F = 0$.

Uniqueness is achieved by using the usual Gronwall argument.

The nonnegative definiteness property follows from the invariance of \mathcal{N}_S^+ with respect to the action of $\mathcal{S}_U(t, S)$. \square

The previous lemma explains the relationship between the effects of perturbations on semigroups T defined on H , and the corresponding \mathcal{R}_T induced on \mathcal{N}_S , respectively.

THEOREM 2. *The Riccati equation (2.5) has a unique solution in $C(0, T; \mathcal{N}_S^+)$.*

Proof. Let us define inductively the sequence $\{P_n\}$ in $C(0, T; \mathcal{N}_S^+)$ by means of the equations

$$\begin{aligned} P_0(t) &= \mathcal{R}_T(t)P_0, \\ (2.9) \quad P_{n+1}(t) &= \mathcal{R}_T(t)P_0 + \int_0^t \mathcal{R}_T(t-s) \\ &\quad \cdot [\Lambda + P_n(s)\Sigma P_n(s) - P_{n+1}(s)\Sigma P_n(s) - P_n(s)\Sigma P_{n+1}(s)] ds, \quad n = 0, 1, \dots \end{aligned}$$

which is of the type (2.7) with $X_0 = P_0$, $F(s) = \Lambda + P_n(s)\Sigma P_n(s)$, $D(s) = P_n(s)\Sigma$, so that it has a unique solution P_{n+1} in $C(0, T; \mathcal{N}_S^+)$. By setting $Q_n(t) = P_n(t) - P_{n+1}(t)$, we obtain from (2.9), for $n = 1, 2, \dots$,

$$Q_n(t) = \int_0^t \mathcal{R}_T(t-s)(Q_{n-1}(s)\Sigma Q_{n-1}(s) - Q_n(s)\Sigma P_n(s) - P_n(s)\Sigma Q_n(s)) ds$$

which is again of type (2.7) and whose solution is

$$(2.10) \quad Q_n(t) = \int_0^t S_{U_n}(t, s)(Q_{n-1}(s)\Sigma Q_{n-1}(s)) ds$$

where

$$U_n(t, s) = T(t-s) - \int_s^t T(t-u)P_n(u)\Sigma U_n(u, s) du.$$

From (2.10) we obtain $Q_n(\cdot) \in C(0, T; \mathcal{N}_S^+)$. It is an easy consequence, given $\|P_{n+1}(t)\|_{\text{HS}} \leq \|P_n(t)\|_{\text{HS}} \leq \dots \leq \|P_1(t)\|_{\text{HS}}$, that the $P_n(\cdot)$'s lie in a bounded subset of $C(0, T; \mathcal{N}_S^+)$. In consequence there exists a constant K such that

$$\|\mathcal{S}_{U_n}(t, s)\| \leq K, \quad 0 \leq s \leq t \leq T, \quad n = 1, 2, \dots;$$

furthermore, by (2.10), there exists a constant H such that

$$\|Q_n(t)\|_{\text{HS}} \leq H \int_0^t \|Q_{n-1}(s)\| ds \leq H^n \|Q_0\|_{C(0, T; \mathcal{N}_S)} \frac{T^n}{n!}.$$

By consequence $\{P_n\}$ is Cauchy in $C(0, T; \mathcal{N}_S)$, since

$$\|P_{n+m} - P_n\|_{C(0, T; \mathcal{N}_S)} \leq \sum_{i=n}^{n+m-1} \|Q_i\|_{C(0, T; \mathcal{N}_S)} \leq \|Q_0\| \sum_{i=n}^{\infty} \frac{H^i T^i}{i!}$$

which goes to zero with n , uniformly with respect to m .

Let $P = \lim_n P_n$ and let n go to infinity in (2.9); thus we obtain that P is a solution of the Riccati equation (2.5).

As for the uniqueness, note that any solution Q of (2.4) in $C(0, T; \mathcal{N}_S^+)$ satisfies

$$\|Q(t)\|_{\text{HS}} \leq \left\| \mathcal{R}_T(t)P_0 + \int_0^t \mathcal{R}_T(t-s)\Lambda ds \right\|_{\text{HS}} \leq G$$

so that its sup-norm is bounded. Now let $E(t) = P(t) - Q(t)$. Then

$$\begin{aligned} E(t) &= \int_0^t \mathcal{R}_T(t-s)[Q(s)\Sigma Q(s) - P(s)\Sigma P(s)] ds \\ &= - \int_0^t \mathcal{R}_T(t-s)(E(s)\Sigma Q(s) + P(s)\Sigma E(s)) ds \end{aligned}$$

from which, for some constant M ,

$$\|E(t)\|_{\text{HS}} \leq M \int_0^t \|E(s)\|_{\text{HS}} ds.$$

This in turn implies that $E(t) \equiv 0$, given $E(0) = 0$. \square

3. The approximate finite-dimensional filter. In this section we will consider an arbitrary linear infinite-dimensional system on the Hilbert space H , in the interval $[0, T]$

$$(3.1) \quad \dot{x}(t) = Ax(t) + B\omega(t), \quad x(0) = x_0,$$

$$(3.2) \quad y(t) = Cx(t) + G\omega(t)$$

where the following hold:

(a) A is the infinitesimal generator of a strongly continuous semigroup $\{T(t)\}$ on H , such that $\|T(t)\| \leq e^{\gamma t}$;

(b) $\omega \in L^2(0, T; H_n)$. H_n is the Hilbert space where the noise takes values;

(c) $B: H_n \rightarrow H$, $C: H \rightarrow H_0$ (where H_0 is the Hilbert observation space), and $G: H_n \rightarrow H_0$ are bounded linear operators, with B Hilbert-Schmidt, such that $GB^* = 0$, $GG^* = I_{H_0}$.

The precise meaning of (3.1) is

$$(3.3) \quad x(t) = T(t)x_0 + \int_0^t T(t-s)B\omega(s) ds.$$

If $L^2(0, T; H_n)$ is equipped with the standard Gauss cylinder measure (which corresponds to model ω as a white noise process) and x_0 is a Gaussian random variable with mean vector m_0 and nuclear covariance operator P_0 , we can define the best linear estimate $\hat{x}(t)$ of $x(t)$ given $\{y(s); 0 \leq s \leq t\}$ and prove that it evolves according to the following equations [1]:

$$(3.4) \quad \begin{aligned} \dot{\hat{x}}(t) &= A\hat{x}(t) + P(t)C^*(y(t) - C\hat{x}(t)), \\ \hat{x}(0) &= m_0 \end{aligned}$$

where $P(t)$ is the unique self-adjoint, nonnegative definite, strongly continuous solution of the following Riccati equation [6]:

$$(3.5) \quad P(t)x = T(t)P_0T^*(t)x + \int_0^t T(t-s)(BB^* - P(s)CC^*P(s))T^*(t-s)x ds, \quad x \in H.$$

Actually in [1] only the case of fixed initial state is considered, but obviously it is possible to also deal with random initial conditions. Of course, (3.4) must be presented as

$$(3.6) \quad \hat{x}(t) = T(t)m_0 + \int_0^t T(t-s)P(s)C^*(y(s) - C\hat{x}(s)) ds$$

and the unique solution of the Riccati equation (2.2) in $C(0, T; \mathcal{N}_S^+)$ with $\Lambda = BB^*$, $\Sigma = C^*C$ will also solve the “strong” version (3.5). Therefore from now on we will refer only to the Riccati equation in $C(0, T; \mathcal{N}_S^+)$.

The system (3.5), (3.6), which we can accurately call an infinite-dimensional Kalman filter, cannot be implemented directly because of the infinite dimensionality of both the filter and the Riccati equation.

We now introduce the proposed approximation. For each n let Π_n be a linear mapping of H onto a finite-dimensional Hilbert space V_n , such that $\Pi_n^* \Pi_n$ converges strongly to the identity and $[N(\Pi_n)]^\perp \subset D(A)$ for each n . Then $\Pi_n A \Pi_n^*$ is a bounded operator which generates the semigroup $T_n(t) = \exp(\Pi_n A \Pi_n^* t)$ on V_n . Since $\|\Pi_n\| \leq M$ for each n , for some M , it is easily obtained that there exists $\bar{\gamma} \geq \gamma$ such that

$$(3.7) \quad \|T_n(t)\| \leq e^{\bar{\gamma}t}.$$

For practical applications Π_n will be a projection onto a subspace of H contained in the domain of A . However, considering V_n as a Hilbert space by itself allows us to make the proofs clearer.

We suppose that for each $x \in H$

$$(3.8) \quad \lim_{n \rightarrow \infty} \sup_{t \in [0, T]} \|\Pi_n T(t)x - T_n(t)\Pi_n x\| = 0.$$

Necessary and sufficient conditions for (3.8) to hold can be deduced from the general Trotter-Kato Theorem [15].

The mapping $\Pi_n: H \rightarrow V_n$ induces the following mapping on operators:

$$\mathcal{J}_n: \mathcal{N}_S(H) \rightarrow \mathcal{N}_S(V_n), \quad \mathcal{J}_n X = \Pi_n X \Pi_n^*.$$

The following lemma ensures that the same type of convergence is obtained for the following sequence $\{\mathcal{R}_{T_n}(t)\}$ of strongly continuous semigroups:

$$(3.9) \quad \mathcal{R}_{T_n}(t)X = T_n(t)X T_n^*(t), \quad X \in \mathcal{N}_S(V_n).$$

LEMMA 3. *Under hypothesis (3.8), we have that*

$$\lim_{n \rightarrow \infty} \sup_{t \in [0, T]} \|\mathcal{R}_{T_n}(t)(\mathcal{J}_n X) - \mathcal{J}_n(\mathcal{R}_T(t)X)\|_{\text{HS}} = 0$$

for any $X \in \mathcal{N}_S(H)$.

Proof. By applying Lemma 1, we have that

$$\begin{aligned} & \sup_{t \in [0, T]} \|T_n(t)\Pi_n X \Pi_n^* T_n^*(t) - \Pi_n T(t)X T^*(t)\Pi_n^*\|_{\text{HS}} \\ & \leq \sup_{t \in [0, T]} \|T_n(t)\Pi_n X \Pi_n^* T_n^*(t) - T_n(t)\Pi_n X T^*(t)\Pi_n^*\|_{\text{HS}} \\ & \quad + \sup_{t \in [0, T]} \|T_n(t)\Pi_n X T^*(t)\Pi_n^* - \Pi_n T(t)X T^*(t)\Pi_n^*\|_{\text{HS}} \\ & \leq 2M(e^{\bar{\gamma}} \vee 1) \sup_{t \in [0, T]} \|(T_n(t)\Pi_n - \Pi_n T(t))X\|_{\text{HS}}, \end{aligned}$$

which proves the lemma. \square

For each n we will now consider the equation on $C(0, T; \mathcal{N}_S^+(V_n))$:

$$(3.10) \quad \begin{aligned} P^{[n]}(t) &= T_n(t)\Pi_n P_0 \Pi_n^* T_n^*(t) + \int_0^t T_n(t-s) \\ & \quad \cdot [\Pi_n B B^* \Pi_n^* - P^{[n]}(s)\Pi_n C^* C \Pi_n^* P^{[n]}(s)] T_n^*(t-s) ds, \end{aligned}$$

which can be rewritten as

$$(3.11) \quad \begin{aligned} P^{[n]}(t) &= \mathcal{R}_{T_n}(t)\mathcal{J}_n(P_0) + \int_0^t \mathcal{R}_{T_n}(t-s) \\ & \quad \cdot (\mathcal{J}_n(B B^*) - P^{[n]}(s)\mathcal{J}_n(C^* C)P^{[n]}(s)) ds. \end{aligned}$$

THEOREM 3. Let $P(t)$ be the solution of the Riccati equation $C(0, T; \mathcal{N}_S^+(H))$:

$$(3.12) \quad P(t) = \mathcal{R}_T(t)P_0 + \int_0^t \mathcal{R}_T(t-s)[BB^* - P(s)C^*CP(s)] ds.$$

Then if (3.8) holds,

$$\lim_{n \rightarrow \infty} \sup_{t \in [0, T]} \|P^{[n]}(t) - \mathcal{J}_n P(t)\|_{\text{HS}} = 0.$$

Proof. Let $\Lambda = BB^*$, $\Sigma = C^*C$. By a direct computation

$$(3.13) \quad \begin{aligned} \|P^{[n]}(t) - \mathcal{J}_n P(t)\|_{\text{HS}} &\leq \|\mathcal{R}_{T_n}(t)\mathcal{J}_n P_0 - \mathcal{J}_n(\mathcal{R}_T(t)P_0)\|_{\text{HS}} \\ &\quad + \int_0^t \|\mathcal{R}_{T_n}(t-s)\mathcal{J}_n \Lambda - \mathcal{J}_n(\mathcal{R}_T(t-s)\Lambda)\|_{\text{HS}} ds \\ &\quad + \int_0^t \|\mathcal{R}_{T_n}(t-s)(P^{[n]}(s)\mathcal{J}_n(\Sigma)P^{[n]}(s)) \\ &\quad - \mathcal{J}_n(\mathcal{R}_T(t-s)(P(s)\Sigma P(s)))\|_{\text{HS}} ds. \end{aligned}$$

The first two terms on the right-hand side are bounded by a constant $\varepsilon'_n > 0$ which goes to zero as n goes to infinity by Lemma 3. For the latter term note that, given $\|\mathcal{R}_{T_n}(t)\|$ uniformly bounded both in n and t , by using the same kind of estimate as in the uniqueness proof of Theorem 2, it also results that $\|P^{[n]}(t)\|_{\text{HS}}$ is uniformly bounded. Moreover,

$$\begin{aligned} &\|\mathcal{R}_{T_n}(t-s)P^{[n]}(s)\Pi_n \Sigma \Pi_n^* P^{[n]}(s) - \Pi_n \mathcal{R}_T(t-s)P(s)\Sigma P(s)\Pi_n^*\|_{\text{HS}} \\ &\leq \|\mathcal{R}_{T_n}(t-s)P^{[n]}(s)\Pi_n \Sigma \Pi_n^* P^{[n]}(s) - \mathcal{R}_{T_n}(t-s)P^{[n]}(s)\Pi_n \Sigma \Pi_n^* \Pi_n P(s)\Pi_n^*\|_{\text{HS}} \\ &\quad + \|\mathcal{R}_{T_n}(t-s)P^{[n]}(s)\Pi_n \Sigma \Pi_n^* \Pi_n P(s)\Pi_n^* - \mathcal{R}_{T_n}(t-s)P^{[n]}(s)\Pi_n \Sigma P(s)\Pi_n^*\|_{\text{HS}} \\ &\quad + \|\mathcal{R}_{T_n}(t-s)P^{[n]}(s)\Pi_n \Sigma P(s)\Pi_n^* - \mathcal{R}_{T_n}(t-s)\Pi_n P(s)\Pi_n^* \Pi_n \Sigma P(s)\Pi_n^*\|_{\text{HS}} \\ &\quad + \|\mathcal{R}_{T_n}(t-s)\Pi_n P(s)\Pi_n^* \Pi_n \Sigma P(s)\Pi_n^* - \Pi_n \mathcal{R}_T(t-s)P(s)\Pi_n^* \Pi_n \Sigma P(s)\Pi_n^*\|_{\text{HS}} \\ &\quad + \|\Pi_n \mathcal{R}_T(t-s)P(s)\Pi_n^* \Pi_n \Sigma P(s)\Pi_n^* - \Pi_n \mathcal{R}_T(t-s)P(s)\Sigma P(s)\Pi_n^*\|_{\text{HS}} \\ &\leq C_1 \|P^{[n]}(s) - \Pi_n P(s)\Pi_n^*\|_{\text{HS}} + C_2 \|\Pi_n^* \Pi_n P(s) - P(s)\|_{\text{HS}} \\ &\quad + C_3 \|\mathcal{R}_{T_n}(t-s)\mathcal{J}_n P(s) - \mathcal{J}_n(\mathcal{R}_T(t-s)P(s))\|_{\text{HS}} \end{aligned}$$

for suitable constants C_1 , C_2 , and C_3 . By Lemma 1 the third term on the right-hand side is bounded by a constant $\varepsilon''_n > 0$ which goes to zero as n goes to infinity, because $\{P(s), s \in [0, T]\}$ is a compact set in \mathcal{N}_S . Now let $\varepsilon_n = \varepsilon'_n + \varepsilon''_n T$, so that by (3.13)

$$\|P^{[n]}(t) - \Pi_n P(t)\Pi_n^*\|_{\text{HS}} \leq \varepsilon_n + (C_1 + C_2) \int_0^t \|P^{[n]}(s) - \Pi_n P(s)\Pi_n^*\|_{\text{HS}} ds$$

which yields the theorem by Gronwall's inequality. \square

Now the approximate solution $P^{[n]}$ of the Riccati equation (3.10) is used as a gain operator in an approximate filter equation. For each n let Q_n be a linear finite-rank operator on H_0 , such that Q_n^*C converges strongly to C . The filter equation (3.4) is replaced by

$$(3.14) \quad \begin{aligned} \dot{\hat{x}}^{[n]}(t) &= \Pi_n A \Pi_n^* \hat{x}^{[n]}(t) + P^{[n]}(t) \Pi_n C^* (Q_n y(t) - C \Pi_n^* \hat{x}^{[n]}(t)), \\ \hat{x}^{[n]}(0) &= \Pi_n m_0 \end{aligned}$$

with $P^{[n]}$ given by (3.10), which is equivalent to the mild equation

$$(3.15) \quad \hat{x}^{[n]}(t) = T_n(t)\Pi_n m_0 + \int_0^t T_n(t-s)P^{[n]}(s)\Pi_n C^*(Q_n y(s) - C\Pi_n^* \hat{x}^{[n]}(s)) ds.$$

Next let $U(t, s)$ be the mild evolution operator corresponding to $T(t)$ by applying (3.6) with $D(s) = P(s)C^*C$ and let $U_n(t, s)$ be the analogous one obtained with respect to $T_n(t)$ by letting $D_n(s) = P^{[n]}(s)\Pi_n C^*C\Pi_n^*$.

Then (3.6) and (3.15) have the following solutions:

$$(3.16) \quad \hat{x}(t) = U(t, 0)m_0 + \int_0^t U(t, s)P(s)C^*y(s) ds,$$

$$(3.17) \quad \hat{x}^{[n]}(t) = U_n(t, 0)\Pi_n m_0 + \int_0^t U_n(t, s)P^{[n]}(s)\Pi_n C^*Q_n y(s) ds.$$

Note that

$$\begin{aligned} \|\Pi_n D(s)\Pi_n^* - D_n(s)\|_{\text{HS}} &\leq \|\Pi_n P(s)C^*C\Pi_n^* - \Pi_n P(s)\Pi_n^* \Pi_n C^*C\Pi_n^*\|_{\text{HS}} \\ &\quad + \|\Pi_n P(s)\Pi_n^* \Pi_n C^*C\Pi_n^* - P^{[n]}(s)\Pi_n C^*C\Pi_n^*\|_{\text{HS}} \\ &\leq C_1 \|(\Pi_n^* \Pi_n - I)P(s)\|_{\text{HS}} + C_2 \|P^{[n]}(s) - \mathcal{J}_n P(s)\|_{\text{HS}} \end{aligned}$$

which goes to zero by (3.13) and Corollary 1. Before proving the final theorem we need the following lemma.

LEMMA 4. *We have*

$$\lim_n \sup_{0 \leq s \leq t \leq T} \|U_n(t, s)\Pi_n x - \Pi_n U(t, s)x\| = 0 \quad \forall x \in H.$$

Proof. Let $U(t, s)x = z(t; s)$ and $U_n(t, s)x = z^{[n]}(t; s)$. Then

$$z^{[n]}(t; s) = T_n(t-s)x - \int_s^t T_n(t-u)D_n(u)z^{[n]}(u; s) du,$$

$$z(t; s) = T(t-s)x - \int_s^t T(t-u)D(u)z(u; s) du.$$

Since

$$\begin{aligned} &\|T_n(t-u)D_n(u)z^{[n]}(u; s) - \Pi_n T(t-u)D(u)z(u; s)\| \\ &\leq \|T_n(t-u)D_n(u)z^{[n]}(u; s) - T_n(t-u)D_n(u)\Pi_n z(u; s)\| \\ &\quad + \|T_n(t-u)D_n(u)\Pi_n z(u; s) - T_n(t-u)\Pi_n D(u)\Pi_n^* \Pi_n z(u; s)\| \\ &\quad + \|T_n(t-u)\Pi_n D(u)\Pi_n^* \Pi_n z(u; s) - T_n(t-u)\Pi_n D(u)z(u; s)\| \\ &\quad + \|T_n(t-u)\Pi_n D(u)z(u; s) - \Pi_n T(t-u)D(u)z(u; s)\| \end{aligned}$$

from which, by arguing as in Theorem 3, we obtain

$$\|z^{[n]}(t; s) - \Pi_n z(t; s)\| \leq \delta_n + K \int_0^t \|z^{[n]}(u; s) - \Pi_n z(u; s)\| du$$

for some constants $K \geq 0$ and $\delta_n \geq 0$ (which goes to zero as n goes to infinity). The strong uniform convergence in $0 \leq s \leq t \leq T$ is then proved by Gronwall's inequality. \square

Now the final result can be obtained.

THEOREM 4. *For each $y \in L^2(0, T; H_0)$, and uniformly on bounded subsets, the sequence $\Pi_n^* \hat{x}^{[n]}$, $\hat{x}^{[n]}$ being the solution of (3.17), converges to the solution \hat{x} of (3.16) in $C(0, T; H)$, provided (3.8) is assumed.*

Proof. From (3.16) and (3.17) we have

$$(3.18) \quad \begin{aligned} \|\hat{x}^{[n]}(t) - \Pi_n \hat{x}(t)\| &\leq \|U_n(t, 0) \Pi_n m_0 - \Pi_n U(t, 0) m_0\| \\ &\quad + \int_0^t \|U_n(t, s) P^{[n]}(s) \Pi_n C^* Q_n - \Pi_n U(t, s) P(s) C^*\|_{\text{HS}} \|y(s)\| \, ds. \end{aligned}$$

By Lemma 3 the first term is bounded by a constant σ_n which goes to zero as n goes to infinity. Moreover,

$$\begin{aligned} &\|U_n(t, s) P^{[n]}(s) \Pi_n C^* Q_n - \Pi_n U(t, s) P(s) C^*\|_{\text{HS}} \\ &\leq \|U_n(t, s) P^{[n]}(s) \Pi_n C^* Q_n - U_n(t, s) \Pi_n P(s) C^*\|_{\text{HS}} \\ &\quad + \|U_n(t, s) \Pi_n P(s) C^* - \Pi_n U(t, s) P(s) C^*\|_{\text{HS}} \\ &\leq C_1 (\|P^{[n]}(s) \Pi_n C^* Q_n - \Pi_n P(s) \Pi_n^* \Pi_n C^* Q_n\|_{\text{HS}} \\ &\quad + \|\Pi_n P(s) \Pi_n^* \Pi_n C^* Q_n - \Pi_n P(s) C^* Q_n\|_{\text{HS}} \\ &\quad + \|\Pi_n P(s) C^* Q_n - \Pi_n P(s) C^*\|_{\text{HS}}) \\ &\quad + C_2 \|U_n(t, s) \Pi_n P(s) - \Pi_n U(t, s) P(s)\|_{\text{HS}} \\ &\leq C_3 (\|P^{[n]}(s) - \Pi_n P(s) \Pi_n^*\|_{\text{HS}} + \|(\Pi_n^* \Pi_n - I) P(s)\|_{\text{HS}} + \|(Q_n^* C - C) P(s)\|_{\text{HS}} \\ &\quad + \|U_n(t, s) \Pi_n P(s) - \Pi_n U(t, s) P(s)\|_{\text{HS}}) \end{aligned}$$

can be bounded uniformly in $0 \leq s \leq t \leq T$ by a constant ψ_n which goes to zero as n goes to infinity by Theorem 3, Lemma 4, the hypotheses on Π_n and Q_n , and the compactness of $\{P(s), 0 \leq s \leq T\}$ in $\mathcal{N}_S(H)$. By substituting in (3.18) and using the Schwartz inequality, we have that

$$\sup_{t \in [0, T]} \|\hat{x}^{[n]}(t) - \Pi_n \hat{x}(t)\| \leq \delta_n + \psi_n \sqrt{T} \|y\|_{L^2(0, T; H_0)}.$$

Moreover, since $\|\Pi_n^* \Pi_n \hat{x}(t) - \hat{x}(t)\|$ goes to zero uniformly on $t \in [0, T]$ and on bounded subsets of y 's by application of $\Pi_n^* \Pi_n$ to both sides of (3.26), and $\|\Pi_n^*\| \leq M$, the theorem is proved. \square

Now let us write the proposed approximate filter (3.10)–(3.15) coordinate-wise. By simplicity of notation we suppose V_n to be n -dimensional. Let us choose a basis $\{v_1 \cdots v_n\}$ in V_n ; this induces an isomorphism ζ_n between V_n and R^n

$$\zeta_n: \sum_{i=1}^n \alpha_i v_i \mapsto (\alpha_1, \dots, \alpha_n)'$$

and a corresponding isomorphism η_n between the algebra μ_n of operators on V_n and the algebra M_n of matrices of order n , defined by

$$\eta_n(S) = W_n^{-1} \tilde{S}$$

where $(W_n)_{i,j} = (v_i, v_j)$ and $(\tilde{S})_{i,j} = (Sv_j, v_i)$. The two isomorphisms are related by

$$\eta_n(S) \zeta_n(x) = \zeta_n(Sx), \quad x \in V_n, \quad S \in \mu_n.$$

Since V_n is finite-dimensional, (3.10) is rewritten in differential form as

$$(3.19) \quad \begin{aligned} \dot{P}^{[n]}(t) &= A_n P^{[n]}(t) + P^{[n]}(t) A_n^* + \Lambda_n - P^{[n]}(t) \Sigma_n P^{[n]}(t), \\ P^{[n]}(0) &= P_0^{[n]} \end{aligned}$$

where $A_n = \Pi_n A \Pi_n^*$, $\Lambda_n = \Pi_n B B^* \Pi_n^*$, $\Sigma_n = \Pi_n C C^* \Pi_n^*$, $P_0^{[n]} = \Pi_n P_0 \Pi_n^*$.

Note that $A_n^* \supset \Pi_n A^* \Pi_n$, which is not, in general, defined on the whole V_n . By applying η_n to both sides we obtain the following matrix equation;

$$(3.20) \quad \begin{aligned} \dot{\tilde{P}}^{[n]}(t) &= \tilde{A}_n W_n^{-1} \tilde{P}^{[n]}(t) + \tilde{P}^{[n]}(t) W_n^{-1} \tilde{A}_n^T + \tilde{A}_n - \tilde{P}^{[n]}(t) W_n^{-1} \tilde{\Sigma}_n W_n^{-1} \tilde{P}^{[n]}(t), \\ \tilde{P}^{[n]}(0) &= \tilde{P}_0^{[n]}. \end{aligned}$$

To express the filter coordinate-wise let us choose a basis $\{\psi_1, \dots, \psi_{m_n}\}$ in range Q_n , so that another isomorphism ζ'_n can be defined between Z_n and R :

$$\zeta'_n: \sum_{j=1}^{m_n} \beta_j v_j \mapsto (\beta_1, \dots, \beta_{m_n})'$$

and the corresponding η'_n between the linear transformation of Z_n into V_n and the $(n \times m_n)$ -matrices, defined by

$$\eta'_n(F) = W_n^{-1} \bar{F}$$

where $(\bar{F})_{i,j} = (F\psi_j, v_i)$, $i = 1, \dots, n$, $j = 1, \dots, m_n$. We have that $\eta'_n(F)\zeta'_n(z) = \zeta_n(Fz)$, $z \in Z_n$, $F \in \mathcal{L}(Z_n, V_n)$. Now let $\tilde{y}_n(t) = \eta'_n(Q_n y(t))$ and $\Gamma_n = \Pi_n C^*$; then by applying η_n to the filter equation and differentiating, we get the approximate Kalman filter for $\hat{x}_c^{[n]}(t) = \eta_n(\hat{x}^{[n]}(t))$:

$$(3.21) \quad \begin{aligned} \dot{\hat{x}}_c^{[n]}(t) &= W_n^{-1}[(\tilde{A}_n - \tilde{P}^{[n]}(t)\tilde{\Sigma}_n)\hat{x}_c^{[n]}(t) + \tilde{P}^{[n]}(t)W_n^{-1}\bar{\Gamma}_n\tilde{y}_n(t)], \\ \hat{x}_c^{[n]}(0) &= (\Pi_n m_0)_c. \end{aligned}$$

In view of Theorem 4, the required approximation for $\hat{x}(t)$ is

$$\sum_{i=1}^n \hat{x}_{c,i}^{[n]}(t) \Pi_n^* v_i.$$

We conclude by observing that in the case of R^m -valued observations we can choose Q_n to be the identity, eventually with n . In this case the proposed filter (3.20), (3.21) reduces, for each $\omega \in L^2(0, T; H_n)$, to the standard Kalman filter for the finite-dimensional system

$$\begin{aligned} \dot{x}^{[n]}(t) &= \Pi_n A \Pi_n^* x^{[n]}(t) + \Pi_n B \omega(t), \quad x^{[n]}(0) = \Pi_n x_0, \\ y^{[n]}(t) &= C \Pi_n^* x^{[n]}(t) + G \omega(t) \end{aligned}$$

which is recognized to be an approximation of the original system (3.1), (3.2), provided that the unavailable measurement $y^{[n]}$ is replaced by the actual measurement y .

A major problem remains the extension of the previous results to the steady-state filter. Theorem 5 gives an answer, under a rather restrictive stability condition.

The steady-state covariance operator P_∞ for the system (3.1)–(3.2) satisfies the algebraic Riccati equation in $\mathcal{L}(H)$ [1]

$$(3.22) \quad A^* P_\infty + P_\infty A = P_\infty \Sigma P_\infty - \Lambda$$

in the sense that the left-hand side has a bounded extension equal to that of the right-hand side. Such an operator will be approximated by means of the solution of the following algebraic Riccati equation in $\mathcal{L}(V_n)$:

$$(3.23) \quad (\Pi^n A \Pi_n^*)^* P_\infty^{[n]} + P_\infty^{[n]} (\Pi_n A \Pi_n^*) = P_\infty^{[n]} \Pi_n \Sigma \Pi_n^* P_\infty^{[n]} - \Pi_n \Lambda \Pi_n^*.$$

THEOREM 5. Assume that Π_n is a projection on the finite-dimensional subspace V_n in $D(A)$ converging strongly to the identity in H , such that (3.8) holds and

$$(3.24) \quad \|T(t)\| \leq e^{-\omega t}$$

where

$$(3.25) \quad \omega > \|C\| \|B\|.$$

Then (3.22) and (3.23) have a unique nonnegative self-adjoint solution and as $n \rightarrow \infty$,

$$(3.26) \quad \|P_\infty^{[n]} - \Pi_n P_\infty \Pi_n\|_{\text{HS}} \rightarrow 0.$$

Proof. Since T satisfies (3.24), (3.22) has a unique solution which satisfies [12]:

$$(3.27) \quad P_\infty x = \int_0^\infty T(s) [\Lambda - P_\infty \Sigma P_\infty] T^*(s) x \, ds$$

so that

$$(3.28) \quad P_\infty \leq \int_0^\infty T(s) \Lambda T^*(s) \, ds$$

which implies, by (3.24),

$$(3.29) \quad \|P_\infty\|_{\text{HS}} \leq (2\omega)^{-1} \|\Lambda\|_{\text{HS}}.$$

Equation (3.24) is equivalent to $\omega I - A$ being dissipative; by taking projections, we obtain the same for $\{T_n(t)\}$, $n = 1, 2, \dots$ so that (3.23) has a unique solution, and we also establish the bounds (3.28) and (3.29) for $P_\infty^{[n]}$.

Moreover, by duality with optimal control on an infinite interval [12], we obtain that if $P(t; \theta)$ is the solution to (3.5) corresponding to the initial covariance $\theta \geq P_\infty$,

$$(3.30) \quad 0 \leq P_\infty - P(t; \theta) \leq Z(t) \theta Z^*(t)$$

where $Z(t)$ is the closed-loop semigroup of operators on H generated by $A - P_\infty \Sigma$, with the bound

$$(3.31) \quad \|Z(t)\| \leq e^{(-\omega + \|\Sigma\| \|P_\infty\|)t} \leq e^{-(\omega - \|\Sigma\| \|\Lambda\|/2\omega)t} \leq e^{-\sigma t}$$

where $\sigma > 0$ by our assumption on ω . Moreover, from (3.30), (3.31), by choosing

$$\theta = \int_0^\infty T(s) \Lambda T^*(s) \, ds,$$

we get

$$(3.32) \quad \|P_\infty - P(t)\|_{\text{HS}} \leq \|\theta\|_{\text{HS}} e^{-2\sigma t} \leq (2\omega)^{-1} \|\Lambda\|_{\text{HS}} e^{-2\sigma t}.$$

An analogous bound is established for the solution $P^{[n]}(t)$ of (3.10) with respect to the approximate semigroups and perturbations by using

$$\theta_n = \int_0^\infty T_n(s) \Pi_n \Lambda \Pi_n T_n^*(s) \, ds.$$

Now we have

$$(3.33) \quad \begin{aligned} \|P_\infty^{[n]} - \mathcal{J}_n P_\infty\| &\leq \|P_\infty^{[n]} - P_n^{[n]}(t; \theta_n)\|_{\text{HS}} \\ &+ \|P_n^{[n]}(t; \theta_n) - \mathcal{J}_n(P(t; \theta))\|_{\text{HS}} + \|P(t; \theta) - P_\infty\|_{\text{HS}} \end{aligned}$$

so that the first and the last term can be made arbitrarily small uniformly in n by choosing t sufficiently large. The problem is therefore reduced to a bounded interval. But

$$(3.34) \quad \begin{aligned} \|P^{[n]}(t; \theta_n) - \mathcal{J}_n(P(t; \theta))\|_{\text{HS}} &\leq \|P^{[n]}(t; \theta_n) - P^{[n]}(t; \mathcal{J}_n(\theta))\|_{\text{HS}} \\ &+ \|P^{[n]}(t; \mathcal{J}_n(\theta)) - \mathcal{J}_n(P(t; \theta))\|_{\text{HS}}. \end{aligned}$$

The latter term goes to zero as $n \rightarrow \infty$ by Theorem 3; the same result is obtained for the former once it is shown that $\|\theta_n - \mathcal{J}_n(\theta)\|_{\text{HS}} \rightarrow 0$, as $n \rightarrow \infty$.

For this we observe that for any $\varepsilon > 0$ there exists L such that

$$(3.35) \quad \|\theta_n - \mathcal{J}_n(\theta)\|_{\text{HS}} \leq \frac{\varepsilon}{2} + \int_0^L \|\mathcal{R}_{T_n}(s)(\mathcal{J}_n \Lambda) - \mathcal{J}_n(\mathcal{R}_T(s)\Lambda)\|_{\text{HS}}$$

which can be made smaller than $\varepsilon/2$ by choosing n large enough according to Lemma 3. \square

We do not know whether the previous result can be extended, at least to the simply stable case. This is because it seems difficult to establish a uniform bound like (3.31) on the closed-loop semigroups in our generality (some cases are trivial, e.g., $\Sigma = I$). The problem still remains difficult with reference to specific classes of systems and approximation schemes, as [12] shows in the case of averaging splines for hereditary systems.

4. Numerical results. This last section is devoted to the numerical study of the following example.

Let us consider the hereditary system

$$(4.1) \quad \dot{z}(t) = \sum_{j=0}^d A_j z(t - h_j) + \int_{-r}^0 A_{01}(s) z(t+s) ds + B_0 \omega_1(t), \quad t \in [0, T],$$

$$(4.2) \quad \begin{aligned} z(0) &= z_0, \quad z(s) = z_1(s), \quad -1 < s < 0, \\ y(t) &= C_0 z(t) + G_0 \omega_2(t), \quad t \in [0, T] \end{aligned}$$

where $z(t) \in \mathbb{R}^N$, A_j , $j = 1, \dots, d$ are $N \times N$ matrices, $0 = h_0 < \dots < h_d = 1$, A_0 is a $N \times N$ matrix of functions in $L^2(-r, 0)$, B_0 , C_0 , and G_0 are matrices of dimensions $N \times p$, $q \times N$, and $q \times r$, respectively, whereas ω_1 and ω_2 are assumed to be standard \mathbb{R}^p - and \mathbb{R}^r -valued white noises, respectively. For any $(z_0, z_1) \in M^2 = \mathbb{R}^N \times L^2(-r, 0; \mathbb{R}^N)$ there exists a unique absolutely continuous function z such that (4.1) holds almost everywhere [11], whose evolution can be described in the following way. Let us first define the operator A on M^2 by

$$(4.3) \quad \begin{aligned} D(A) &= \{(z_0, z_1) \in M^2: z_1 \in H^1(-r, 0; \mathbb{R}^N), z_0 = z_1(0)\}, \\ A(z_0, z_1) &= \left(A_0 z_0 + \sum_{j=1}^d A_j z_1(-h_j) + \int_{-r}^0 A_{01}(s) z_1(s) ds, \dot{z}_1 \right). \end{aligned}$$

Then A is the generator of a strongly continuous semigroup $\{T(t)\}$ on M^2 and the solution $z(t)$ of (4.1) is given by the first component of

$$(4.4) \quad x(t) = T(t)x_0 + \int_0^t T(t-s)B\omega(s) ds, \quad t \in [0, T]$$

where

$$B = \begin{bmatrix} B_0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \omega = (\omega_1, \omega_2)' \in L^2(0, T; \mathbb{R}^{p+r}), \quad x = (z_0, z_1).$$

We can write (4.2) in a similar form:

$$(4.5) \quad y(t) = Cx(t) + G\omega(t), \quad t \in [0, T]$$

with $C = [C_0 \ 0]$, $G = [0 \ G_0]$. It is clear that we can suppose that $G_0 G_0^T = I$ without loss of generality and that B is Hilbert-Schmidt, having finite-dimensional range.

Therefore we have a system of the type described in § 3, and given the mean vector and the covariance operator of the initial state x_0 , the best linear estimate of $x(t)$ (whose first component is the best linear estimate of $z(t)$ defined in (4.1)) can be obtained by the infinite-dimensional Kalman filter.

As far as approximations are concerned, we note that estimate (3.7) holds for $\{T(t)\}$, provided the following equivalent scalar product $[\cdot, \cdot]$ is defined on $M^2[2]$, [4]:

$$[(z_0, z_1), (w_0, w_1)]_0 = z_0^T w_0 + \int_{-1}^0 z_1^T(s) w_1(s) g(s) ds,$$

$$g(s) = \begin{cases} 1, & -1 \leq s < -h_{d-1}, \\ 2, & -h_{d-1} \leq s < -h_{d-2}, \\ \vdots & \\ d, & -h_1 \leq s \leq 0. \end{cases}$$

We suppose that, for each n , Π_n is the corresponding orthogonal projection on the subspace V_n generated by the following piecewise linear splines for $i = 1, \dots, N$:

$$v_{i,0}^{(n)}(t) = 1_{[-r/n, 0]}(t) \left\{ 1 + \frac{n}{r} t \right\} e_i,$$

$$v_{i,m}^{(n)}(t) = -1_{[-mr/n, -(m-1)r/n]}(t) \{m-1 + (nx)/r\} 1_{[-(m+1)r/n, -mr/n]}(t) \\ \cdot \{m+1 + (nx)/r\} e_i, \quad m = 1, \dots, 2^n - 1,$$

$$v_{i,n}^{(n)}(t) = -1_{[-1, -1+r/n]}(t) \{n-1 + (nx)/r\} e_i$$

where e_i , $i = 1, \dots, N$, is the canonical basis of R^N and we are using the identification of $H^1(-1, 0; R^N)$ with $D(A)$. For such a scheme the convergence condition (3.8) holds [2] so that the approximation theorems in § 3 are valid.

Moreover, we remark that, in general, piecewise linear splines will not be in $D(A^*)$ and this will cause the problems addressed in the Introduction, as far as approximations of optimal controls are concerned. In [13] these linear splines are modified to overcome these problems.

To implement the approximate filter we need to compute the matrix representation of the operators with respect to the spline basis above. (For the elements of the matrices W_n and \tilde{A}_n we refer to [2].) Furthermore, we note that the form of the matrices $\tilde{\Lambda}_n$, $\tilde{\Sigma}_n$, $\tilde{\Gamma}_n$ (where in the latter the natural basis $\{\psi_j\}$ is taken on the output space R^q) does not depend on n . In fact, if the spline basis $\{v_{i,l}^{(n)}\}$ is ordered with respect to i , for each value of l in increasing order, such matrices have the form

$$\tilde{\Lambda}_n = \begin{bmatrix} B_0 B_0^T & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{\Sigma}_n = \begin{bmatrix} C_0^T C_0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{\Gamma}_n = \begin{bmatrix} C_0^T \\ 0 \end{bmatrix}.$$

In our numerical example we have considered the problem of estimating the position and velocity of an oscillator with retarded damping and restoring forces. The system is described by the equation

$$\ddot{x}(t) + 16x(t) + 2\dot{x}(t-0.2) - 10x(t-0.2) = 0.01\omega^1(t)$$

with initial conditions

$$x(s) = 1, \quad \dot{x}(s) = 0, \quad -0.2 \leq s \leq 0$$

and the output is given by

$$y(t) = x(t) + 0.3\omega^2(t)$$

where $\omega^1(t)$ and $\omega^2(t)$ are independent standard Gaussian white noises. The previous equation can be rewritten as the following first-order system:

$$\dot{x}(t) = A_0 x(t) + A_1 x(t-0.2) + B_0 \omega^1(t)$$

with

$$x(t) = \begin{bmatrix} x(t) \\ \dot{x}(t) \end{bmatrix}, \quad A_0 = \begin{bmatrix} 0 & 1 \\ -16 & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 0 \\ 10 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0.01 \end{bmatrix}.$$

The system dynamics were simulated according to the scheme

$$x(k\Delta) = e^{A_0\Delta} x((k-1)\Delta) + \frac{\Delta}{2} (e^{A_0\Delta} A_1 x((k-1)\Delta - 0.2) + A_1 x(k\Delta - 0.2)) + z_k,$$

$$y(k\Delta) = x(k\Delta) + 0.3\omega_k^2$$

where $\Delta = 0.01$, $k = 1, 2, \dots, 240$, and $\{z_k\}, \{\omega_k^2\}$ are independent zero mean Gaussian sequences with covariances $\int_0^\Delta a^{A_0\tau} B_0 B_0^T e^{A_0^T \tau} d\tau$ and 1, respectively, which have been generated using the NAG FORTRAN subroutine G05DDF.

The filter was initialized with complete information about the state, i.e., $m_0 = (1, 0)$, $P_0 = 0$, so that $\hat{x}_{c(i,l)}^{[n]} = \delta_{1i}$ and $\tilde{P}_0^{[n]} = 0_{2(n+1)}$ (the zero matrix of order $2(n+1)$) for $n = 3, 5$ ($n+1$ is the number of splines for each state dimension).

The state estimate $\hat{x}_c^{[n]}(t)$ and its error covariance matrix $\tilde{P}^{[n]}(t)$ were numerically computed by integrating (3.20) and (3.21) by means of the NAG FORTRAN subroutine D02EAF which uses a variable-order, variable-step Gear method. The whole numerical example was carried out on a CDC 7600 computer.

To evaluate the filter performance we have computed the following error statistics:

$$\sigma_p^2 = \frac{1}{240} \sum_{k=1}^{240} [x(k\Delta) - \hat{x}^{[n]}(k\Delta)],$$

$$\sigma_v^2 = \frac{1}{240} \sum_{k=1}^{240} [\dot{x}(k\Delta) - \hat{\dot{x}}^{[n]}(k\Delta)]^2$$

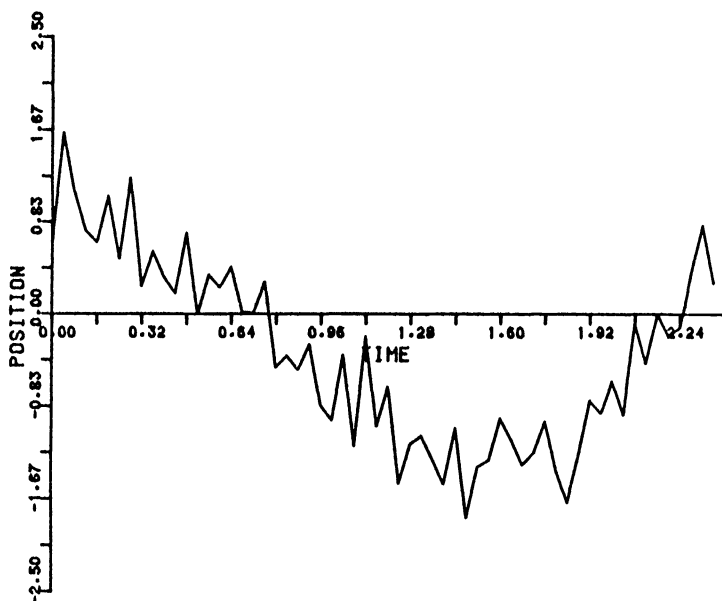


FIG. 1. Computer-produced plot. Noisy position of the oscillator.

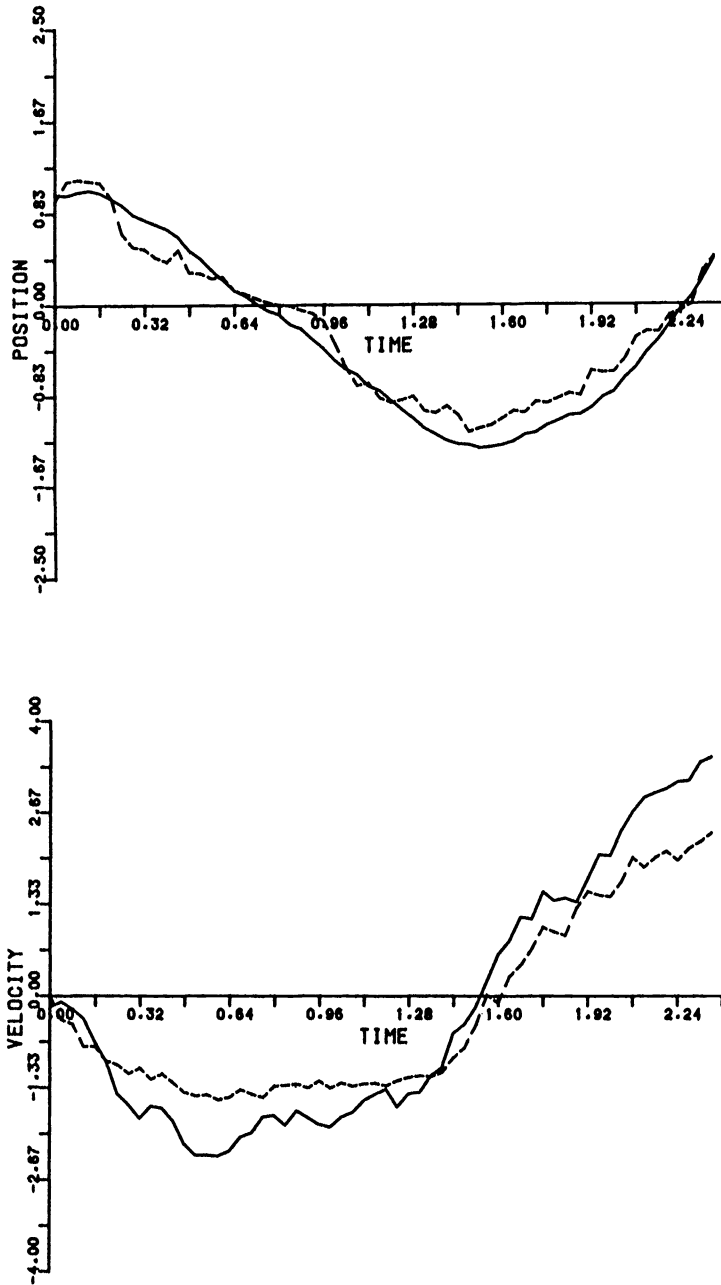


FIG. 2. Computer-produced plots. Diagrams showing the real (continuous line) and the estimated (dashed line) position and velocity obtained considering $N = 4$ splines. The signal-to-noise ratio improvement for the position is 5.12 db, the residual variance for the velocity is 0.34.

and the signal-to-noise ratio improvement:

$$\eta = 10 \log_{10} \left(\frac{\text{variance of observation noise}}{\sigma_p^2} \right) = 10 \log_{10} \left(\frac{0.09}{\sigma_p^2} \right).$$

The results of the simulation are reported in Figs. 1-3.

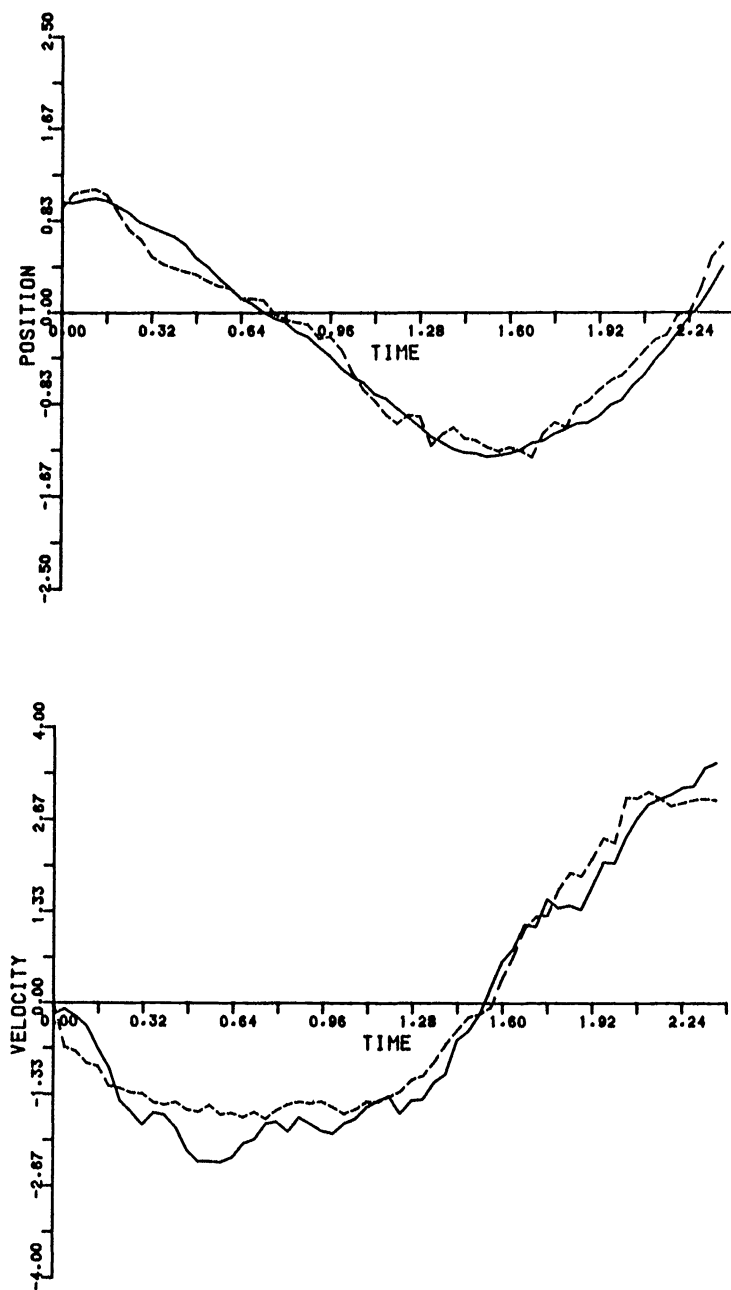


FIG. 3. Computer-produced plots. Diagrams showing the real (continuous line) and the estimated (dashed line) position and velocity obtained considering $N=6$ splines. The signal-to noise ratio improvement for the position is 7.01 db, the residual variance for the velocity is 0.13.

REFERENCES

- [1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, Berlin, New York, 1976.
- [2] H. T. BANKS AND F. KAPPEL, *Spline approximations for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496-522.
- [3] H. T. BANKS AND K. KUNISCH, *An approximation theory for nonlinear partial differential equations with application to identification and control*, SIAM J. Control Optim., 20 (1982), pp. 815-849.

- [4] H. T. BANKS, G. J. ROSEN, AND K. ITO, *A spline-based technique for computing Riccati operators and feedback controls in regulator problems for delay equations*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 830–855.
- [5] A. BENSOUSSAN, *Filtrage optimal des systèmes linéaires*, Dunod, Paris, 1971.
- [6] R. F. CURTAIN AND A. J. PRITCHARD, *The infinite dimensional Riccati equation*, J. Math. Anal. Appl., 47 (1974), pp. 43–57.
- [7] ———, *Infinite Dimensional Linear System Theory*, Springer-Verlag, Berlin, New York, 1978.
- [8] P. D'ALESSANDRO, A. GERMANI, AND M. PICCIONI, *Relationships between measures induced by Itô and white noise linear equations*, Math. Comput. Simulation 26 (1984), pp. 368–372.
- [9] M. C. DELFOUR, *The linear quadratic optimal control problem for hereditary differential systems: theory and numerical solution*, Appl. Math. Optim., 3 (1977), pp. 101–162.
- [10] M. C. DELFOUR AND R. S. K. MITTER, *Hereditary differential systems with constant delays, I—General case*, J. Differential Equations, 12 (1972), pp. 213–235.
- [11] J. S. GIBSON, *An analysis of optimal model regulation: convergence and stability*, SIAM J. Control Optim., 19 (1981), pp. 686–707.
- [12] ———, *Linear quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 31 (1983), pp. 95–139.
- [13] F. KAPPEL AND D. SALAMON, *Splines approximation for retarded systems and the Riccati equation*, Tech. Summary Report 2680, University of Wisconsin, Madison, WI, 1984.
- [14] J. L. KELLEY, *General Topology*, Springer-Verlag, Berlin, New York, 1955.
- [15] T. G. KURTZ, *Extension of Trotter's operator semigroup approximation theorems*, J. Funct. Anal., 3 (1969), pp. 111–132.

ON THE OPTIMAL DESIGN OF STRUCTURES SUBJECTED TO PERIODIC BASE EXCITATIONS*

F. E. UDWADIA[†], F. MANTEL[‡], AND M. DRYJA[§]

Abstract. This paper considers the optimal apportionment of the stiffness of a building structure modeled as an undamped shear beam when subjected to a periodic base excitation of frequency ω . A suitable "cost" associated with the stiffness distribution is minimized subject to constraints on the lowest natural vibration frequency of the system, the base shear generated by the periodic excitations, and the given lower-bound stiffness distribution. Closed form solutions for such constrained optimization problems are generally very difficult to obtain; numerical techniques suffer from a host of problems. The paper uses Farkas's Theorem to investigate the underlying structure of the optimization problem and obtains closed-form solutions in several cases of engineering importance.

Key words. optimal design, continuous system, stiffness distribution, frequency and base shear constraints, closed-form solutions, Farkas's theorem

AMS(MOS) subject classification. 49

Introduction. In the field of earthquake engineering we often model tall building structures by one-dimensional shear of bending beams. Such beams are assumed to be fixed at one end (ground level) and free at the other. The effect of strong earthquake ground shaking is then modeled by a suitable base excitation of the system. In most structural systems, the mass distribution of the structure with height is reasonably well prescribed by the building codes and depends on the use to which the structure will be put. The structural designer is then left with the problem of apportioning the stiffness of the structure as a function of height, so that the system is safe, not only under the static gravity loading created by the assumed mass distribution throughout the structure, but also under the dynamic loads induced by the base excitations from earthquake ground shaking that the structure is likely to experience during its useful life. In the design for earthquake safety, a critical parameter that the analyst must consider is the shear force induced at the base of the structure by the ground shaking. Most building codes around the world require this base shear to be calculated and within safe limits for any particular structural configuration on which the designer settles.

In this paper we broach this problem of optimally distributing the stiffness (with height) of a structure so that the base shear generated is less than some fixed value F_0 . We assume that the base motion is harmonic with frequency ω and amplitude A_0 , and that the system is required to have its lowest fundamental frequency ω_0 , to be greater than the base excitation frequency ω . This prevents resonance. In addition, we take the stiffness distribution to be constrained from below, by the continuous function $k^0(x)$. This function is presumably obtained by considering the safety requirements pertinent to the structure under the static loading condition.

The optimal design of structures subjected to harmonic excitations was first studied by Icerman [1] when he considered a "response constraint" in the form of the virtual work of the load amplitude on the displacement amplitude at its point of application.

* Received by the editors December 30, 1985; accepted for publication (in revised form) November 15, 1987. This work was funded in part by the Rocket Propulsion Laboratory, Edwards Air Force Base, California, and in part by Design, Analysis, and Software, Inc., Pasadena, California.

[†] University of Southern California, Los Angeles, California 90089-1114.

[‡] University of California, Irvine, California 92717.

[§] University of Southern California, Los Angeles, California 90089.

Plaut [2] and Huang [3] extended the minimum-weight design to structures subjected to periodic excitations by using constraints on the prescribed deflections at a specified point of the structure. Johnson and Rizzi [4] studied a cantilever bar excited by a harmonic tip force and subject to a maximum allowable stress constraint. They discretized the bar into finite elements, and used variational methods to obtain numerical solutions to their problem. They show that the global minimum of the objective function may be difficult to obtain. In a previous study, Johnson [5] used a displacement constraint to obtain the optimal weight design, again numerically. Recently, Ivanova [6] has used the variational method with constraints on the stiffness to obtain the gradient of the objective function with respect to the weight distribution. This method, which is most commonly used in such problems, leads to a nonlinear programming problem that needs to be worked out numerically. Here, we study a different variety of optimum design problems where the mass distribution is given a priori, and a suitably defined "cost" associated with the stiffness distribution must be minimized. Constraints are imposed on the values of the base shear, the natural vibration frequency of the system, and on the stiffness distribution of the system. Such optimization problems, as indicated above, have been attacked in the past by the use of variational calculus, which in turn transforms the problem to one involving nonlinear programming. Numerical results are then obtained by one of several standard computational procedures. Often the nature of the constraints make the problem notoriously difficult to solve numerically and lead to numerical results which may, at best, be uncertain. Comparing the characteristics of the problem addressed in this paper and our method of approach with some of the work done in the past, we note the following.

(1) The number of constraints that we deal with here are more than those considered by, for example, Ivanova. We obtain closed-form solutions rather than expressions for the gradients which then require numerical computations, in general.

(2) The variational calculus method is difficult to implement when there are frequency constraints. Even if it were to be implemented in an approximate manner, the resulting equations for the adjoint variables, in our problem, would lead to a nonconstant coefficient, nonlinear differential boundary value problem. At each step in the optimization process, these adjoint variables would have to be numerically solved for, to obtain the gradient of the objective function. This gradient would then be used in the numerical minimization scheme. The stiffness distribution would need to be discretized. The handling of the nonlinear constraints would lead to a difficult numerical optimization problem, whose results may leave us yet unsatisfied because of the vulnerability of the computational method to local minima. Besides, such numerical procedures seldom provide insight (unless we have large computing budgets) into the basic structure of the optimization problem.

(3) We expose the structure of the optimization problem in this paper through the use of the Farkas Theorem, which yields closed-form solutions for the global optimum under most situations of engineering interest. Consequently, no numerical schemes are involved, and we do not discretize the stiffness distribution to obtain the minimum cost. The search is carried out in function space. The method does not require us to solve either the nonconstant coefficient differential boundary value problem characterizing the dynamic response of the system, or the nonlinear adjoint equation.

No doubt the approximation of assuming the base excitation as harmonic may be unrealistic in the seismic environment. We use it here, following common engineering practice, realizing that the ground excitations are perhaps best modeled by a nonstationary stochastic process [7]. Though the problem statement in this paper has been

motivated by an application from the field of aseismic design of structures, it is equally applicable to the design of cantilever beams subjected to harmonic base excitations. The results of this work will find wide application in the area of mechanical, nuclear, and aerospace engineering. An area of particular interest may be the design of machine foundations.

1. Problem definition. Consider a structure whose relative response $u(x, t)$ is modeled by the following differential problem:

$$(1.1) \quad \begin{aligned} \rho(x)u_{tt} &= (k(x)u_x)_x - \rho(x)\ddot{u}_g(t), \quad x \in (0, 1), \quad -\infty < t < \infty, \\ u(0, t) &= 0, \\ k(x)u_x(x, t)|_{x=0} &= 0 \end{aligned}$$

with

$$u_g(t) = A_0 \cos \omega t, \quad -\infty < t < \infty, \quad A_0 > 0, \quad \omega > 0.$$

The end $x = 0$ (is assumed fixed while the end $x = 1$, is assumed stress-free. The positive function $k(x)$ represents the distribution of the shear stiffness along the height of the structure, and $\ddot{u}_g(t)$ represents the periodic base excitation (see Fig. 1).

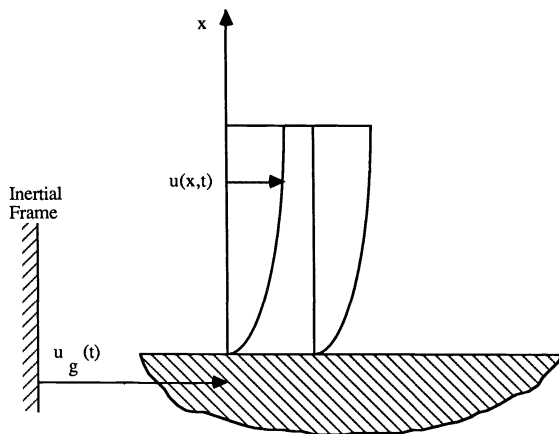


FIG. 1

Given the function $\rho(x) \in C[0, 1]$ with $\rho(x) \geq \rho_0 > 0$, our aim is to find a function $k(x) \in C[0, 1]$ such that the cost of penalty defined by

$$(1.2) \quad J(k) = \int_0^1 g^2(x)k(x) dx$$

is minimized, subject to the following three constraints:

- (1.3a) The lowest natural frequency of the system, ω_0 , which depends on the choice of $k(x)$, is greater than ω , the given forcing frequency. This condition prevents the system from undergoing resonance.
- (1.3b) The base shear induced by the excitation, $u_g(t)$, is less than or equal to a fixed value, F_0 . This constraint is relevant to design procedures followed in assessing the safety of the soil-structural system.

(1.3c) $k(x) \geq k^0(x) \geq \varepsilon > 0$, where $k^0(x)$ is a prescribed function belonging to $C^1[0, 1]$ and ε is a positive number. This lower bound on the stiffness distribution comes about from static design conditions.

The function $g^2(x)$ is a weighting function and $J(k)$ provides the cost of choosing a given stiffness distribution $k(x)$. The function $g^2(x)$ is often a monotone increasing function; it is increasingly more expensive, from a construction standpoint, to add stiffness at higher and higher levels in a structure. The condition that the end, $x = 1$, is stressfree would imply that there is no need to provide stiffness at that location, and therefore it appears appropriate not to penalize the stiffness at $x = 1$. We therefore define the positive function $g(x) \in C[0, 1]$ such that $g^2(x) > 0$, $x \in [0, 1)$ and $g(1) = 0$.

The continuity requirements on $\rho(x)$ and $k(x)$ are imposed on physical grounds and could be relaxed.

It would be worthwhile at this point to express the constraints (1.3) in a more formal fashion. Let \hat{H} be a Sobolev space of the form $\hat{H} := \{v \in H^1(0, 1): v(0) = 0\}$. Then the lowest natural frequency can be expressed as

$$(1.4) \quad \omega_0^2(k) = \min_{z \in \hat{H}} \left[\frac{(kz_x, z_x)}{(\rho z, z)} \right],$$

and constraint (1.3a) can be written as

$$(1.3'a) \quad \omega_0^2(k) > \omega^2.$$

We are thus restricted to considering those designs $k(x)$ which satisfy (1.3'a).

The shear force induced at the base by the base excitation, for a given design $k(x)$ is given by Newton's Law as

$$L(t) = - \int_0^1 \rho(u_{tt} + \ddot{u}_g) dx$$

where $u(x, t)$ is the solution of (1.1) corresponding to the chosen function $k(x)$. For the steady-state response of the system, defined by

$$(1.5a) \quad u(x, t) = U(x) \cos \omega t,$$

and the steady-state base shear force, defined by

$$(1.5b) \quad L(t) = F \cos \omega t,$$

we get

$$(1.6) \quad F = \omega^2(\rho, U) + M_t \omega^2 A_0$$

where $M_t \triangleq \int_0^1 \rho(x) dx$, is the total mass of the structure and $(\rho, U) \triangleq \int_0^1 \rho U dx$.

If the constraint (1.3b) is to be satisfied, we are further restricted to those functions $k(x)$ for which

$$(1.3'b) \quad |F(U(k))| \leq F_0$$

for some given fixed value of F_0 . We note that the shear force F depends on the solution $U(x)$ of (1.1), which in turn depends on the choice of $k(x)$. From here on we shall denote the dependence of $U(x)$ on the choice of $k(x)$ by $U(k)$. Let us introduce the set \mathcal{H} of functions $k(x) \geq k^0(x) \in C^1(0, 1)$ such that the constraints (1.3'a) and (1.3'b) are satisfied by each element of the set. The optimization problem may now be stated as follows:

Find $\hat{k}(x) \in \mathcal{H}$ such that

$$(1.7) \quad J(\hat{k}) = \min_{k \in \mathcal{H}} J(k).$$

We next provide three useful results pertinent to the steady-state response of the shear beam system. In the interest of brevity the results have been provided without proof. They can be derived using [8]. Physical interpretation of the results will be provided as we go along.

LEMMA 1. *Consider the shear beam problem represented by the differential problem (1.1). Let $k(x)$ be such that the lowest fundamental natural vibration frequency of the system ω_0 is greater than the given frequency of the base excitation ω . Then the steady-state response $U(x)$ defined by relation (1.5a) is such that $U_x(x) \neq 0$ for $x \in [0, 1]$.*

Lemma 1 implies the following: Consider a shear beam which is subjected to a base excitation whose frequency, ω , is less than the lowest natural frequency of vibration of the beam. Then there is no point in the interior of the beam at which the stresses, induced by the steady-state vibratory response of the beam to this base excitation, are zero.

LEMMA 2. *Consider the steady-state response of the system described by (1.1). Then as long as the forcing frequency ω is less than ω_0 , where ω_0 is the lowest natural frequency of the system, the inner product (ρ, U) is always positive.*

This result implies that the base shear force F defined in relation (1.6) is always positive. We note that the shear force induced at the base is a consequence of (a) the rigid body motion induced in the entire beam, given by the term $M_t A_0 \omega^2$; and (b) the vibratory response of the beam, given by $\omega^2(\rho, U)$. We have thus shown that as long as the base excitation frequency ω is less than the lowest fundamental vibration frequency of the system, these two contributions are always in phase and augment each other. Further, the base shear is always in phase with the base displacement.

Thus if condition (1.3'a) is satisfied, then condition (1.3'b) can be restated as

$$(1.8) \quad |F(U(k))| = F(U(k)) = \omega^2 M_t A_0 + \omega^2(\rho, U) \leq F_0.$$

LEMMA 3. *Consider the system represented by (1.1). If the forcing frequency of the base excitation $\omega < \omega_0$, then the steady-state response function $U(x)$ is a strictly positive monotone increasing function in $(0, 1)$.*

2. Inequality constraints.

LEMMA 4. *Let $k(x)$ be a candidate design which satisfies condition (1.3'a). Let $u(x, t) = U(x) \cos \omega t$ be the solution of (1.1) for this candidate design. Then for any function $W(x) \in \hat{H}(0, 1)$, we have*

$$(2.1) \quad (kW_x, W_x) - \omega^2(\rho W, W) - 2A_0\omega^2(\rho, W) + A_0\omega^2(\rho, U) \geq 0.$$

Moreover, the equality holds for $W(x) = U(x)$.

Proof. Since $u = U(x) \cos \omega t$ is a solution of (1.1) we have

$$(2.2) \quad (k(x)U_x(x))_x + (A_0 + U)\rho\omega^2 = 0, \quad U(0) = 0, \quad U_x(1) = 0.$$

Then $U(x) \in \hat{H}(0, 1)$. Let $W = U + V$. Inserting $U = W - V$ in (2.2) we get

$$(2.3) \quad (k(x)W_x(x))_x + (A_0 + W)\rho\omega^2 = (k(x)V_x(x))_x + \omega^2\rho V.$$

Multiplying by $W(x)$ in the $L_2(0, 1)$ sense and integrating by parts, we get

$$(2.4) \quad (k(x)W_x, W_x) - \omega^2(\rho W, W) - A_0\omega^2(\rho, W) = (kV_x, W_x) - \omega^2(\rho V, W).$$

Denoting the right-hand side of (2.4) by P we can restate it, using $W = U + V$, as

$$P = P_1 + P_2$$

where

$$P_1 = (kV_x, V_x) - \omega^2(\rho V, V),$$

$$P_2 = (kV_x, U_x) - \omega^2(\rho V, U).$$

Noting that $V(x) \in \hat{H}(0, 1)$, $P_1 \geq 0$ because of (1.3'a) and (1.4). Also taking the inner product of (2.2) with $V(x)$ we get

$$(2.5) \quad (kU_x, V_x) - \omega^2(\rho U, V) = A_0 \omega^2(\rho, V).$$

Thus, $P_2 = A_0 \omega^2(\rho, V) = A_0 \omega^2(\rho, W) - A_0 \omega^2(\rho, U)$. Relation (2.4) now yields the result. The equality, which holds for $W(x) = U(x)$, is obvious when (2.2) is multiplied by $U(x)$ in $L_2(0, 1)$.

LEMMA 5. Let $k(x)$ be such that it satisfies conditions (1.3'a) and (1.8), i.e.,

$$\omega_0^2(k) > \omega^2$$

and

$$F(U(k)) \leq F_0.$$

Then for any $W \in \hat{H}(0, 1)$,

$$(2.6) \quad (kW_x, W_x) - \omega^2(\rho W, W) - 2A_0 \omega^2(\rho, W) + A_0 F_0 - M_t \omega^2 A_0^2 \geq 0.$$

Proof. When we use (1.6) and (1.8) in (2.1), the result follows.

LEMMA 6. If (2.6) is satisfied for all $W \in \hat{H}(0, 1)$ with $F(k) \leq F_0$, then $k(x)$ is such that

$$(2.7) \quad \omega_0^2 \geq \omega^2 - \Delta$$

where

$$(2.8) \quad \Delta = \frac{A_0[F_0 - F(k)]}{(\rho V^0, V^0)}$$

where V^0 is the eigenfunction corresponding to the lowest eigenvalue ω_0^2 of

$$(2.9) \quad \begin{aligned} (k(x) V_x(x))_x + \lambda^2 \rho(x) V(x) &= 0, \quad x \in (0, 1), \\ V(0) &= V_x(1) = 0. \end{aligned}$$

Proof. Let $W = U + V$ where $U(x)$ is the solution of

$$\begin{aligned} (k(x) U_x(x))_x + (A_0 + U(x)) \rho \omega^2 &= 0, \quad x \in (0, 1), \\ U(0) &= 0, \quad U_x(1) = 0 \end{aligned}$$

for the chosen $k(x)$. Then by (2.6), and using (1.6), we get

$$(2.10) \quad \begin{aligned} (k(U_x + V_x), U_x + V_x) - \omega^2(\rho U + \rho V, U + V) - 2A_0 \omega^2(\rho, U + V) \\ + A_0 \omega^2(\rho, U) + A_0[F_0 - F(k)] \geq 0. \end{aligned}$$

Since U is the solution of (1.12) we have

$$(kU_x, U_x + V_x) - A_0(\rho, U + V) \omega^2 - \omega^2(\rho U, U + V) = 0.$$

Using this in (2.10) and again noting that U is a solution of (2.2) we get

$$(kV_x, V_x) - \omega^2(\rho V, V) \geq -A_0[F_0 - F(k)].$$

Thus

$$\frac{(kV_x V_x)}{(\rho V, V)} \geq \omega^2 - \frac{A_0[F_0 - F(k)]}{(\rho V, V)} \quad \text{for all } V \in \hat{H}(0, 1).$$

Choosing V^0 to be the eigenfunction corresponding to $\lambda = \omega_0$ for (2.9) we get

$$\omega_0^2 \geq \omega^2 - \Delta.$$

COROLLARY 1. *If (2.6) is satisfied for all $W \in \hat{H}(0, 1)$ with $F(k) = F_0$, then $k(x)$ must be such that $\omega_0^2 \geq \omega^2$.*

Proof. Since $k(x)$ is such that $F_0 = F(k)$, $\Delta = 0$.

THEOREM 1. *Let $k(x)$ belong to the set \mathcal{K} . If an optimal design $\hat{k}(x)$ exists such that it minimizes $J(k)$ as defined in (1.7), then $\hat{k}(x)$ must satisfy the following relations:*

$$(2.11a) \quad (\hat{U}_x, k_1(x) \hat{U}_x) + A_0 \alpha(\hat{k}) \geq 0,$$

$$(2.11b) \quad k_1 + \hat{k}(x) - k^0(x) \geq 0,$$

$$(2.11c) \quad (g^2(x), k_1(x)) \geq 0$$

where $k_1(x) = k(x) - \hat{k}(x)$, $\hat{U}(x)$ is the solution of (2.2) corresponding to $\hat{k}(x)$, and $\alpha(k) \triangleq F_0 - F(\hat{k}) \geq 0$.

Proof. The second and third inequalities follow directly from relation (1.2) and condition (1.3c). We prove the first inequality as follows: Since $\hat{U}(x) \in \hat{H}(0, 1)$, let $W(x) = \hat{U}(x)$ in (2.6). Thus

$$(2.12) \quad (k \hat{U}_x, \hat{U}_x) - \omega^2(\rho \hat{U}, \hat{U}) - 2A_0 \omega^2(\rho, \hat{U}) + A_0 F_0 - M \omega^2 A_0^2 \geq 0.$$

Since $\hat{U}(x)$ is the solution of (2.2) for $k(x) = \hat{k}(x)$, by taking the inner product of (2.2) with $\hat{U}(x)$, we get

$$(2.13) \quad (\hat{k} \hat{U}_x, \hat{U}_x) - \omega^2(\rho \hat{U}, \hat{U}) - 2A_0 \omega^2(\rho, \hat{U}) + A_0 F(\hat{k}) - M \omega^2 A_0^2 = 0.$$

When we subtract (2.13) from (2.12) the result follows.

In order to find the optimal stiffness distribution $\hat{k}(x)$, we shall use the Farkas Theorem (see [9]). We begin by establishing the following nomenclature. In what follows we shall assume that an element $\hat{k}(x) \in \mathcal{K}$ exists such that $J(\hat{k})$ is a minimum.

3. Nomenclature and the Farkas Theorem. Let V_1 and V_2 be two Hilbert spaces and let V_1^* and V_2^* be their corresponding duals. For $x \in V_i$, $f \in V_i^*$, let $f(x) = \langle f | x \rangle_i$ denote the duality pairing bilinear form on $V_i^* \times V_i$, $i = 1, 2$. Let A be a linear bounded operator from V_1 to V_2 , and denote by A^* its transpose. Then $\langle A^* f_2 | x_1 \rangle_1 = \langle f_2 | A x_1 \rangle_2$ for any $f_2 \in V_2^*$ and any $x_1 \in V_1$.

Let M be a cone in V_1 . We define the positive polar cone of M as

$$(3.1) \quad M^+ = \{f \in V_1^*: \langle f | x_1 \rangle_1 \geq 0 \text{ for all } x_1 \in M\}.$$

Note that in [9] the set M^+ is denoted $-M^-$. Using these definitions, the Farkas Theorem states the following [9].

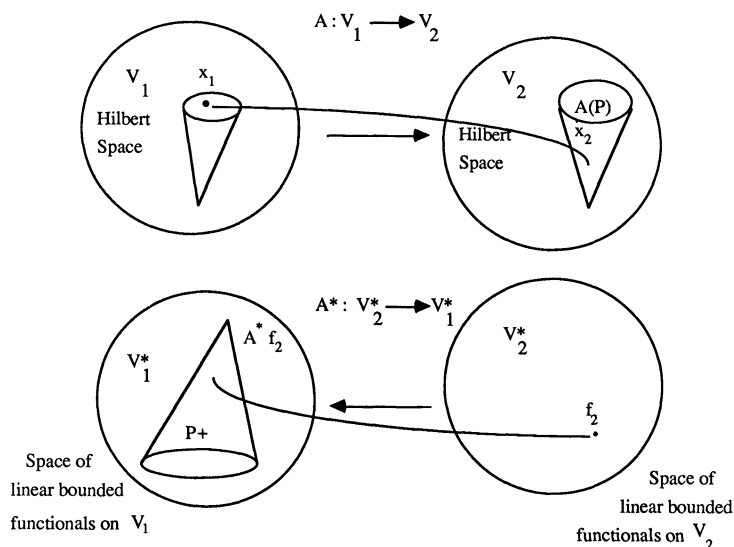
Let V_1 and V_2 be Hilbert spaces and let A be a bounded linear operator from V_1 to V_2 . Let P be a closed convex cone in V_1 . Then, the following two statements are equivalent:

$$(3.2a) \quad \text{For all } f_2 \in V_2^*, \quad A^* f_2 \in P^+ \text{ implies } \langle f_2 | x_2 \rangle_2 \geq 0,$$

$$(3.2b) \quad \exists x_1 \in P \text{ such that } A x_1 = x_2 \text{ for given } x_2 \in V_2.$$

Figure 2 presents the “geometry of the Farkas Theorem.” We note that the closed convex cone P is in the space V_1 and thus its positive polar cone P^+ belongs to the space V_1^* . $A(P)$ is a cone in the space V_2 . If we can show that $A(P)$ is a closed set in V_2 , then the Farkas Theorem states that (3.2a) and (3.2b) are equivalent assertions.

This means that an element x_2 in V_2 is in $A(P)$, i.e., x_2 is in the range of the restriction of A to P , if and only if, for arbitrary bounded linear functionals f_2 acting on V_2 , we have $f_2(x_2) = \langle f_2 | x_2 \rangle_2 \geq 0$ whenever the functional $A^* f_2$ is in P^+ of V_1^* . In turn, $A^* f_2$ is in P^+ when the image of P under $A^* f_2$ is a subset of $[0, \infty)$.



P is a closed Convex Cone in V_1 ; $A(P)$ is a closed cone in V_2
 $(\forall f_2 \in V_2^*), (A^* f_2 \in P^+ \Rightarrow f_2(x_2) \geq 0) \iff (\exists x_1 \in P \ni Ax_1 = x_2 \text{ for given } x_2 \in V_2)$

FIG. 2. Geometry of Farkas' theorem.

In order to use the theorem we shall need to define the spaces V_1 and V_2 appropriately, and also the operator A . We specify a set P in V_1 and an element $x_2 \in V_2$. We then prove several assertions, stating the following: P is a closed convex cone in V_1 (Lemma 7); A is a bounded linear operator (Lemma 8); $A(P)$ is a closed cone in V_2 (Lemma 9); $A^* f_2 \in P^+ \subseteq V_1^*$ (Lemma 10); $f_2(x_2) \geq 0$ (Lemma 11); and finally we apply Farkas's result (in Theorem 2).

DEFINITIONS OF V_1 AND V_2 . Let $V_1 = R \times H^1(0, 1)$ and define the inner product and linear space operations for $x_1 = (\nu_0, \nu(x))$, $y_1 = (\mu_0, \mu(x))$ in V_1 by

$$\begin{aligned}
 (3.3) \quad & x_1 + y_1 = (\nu_0 + \mu_0, \nu(x) + \mu(x)), \\
 & \alpha x_1 = (\alpha \nu_0, \alpha \nu(x)), \\
 & \langle x_1, y_1 \rangle_1 = \nu_0 \mu_0 + \int_0^1 \nu(x) \mu(x) dx + \int_0^1 \nu'(x) \mu'(x) dx.
 \end{aligned}$$

It is clear that V_1 is a Hilbert space. Let V_2 be the Hilbert space $L_2(0, 2)$.

DEFINITION OF P . Let

$$(3.4) \quad P = \{x_1 = (\nu_0, \nu(x)) \in V_1: \nu_0 \geq 0 \text{ and } \forall x \in (0, 1), \nu(x) \geq 0\}.$$

LEMMA 7. P is a closed convex cone in V_1 .

Proof. The proof is straightforward and is therefore omitted.

FUNCTION AND OPERATOR DEFINITION. Consider the functions $k_1(x)$, $\hat{U}_x^2(x)$, and $g^2(x) \in C[0, 1)$. We extend these functions as follows:

$$(3.5) \quad K_1(x) = \begin{cases} k_1(x), & 0 \leq x < 1, \\ 1, & 1 \leq x \leq 2, \end{cases}$$

$$(3.6) \quad \mathcal{U}_x^2(x) = \begin{cases} \hat{U}_x^2(x), & 0 \leq x < 1, \\ A_0 \alpha(\hat{k}), & 1 \leq x \leq 2, \end{cases}$$

$$(3.7) \quad G(x) = \begin{cases} g^2(x), & 0 \leq x < 1, \\ 0, & 1 \leq x \leq 2. \end{cases}$$

Define the function $D(x, \xi)$ on the rectangle $[0, 2] \times [0, 1]$ as

$$(3.8) \quad D(x, \xi) = \delta(\xi - x) + [\hat{k}(x-1) - k^0(x-1)]H(x-1)\delta(\xi - (x-1))$$

where δ is the Dirac delta distribution, and H is the unit-step function defined by

$$(3.9) \quad H(x-1) = \begin{cases} 0, & 0 \leq x < 1, \\ 1, & 1 \leq x \leq 2. \end{cases}$$

Let $A: V_1 \rightarrow V_2$ be defined by the relation

$$(3.10) \quad Ax_1 = A(\nu_0, \nu(x)) = \mathcal{U}_x^2(x) \cdot \nu_0 + \int_0^1 D(x, \xi) \nu(\xi) d\xi = x_2.$$

LEMMA 8. *The operator A defined by (3.10) is a bounded linear operator from V_1 to V_2 .*

Proof. Using (3.8) we can expand (3.10) to read

$$(3.11) \quad Ax_1 = \begin{cases} \hat{U}_x^2(x) \nu_0 + \nu(x), & x \in [0, 1), \\ A_0 \alpha(\hat{k}) \nu_0 + \nu(1) + [\hat{k}(0) - k^0(0)] \nu(0), & x = 1, \\ A_0 \alpha(\hat{k}) \nu_0 + [\hat{k}(x-1) - k^0(x-1)] \nu(x-1), & x \in (1, 2]. \end{cases}$$

It is clear that $Ax_1 \in L_2[0, 2]$ and the linearity of A can be easily shown. We show that it is bounded.

Let $x_1 = (\nu_0, \nu(x)) \in V_1$ such that

$$(3.12) \quad \|x_1\|_{V_1}^2 = \nu_0^2 + \int_0^1 \nu^2(x) dx + \int_0^1 \nu'^2(x) dx \leq 1.$$

Using (3.11), $\|x_2\|_{V_2}^2 = \|Ax_1\|_{L_2[0,2]}^2$ can be written as follows:

$$(3.13) \quad \begin{aligned} \|x_2\|_{V_2}^2 = & \nu_0^2 \int_0^1 \{ \hat{U}_x^4(x) + [\alpha(\hat{k}) A_0]^2 \} dx \\ & + \int_0^1 \{ 1 + [\hat{k}(x) - k^0(x)]^2 \} \nu^2(x) dx \\ & + 2\nu_0 \int_0^1 [\hat{U}_x^2(x) + \alpha(\hat{k}) A_0 (\hat{k}(x) - k^0(x))] \nu(x) dx. \end{aligned}$$

Since $\hat{U}_x^2(x)$, $\hat{k}(x)$, $k^0(x) \in C[0, 1)$, $\alpha(\hat{k}) A \in \mathbf{R}$, inequality (3.12) implies that $\sup_{\|x_1\|_{V_1} < 1} \|Ax_1\|_{V_2} < \infty$ and therefore A is bounded.

LEMMA 9. *If P is the closed convex cone defined by (3.4) in V_1 , then $A(P)$ is a closed cone in V_2 .*

Proof. If $x_2 \in A(P) \subseteq V_2$, then there is an $x_1 = (\nu_0, \nu(x)) \in P \subseteq V_1$ such that $Ax_1 = x_2$. If we note the definition of the operator A , it is obvious that $A(P)$ is a cone. Since A is a continuous operator, $A(P)$ is closed.

Let f_2 be a bounded linear functional on $V_2 = L_2[0, 2]$ whose action on the elements of V_2 can be expressed by the Riesz Representation Theorem as

$$(3.14) \quad f_2(*) = \langle *, K_1(x) \rangle_{V_2}$$

where $K_1(x)$ is defined in (3.5).

LEMMA 10. *Let relations (2.11a) and (2.11b) be satisfied. Let $A^*: V_2^* \rightarrow V_1^*$ be the transpose operator of A , and let P^+ be the positive polar cone of P . Then $A^*f_2 \in P^+ \subseteq V_1^*$, where f_2 is given by (3.14).*

Proof. Let $x_1 = (\nu_0, \nu(x)) \in P$. We have $\nu(\xi) \geq 0$ for any $\xi \in (0, 1)$. By (3.5) and (3.6), (2.11a) and (2.11b) become, respectively,

$$(3.15) \quad \int_0^2 K_1(x) \mathcal{U}_x^2(x) dx \geq 0$$

and

$$(3.16) \quad \int_0^2 D(x, \xi) K_1(x) dx \geq 0 \quad \text{for any } \xi \in [0, 1].$$

Multiplying both sides of (3.16) by $\nu(\xi)$ and integrating from $\xi = 0$ to $\xi = 1$, we get

$$\int_0^2 K_1(x) \left[\int_0^1 D(x, \xi) \nu(\xi) d\xi \right] dx \geq 0.$$

Noting that $\langle A^*f_2 | x_1 \rangle_1 = \langle f_2 | Ax_1 \rangle_2 = f_2(Ax_1) = \langle Ax_1, K_1(x) \rangle_{V_2}$ and using the linearity of the inner product, we get

$$(3.17) \quad \langle A^*f_2 | x_1 \rangle_1 = \nu_0 \int_0^2 K_1(x) \mathcal{U}_x^2(x) dx + \int_0^2 K_1(x) \left[\int_0^1 D(x, \xi) \nu(\xi) d\xi \right] dx.$$

Using (3.15) and (3.16) and remembering that $x_1 \in P$, so $\nu_0 \geq 0$, we have $A^*f_2 \in P^+ \subseteq V_1^*$ because $\langle A^*f_2 | x_1 \rangle_1 \geq 0$.

LEMMA 11. *The relation (2.11c) can be expressed as $\langle f_2 | x_2 \rangle_2 \geq 0$, where $x_2 = G(x)$ is defined in (3.7).*

Proof. Because $f_2 \in V_2^*$,

$$(3.18) \quad \langle f_2 | x_2 \rangle_2 = \langle x_2, K_1(x) \rangle_{V_2} = \langle G(x), K_1(x) \rangle_{V_2} = \int_0^1 k_1(x) g^2(x) dx.$$

Hence we have the result.

4. The optimal design $\hat{k}(x)$.

THEOREM 2. *If relations (2.11a) and (2.11b) imply relation (2.11c), then*

$$(4.1) \quad \nu_0 \hat{U}_x^2(x) + \nu(x) = g^2(x), \quad x \in [0, 1]$$

and

$$(4.2) \quad \nu_0 A_0 \alpha(\hat{k}) + \nu(x) [\hat{k}(x) - k^0(x)] = 0, \quad x \in (0, 1].$$

Proof. By Lemmas 7–11, we see that the Farkas Theorem is applicable. Thus there exists an $x_1 \in P$ such that

$$Ax_1 = x_2$$

where $x_2 = G(x)$, the operator A is defined in relation (3.11), and the equality holds in the L_2 sense. Noting that $\nu(x)$, $\hat{k}(x)$, $k^0(x)$, and $\hat{U}_x(x)$ are continuous in $[0, 1]$, the result follows pointwise in the respective regions.

THEOREM 3. When the optimal design is such that the base shear constraint is binding, i.e., $F_0 = F(\hat{k})$, then we have the following:

- (1) If $\nu_0 = 0$, $\hat{k}(x) = k^0(x)$, $x \in [0, 1]$;
 (2) If $\nu_0 > 0$, and $\hat{k}(x) > k^0(x)$, $x \in [0, 1]$, then

$$(4.4) \quad \hat{U}_x^2 = \frac{g^2(x)}{\nu_0}, \quad x \in [0, 1).$$

Proof. (1) If $\nu_0 = 0$, (4.1) gives

$$\nu(x) = g^2(x), \quad x \in (0, 1).$$

Equation (4.2), for $\alpha(\hat{k}) = 0$, gives

$$\hat{k}(x) = k^0(x), \quad x \in (0, 1).$$

the result follows from the assumed continuity of $\hat{k}(x)$ and $k^0(x)$ in $[0, 1]$.

(2) When we use (4.2), $\nu(x) = 0$, $x \in (0, 1)$. Then relation (4.1) gives

$$\hat{U}_x^2 = g^2 \frac{(x)}{\nu_0}, \quad x \in [0, 1).$$

COROLLARY 2. If the set $B = \{x: \hat{k}(x) - k^0(x) > 0\}$ is dense in $[0, 1]$, then relation (4.4) is valid.

Proof. For each $x \in B$, by (4.2), $\nu(x) = 0$. But $\nu(x)$ is continuous in $[0, 1]$, and since B is dense in $[0, 1]$, $\nu(x) = 0$ on $[0, 1]$. Hence we have the result.

In particular, if the optimal stiffness \hat{k} coincides with the lower bound k^0 at only a finite number of points, then B is dense in $[0, 1]$ and relation (4.4) holds.

THEOREM 4. When the shear force constraint is binding and relation (4.4) is valid, the optimal stiffness distribution is given by

$$(4.5) \quad k(x) = \frac{\int_x^1 \omega^2 \rho(x) \int_0^x g(\alpha) d\alpha}{g(x)} + \frac{\sqrt{\nu_0} A_0 \omega^2 \int_x^1 \rho(x) dx}{g(x)}, \quad x \in [0, 1)$$

where

$$\sqrt{\nu_0} = \omega^2 \frac{\int_0^1 \rho(y) [\int_0^y g(x) dx] dy}{F_0 - \omega^2 M_t A_0}.$$

Proof. Using (2.2), we have

$$k(x) = \frac{\int_x^1 \omega^2 \rho(x) U(x) dx}{U_x(x)} + \frac{A_0 \omega^2 \int_x^1 \rho(x) dx}{U_x(x)}.$$

Since $\alpha(\hat{k}) = 0$, $\hat{F} = F_0 = \omega^2(\rho, U) + \omega^2 M_t A_0$.

But by Lemma 3 and (4.4), $\hat{U}(x) = \int_0^x (g(x)/\sqrt{\nu_0}) dx$. Using this in the expression for F , we get the result.

THEOREM 5. If $\alpha(\hat{k}) > 0$, then if an optimal solution exists satisfying the constraints, it must be $\hat{k}(x) = k^0(x)$ for $x \in [0, 1]$.

Proof. If $\alpha(\hat{k}) > 0$ then by (4.2), $\nu_0 = 0$. Then (4.1) yields

$$\nu(x) = g^2(x) > 0,$$

and so, by (4.2),

$$\hat{k}(x) = k^0(x), \quad x \in [0, 1].$$

From the continuity of $\hat{k}(x)$ and $k^0(x)$ in $[0, 1]$, the result follows.

It should be noted that when the shear force constraint is binding, conditions (1.2) and (1.3) are equivalent to conditions (2.11). However, when the shear force $F(\hat{k}) < F_0$, while conditions (1.2) and (1.3) imply (2.11), conditions (2.11) do not imply (1.3'a) as shown in Lemma 6. The set, \mathcal{H}_1 , of elements $k(x)$ that satisfy (2.11) with $\alpha(k) > 0$ is such that $\mathcal{H}_1 \supseteq \mathcal{H}$. Thus we need to check that the solution $\hat{k}(x) = k^0(x)$ obtained in Theorem 5 satisfies condition (1.3'a). If it does not, no solution to our constrained optimization problem exists, since $\mathcal{H}_1 \supseteq \mathcal{H}$.

5. Physical interpretation of the results. The results obtained in Theorems 2-5 can best be described by the flowchart shown in Fig. 3. The novelty of this paper's suggested approach to solving our constrained optimization problem lies in obtaining, in closed form, the optimal stiffness distribution without ever having to solve the nonconstant coefficient partial differential equation in (1.1). The penalty paid for this is that we do not obtain the optimal solution for all possible situations. However, we shall show that, from a practical standpoint, those situations for which we do not obtain the stiffness distributions in closed form, hardly ever occur.

The optimization problem is related to two constraints: (a) the base shear constraint and (b) the stiffness distribution constraint. They are the following:

$$(5.1) \quad \alpha(\hat{k}) = F_0 - F(\hat{k}) \geq 0,$$

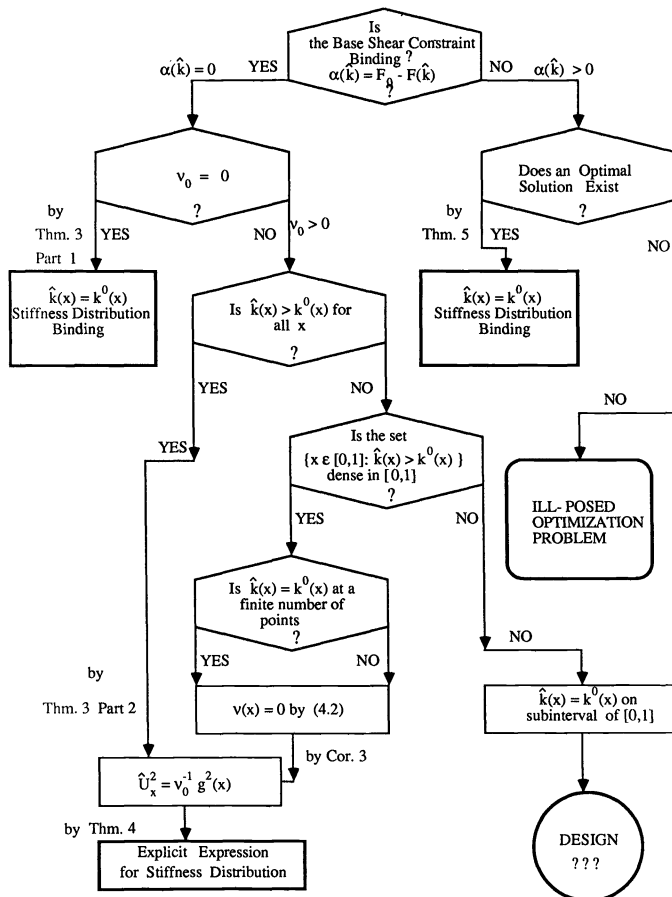


FIG. 3. Design flowchart.

$$(5.2) \quad h(x) = \hat{k}(x) - k^0(x) \geq 0 \quad \text{for all } x \in [0, 1].$$

The solution x_1 of the operator equation $Ax_1 = x_2$ is in the cone P of the space V_1 . That means the following:

$$(5.3) \quad \nu_0 \geq 0,$$

$$(5.4) \quad \nu(x) \geq 0 \quad \text{for all } x \in [0, 1].$$

Inequalities (5.1) and (5.3) can be satisfied either by equality (base shear binding) or by strict inequality. Relation (5.2) can be satisfied in different ways. We have here a nonnegative continuous function defined on $[0, 1]$. Its graph is a subset of the band $[0, 1] \times [0, \infty)$ of R^2 . We describe the different possibilities in Fig. 4. The different possible cases are as follows:

$$(5.2a) \quad h(x) \equiv 0 \quad \text{on } [0, 1],$$

$$(5.2b) \quad h(x) > 0 \quad \text{on } [0, 1],$$

$$(5.2c) \quad h(x) > 0 \quad \text{on a set dense in } [0, 1],$$

$$(5.2d) \quad h(x) = 0 \quad \text{on a proper subinterval of } [0, 1].$$

Case (5.2c) includes the possibility that the equation $h(x) = 0$ has finitely many solutions in $[0, 1]$.

The flowchart shows that the only case in which we cannot explicitly obtain $\hat{k}(x)$, although the optimization problem may have a solution, is the case (5.2d). However, from a practical point of view, the probability that the optimal stiffness distribution will coincide with its lower bound on a proper subinterval of $[0, 1]$ is extremely small.

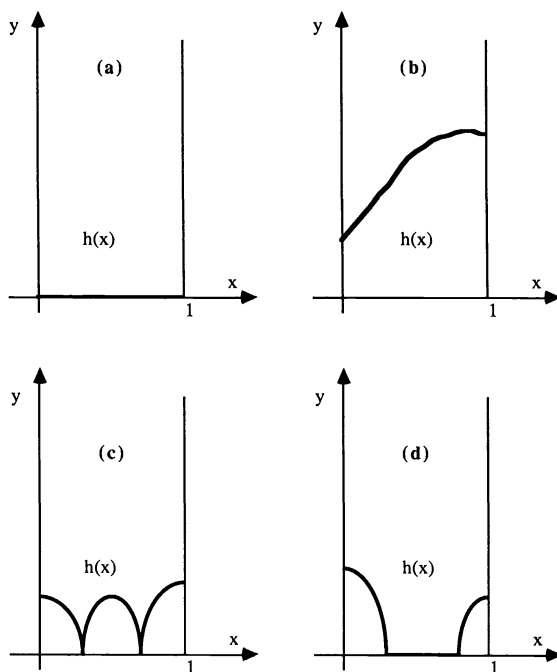


FIG. 4. Graph of $h(x) = \hat{k}(x) - k^0(x)$.

Therefore this case is very unlikely to occur. This, in turn, means that whenever a solution to the optimization problem exists, in most cases we succeed in finding the solution in closed form.

Now we describe Fig. 3 in detail.

Case 1. Base shear not binding. In this case we have to check whether the solution to (2.11) satisfies (1.3). If not, then the optimization problem is ill-posed and has no solution. Otherwise, Theorem 5 proves that we are in situation (5.2a), which means that the optimal stiffness distribution is the same as the lower bound $k^0(x)$ for all x .

This describes the right branch of the flowchart. Next we describe the left branch of the flowchart.

Case 2. Base shear binding. For the next branching, we ask whether (5.3) is satisfied by equality. If so, then Theorem 3(1) shows that we are again in situation (5.2a). Otherwise, we ask if (5.2b) happens. If so, then by Theorem 3(2) we obtain explicitly $\hat{U}_x^2(x)$, and then Theorem 4 gives us the optimal stiffness distribution in closed form. If (5.2b) is not the case, we ask whether (5.2c) occurs. (Let us emphasize again that, from a practical point of view, the subcase of importance here is the situation where graph $h(x)$ hits the x -axis at a finite number of points). If so, then (5.4) is satisfied by identity; by Corollary 2, we again get $\hat{U}_x^2(x)$ explicitly; and Theorem 4 gives us the optimal stiffness distribution in closed form.

However, if (5.2c) is not the case, but $h(x)$ is equal to zero on a subinterval of $[0, 1]$, then we cannot find $\hat{k}(x)$ in closed form on $[0, 1]$. Thus we only know $\hat{k}(x)$ on the subinterval where it is equal to k^0 . As mentioned, this is unlikely to happen in real-life situations.

As a final comment, we observe that whenever the optimization problem has a solution, in most cases, that solution is obtained, in closed form, via the Farkas Theorem, and this does not require the explicit solution of the differential problem (1.1).

6. Numerical example. The results of the previous section can be illustrated for the case when the shear force constraint is taken to be binding with $g(x) = 1$, $\rho(x) = 1$, $\omega = 1$, $F_0 = 2$, $A_0 = 1$, and $k^0(x) = 0$. We then obtain

$$(6.1) \quad \nu_0 = \frac{1}{4} \quad \text{and} \quad k(x) = 1 - \frac{x}{2} - \frac{x^2}{2}.$$

Using the same parameters with $g(x) = 1 + x$, we get

$$(6.2) \quad \nu_0 = \frac{4}{9} \quad \text{and} \quad k(x) = \left(\frac{4}{3} - \frac{2}{3}x - \frac{x^2}{2} - \frac{x^3}{6} \right) / (1 + x).$$

The results (6.1) and (6.2) are indicated in the Fig. 5, along with the result for $g(x) = (1 - x/2)$.

7. Conclusions and discussion. This paper attempts to study the structure of the optimal design problem for a building structure subjected to harmonic base excitation. The aim is to find a stiffness distribution which minimizes a suitable cost function subject to constraints on the base shear, the lowest fundamental frequency, and a lower bound function for the stiffness distribution. The Farkas Theorem is used to study the underlying structure of the optimization problem. It is shown that when the base shear constraint is nonbinding, the optimal stiffness, if it exists, is given by its lower bound, $k^0(x)$. The Farkas Theorem is also used to find the optimal stiffness when the shear force constraint is binding and $\hat{k}(x) > k^0(x)$ for all x in $[0, 1]$ except perhaps at a discrete set of points. The closed-form solution of the optimal mode shape $\hat{U}(x)$ is first obtained and then $\hat{k}(x)$. The global optimum is thus analytically determined.

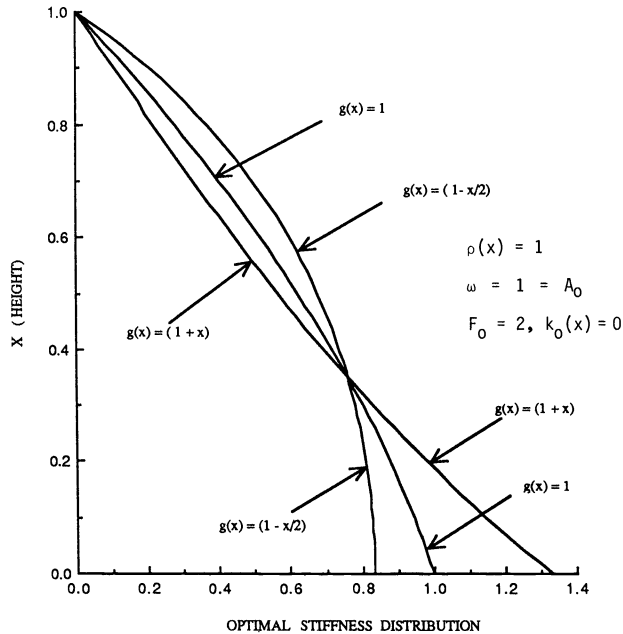


FIG. 5

These analytical solutions become all the more useful when it is realized that computational nonlinear programming methods for obtaining the optimal solution for such problems are fraught with numerical difficulties, even when the number of unknowns (the values of $\hat{k}(x)$ at a discrete set of points) is small.

This paper gives only simplistic modeling of a tall building structure subjected to ground shaking. However, such models are used commonly in the design and analysis of buildings in seismic areas, and therefore the results obtained here are of practical interest to the designer. It is anticipated that the results will find use in other application areas, such as the optimal design of space structures and the design of machine foundations.

REFERENCES

- [1] L. J. ICERMAN, *Optimal structural design for given dynamic deflection*, Internat. J. Solids and Structures, 5 (1969), pp. 473-490.
- [2] R. H. PLAUT, *Optimal structural design for given deflection under periodic loading*, Quart. Appl. Math., 29 (1971), pp. 315-318.
- [3] N. C. HUANG, *Minimum weight design on vibrating elastic structures with dynamic deflection constraint*, J. Appl. Mech., (1976), pp. 178-180.
- [4] E. H. JOHNSON AND A. RIZZI, *Optimization of continuous one-dimensional structures under steady harmonic excitation*, in Proc. 17th Structures, Dynamics, and Materials Congress, AIAA J., 14 (1976), pp. 1690-1698.
- [5] E. H. JOHNSON, *Disjoint design spaces in the optimization of harmonically excited structures*, AIAA J., 14 (1976), pp. 259-261.
- [6] S. YU. IVANOVA, *Some problems of reducing the weight of structures in a mode of forced harmonic vibrations*, Mech. Solids, 19 (1984), pp. 141-147.
- [7] F. E. UDWADIA AND M. D. TRIFUNAC, *Damped Fourier transforms and statistics of oscillator response*, Seismol. Soc. Amer., 63 (1973), pp. 1775-1783.
- [8] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. 1, Interscience, New York, 1953.
- [9] J. P. AUBIN, *Applied Functional Analysis*, John Wiley, New York, 1979.

GLOBAL OUTPUT TRACKING FOR NONLINEAR SYSTEMS*

R. M. HIRSCHORN† AND J. H. DAVIS†

Abstract. The well-known local results on output tracking makes it possible to control the state so that the output follows some desired path y_d over some time interval $[t_0, t_0 + \varepsilon)$. These results have been extended to give global tracking, so that the output follows y_d over a given interval $[t_0, t_1]$, for a class of systems that are sufficiently “observable.” An example is presented to illustrate how these results can be used to steer the state of the system from a given initial state to a desired final state.

Key words. nonlinear systems, output tracking, singularities

AMS(MOS) subject classification. 93B05

1. Introduction. The problem of controlling a system so that its output follows some desired path was first solved by Brockett and Mesarovic [1] in 1965 for time-invariant linear control systems. These results have been generalized to include affine nonlinear control system models provided the state trajectories avoid “singular states” (cf. [2]–[7]). For linear systems there are no singular states, but for many nonlinear systems the set of singular states M_s forms a codimension-one submanifold of the state space. If the initial state x_0 is not a singular state, then the output can be made to follow a desired path for some time interval over which the state trajectory avoids M_s . Thus, for nonlinear systems, output tracking results are generally local in nature. In [10] the local tracking problem is solved under the assumption that $x_0 \in M_s$ but that the state never returns to M_s . The purpose of this paper is to identify a class of outputs for which global output tracking results can be obtained. These are output functions with “observability” properties that permit us to accommodate transversals of the singular set by use of the available output data.

In § 2, the main results, Theorems 2.2 and 2.3, are derived and two examples are presented. The use of these global tracking results in controlling a system to transfer the states from $x(t_0)$ to $x(t_1)$ is illustrated.

2. Tracking through singular points. Consider the single-input affine nonlinear system model

$$(2.1) \quad \begin{aligned} \dot{x}(t) &= f(x(t)) + u(t)g(x(t)), & x(t_0) &= x_0 \in M, \\ y(t) &= h(x(t)) \end{aligned}$$

where M is a connected C^∞ manifold, f and g are C^∞ vector fields on M , and h is a C^∞ function of M . The controls $u: [t_0, \infty) \rightarrow \mathbb{R}$ are continuous. For each such admissible u , let $x(t, u, x_0)$ denote the corresponding solution to the differential equation (2.1) and $y(t, u, x_0)$ denote the resulting output. As in [2]–[7] the *relative order* α of (2.1) is defined as the least nonnegative integer k such that $gf^{k-1}h \neq 0$ on M . We take $\alpha = \infty$ if $gf^k h = 0$ for all $k \geq 0$. If $x \in M$ is such that $gf^{\alpha-1}h(x) = 0$, then x is called a *singular point for output tracking*, and the set of all singular points is called the *singular set*

$$M_s = \{x \in M \mid gf^{\alpha-1}h(x) = 0\}.$$

Generically M_s is a codimension-one submanifold of M (cf. [11, Cor. 4.12]).

* Received by the editors March 23, 1987; accepted for publication (in revised form) February 17, 1988.

† Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, K7L 3N6, Canada.

In the conventional approach to output tracking, to steer the state so that the output y follows some desired path y_d we solve for u as a function of y and its derivatives. Computing derivatives of y we have (since α is the system relative order)

$$\begin{aligned} y(t) &= h(x(t)), \\ y^{(1)}(t) &= dh_{x(t)}\dot{x}(t) \\ &= dh_{x(t)}f(x) + u(t) dh_{x(t)}g(x(t)) \\ &= fh(x(t)), \\ y^{(2)}(t) &= f^2h(x(t)), \\ &\vdots \\ y^{(\alpha-1)}(t) &= f^{\alpha-1}h(x(t)), \\ y^{(\alpha)}(t) &= f^\alpha h(x(t)) + u(t)gf^{\alpha-1}h(x(t)) \end{aligned}$$

where $gf^{\alpha-1}h(\cdot) \neq 0$. When we set

$$\begin{aligned} y^\alpha(t) &= (y(t), y^{(1)}(t), \dots, y^{(\alpha-1)}(t)), \\ h^\alpha(x) &= (h(x), fh(x), \dots, f^{\alpha-1}h(x)), \\ a(x) &= f^\alpha h(x), \\ b(x) &= gf^{\alpha-1}h(x), \end{aligned}$$

it follows that attainable output paths are constrained by $y^\alpha(t) = h^\alpha(x(t))$, and that a necessary condition for y_d to be a (C^∞) output of (2.1) is that $y_d^\alpha(t_0) = h^\alpha(x_0)$. If $x_0 \notin M_s$ (a reasonable assumption since M_s will be of measure zero in M) then this condition is also (locally) sufficient (cf. [2]–[7]). In fact, since $y^{(\alpha)}(t) = a(x(t)) + u(t)b(x(t))$, the required control is simply

$$(2.2) \quad u_d(t, x) = \frac{y_d^{(\alpha)}(t) - a(x)}{b(x)}.$$

Thus if $u(t) = u_d(t, x(t))$ it follows that $y(t, u, x_0) = y_d(t)$. For linear systems $b(x(t))$ is constant, and thus M_s is empty and $y(t, u, x_0) = y_d(t)$ for all $t \geq t_0$. For nonlinear systems M_s is usually not empty, and the control scheme (2.2) can break down when $x(t)$ approaches M_s (i.e., when $b(x(t))$ approaches zero). The nonlinear results therefore are local, i.e., given y_d there exists $\varepsilon_{y_d} > 0$ such that $y(t, u, x_0) = y_d(t)$ for $t_0 \leq t < t_0 + \varepsilon_{y_d}$. Since trajectories are not allowed to approach M_s and generically M_s “divides the state space in half,” this seriously limits the applications of the current theory.

To develop a practical global theory for nonlinear output tracking, the state trajectory must be allowed to pass through the singular set. This means that we must identify those paths y_d for which the control u_d described by (2.2) is well-behaved at those times when the state enters M_s and the denominator $b(x)$ vanishes. In particular, if we set $u(t) = u_d(t, x(t))$ so that $y(t) = y_d(t)$ and $x(t) = x(t, u, x_0)$ approaches the singular set at time t_s (i.e., $b(x(t_s)) = 0$), then the tracking control is given by

$$u(t) = \frac{y_d^{(\alpha)}(t) - a(x(t))}{b(x(t))} \quad \text{for } t < t_s,$$

and thus an evident necessary condition for u to be well defined at time t_s is that $y_d^{(\alpha)}(t_s) = a(x(t_s))$. Thus, in order to make tracking practical, we must identify, in terms

of y_d , those times t_s when $b(x(t_s))=0$ (i.e., $x(t_s) \in M_s$), as well as the corresponding constraint value $a(x(t_s))$ so that $y_d(t_s) = a(x(t_s))$ translates into a restriction on the allowable desired paths. That is, in order to be able to determine this from the available observations, we want to ensure that there exist functions $a_0, b_0: R^\alpha \rightarrow R$ such that

$$(2.3) \quad a(x) = a_0(h^\alpha(x)) \quad \text{and} \quad b(x) = b_0(h^\alpha(x)).$$

This means that along trajectories $t \rightarrow x(t)$ of (2.1)

$$a(x(t)) = a_0(y^\alpha(t)) \quad \text{and} \quad b(x(t)) = b_0(y^\alpha(t)).$$

Condition (2.3) has implications for $y(t)$ at times t_s when $x(t_s) \in M_s$. For example,

$$\begin{aligned} y^{(\alpha)}(t_s) &= a(x(t_s)) + u(t_s)b(x(t_s)) \\ &= a_0(y^\alpha(t_s)) + u(t_s) \cdot 0 \end{aligned}$$

so that $y^{(\alpha)}(t_s) = a_0(y^\alpha(t_s))$, and similar restrictions are placed on higher-order derivatives of y at time t_s . To describe the extent of these restrictions some notation will be introduced.

If $x(t)$ is any solution to (2.1) then set

$$b_1(x(t), u(t)) = (d/dt)b(x(t)) = fb(x(t)) + u(t)gb(x(t)),$$

and similarly we can define $b_k(x(t), u(t), u^{(1)}(t), \dots, u^{(k-1)}(t)) = (d^k/dt^k)b(x(t))$. As in [10] the *degree of singularity of a state* $x \in M$, $\beta(x)$, is defined to be the least integer k such that $(r_1, \dots, r_k) \rightarrow b_k(x, r_1, \dots, r_k)$ is not the zero function. Thus for $x \notin M_s$, $\beta(x) = 0$, and for $x \in M_s$, $\beta(x) > 0$. If $\beta(x) = \infty$, then (in the real analytic case) the input-output map defined by (2.1) will be trivial (cf. [10]). To save accounting assume $\beta(x) = \beta < \infty$ for all $x \in M_s$.

Thus

$$\begin{aligned} y^{(\alpha)}(t) &= a(x(t)) + u(t)b(x(t)), \\ y^{(\alpha+1)}(t) &= \frac{d}{dt} a(x(t)) + u(t) \frac{d}{dt} b(x(t)) + \dot{u}(t)b(x(t)), \\ &\vdots \\ y^{(\alpha+\beta)}(t) &= \frac{d^\beta}{dt^\beta} a(x(t)) + u(t) \frac{d^\beta}{dt^\beta} b(x(t)) + \dots + u^{(\beta)}(t)b(x(t)) \end{aligned}$$

and from the definition of β it follows that at any time t_s when $x(t_s) \in M_s$, we have

$$\begin{aligned} y^{(\alpha)}(t_s) &= a(x(t_s)), \\ y^{(\alpha+1)}(t_s) &= \frac{d}{dt} a(x(t_s)), \\ &\vdots \\ y^{(\alpha+\beta-1)}(t_s) &= \frac{d^{\beta-1}}{dt^{\beta-1}} a(x(t_s)), \\ y^{(\alpha+\beta)}(t_s) &= \frac{d^\beta}{dt^\beta} a(x(t_s)) + u(t_s) \frac{d^\beta}{dt^\beta} b(x(t_s)). \end{aligned} \tag{2.4}$$

The following lemma simplifies notation when condition (2.3) holds.

LEMMA 2.1. Consider the system (2.1) and suppose that condition (2.3) holds. Then there exist functions $a_k, b_k: R^{\alpha+k} \rightarrow R$ such that

$$\frac{d^k}{dt^k} a(x(t)) = a_k(y^{\alpha+k}(t)), \quad \frac{d^k}{dt^k} b(x(t)) = b_k(y^{\alpha+k}(t))$$

for $k=0, 1, \dots, \beta$. In particular, $a_0(y^\alpha)$ and $b_0(y^\alpha)$ are the functions described in (2.3) and

$$a_{k+1}(r_1, \dots, r_{\alpha+k}, r_{\alpha+k+1}) = (da_k)_{(r_1, \dots, r_{\alpha+k})}(r_2, \dots, r_{\alpha+k+1}),$$

$$b_{k+1}(r_1, \dots, r_{\alpha+k+1}) = (db_k)_{(r_1, \dots, r_{\alpha+k})}(r_2, \dots, r_{\alpha+k+1}).$$

Proof. From (2.3) $a(x(t)) = a_0(y^\alpha(t))$ so that $(d/dt)a(x(t)) = (da_0)_{y^\alpha(t)}((d/dt)y^\alpha(t)) = (da_0)_{y^\alpha(t)}(y^{(1)}(t), \dots, y^{(\alpha)}(t)) = a_1(y^{\alpha+1}(t))$ by definition of a_1 . Continuing, we have $(d^2/dt^2)a(x(t)) = (d/dt)a_1(y^{\alpha+1}(t)) = (da_1)_{y^{\alpha+1}(t)}(y^{(1)}, y^{(2)}, \dots, y^{(\alpha+1)}) = a_2(y^{\alpha+2}(t))$ from the definition of a_2 . Similar relations hold for b_1, b_2 . The rest of the proof follows from these observations.

Thus for systems that satisfy condition (2.3), Lemma 2.1 yields functions $a_k, b_k: R^{\alpha+k} \rightarrow R$ with $(d^k/dt^k)a(x(t)) = a_k(y^{\alpha+k}(t))$ and $(d^k/dt^k)b(x(t)) = b_k(y^{\alpha+k}(t))$ (assuming a_0, b_0 are k -times differentiable at $y^\alpha(t)$).

THEOREM 2.2. Consider system (2.1) and suppose (2.3) holds. A necessary condition for a C^∞ function y_d to be tracked by the output (using an input that is $(\beta-2)$ times differentiable) is that

$$y_d^\alpha(t_0) = h^\alpha(x_0)$$

and for each time t_s when $b_0(y(\alpha/d)(t_s)) = 0$ (i.e., $x(t_s) \in M_s$)

$$y_d^{(\alpha+k)}(t_s) = a_k(y_d^{\alpha+k}(t_s)) \quad \text{for } k=0, 1, \dots, \beta-1.$$

Proof. Suppose that y_d is a C^∞ function that can be tracked by the output y of (2.1) using a $(\beta-2)$ times differentiable control u . When we use Lemma 2.1 and (2.4), it follows that at each time t_s

$$y_d^{(\alpha+k)}(t_s) = \frac{d}{dt^k} a(x(t)) \Big|_{t=t_s}$$

$$= a_k(y_d^{\alpha+k}(t_s))$$

for $k=0, 1, \dots, \beta-1$. At time t_0 ,

$$y(t_0) = h(x_0) = y_d(t_0)$$

$$y^{(1)}(t_0) = fh(x_0) = y_d^{(1)}(t_0),$$

$$\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots$$

$$y^{(\alpha-1)}(t_0) = f^{\alpha-1}h(x_0) = y_d^{(\alpha-1)}(t_0),$$

so that $y_d^\alpha(t_0) = h^\alpha(x_0)$ and the proof is complete.

Theorem 2.2 comes close to identifying those functions y_d that can be tracked by the output of a nonlinear system for which the “observability condition” (2.3) holds. By slightly reducing the set of outputs y_d described in Theorem 2.2 we can find a continuous control u_d that results in $y \equiv y_d$. From the definition of β ,

$$\frac{d^\beta b}{dt^\beta}(x(t)) = b^\beta(x(t), u(t), \dots, u^{(\beta-1)}(t))$$

and is nonzero (for “most” controls u). From Lemma 2.1

$$\frac{d^\beta b(x(t))}{dt^\beta} = b_\beta(y^{\alpha+\beta}(t)).$$

A further restriction is placed on y_d : at each time t_s , when $b_0(y_d^\alpha(t_s)) = 0$, we require that $b_\beta(y_d^{\alpha+\beta}(t_s)) \neq 0$ (which will be the case for “most” $y_d^{\alpha+\beta}(t_s)$). This assumption amounts to the condition that the singular set be traversed with the minimum possible degree of tangency.

THEOREM 2.3. *Consider the system (2.1) and suppose (2.3) holds. A sufficient condition for a C^∞ function y_d to be tracked by the output is that*

$$y_d^\alpha(t_0) = h^\alpha(x_0)$$

and for each time t_s such that $b_0(y_d^\alpha(t_s)) = 0$

$$y_d^{(\alpha+k)}(t_s) = a_k(y_d^{\alpha+k}(t_s)) \quad \text{for } k = 0, 1, \dots, \beta - 1,$$

and $b_\beta(y_d^{\alpha+\beta}(t_s)) \neq 0$.

Proof. Let $y(t)$ denote the output corresponding to a control u . If $b_0(x(t_s)) = 0$, then from (2.4) and Lemma 2.1,

$$y^{(\alpha+\beta)}(t_s) = a_\beta(y^{\alpha+\beta}(t_s)) + u(t_s)b_\beta(y^{\alpha+\beta}(t_s))$$

and if $b_\beta(y^{\alpha+\beta}(t_s)) \neq 0$ then

$$u(t_s) = \frac{y^{(\alpha+\beta)}(t_s) - a_\beta(y^{\alpha+\beta}(t_s))}{b_\beta(y^{\alpha+\beta}(t_s))}.$$

This motivates the following control scheme. Suppose that y_d satisfies the requirements of Theorem 2.3. Define the control u_d as follows:

$$u_d(t, x) = \begin{cases} \frac{y_d^{(\alpha)}(t) - a(x)}{b(x)} & \text{when } b_0(y_d^\alpha(t)) \neq 0, \\ r(t_s) & \text{when } b_0(y_d^\alpha(t_s)) = 0 \end{cases}$$

where $r(t_s) = (y_d^{(\alpha+\beta)}(t_s) - a_\beta(y_d^{\alpha+\beta}(t_s))) / b_\beta(y_d^{\alpha+\beta}(t_s))$. Since $y^\alpha(t_0) = h^\alpha(x_0) = y_d^\alpha(t_0)$ it suffices to verify that $u_d(t, x(t))$ is continuous and $y^{(\alpha)}(t) = y_d^{(\alpha)}(t)$ for all $t \geq t_0$ to complete the proof. If $t \neq t_s$ (i.e., $b(x(t)) \neq 0$) then

$$\begin{aligned} y^{(\alpha)}(t) &= a(x(t)) + u_d(t, x(t))b(x(t)) \\ &= a(x(t)) + \frac{y_d^{(\alpha)}(t) - a(x(t))}{b(x(t))} b(x(t)) \\ &= y_d^{(\alpha)}(t), \end{aligned}$$

and so it suffices to show that $u_d(t, x(t))$ is continuous when $t = t_s$ (i.e., $b(x(t_s)) = b_0(y_d^\alpha(t_s)) = 0$). Assume $y \equiv y_d$ for $t_0 \leq t < t_s$ (this is the case, since the known sufficient conditions for local tracking are satisfied by y_d cf. [10]).

Set $L = \lim_{t \rightarrow t_s^-} ((y_d^{(\alpha)}(t) - a(x(t))) / b(x(t)))$. Now $b(x(t)) \rightarrow 0$ and since $y_d^{(\alpha)}(t_s) = a_0(y_d^\alpha(t_s))$, l'Hôpital's rule may be invoked in an attempt to evaluate the limit

$$L = \lim_{t \rightarrow t_s^-} \frac{y_d^{(\alpha+1)}(t) - a_1(y_d^{\alpha+1}(t))}{b_1(y_d^{\alpha+1}(t))}.$$

Once again, since $y_d^{(\alpha+1)}(t_s) = a_1(y_d^{\alpha+1}(t_s))$ by hypothesis and $b_1(y_d^{\alpha+1}(t_s)) = 0$ if $\beta > 1$, the limit is indeterminate, so l'Hôpital's rule is successively applied. These steps are repeated until we finally obtain the determinate limit

$$L = \frac{y_d^{(\alpha+\beta)}(t_s) - a_\beta(y_d^{\alpha+\beta}(t_s))}{b_\beta(y_d^{\alpha+\beta}(t_s))} = r(t_s).$$

Hence by l'Hôpital's rule we find that

$$\lim_{t \rightarrow t_s} \frac{y_d^{(\alpha)}(t) - a(x(t))}{b(x(t))} = r(t_s).$$

Thus u_d is continuous and the proof is complete.

COROLLARY 1. *Suppose that $y_d \in C^\infty(R)$ satisfies the hypothesis of Theorem 2.3 so that at each time t_s when $b_0(y_d^\alpha(t_s)) = 0$ we can choose $r(t_s)$ so that*

$$y_d^{(\alpha+\beta)}(t_s) = a_\beta(y_d^{\alpha+\beta}(t_s)) + r(t_s)b_\beta(y_d^{\alpha+\beta}(t_s)).$$

Then the output y of system (2.1) can be controlled so that $y \equiv y_d$ by using

$$u_d(t, x) = \begin{cases} r(t_s) & \text{for each time } t_s \text{ when } b_0(y_d^\alpha(t_s)) = 0, \\ \frac{y_d^{(\alpha)}(t) - a(x)}{b(x)} & \text{when } t \neq t_s. \end{cases}$$

In open-loop form

$$u_d(t) = u_d(t, x(t)) = \begin{cases} r(t_s) & \text{for each time when } b_0(y_d^\alpha(t_s)) = 0, \\ \frac{y_d^{(\alpha)}(t) - a_0(y_d^\alpha(t))}{b_0(y_d^\alpha(t))} & \text{when } t \neq t_s. \end{cases}$$

Proof. The proof follows directly from the proof of Theorem 2.3.

Example 2.1. Consider the system model

$$\dot{x}_1(t) = x_1(t)x_2(t) + x_2^2(t)u(t),$$

$$\dot{x}_2(t) = x_1(t),$$

$$y(t) = x_2(t)$$

where $M = R^2 \sim \{(0, 0)\}$ (since the system cannot be steered to the origin from any nonzero initial state). Here

$$y = x_2,$$

$$\dot{y} = x_1,$$

$$\ddot{y} = x_1x_2 + x_2^2u,$$

so that $\alpha = 2$, $a(x) = x_1x_2$, $b(x) = x_2^2$, and M_s is $\{x_2 = 0\}$. Also

$$a(x(t)) = \dot{y}(t)y(t) = a_0(y(t), \dot{y}(t)) = a_0(y^\alpha(t))$$

with $a_0(r_1, r_1) = r_1r_2$ and

$$b(x(t)) = y(t)^2 = b_0(y(t), \dot{y}(t)) = b_0(y^\alpha(t))$$

with $b_0(r_1, r_2) = r_1^2$. To determine β notice that $(d/dt)b(x(t)) = 2x_2(t)x_1(t)$, which vanishes on M_s , and $(d^2/dt^2)b(x(t)) = 2x_2^2(t) + 2x_1(t)x_2^2(t) + 2x_2^3(t)u(t)$, which is not identically zero on M_s , so $\beta = 2$. Following the algorithm in Lemma 2.1, we have

$$a_1(r_1, r_2, r_3) = (da_0)_{(r_1, r_2)}(r_2, r_3) = [r_2 r_1] \begin{bmatrix} r_2 \\ r_3 \end{bmatrix} = r_2^2 + r_1 r_3,$$

$$a_2(r_1, r_2, r_3, r_4) = (da_1)_{(r_1, r_2, r_3)}(r_2, r_3, r_4) = [r_3 2r_2 r_1] \begin{bmatrix} r_2 \\ r_3 \\ r_4 \end{bmatrix} = 3r_2 r_3 + r_1 r_4,$$

$$b_1(r_1, r_2, r_3) = (db_0)_{(r_1, r_2)}(r_2, r_3) = 2r_1 r_2,$$

$$b_2(r_1, r_2, r_3, r_4) = (db_1)_{(r_1, r_2, r_3)}(r_2, r_3, r_4) = 2r_2^2 + 2r_1 r_3.$$

Suppose that the initial state is $x(t_0) = (x_{10}, x_{20})$. From Theorem 2.2 a necessary condition for y_d to appear as an output function is

$$y_d^\alpha(t_0) = \begin{bmatrix} y_d(t_0) \\ \dot{y}_d(t_0) \end{bmatrix} = \begin{bmatrix} x_{20} \\ x_{10} \end{bmatrix}$$

and for each time t_s , when $b_0(y_d^\alpha(t_s)) = y_d^2(t_s) = 0$

$$\ddot{y}_d(t_s) = \alpha_0(y_d^\alpha(t_s)) = y_d(t_s)\dot{y}_d(t_s),$$

$$\ddot{y}_d(t_s) = a_1(y_d^{\alpha+1}(t_s)) = \dot{y}_d^2(t_s) + y_d(t_s)\ddot{y}_d(t_s).$$

Thus a necessary condition for y_d to appear as the output when $x(t_0) = (x_{10}, x_{20})$ (and u differentiable) is

$$y_d(t_0) = x_{20}, \quad \dot{y}_d(t_0) = x_{10}$$

and when $y_d(t_s) = 0$,

$$\ddot{y}_d(t_s) = 0, \quad \ddot{y}_d(t_s) = \dot{y}_d^2(t_s).$$

Note that $\dot{y}(t_s) = x_1(t_s) \neq 0$ as $y(t_s) = x_2(t_s) = 0$ and that the origin is not in the state space (it is not in leaf of the distribution generated by the vector fields f and g containing \underline{x}_0). Thus a further necessary condition is $\dot{y}_d(t_s) \neq 0$.

From Theorem 2.3 this condition is sufficient for y_d to appear as an output using a continuous control, provided that $b_\beta(y_d^{\alpha+\beta}(t_s)) = b_2(y_d^4(t_s)) = 2\dot{y}_d^2(t_s) + 2y_d(t_s)\ddot{y}_d(t_s) = 2\dot{y}_d^2(t_s) \neq 0$. From the above this is expected. Thus the singular set $M_s = \{x_2 = 0\}$ can be crossed by the state trajectory, provided that when $y_d(t_s) = 0$ we ensure that $\ddot{y}_d(t_s) = 0$ and $\ddot{y}_d(t_s) = \dot{y}_d^2(t_s) \neq 0$ (when the state approaches M_s the local theory breaks down).

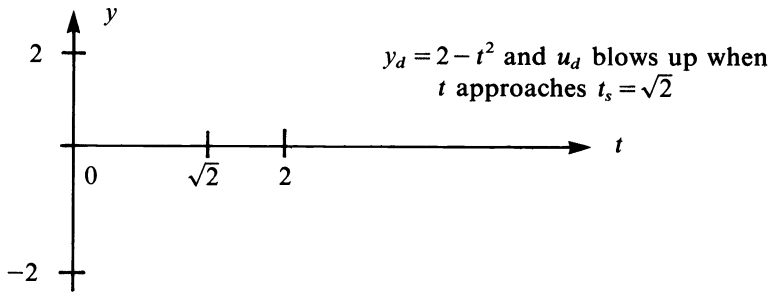
When we take $t_0 = 0$, $x_0 = (0, 2)$ as the necessary condition from Theorem 2.2, namely

$$y_d(0) = 2, \quad \dot{y}_d(0) = 0$$

and $\ddot{y}_d(t_s) = 0$, $\ddot{y}_d(t_s) = \dot{y}_d^2(t_s)$ whenever $y_d(t_s) = 0$, it is sufficient to track y_d (Theorem 2.3) provided $\dot{y}_d(t_s) \neq 0$. Thus, to transfer the output from $y(0) = 2$ to $y(2) = -2$ we could use $y_d(t) = 2 - t^2$, which can be locally tracked on $[0, 2]$ using $u_d(t, x) = (\ddot{y}_d(t) - x_1 x_2)/x_2^2 = -(2 + x_1 x_2)/x_2^2$ (cf. [2]–[7]). Here $y_d(t_s) = 0$ when $t_s = \sqrt{2}$ and in this case we can solve for x_1 and x_2 explicitly when $u = u_d$, i.e., $\dot{x}_1 = x_1 x_2 - ((2 + x_1 x_2)/x_2^2)x_2^2 = -2$ so that $x_1(t) = -2t$, and $\dot{x}_2 = x_1$ so that $x_2(t) = -t^2 + 2$. Thus

$$u_d(t, x(t)) = \frac{-2 - (-2t)(-t^2 + 2)}{(-t^2 + 2)} = \frac{4t - 2t^3 - 2}{(2 - t^2)^2}$$

and $\lim_{t \rightarrow t_s} u_d(t, x(t)) = -\infty$.



To transfer y from $+2$ to -2 using Theorem 2.3, decide on a time t_s for y_d to vanish and ensure that at this time the additional conditions $\ddot{y}_d(t_s) = 0$ and $\ddot{y}_d(t_s) = \dot{y}_d^2(t_s) \neq 0$ hold. If we want $t_s = 1$, for example, then guess

$$y_d(t) = a + b(t-1) + \frac{c}{2!}(t-1)^2 + \frac{d}{3!}(t-1)^3 + \frac{e}{4!}(t-1)^4 + \frac{f}{5!}(t-1)^5$$

and require

$$y_d(0) = 2, \quad \dot{y}_d(0) = 0$$

and, in addition, at time $t_s = 1$

$$y_d(t_s) = y_d(1) = a = 0,$$

$$\ddot{y}_d(t_s) = \ddot{y}_d(1) = c = 0,$$

$$\ddot{y}_d(t_s) = d = \dot{y}_d^2(t_s) = b^2,$$

$$y_d(2) = -2$$

we obtain

$$y_d(t) = (-8.449)(t-1) + (11.899)(t-1)^3 + (-5.449)(t-1)^5.$$

One application of output tracking is controlling a system to move the state from x_0 at time t_0 to x_1 at time t_1 . For linear systems the variation of constants formula lets us find the required control. For nonlinear systems, results tend to be qualitative and there is no explicit input-output relationship available. To transfer states using output tracking, one possible method is to find an output function $h(x)$ such that $x \rightarrow h^\alpha(x)$ distinguishes between states. Then condition (2.3) is automatically satisfied and the corollary to Theorem 2.3 can be used to generate u . That is, we look for a path $y_d(t)$ such that $y_d^\alpha(t_0) = h^\alpha(x_0)$ and $y_d^\alpha(t_1) = h^\alpha(x_1)$. Theorem 2.3 enables us to choose y_d so that u_d is well defined on all of $[t_0, t_1]$. Of course, as in the Lyapunov theory, there is no easy way to choose a suitable $h(x)$.

Example 2.2. Consider the system model

$$\dot{x}_1(t) = x_1(t)x_2(t)u(t),$$

$$\dot{x}_2(t) = x_1(t) + x_2(t)u(t)$$

where $x \in M = \{(x_1, x_2) | x_1 > 0\}$, i.e., the system has the accessibility property (cf. [8]). We want α as large as possible so that $h^\alpha(x)$ separates states. We are led to try

$$y = h(x) = x_1 e^{-x_2}$$

so that

$$\dot{y} = -x_1^2 e^{-x_2}, \quad \ddot{y} = -x_1^3 e^{-x_2} - x_1^2 x_2 e^{-x_2} u,$$

and thus $a(x) = -x_1^3 e^{-x_2}$, $b(x) = -x_1^2 x_2 e^{-x_2}$, $\alpha = 2$, and $M_s = \{x_2 = 0\}$. Here $\dot{b}(x(t)) \neq 0$ so that $\beta = 1$ and

$$h^\alpha(x) = \begin{bmatrix} x_1 e^{-x_2} \\ -x_1^2 e^{-x_2} \end{bmatrix}$$

so that $x \rightarrow h^\alpha(x)$ is one to one on M .

To find a_0, b_0 in condition (2.3), simply solve for x_1, x_2 in terms of $h^\alpha(x)$. Since

$$y = h(x) = x_1 e^{-x_2} \quad \text{and} \quad \dot{y} = fh(x) = -x_1^2 e^{-x_2},$$

it follows that

$$a(x) = a_0(y, \dot{y}) = (\dot{y}^2/y) \quad \text{and} \quad b(x) = b_0(y, \dot{y}) = -\ln(-y^2/\dot{y}).$$

Using Lemma 2.1, we have that $b_1(y, \dot{y}, \ddot{y}) = (y\ddot{y} - 2\dot{y}^2)/(y\dot{y})$. Thus $x(t_s) \in M_s$ if and only if $b_0(y(t_s), \dot{y}(t_s)) = -\ln(-y^2(t_s)/\dot{y}(t_s)) = 0$, i.e., $\dot{y}(t_s) + y^2(t_s) = 0$. Set $z(t) = \dot{y}(t) + y^2(t)$. Then $x(t_s) \in M_s$ if and only if $z(t_s) = 0$.

A necessary condition for y_d to be tracked by the output y of this system when $x(t_0) = (x_{10}, x_{20})$ is

$$y_d(t_0) = h(x_0) = x_{10} e^{-x_{20}},$$

$$\dot{y}_d(t_0) = fh(x_0) = -x_{10}^2 e^{-x_{20}} = -x_{10} y_d(t_0)$$

and for each time t_s such that $z(t_s) = \dot{y}_d(t_s) + y_d^2(t_s) = 0$, $\ddot{y}_d(t_s) = a_0(y_d(t_s), \dot{y}_d(t_s)) = \dot{y}_d^2(t_s)/y_d(t_s)$. When we use Theorem 2.3, a sufficient condition for $y \equiv y_d$ is that

$$(2.5) \quad y_d(t_0) = x_{10} e^{-x_{20}} \quad (>0),$$

$$\dot{y}_d(t_0) = -x_{10} y_d(t_0)$$

and for each time t_s such that $z(t_s) = \dot{y}_d(t_s) + y_d^2(t_s) = 0$

$$\ddot{y}_d(t_s) = \dot{y}_d^2(t_s)/y_d(t_s),$$

and $b_1(y_d^3(t_s)) = -y_d(t_s) \neq 0$. Of course this last restriction is to be expected since $y(t) > 0$ for all t .

Thus to transfer the state from $x(t_0) = (x_{10}, x_{20})$ to $x(t_1) = (x_{11}, x_{21})$ we need only choose y_d that satisfies the above sufficient condition and also has $y_d(t_1) = h(x(t_1)) = x_{11} e^{-x_{21}} > 0$ and $\dot{y}_d(t_1) = -x_{11} y_d(t_1)$.

In particular, if $z_d = \dot{y}_d + y_d^2$ then $z_d(t_0) = -x_{10} y_d(t_0) + y_d(t_0)^2 = y_d(t_0)[y_d(t_0) - x_{10}]$ and $z_d(t_1) = y_d(t_1)[y_d(t_1) - x_{11}]$. Here $y_d(t_0) > 0$, $y_d(t_1) > 0$, and $y_d(t_0) - x_{10} = x_{10} e^{-x_{20}} - x_{10} = x_{10}(e^{-x_{20}} - 1)$ so that

$$z_d(t_0) = y_d(t_0) x_{10} (e^{-x_{20}} - 1) \quad \text{and} \quad z_d(t_1) = y_d(t_1) x_{11} (e^{-x_{21}} - 1).$$

Thus

$$z_d(t_i) \text{ is } \begin{cases} \text{positive} & \text{for } x_{2i} > 0, \\ 0 & \text{for } x_{2i} = 0 \quad \text{for } i = 0, 1, \\ \text{negative} & \text{for } x_{2i} < 0. \end{cases}$$

and times t_s when $z_d(t_s) = 0$ ($x(t_s) \in M_s$) correspond to times when $x_2(t_s) = 0$. Also

$$\begin{aligned} \dot{z}_d(t_s) &= \ddot{y}_d(t_s) + 2y_d(t_s)\dot{y}_d(t_s) \\ &= \frac{\dot{y}_d^2(t_s)}{y_d(t_s)} + 2y_d(t_s)(-y_d^2(t_s)) \\ &= \frac{y_d^4(t_s) - 2y_d^4(t_s)}{y_d(t_s)} = -y_d^3(t_s) \neq 0. \end{aligned}$$

Since $y \equiv y_d$ until time t_s , $y_d > 0$, and $\dot{z}_d(t_s) < 0$.

In conclusion, a trajectory that crosses M_s at time t_s has z_d decreasing, so that if $z_d(t_0) > 0$ we can control the system to make $z_d(t_1) < 0$, but if $z_d(t_0) < 0$ then $z_d(t) < 0$ for all $t \geq t_0$. This means that for initial states with x_2 -coordinate negative the system can be steered to *any* final state in M . If the initial x_2 coordinate is positive, then we can only steer the states to other states with positive x_2 -coordinate.

REFERENCES

- [1] R. W. BROCKETT AND M. D. MESAROVIC, *The reproducibility of multivariable systems*, J. Math. Anal. Appl., 11 (1965), pp. 548–563.
- [2] R. M. HIRSCHORN, *Invertibility of nonlinear control systems*, SIAM J. Control Optim., 17 (1979), pp. 289–297.
- [3] ———, *Output tracking in multivariable nonlinear systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 593–595.
- [4] S. N. SINGH, *Reproducibility in nonlinear systems using dynamic compensation and output feedback*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 955–958.
- [5] ———, *Generalized functional reproducibility condition for nonlinear systems*, IEEE Trans. Automat. Control, 27 (1982), pp. 958–960.
- [6] H. NUMEIJER, *Invertibility of affine nonlinear control systems: a geometric approach*, Systems Control Lett., 2 (1982), pp. 163–168.
- [7] S. N. SINGH AND A. A. SCKY, *Invertibility and robust nonlinear control of robotic systems*, in Proc. 23rd IEEE Conference on Decision and Control, Las Vegas, NV, 1984.
- [8] H. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
- [9] H. SUSSMANN, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Systems Theory (1979), pp. 263–284.
- [10] R. M. HIRSCHORN AND J. H. DAVIS, *Output tracking for nonlinear systems with singular points*, SIAM J. Control and Optim., 25 (1987), pp. 547–557.
- [11] M. GOLUBITSKY AND V. GUILLEMIN, *Stable Mappings and their Singularities*, Springer-Verlag, Berlin, New York, 1973.

REALIZATION AND CANONICITY FOR IMPLICIT SYSTEMS*

JOSÉ GRIMM†

Abstract. This paper studies realization theory and canonicity conditions for implicit systems. Here an implicit system is any system defined by an input-output finite-dimensional, linear, time-invariant relation, in discrete or continuous time (they are also called “generalized” or “descriptor” systems in the literature).

The familiar definitions of canonicity and minimality are extended to such systems, and a so-called “reduced form” proves that, as for classical systems, minimality and canonicity are two equivalent concepts. We replace the notion of transfer function by a notion of external form, and prove that minimal realization can always be achieved.

Key words. implicit systems, canonicity, minimality, transfer function, realization

AMS(MOS) subject classification. 93B

1. Introduction. What we give, in this paper, is a generalization to implicit systems of the (linear time-invariant) state-space theory, that is, to define canonicity, minimality, equivalence, and strong equivalence for implicit systems, in such a way that two canonical equivalent systems are strongly equivalent, and that a system is canonical if and only if it is minimal.

This work extends, and (we hope) improves upon, earlier work concerning so-called “singular” systems [5], or “generalized” systems [13], or “descriptor” systems [10]. The improvement is in two directions.

On the one hand, our definition of canonicity, although strongly related to that of [5] or [13], seems to be more appropriate. An indication of this is that it allows us to derive the fundamental theorem that canonicity is equivalent to minimality, and that the minimal realization is unique (up to strong equivalence). Other remarks in this direction will be given later.

On the other hand, our work holds for singular implicit systems, i.e., those whose pencil $zE - F$ is singular, e.g., not square. This is nice on mathematical grounds, but also from a control-theoretic perspective.

It is well known that such systems may not have a solution for every input function u . However, phenomena exist whose “natural” models exhibit this (bad) property, be it in economics, where the number of variables and equations do not always agree, or in the physical sciences.

Another point is that when the solution exists, it is not generally unique. There is no way around this if the number of equations provided by the “physics” of the phenomenon, for instance, for econometric reasoning, is not large enough, and investigating the possible solutions may then be of interest. But note, anyway, that this is not such a new situation in control theory. As a matter of fact, a stochastic system is a system whose state history is not uniquely defined by the control function. Adding an extra “noise” variable to account for this nonuniqueness must be considered only as a clever device, which might also be used for implicit systems as shown in [2].

As a last instance, observe that a (multivariable) PID controller induces an implicit system and we must know what happens if it turns out to be (close to) singular.

* Received by the editors November 13, 1985; accepted for publication (in revised form) January 4, 1988.

† Institut National de Recherche en Informatique et Automatique, Avenue Emile Hugues, Sophia-Antipolis, 06560 Valbonne, France.

We shall not worry here about the smoothness of inputs or outputs. It should be clear that our input-output relations will be completely determined by their effect on C^∞ pairs whose growth is exponentially bounded. Hence, the reader can always assume smoothness if he or she wishes. However, for discussion of difficulties that may arise when dealing with more general inputs, we refer the reader to [3], [4].

The remark above also allows us to use Laplace transforms in a purely formal manner, as has become customary in system theory. That is, a rational relation between Laplace transforms is understood to hold in some open subset of the complex plane where all functions involved are defined.

A basic knowledge of classical state-space theory (see, e.g., [8]) is assumed.

2. Definitions. In the sequel we shall denote by lower-case letters (e.g., u) time-dependent vectors and matrices, by uppercase boldface letters (e.g., \mathbf{U}) their Laplace transform, by script letters (e.g., \mathcal{H}) rational functions of z (the Laplace variable), and by uppercase letters (e.g., H) constant matrices.

We shall denote by “system” any functional relation between a function $u(t)$ (called the input) with values in \mathbf{R}^m and a function $y(t)$ (called the output) with values in \mathbf{R}^p . Some systems may be defined by linear differential or recurrent equations involving an auxiliary function $x(t)$ with values in \mathbf{R}^n (the state), and we call such systems “implicit systems in internal form,” or internal systems. On the other hand, some other systems may be defined by rational relations linking the Laplace transforms of u and y , and we call them “implicit systems in external form” or external systems. The precise definition of implicit systems will be given later.

For the sake of simplicity, we shall consider only continuous time systems (those defined by differential equations), but everything applies as well to discrete time systems (those defined by recurrent equations).

Two systems are said to be *equivalent* if they define the same input-output relation. In accordance with the introduction, this means that two systems are equivalent if they possess the same respective sets of input-output pairs (u, y) which are smooth, whose growth is exponentially bounded, and which satisfy the equations of the system.

Note that the following definition is the classical definition of an implicit system in the literature [2], [5], [9], [10], [13], [14], [16].

DEFINITION 1. An *implicit system in internal form* is one given by the following equations:

$$\begin{aligned} (1) \quad & E\dot{x} = Fx + Gu, \\ & y = Hx + Ju \end{aligned}$$

where the state x lies in \mathbf{R}^n , E , and F have r rows (i.e., there are r differential equations, and r need not be equal to n) and all the matrices are constant. The matrix E will be called the *kinetic matrix* of the system.

DEFINITION 2. An *implicit system in external form* is one given by the following equations:

$$\begin{aligned} (2) \quad & \mathcal{M}\mathbf{U} = 0, \\ & \mathbf{Y} - \mathcal{H}\mathbf{U} \in \text{Im } \mathcal{L} \end{aligned}$$

where \mathcal{H} , \mathcal{L} , \mathcal{M} are rational matrices, and \mathbf{U} and \mathbf{Y} are the Laplace transforms of the input u and the output y .

Note that ordinary systems are those for which the kinetic matrix E is regular (Definition 1), or $\mathcal{L} = 0$, $\mathcal{M} = 0$, and \mathcal{H} proper (Definition 2).

To begin with, let us list equivalent definitions.

DEFINITION 1a. An implicit system in internal form is a system where the input u and the output y are related by linear time-invariant differential equations involving an auxiliary variable x .

DEFINITION 2a. An implicit system in external form is a system where the Laplace transforms of the input u and the output y are related by a linear relation over the field of rational fractions of the Laplace variable.

As we shall see, all these definitions define the same objects. We first show the equivalence of Definitions 2 and 2a.

Consider a system as in Definition 2a. The system is

$$(3) \quad \mathcal{A} \begin{pmatrix} \mathbf{U} \\ \mathbf{Y} \end{pmatrix} = 0$$

where \mathcal{A} is rational. Let $\mathcal{A} = (\mathcal{B} \quad \mathcal{C})$ and \mathcal{P} be a regular rational matrix such that $\mathcal{P}\mathcal{C} = \begin{pmatrix} \mathcal{D} \\ 0 \end{pmatrix}$ and \mathcal{D} is surjective. Write $\mathcal{P}\mathcal{B} = \begin{pmatrix} \mathcal{E} \\ \mathcal{M} \end{pmatrix}$. Since \mathcal{D} is surjective, there exists \mathcal{K} such that $\mathcal{E} = -\mathcal{D}\mathcal{K}$. Hence (3) is equivalent to

$$\begin{aligned} \mathcal{D}(\mathbf{Y} - \mathcal{K}\mathbf{U}) &= 0, \\ \mathcal{M}\mathbf{U} &= 0. \end{aligned}$$

Taking \mathcal{L} such that $\text{Ker } \mathcal{D} = \text{Im } \mathcal{L}$ we get the form of Definition 2. On the other hand, if we take \mathcal{D} such that $\text{Ker } \mathcal{D} = \text{Im } \mathcal{L}$ in Definition 2, this yields

$$\begin{pmatrix} -\mathcal{D}\mathcal{K} & \mathcal{D} \\ \mathcal{M} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{U} \\ \mathbf{Y} \end{pmatrix} = 0.$$

We now prove the equivalence of Definition 1 and Definition 1a. According to Definition 1a, a system is described by

$$(4) \quad \sum_{i=0}^j A_i x^{(i)} + \sum_{i=0}^k B_i y^{(i)} + \sum_{i=0}^l C_i u^{(i)} = 0$$

where the superscript (i) stands for differentiation. By adding zero matrices, we may assume that $j = k = l$.

We introduce the vector w whose components are $x, x^{(1)}, \dots, x^{(j)}, y, y^{(1)}, \dots, y^{(j)}, u, u^{(1)}, \dots, u^{(j)}$. For $0 \leq i \leq j$, let H_i be the matrix such that $y^{(i)} = H_i w$, G_i such that $u^{(i)} = G_i w$, and F_i such that $x^{(i)} = F_i w$. The following relations, which are valid for $0 \leq i \leq j-1$, can be summarized as $D_0 w' = D_1 w$:

$$\begin{pmatrix} H_i \\ G_i \\ F_i \end{pmatrix} w' = \begin{pmatrix} H_{i+1} \\ G_{i+1} \\ F_{i+1} \end{pmatrix} w.$$

Let $D = \sum_{i=0}^j (A_i F_i + B_i H_i + C_i G_i)$. Then (4) is equivalent to

$$\begin{pmatrix} D_0 \\ 0 \end{pmatrix} w' = \begin{pmatrix} D_1 \\ D \end{pmatrix} w.$$

We get (1) by letting

$$E = \begin{pmatrix} D_0 \\ 0 \end{pmatrix}, \quad F = \begin{pmatrix} D_1 \\ D \\ -G_0 \end{pmatrix}, \quad G = \begin{pmatrix} 0 \\ 0 \\ I \end{pmatrix}, \quad H = H_0, \quad J = 0$$

and by renaming w as x . Conversely it is trivial that Definition 1 is a special case of Definition 1a.

We now prove that Definition 2a is not less general than Definition 1. Using the Laplace transform, (1) may be converted into

$$(zE - F)X = GU,$$

$$Y = HX + JU.$$

This equation is of the form $\mathcal{B}(\mathcal{Y}) + CX = 0$. If \mathcal{D} is such that $\text{Ker } \mathcal{D} = \text{Im } \mathcal{C}$ and $\mathcal{A} = \mathcal{D}\mathcal{B}$, it is equivalent to $\mathcal{A}(\mathcal{Y}) = 0$.

To prove that Definition 2 is, in fact, equivalent to Definition 1 will require some more work, i.e., realization theory, which is the main objective of this paper. For the moment, we content ourselves by introducing the following definition.

DEFINITION 3. An internal system equivalent to an external system is called a *realization* of the external system.

Note. Contrary to the classical case, the matrices \mathcal{K} , \mathcal{L} , \mathcal{M} in Definition 2 or the matrix \mathcal{A} in Definition 2a are not uniquely determined by the input-output relation. If we exclude the two cases where (3) can be written as $U = \mathcal{K}Y$ or $Y = \mathcal{L}U$, there is no natural way of associating to an external representation a well-defined matrix \mathcal{K} that could be considered a transfer function. This explains why we do not define transfer functions, but only external forms.

DEFINITION 4. An implicit system in internal form (Definition 1) is called *minimal* if the size of the system, which is by definition the size of its kinetic matrix E , is minimal (among systems having the same input-output relation), that is, both the number of rows and columns of E are minimal. An implicit system in internal form as given by (1) is called *canonical* if the following relations hold:

$$(CA1) \quad F \text{ Ker } E \subset \text{Im } E,$$

$$(CA2) \quad \forall (\lambda, \mu) \in \mathbb{C} \times \mathbb{C} \quad (\lambda, \mu) \neq (0, 0) \quad \begin{pmatrix} \lambda E - \mu F \\ H \end{pmatrix} \quad \text{injective},$$

$$(CA3) \quad \forall (\lambda, \mu) \in \mathbb{C} \times \mathbb{C} \quad (\lambda, \mu) \neq (0, 0) \quad (\lambda E - \mu F \quad G) \quad \text{surjective}.$$

Note. Contrary to the classical case, it is not clear that a given system is always equivalent to some minimal one, or even that minimal systems exist at all, so we shall have to prove the existence of both minimal and canonical systems.

At the end of the paper we shall return to the above definitions, whose meaning should then become clear to the reader. Note that (CA2) and (CA3) are dual to each other. A moment's consideration will convince the reader that (CA1) is self-dual. Note also that if the kinetic matrix is regular, these are the classical canonicity conditions.

Our main objective is to prove that "minimal" and "canonical" are equivalent notions. To achieve this, we shall need a special type of system, called "reduced," which constitutes the main technical tool in our theory. In the following two definitions, we assume that there exist two regular constant matrices S and T such that

$$u = S \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad y = T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

Such matrices are called input (respectively, output) partition matrices.

DEFINITION 5a. An internal system is said to be *in internal reduced form* if there exist input and output partition matrices such that its equations are:

$$\begin{aligned} \dot{x}_1 &= Fx_1 + Kx_2 + Gu_2, \\ 0 &= Lx_1 + u_1, \\ y_1 &= Hx_1 + Ju_2, \\ y_2 &= x_2, \end{aligned} \tag{5a}$$

where

$$\left. \begin{array}{l} \left(\begin{pmatrix} L \\ H \end{pmatrix}, F \right) \text{ is completely observable} \\ (F, (G \ K)) \text{ is completely reachable} \end{array} \right\} \text{ in the classical sense.}$$

When U and Y are the Laplace transforms of u and y , and u_1, u_2, y_1, y_2 are defined via some input and output partition matrices, we shall often consider a relation such as

$$(5b) \quad \begin{pmatrix} U_1 \\ Y_1 \end{pmatrix} = \begin{pmatrix} \mathcal{P} & \mathcal{Q} \\ \mathcal{R} & \mathcal{S} \end{pmatrix} \begin{pmatrix} U_2 \\ Y_2 \end{pmatrix}$$

and, when referencing (5b), the condition “ $\mathcal{P}, \mathcal{Q}, \mathcal{S}$ are strictly proper rational matrices, \mathcal{R} is proper rational,” will be called condition (R). We can now state the following definition.

DEFINITION 5b. An external system is said to be *in external reduced form* if there exist input and output partition matrices such that its equations are of the form (5b) and condition (R) holds.

When referencing (5b), \mathcal{H} will always mean the matrix $\begin{pmatrix} \mathcal{P} & \mathcal{Q} \\ \mathcal{R} & \mathcal{S} \end{pmatrix}$ and this matrix will be called the *pseudotransfer function* of the reduced system. Note that the pseudotransfer function depends on the system (via S and T) and not only on the input-output relation; also, it is not defined for every system, but only for reduced ones.

Most of the subsequent work is devoted to bringing a system into reduced form, because once this is achieved, classical state-space theory can be applied directly. For instance, the following proposition is obvious.

PROPOSITION 1. A system in internal reduced form is equivalent to a system in external reduced form via the following formula:

$$(6) \quad \begin{pmatrix} \mathcal{P} & \mathcal{Q} \\ \mathcal{R} & \mathcal{S} \end{pmatrix} = \begin{pmatrix} -L \\ H \end{pmatrix} (zI - F)^{-1} (G \ K) + \begin{pmatrix} 0 & 0 \\ J & 0 \end{pmatrix}$$

where the right-hand side is a (classical) minimal factorization of the left-hand side, and the input and output partition matrices are the same.

It is easy to see that an internal reduced system as in Definition 5a is canonical since (CA1) is clearly satisfied, while (CA2) and (CA3) are equivalent to the reachability and observability conditions required in the definition.

Note that a classical system $\dot{x} = Fx + Gu, y = Hx + Ju$ is reduced if and only if it is canonical (in the classical sense) and that a system defined by $Y = \mathcal{H}U$, where $\mathcal{H} = H(zI - F)^{-1}G + J$, is also reduced in external form. In the two cases, the input and output partition matrices are the trivial ones, i.e., $u = u_2$ and $y = y_1$.

DEFINITION 6 (Rosenbrock [11]). For an internal system as given by Definition 1, we define the system matrix as

$$\mathcal{S}(z) = \begin{pmatrix} zE - F & -G \\ H & J \end{pmatrix}.$$

We say that two internal systems are *strongly equivalent*, if their system matrices $\mathcal{S}_1, \mathcal{S}_2$, are related by

$$\mathcal{S}_2 = \begin{pmatrix} U & 0 \\ N & I \end{pmatrix} \mathcal{S}_1 \begin{pmatrix} V & M \\ 0 & I \end{pmatrix}$$

where U and V are constant regular matrices, N , and M are constant.

If the subscripts 1 and 2 refer to the first and second system, respectively, this last equation can be written as

$$\begin{aligned} E_2 &= UE_1 V, & J_2 &= J_1 - NG_1 + H_1 M - NF_1 M, \\ F_2 &= UF_1 V, & E_1 M &= 0, \\ G_2 &= U(G_1 + F_1 M), & NE_1 &= 0, \\ H_2 &= (H_1 - NF_1) V. \end{aligned}$$

PROPOSITION 2. *Two strongly equivalent internal systems are equivalent. Strong equivalence does not change the canonicity or the minimality of an internal system.*

Proof. Adopt the notation of Definition 6 and consider the first system:

$$\begin{aligned} E_1 \dot{x}_1 &= F_1 x_1 + G_1 u, \\ y &= H_1 x_1 + J_1 u. \end{aligned}$$

Let $x = V^{-1}(x_1 - Mu)$. Writing x_1 in terms of x and u , we obtain

$$\begin{aligned} E_1 V \dot{x} &= F_1 Vx + (G_1 + F_1 M)u - E_1 M \dot{u}, \\ y &= (H_1 - NF_1) Vx + (J_1 + H_1 M - NG_1 - NF_1 M)u + N(G_1 u + F_1 Mu + F_1 Vx). \end{aligned}$$

Add the first equation multiplied by $-N$ to the second, and then multiply the first equation on the left by U . Because of $E_1 M = 0$ and $NE_1 = 0$ we get

$$\begin{aligned} E_2 \dot{x} &= F_2 x + G_2 u, \\ y &= H_2 x + J_2 u. \end{aligned}$$

This is the second system. Hence strong equivalence does not change the input-output relation. Note that strong equivalence is a generalization of the notion of change of basis in the state-space for classical systems.

Since strong equivalence does not change the size of the kinetic matrix, it is obvious that minimality is preserved. It is clear that condition (CA1) is preserved. Let us prove that (CA2) is also preserved. Suppose that the first system is canonical. We have to prove that

$$\begin{pmatrix} \lambda E_1 - \mu F_1 \\ H_1 - NF_1 \end{pmatrix} x = 0 \quad \text{implies } x = 0.$$

In fact we need only show that $NF_1 x = 0$. If $\mu \neq 0$ we deduce it from $N(\lambda E_1 - \mu F_1)x = 0$ and $NE_1 = 0$. If $\mu = 0$, since $E_1 x = 0$, by (CA1) there exists y such that $F_1 x = E_1 y$ and $NF_1 x = NE_1 y = 0$. The proof of (CA3) is similar.

We have used (CA1) to prove that (CA2) and (CA3) are preserved by strong equivalence. Later, we shall see the real importance of this condition.

The difference between implicit and ordinary systems lies in (5a) and (5b): part of the input and part of the output (namely u_2 and y_2) may be chosen freely, and the remainder (u_1 and y_1) is determined by a proper rational transformation (in the frequency domain), that is, by state-space equations. It is well known that implicit systems may not have a solution if the input is not chosen in a set of "admissible inputs" (that is, the set $\text{Ker } \mathcal{M}$ of Definition 2), and that if the input is admissible there may be many solutions: in fact, each solution is the sum of a particular solution depending on the input, and an arbitrary element of a certain space ($\text{Im } \mathcal{L}$ of Definition 2) which does not depend on the input.

One way to state this is to say that u_2 and y_2 are natural inputs for the system, while u_1 and y_1 are natural outputs. Going further, we could merge u and y into a single variable w , split it in two parts, w_1 and w_2 , define w_2 as the input and w_1 as the output of the system, and never talk about u and y . Such a theory is developed in [18]. The forthcoming proof of Theorem 2 is essentially that of the corresponding result in [18], but it is a bit more complicated since we have to keep track of our variables u and y .

Comparing classical work on implicit systems, we find that Definitions 5a and b give a completely different, and as far as we know, new point of view. A case of interest is when the system is regular, that is, $Y = \mathcal{H}U$, where \mathcal{H} is rational but not necessarily proper. This case was studied by several authors including those previously mentioned: the general approach was to consider the polynomial and the strictly proper parts of \mathcal{H} , and study them separately.

Our goal is to establish the following theorem, whose proof will require several steps and is carried out in the next three sections.

THEOREM 1. *Each implicit system (internal or external) is equivalent to a minimal one (unique up to strong equivalence). Moreover an internal system is minimal if and only if it is canonical.*

3. Realization. The main objective of this section is to prove the following theorem.

THEOREM 2. *Each implicit system is equivalent to one in reduced form.*

Because of Proposition 1 we have only to prove that, given an implicit system in external form, there exists an equivalent system which is in external reduced form.

We consider a system under external form, written as $\mathcal{A}(\frac{U}{Y}) = 0$, where \mathcal{A} is a rational matrix. Let \mathcal{P} be a regular matrix such that $\mathcal{P}\mathcal{A} = (\frac{\mathcal{A}_0}{0})$, where \mathcal{A}_0 is surjective. The system is then equivalent to $\mathcal{A}_0(\frac{U}{Y}) = 0$. Let $q(z)$ be a common denominator of the entries of \mathcal{A}_0 such that $\mathcal{A}_1 = q(z)\mathcal{A}_0$ is a polynomial matrix. There exists a polynomial regular matrix \mathcal{R} such that $\mathcal{A}_2 = \mathcal{R}\mathcal{A}_1$ is row reduced. Let α_i be the degree of the i th row of \mathcal{A}_2 and $\mathcal{S}(z) = \text{diag}(z^{-\alpha_i})$. Then $\mathcal{A}_3 = \mathcal{S}\mathcal{A}_2$ is proper with surjective constant term, and the system is equivalent to $\mathcal{A}_3(\frac{U}{Y}) = 0$.

We may therefore suppose that the matrix \mathcal{A} is proper with surjective constant term.

Write $\mathcal{A} = (\mathcal{B} \quad \mathcal{C})$ according to the partition $(\frac{U}{Y})$. Let $(B_0 \quad C_0)$ be the constant term of $(\mathcal{B} \quad \mathcal{C})$. There exist regular constant matrices P and T such that $PC_0T = (\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix})$. Write $PB_0 = (\begin{smallmatrix} B_1 \\ B_2 \end{smallmatrix})$. Since $(B_0 \quad C_0)$ is surjective, the matrix $(\begin{smallmatrix} B_1 & I & 0 \\ B_2 & 0 & 0 \end{smallmatrix})$ is surjective, hence B_2 is surjective. Therefore there exists a regular matrix S such that $B_2S = (I \quad 0)$. Let $B_1S = (B_3 \quad B_4)$. We now partition y and u by $T^{-1}y = (\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix})$ and $S^{-1}u = (\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix})$. The system is then equivalent to

$$(P\mathcal{B}S \quad P\mathcal{C}T) \begin{pmatrix} U_1 \\ U_2 \\ Y_1 \\ Y_2 \end{pmatrix} = 0.$$

Note that $(P\mathcal{B}S \quad P\mathcal{C}T)$ is of the form

$$\begin{pmatrix} \mathcal{D}_1 & \mathcal{D}_2 & \mathcal{D}_3 & \mathcal{D}_4 \\ \mathcal{D}_5 & \mathcal{D}_6 & \mathcal{D}_7 & \mathcal{D}_8 \end{pmatrix}$$

which is proper and has constant term

$$\begin{pmatrix} B_3 & B_4 & I & 0 \\ I & 0 & 0 & 0 \end{pmatrix}.$$

The matrix $\begin{pmatrix} \mathcal{P}_1 & \mathcal{Q}_1 \\ \mathcal{P}_2 & \mathcal{Q}_2 \end{pmatrix}$ is therefore regular and, in fact, bicausal. Premultiply our system by the inverse of this matrix. We get

$$\begin{pmatrix} I & -\mathcal{P} & 0 & -\mathcal{Q} \\ 0 & -\mathcal{R} & I & -\mathcal{S} \end{pmatrix} \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \\ \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = 0,$$

that is,

$$\begin{pmatrix} \mathbf{U}_1 \\ \mathbf{Y}_1 \end{pmatrix} = \begin{pmatrix} \mathcal{P} & \mathcal{Q} \\ \mathcal{R} & \mathcal{S} \end{pmatrix} \begin{pmatrix} \mathbf{U}_2 \\ \mathbf{Y}_2 \end{pmatrix}.$$

Note that the matrix $\begin{pmatrix} \mathcal{P} & \mathcal{Q} \\ \mathcal{R} & \mathcal{S} \end{pmatrix}$ is proper and has constant term $\begin{pmatrix} 0 & 0 \\ B_4 & 0 \end{pmatrix}$. This proves the theorem.

We now examine the extent to which such a form is unique. Suppose that we have two such forms:

$$\begin{pmatrix} I & -\mathcal{P} & 0 & -\mathcal{Q} \\ 0 & -\mathcal{R} & I & -\mathcal{S} \end{pmatrix} \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \\ \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = 0, \quad \begin{pmatrix} I & -\bar{\mathcal{P}} & 0 & -\bar{\mathcal{Q}} \\ 0 & -\bar{\mathcal{R}} & I & -\bar{\mathcal{S}} \end{pmatrix} \begin{pmatrix} \bar{\mathbf{U}}_1 \\ \bar{\mathbf{U}}_2 \\ \bar{\mathbf{Y}}_1 \\ \bar{\mathbf{Y}}_2 \end{pmatrix} = 0$$

with

$$y = T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad u = S \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad \bar{y} = \bar{T} \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix}, \quad \bar{u} = \bar{S} \begin{pmatrix} \bar{u}_1 \\ \bar{u}_2 \end{pmatrix}.$$

We know that T and \bar{T} , S and \bar{S} are regular. Write

$$\bar{T}^{-1}T = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix}, \quad \bar{S}^{-1}S = \begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix};$$

then

$$\begin{pmatrix} \bar{\mathbf{U}}_1 \\ \bar{\mathbf{U}}_2 \\ \bar{\mathbf{Y}}_1 \\ \bar{\mathbf{Y}}_2 \end{pmatrix} = \begin{pmatrix} B_1 & B_2 & 0 & 0 \\ B_3 & B_4 & 0 & 0 \\ 0 & 0 & A_1 & A_1 \\ 0 & 0 & A_3 & A_4 \end{pmatrix} \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \\ \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}.$$

It follows that the two matrices

$$\begin{pmatrix} I & -\mathcal{P} & 0 & -\mathcal{Q} \\ 0 & -\mathcal{R} & I & -\mathcal{S} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} I & -\bar{\mathcal{P}} & 0 & -\bar{\mathcal{Q}} \\ 0 & -\bar{\mathcal{R}} & I & -\bar{\mathcal{S}} \end{pmatrix} \begin{pmatrix} B_1 & B_2 & 0 & 0 \\ B_3 & B_4 & 0 & 0 \\ 0 & 0 & A_1 & A_2 \\ 0 & 0 & A_3 & A_4 \end{pmatrix}$$

have the same kernel.

The latter matrix is computed to be

$$\begin{pmatrix} B_1 - \bar{\mathcal{P}}B_3 & B_2 - \bar{\mathcal{P}}B_4 & -\bar{\mathcal{Q}}A_3 & -\bar{\mathcal{Q}}A_4 \\ -\bar{\mathcal{R}}B_3 & -\bar{\mathcal{R}}B_4 & A_1 - \bar{\mathcal{S}}A_3 & A_2 - \bar{\mathcal{S}}A_4 \end{pmatrix}.$$

Write

$$w_1 = \begin{pmatrix} u_1 \\ y_1 \end{pmatrix}, \quad w_2 = \begin{pmatrix} u_2 \\ y_2 \end{pmatrix}, \quad \mathcal{H} = \begin{pmatrix} \mathcal{P} & \mathcal{Q} \\ \mathcal{R} & \mathcal{S} \end{pmatrix}.$$

The first system becomes $(I - \mathcal{H})(\mathbf{w}_1) = 0$. Also write

$$\mathcal{C}_1 = \begin{pmatrix} B_1 - \bar{\mathcal{P}}B_3 & -\bar{\mathcal{Q}}A_3 \\ -\bar{\mathcal{R}}B_3 & A_1 - \bar{\mathcal{P}}A_3 \end{pmatrix}, \quad \mathcal{C}_2 = \begin{pmatrix} B_2 - \bar{\mathcal{P}}B_4 & -\bar{\mathcal{Q}}A_4 \\ -\bar{\mathcal{R}}B_4 & A_2 - \bar{\mathcal{P}}A_4 \end{pmatrix}.$$

The second system is then $(\mathcal{C}_1 \quad \mathcal{C}_2)(\mathbf{w}_2) = 0$. Hence $\mathbf{W}_1 = \mathcal{H}\mathbf{W}_2$ is equivalent to $\mathcal{C}_1\mathbf{W}_1 + \mathcal{C}_2\mathbf{W}_2 = 0$. This implies

$$(7) \quad \mathcal{C}_2 = -\mathcal{C}_1\mathcal{H}.$$

Hence $\mathbf{W}_1 = \mathcal{H}\mathbf{W}_2$ is equivalent to $\mathcal{C}_1(\mathbf{W}_1 - \mathcal{H}\mathbf{W}_2) = 0$, and hence \mathcal{C}_1 is injective. Since $(\mathcal{C}_1 \quad \mathcal{C}_2) = (\mathcal{C}_1 \quad -\mathcal{C}_1\mathcal{H})$ is surjective, we also have that \mathcal{C}_1 is surjective. This establishes a first result, namely, Proposition 3a.

PROPOSITION 3a. *Two equivalent systems of the form (5b) satisfy (7), where \mathcal{C}_1 is regular.*

Assume now that condition (R) holds for both systems. The matrices \mathcal{C}_1 , \mathcal{C}_2 , and \mathcal{H} are proper. Equating the constant terms in (7) yields

$$\begin{pmatrix} B_2 & 0 \\ -\bar{J}B_4 & A_2 \end{pmatrix} = -\begin{pmatrix} B_1 & 0 \\ -\bar{J}B_3 & A_1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ J & 0 \end{pmatrix}$$

where J and \bar{J} are the constant terms of \mathcal{R} and $\bar{\mathcal{R}}$. We deduce $B_2 = 0$ and $A_2 = 0$. Since $\begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix}$ is regular, and $A_2 = 0$, the matrix A_1 is surjective. In the same manner B_1 is also surjective. The constant term of \mathcal{C}_1 , which is $\begin{pmatrix} B_1 & 0 \\ -\bar{J}B_3 & A_1 \end{pmatrix}$ is therefore surjective. But \mathcal{C}_1 is square; hence A_1 and B_1 are regular and \mathcal{C}_1 is bicausal. From the fact that A_1 is regular and A_2 is zero we deduce that A_4 is also regular. In the same fashion, B_4 is regular. We summarize this as Proposition 3b.

PROPOSITION 3b. *Two equivalent systems of the form (5b) satisfying condition (R) also satisfy $A_2 = 0$, $B_2 = 0$, A_1 , A_4 , B_1 , B_4 regular.*

4. Minimization. The aim of this section is to provide us with an algorithm to obtain a reduced internal system from a given internal one without increasing any dimension of the kinetic matrix.

THEOREM 3. *Given an implicit system in internal form Σ , it can be brought under reduced form, by means of two types of operations: (a) strong equivalence operations; (b) equivalence operations that decrease the number of rows or columns of the kinetic matrix.*

Moreover, if the system is canonical, only strong equivalence operations are necessary.

The proof of this theorem is an algorithm. We leave it to the reader to check that every operation we perform is of type (a) or (b).

We start from

$$E\dot{x} = Fx + Gu,$$

$$y = Hx + Ju.$$

Step I. There exist two regular matrices P and Q such that $PEQ = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$. Putting $Q\tilde{x} = x$, the system is equivalent to

$$PEQ\dot{\tilde{x}} = PFQ\tilde{x} + PGu,$$

$$y = HQ\tilde{x} + Ju.$$

Let

$$PFQ = \begin{pmatrix} F & K \\ L & M \end{pmatrix}, \quad PG = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}, \quad HQ = (H_1 \quad H_2), \quad \tilde{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

(with a slight ambiguity of notation in that we used F again). The system is then

$$(8) \quad \begin{aligned} \dot{x}_1 &= Fx_1 + Kx_2 + G_1u, \\ 0 &= Lx_1 + Mx_2 + G_2u, \\ y &= H_1x_1 + H_2x_2 + Ju. \end{aligned}$$

In the sequel, when dealing with a similar situation, we shall simply say “By change of variable we may suppose $E = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$ ” and it will be understood that G_1 , G_2 , H_1 , H_2 are partitions of the new G and H with respect to the new form of E .

Each time we perform such manipulations, there are some particular cases, when some variables or some equations are missing.

For example, here there are four special cases:

- (a) E is 0.
- (b) E is injective, not surjective: we may write $E = \begin{pmatrix} I \\ 0 \end{pmatrix}$.
- (c) E is surjective, not injective: we may write $E = \begin{pmatrix} I & 0 \end{pmatrix}$.
- (d) E is invertible: we may write $E = I$.

We will not consider these special cases which can be covered by easy modifications and are left to the reader. (Anyway, our proof is complete if we expect to work with matrices with zero rows or columns.)

Step II. For a system defined by (8), condition (CA1) is equivalent to $M = 0$. In this step we make M vanish, if it is not already zero. By change of variable we may suppose $M = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$. The system is

$$\begin{aligned} \dot{x}_1 &= Fx_1 + K_1x_{21} + K_2x_{22} + G_1u, \\ 0 &= L_1x_1 + x_{21} + G_{21}u, \\ 0 &= L_2x_1 + G_{22}u, \\ y &= H_1x_1 + H_{21}x_{21} + H_{22}x_{22} + Ju. \end{aligned}$$

We compute x_{21} from the second equation, and substitute in the others. We get

$$\begin{aligned} \dot{x}_1 &= (F - K_1L_1)x_1 + K_2x_{22} + (G_1 - K_1G_{21})u, \\ 0 &= L_2x_1 + G_{22}u, \\ y &= (H_1 - H_{21}L_1)x_1 + H_{22}x_{22} + (J - H_{21}G_{21})u. \end{aligned}$$

The system is now under form (8) with $M = 0$. Note that the number of rows and columns of the kinetic matrix has decreased by rank M .

Step III. We now introduce an output partition matrix T as in Definitions 5a and b, to get a form which is closer to the reduced form than (8).

There exist regular matrices T and P such that $T^{-1}H_2P = \begin{pmatrix} 0 & 0 \\ I & 0 \end{pmatrix}$. By change of variables we may therefore assume that the system is

$$\begin{aligned} \dot{x}_1 &= Fx_1 + K_1x_{21} + K_2x_{22} + G_1u, \\ 0 &= Lx_1 + G_2u, \\ y_1 &= H_{11}x_1 + J_1u, \\ y_2 &= H_{12}x_1 + x_{21} + J_2u, \\ y &= T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \end{aligned}$$

When we rename $H_{12}x_1 + x_{21} + J_2u$ as x_2 , H_{11} as H , $F - K_1H_{12}$ as F , $G_1 - K_1J_2$ as G_1 , J_1 as J , and x_{22} as x_3 , we get

$$(9) \quad \begin{aligned} \dot{x}_1 &= Fx_1 + K_1x_2 + K_2x_3 + G_1u, \\ 0 &= Lx_1 + G_2u, \\ y_2 &= x_2, \\ y_1 &= Hx_1 + Ju, \\ y &= T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \end{aligned}$$

Note that the size of y_2 is the rank of the matrix H_2 in (8).

Step IV. The observability condition (CA2) for a system such as (9) is: (a) $((\begin{smallmatrix} L \\ H \end{smallmatrix}), F)$ is completely observable; (b) there is no x_3 in the system.

In this step we achieve (a). Suppose $((\begin{smallmatrix} L \\ H \end{smallmatrix}), F)$ is not observable. Put it into nonobservable characteristic form. The system is

$$\begin{aligned} \dot{x}_{11} &= F_{11}x_{11} + F_{12}x_{12} + K_{11}x_2 + K_{21}x_3 + G_{11}u, \\ \dot{x}_{12} &= F_{22}x_{12} + K_{12}x_2 + K_{22}x_3 + G_{12}u, \\ 0 &= L_1x_{12} + G_2u, \\ y_2 &= x_2, \\ y_1 &= H_1x_{12} + Ju, \\ y &= T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \end{aligned}$$

where $((\begin{smallmatrix} L_1 \\ H_1 \end{smallmatrix}), F_{22})$ is observable.

The variable x_{11} enters only in the first equation, and is not needed to compute the output. Moreover, this equation always has a solution for any value of the other variables. Hence removing it does not change the set of admissible inputs (a point that must be checked before dropping an equation in an implicit system). We therefore remove it from the system. Now (a) holds.

Step V. The purpose of this step is to suppress x_3 in (9). We may assume $K_2 \neq 0$; otherwise we are done.

Let $x_3 = \dot{\tilde{x}}_3$ and $\tilde{x}_1 = x_1 - K_2\tilde{x}_3$. We obtain

$$\begin{aligned} \dot{\tilde{x}}_1 &= F\tilde{x}_1 + K_1x_2 + FK_2\tilde{x}_3 + G_1u, \\ 0 &= L\tilde{x}_1 + LK_2\tilde{x}_3 + G_2u, \\ y_2 &= x_2, \\ y_1 &= H\tilde{x}_1 + HK_2\tilde{x}_3 + Ju, \\ y &= T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \end{aligned}$$

We have to consider three cases:

(1) $LK_2 \neq 0$. Here the system is of the form (8) with $M \neq 0$. We continue the algorithm with Step II, hence decreasing the size of the kinetic matrix.

(2) $LK_2 = 0$, $HK_2 \neq 0$. Here the system is of the form (8) with $M = 0$ and $H_2 = T(\begin{smallmatrix} 0 & I \\ HK_2 & 0 \end{smallmatrix})$. We continue the algorithm with Step III, so that the size of y_2 is increased by the rank of HK_2 , which is nonzero.

(3) $LK_2 = 0$, $HK_2 = 0$. Here the system is exactly as before Step V, with K_2 replaced by FK_2 . We again execute Step V.

We now show that the procedure explained so far eventually ends. Indeed, (1) can be performed only a finite number of times, since it decreases the number of rows and columns of the kinetic matrix, while (2) and (3) leave it unchanged. So, eventually, we meet only (2) and (3). But (2) cannot be performed an infinite number of times, since it increases the size of y_2 (which is bounded by that of y), while (3) leaves it unchanged. Thus, eventually, we meet only (3).

Suppose we have processed (3) of Step V, p times consecutively. For $n < p$ we had $LF^n K_2 = 0$ and $HF^n K_2 = 0$. But $((\begin{smallmatrix} L \\ H \end{smallmatrix}), F)$ is observable. Hence $LF^n K_2 = 0$, and $HF^n K_2 = 0$ for all $n < (\text{size } F)$ implies $K_2 = 0$. Since we have assumed $K_2 \neq 0$, we get $p \leq (\text{size } F)$. Thus (3) cannot be executed an infinite number of times. This shows that the procedure terminates in a finite number of steps. We have obtained

$$\dot{x}_1 = Fx_1 + K_1 x_2 + G_1 u,$$

$$0 = Lx_1 + G_2 u,$$

$$y_2 = x_2,$$

$$y_1 = Hx_1 + Ju.$$

Step VI. We introduce here an input partition matrix as follows. Consider two regular matrices P and S such that $PG_2 S = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$ and write $S^{-1}u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$. This leads to

$$\dot{x}_1 = Fx_1 + Kx_2 + G_3 u_1 + Gu_2,$$

$$0 = L_1 x_1 + u_1,$$

$$0 = L_2 x_1,$$

$$y_2 = x_2,$$

$$y_1 = Hx_1 + J_1 u_1 + J_2 u_2.$$

Using the second equation, we may assume $G_3 = 0$ and $J_1 = 0$. We proceed then with the dual of Step IV to make $(F, (K \ G))$ completely reachable, and then perform the dual of Step V, replacing $0 = L_2 x_1$ by its derivative $0 = L_2 Fx_1 + L_2 Kx_2 + L_2 Gu_2$, and consider, as in Step V, three cases: $L_2 K \neq 0$, $L_2 K = 0$ and $L_2 G \neq 0$, and $L_2 K = L_2 G = 0$, and conclude in the same manner.

If the resulting system satisfies (CA2), we are through. Otherwise we go back to Step IV. This decreases the size of the kinetic matrix and proves that the algorithm terminates.

If the original system is canonical, Steps II, IV, or V need never be performed, and only strong equivalence transformations are needed, thanks to Proposition 2. This proves the last assertion. Note that if the original system is minimal, operations of type (b) cannot be used in the algorithm, so that we have the following corollary.

COROLLARY. *A minimal system is canonical.*

5. Proof of the main theorem. We first prove the following theorem.

THEOREM 4. *Two equivalent and reduced internal systems are strongly equivalent.*

Let Σ and $\bar{\Sigma}$ be two reduced internal systems. From Proposition 1, we get their reduced external form. We adopt the notation of Proposition 3. Let T , \bar{T} , S , and \bar{S} be

their input-output partition matrices, $A = \bar{T}^{-1}T$, $B = \bar{S}^{-1}S$. By Proposition 3b, we know that

$$A = \begin{pmatrix} A_1 & 0 \\ A_3 & A_4 \end{pmatrix}, \quad B = \begin{pmatrix} B_1 & 0 \\ B_3 & B_4 \end{pmatrix},$$

with A_1 , A_4 , B_1 , and B_4 regular.

Now perform the strong equivalence transformation defined by

$$U = I, \quad N = 0, \quad V = \begin{pmatrix} I & 0 \\ -A_4^{-1}A_3H & A_4^{-1} \end{pmatrix}, \quad M = \begin{pmatrix} 0 & 0 \\ 0 & -A_4^{-1}A_3J \end{pmatrix} S^{-1}$$

on Σ . The effect of the above is to put the system under a reduced form where the new output partition matrix is \bar{T} . Denote with a hat the matrices of this new system, and perform on it the strong equivalence defined by

$$V = I, \quad M = 0, \quad U = \begin{pmatrix} I & \hat{G}B_4^{-1}B_3 \\ 0 & B_1 \end{pmatrix}, \quad N = \bar{T} \begin{pmatrix} 0 & -\hat{J}B_4^{-1}B_3 \\ 0 & 0 \end{pmatrix}.$$

The effect of the above is to produce a system under reduced form where the new input partition matrix is \bar{S} .

This leads us to a new system $\tilde{\Sigma}$, strongly equivalent to Σ , having the same input-output partition matrices as $\bar{\Sigma}$. We can use Proposition 3a, and since the matrices A and B are now the identity, (7) is just

$$\tilde{\mathcal{P}} = \bar{\mathcal{P}}, \quad \tilde{\mathcal{Q}} = \bar{\mathcal{Q}}, \quad \tilde{\mathcal{R}} = \bar{\mathcal{R}}, \quad \tilde{\mathcal{J}} = \bar{\mathcal{J}},$$

that is

$$\begin{pmatrix} \tilde{L} \\ \tilde{H} \end{pmatrix} (zI - \tilde{F})^{-1} (\tilde{K} \quad \tilde{G}) + \tilde{J} = \begin{pmatrix} \bar{L} \\ \bar{H} \end{pmatrix} (zI - \bar{F})^{-1} (\bar{K} \quad \bar{G}) + \bar{J}.$$

Equating the constant terms yields $\tilde{J} = \bar{J}$. Because of the canonicity conditions, there exists a regular matrix P such that

$$\bar{F} = P\tilde{F}P^{-1}, \quad \begin{pmatrix} \bar{L} \\ \bar{H} \end{pmatrix} = \begin{pmatrix} \tilde{L} \\ \tilde{H} \end{pmatrix} P^{-1}, \quad (\bar{K} \quad \bar{G}) = P(\tilde{K} \quad \tilde{G}).$$

Hence replacing x_1 by Px_1 (i.e., using the strong equivalence defined by $U = V^{-1} = \begin{pmatrix} P & 0 \\ 0 & I \end{pmatrix}$, $M = N = 0$), we have brought Σ into $\tilde{\Sigma}$, thus proving that the two systems are strongly equivalent.

The proof of Theorem 1 is now easy.

By Theorem 3 and its corollary, we know that a minimal system is canonical and strongly equivalent to a reduced system. By Theorem 4, we deduce that two equivalent and minimal systems are strongly equivalent. Since by Theorem 2 each system is equivalent to an internal reduced one, we need only prove that a canonical system is minimal.

For this purpose, consider a canonical system Σ and an equivalent system Σ_1 . By Theorem 3 we can bring them both into reduced form and call the result $\tilde{\Sigma}$ and $\tilde{\Sigma}_1$.

By Theorem 4, $\tilde{\Sigma}$ and $\tilde{\Sigma}_1$ are of the same size (since they are strongly equivalent). By Theorem 3, Σ and $\tilde{\Sigma}$ are of the same size (since Σ is canonical), and $\tilde{\Sigma}_1$ is not of greater size than Σ_1 , so that Σ is not of greater size than Σ_1 . This shows that Σ is minimal, thereby achieving the proof.

In the course of the above we have established a result that is worth stating separately as Proposition 4.

PROPOSITION 4. *Given an implicit system in the form*

$$\begin{pmatrix} \mathbf{U}_1 \\ \mathbf{Y}_1 \end{pmatrix} = \mathcal{H} \begin{pmatrix} \mathbf{U}_2 \\ \mathbf{U}_2 \end{pmatrix}, \quad y = T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad u = S \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

the size of \mathcal{H} depends only on the input-output relation.

Better yet, if the system is reduced (condition R), the sizes of u_1 , u_2 , y_1 , y_2 are uniquely determined.

The first assertion is true because the matrix \mathcal{C}_1 of Proposition 3a is square. The second is true because in Proposition 3b, A_1 , A_4 are square.

6. Concluding remarks.

6.1. Normal form. In the previous literature dealing with regular systems [4], [5], [9], [10], [13], [15], [16], another form has been introduced to handle implicit systems. This form, closely related to the Kronecker form of the pair E , F , has been extended to the general case in [6], where it has been called *normal form*, and goes as follows:

$$\begin{aligned} \dot{x}_1 &= Fx_1 + P_1x_4 + G_1u_2, \\ N\dot{x}_2 &= x_2 + G_2u_2, \\ \dot{x}_3 &= Ax_3 + Bv + P_2x_4 + P_3x_1 + G_3u_2, \\ \dot{x}_4 &= Cx_4 + G_4u_2, \\ 0 &= Dx_4 + u_1, \\ y_1 &= H_1x_1 + H_2x_2 + H_3x_3 + H_4x_4 + Ju_2, \\ y_2 &= v \end{aligned}$$

where

$$y = T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad u = S \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

v is an arbitrary function, and N is nilpotent with no Jordan blocks of size one. Moreover (A, B) and (N, G_2) are completely reachable as well as $((\begin{smallmatrix} F & P_1 \\ 0 & C \end{smallmatrix}), (\begin{smallmatrix} G_1 \\ G_2 \end{smallmatrix}))$, while (D, C) and (H_2, N) are completely observable as well as $((H_1 \ H_2), (\begin{smallmatrix} F & 0 \\ P_3 & A \end{smallmatrix}))$. This form is easily seen to be canonical, and hence minimal.

It is clear that the notion of pseudotransfer function can also be defined for normal systems. If, for instance, we let

$$\mathcal{P} = -D(zI - C)^{-1}G_4,$$

$$\mathcal{Q} = 0,$$

$$\mathcal{R} = \begin{pmatrix} H_1 & H_3 & H_4 \end{pmatrix} \begin{pmatrix} zI - F & 0 & -P_1 \\ -P_3 & zI - A & -P_2 \\ 0 & 0 & zI - C \end{pmatrix}^{-1} \begin{pmatrix} G_1 \\ G_3 \\ G_4 \end{pmatrix} + H_2(zN - I)^{-1}G_2 + J,$$

$$\mathcal{S} = H_3(zI - A)^{-1}B,$$

then (5b) holds.

The normal form in [6] plays the same technical role that the reduced form does in the present paper, and allows us to prove our main theorem, thanks to the fact that any system admits an equivalent normal form.

We have introduced the normal form here because of the following fact. When dealing with implicit systems, we can bring the system in explicit form, such as (5b). From a system theoretic viewpoint, it is then reasonable to try to minimize the size of u_1 , and maximize the size of y_1 . By Proposition 4, these two operations are equivalent, and the result is Proposition 5.

PROPOSITION 5. *If we want \mathcal{H} to be proper, reduced systems give minimal size for u_1 . In the general case, normal systems give minimal size for u_1 .*

Proof. Let $\Sigma, \bar{\Sigma}$ be described by (5b). Assume first that $\bar{\Sigma}$ is normal; hence $\bar{\mathcal{Q}} = 0$. By Proposition 3a, \mathcal{C}_1 is invertible. Hence $B_1 - \bar{\mathcal{P}}B_3$ is surjective; hence the number of its columns (i.e., the size of u_1) is at least equal to its number of rows (i.e., the size of \bar{u}_1).

Now assume $\bar{\Sigma}$ is reduced, and that \mathcal{H} is proper. If P_0 denotes the constant term of \mathcal{P} , equating constant terms in (7) yields $B_1P_0 = -B_2$. Since $\begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix}$ is invertible, it follows that B_1 is surjective. We conclude as before.

6.2. Feedback. Feedback for implicit systems is not yet well understood. For example, it is easy to see that pole-placement can be achieved. Indeed, if the system under reduced form, since $(F, (K \ G))$ is completely reachable, there exist two matrices C_1 and C_2 such that $\bar{F} = F - GC_1 - KC_2$ has preassigned poles.

Consider $\bar{u}_1 = u_1 + Lx_1$, $\bar{u}_2 = u_2 + C_1x_1$, and $\bar{x}_2 = x_2 + C_2x_1$. Then the system is

$$\dot{x}_1 = \bar{F}x_1 + K\bar{x}_2 + G\bar{u}_2,$$

$$0 = \bar{u}_1$$

up to the output equations. However, we cannot say that we control the system even if \bar{F} is stable, because \bar{x}_2 can do anything.

6.3. Inversion of systems. The reduced form is almost symmetric in u and y . Almost, because \mathcal{Q} is supposed to be strictly proper, and \mathcal{R} proper.

Suppose that \mathcal{R} is strictly proper. Then interchanging u and y gives immediately, without any computation, the inverse system. When \mathcal{R} is not strictly proper, this is a little more complicated.

In fact, implicit systems provide the natural framework for the study of invertibility of systems (see [7]).

Related to feedback and inversion of systems is the problem of optimal control; these can be solved within this framework, despite the nonunicity of the state (cf. [17]).

6.4. Parameterization. Despite the fact that we cannot associate to an implicit system a well-defined matrix characterizing the input-output relation, the existence of pseudo-transfer function for reduced systems and its properties allows us to parameterize implicit systems, in the following way.

Given the sizes of the input and the output, by Proposition 4, the sizes of u_1 , u_2 , y_1 , and y_2 are uniquely defined. The proof of Theorem 4 shows that the size of F is also uniquely determined, and so is the size of x_1 , which can be considered as the real state of the system. The set of all implicit systems having a given size for u_1 , y_1 , and x_1 is thus stable under equivalence, and, on the quotient space, a C^∞ manifold structure can be defined, the only difficulty being to take into account the matrices S and T . As a result, the number of parameters to consider is the number of entries of the matrices A_2 and B_2 in Proposition 3b. The other matrices (F, K, G, L, H, J) are handled as in the classical theory (see [1]). This is another indication that the present approach is convenient for handling parameterization problems for implicit systems.

6.5. Interpretation of the canonicity conditions. The condition (CA1) says that the system has no *nondynamic* variables (in the sense of [13]). Indeed, if the system has nondynamic variables, it can be written in the following form:

$$E\dot{x} = Fx + Gu + Kx_1,$$

$$0 = x_1 + G_1u,$$

$$y = \dots$$

Clearly (CA1) is not satisfied. Conversely, if (CA1) is not satisfied, Step II of the algorithm extracts the nondynamic variables $(x_2 + L_1x_1)$ and suppresses them.

Now suppose that (CA1) holds. Execute Steps I-III of the algorithm. In fact, Step II is unnecessary because of (CA1). Hence, all we need is strong equivalence.

We obtain the following:

$$\begin{aligned} \dot{x}_1 &= Fx_1 + K_1x_2 + K_2x_3 + G_1u, \\ 0 &= Lx_1 + G_2u, \\ (*) \quad y_2 &= x_2, \\ y_1 &= Hx_1 + Ju. \end{aligned}$$

If we suppose that $u = 0$ and $y = 0$ we get

$$\begin{aligned} \dot{x}_1 &= Fx_1 + K_2x_3, \\ (**) \quad 0 &= Lx_1, \\ 0 &= x_2, \\ 0 &= Hx_1. \end{aligned}$$

If (CA2) holds, then there is no x_3 , and (x_1, x_2) is the state (because we did not execute Step II). From the observability of $((\frac{L}{H}), F)$, we can deduce $x_1 = 0$. Hence the system is “observable” according to the classical definitions.

Conversely, if the system is observable, clearly $((\frac{L}{H}), F)$ has to be an observable pair. Also, this implies that there is no x_3 in the system. Indeed, for instance, suppose that $x_3 = 0$ on $[0, T[$ and $x_3(T) \neq 0$. Then (**) is satisfied on $[0, T]$ but x_3 is not zero, and the system would not be observable.

In the same manner, (CA3) is equivalent to the “reachability” of the system.

7. Conclusion. If we had defined canonicity only by (CA2) and (CA3), we would have run into serious difficulties. Consider for instance the system

$$0 = x + u, \quad y = x.$$

This strange system would be canonical, but upon introducing $x_1 = x + u$ (strong equivalence transformation), the system becomes

$$x_1 = 0, \quad y = -u \quad !!$$

In fact, (CA2) and (CA3) are equivalent to observability and controllability as defined by Cobb [5]. But in [5], nothing is said about minimality, and the above transformation is forbidden anyway.

If we had defined minimality by minimizing the rank of the kinetic matrix, we would also have encountered difficulties. The associated canonicity conditions are rather involved already in the regular case (see Verghese, Levy, and Kailath [13]), and their generalization seems to be delicate.

In this paper, we approach the problem by changing both the canonicity definition of [5] (adding the condition (CA1)), and the minimality condition of [13] (replacing the minimality of the rank of the kinetic matrix by the minimality of the number of both rows and columns of the kinetic matrix). This allows us to establish the equivalence between our two notions of canonicity and minimality in an enlarged setting.

The following facts, which are easy to prove, show the differences between these frameworks:

- (1) A "Cobb canonical" system is "Verghese equivalent" to a canonical system.
- (2) A "Verghese minimal" system is "Verghese equivalent" to a minimal system.
- (3) A minimal system is "Verghese minimal."
- (4) A canonical system is "Cobb canonical."
- (5) A "Cobb canonical" system is "Verghese minimal."
- (6) A "Verghese minimal" system is NOT always "Cobb canonical."

Acknowledgments. The author is most grateful to Professor P. Bernhard, who introduced him to implicit system theory, and provided him with constant support when supervising his thesis. He also thanks L. Baratchart and Professor J. C. Willems for their suggestions and criticisms.

REFERENCES

- [1] L. BARATCHART AND J. GRIMM, *Une structure différentielle pour les systèmes implicites*, Proc. 7th Internat. Conf. on Analysis and Optimization of Systems, Lecture Notes in Control and Inform. Sci., 83, Springer-Verlag, Berlin, New York, 1986.
- [2] P. BERNHARD, *On singular implicit linear dynamical systems*, SIAM J. Control Optim., 20 (1982), pp. 612-633.
- [3] D. COBB, *Feedback and pole placement in descriptor-variable systems*, Internat. J. Control, 33 (1981), pp. 1135-1146.
- [4] ———, *Descriptor-variable systems and optimal state regulation*, IEEE Trans. Automat. Control, 28 (1983), pp. 601-611.
- [5] ———, *Controllability observability and duality in singular systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 1076-1082.
- [6] J. GRIMM, *Sur les systèmes dynamiques linéaires implicites singuliers*, Thèse de troisième cycle, Université de Paris IX, Paris, 1983.
- [7] ———, *Application de la théorie des systèmes implicites à l'inversion des systèmes*, Proc. 6th Internat. Conf. on Analysis and Optimization of Systems, Lecture Notes in Control and Inform. Sci., 63 (1984), pp. 142-156.
- [8] R. KALMAN, P. FALB, AND M. ARBIB, *Topics in Mathematical Theory*, McGraw-Hill, New York, 1979.
- [9] D. G. LUENBERGER, *Dynamical equations in descriptor form*, IEEE Trans. Automat. Control, 22 (1976), pp. 312-321.
- [10] ———, *Time-invariant descriptor systems*, Automatica, 14 (1978), pp. 473-480.
- [11] H. H. ROSENBRICK, *State-Space and Multivariable Theory*, Nelson, London, 1970.
- [12] L. M. SILVERMAN AND M. L. J. HAUTUS, *System structure and singular control*, Linear Algebra Appl., 50 (1983), pp. 369-402.
- [13] G. C. VERGHESE, B. C. LEVY, AND T. KAILATH, *A generalized state-space for singular systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 811-831.
- [14] G. VERGHESE, *Infinite-frequency behavior in generalized dynamical systems*, Ph.D. thesis, Stanford Univ., Stanford, CA, December 1978.
- [15] J. H. WILKINSON, *The differential system $B\dot{x} = Ax$ and the generalized eigenvalue problem $Au = \lambda Bu$* , Report NAC73, National Physics Laboratory, U.K.
- [16] ———, *Linear differential equations and Kronecker's canonical form*, in Recent Advances in Numerical Analysis, C. deBoor and G. H. Golub, eds., Academic Press, New York, 1979.
- [17] X. M. WANG, P. BERNHARD, AND J. GRIMM, *Commande optimale linéaire quadratique des systèmes implicites discrets*, C.R. Acad. Sci. Paris, Sér. I, 303 (1986), pp. 127-130.
- [18] J. C. WILLEMS, *From time series to linear systems. Part 1. Finite dimensional linear time invariant systems*, Automatica, 22 (1986), pp. 561-580.

ENERGY DECAY PROBLEMS IN THE DESIGN OF A POINT STABILIZER FOR COUPLED STRING VIBRATING SYSTEMS*

KANG-SHENG LIU†

Abstract. This paper solves completely the following problem posed by Chen, Coleman, and West [SIAM J. Appl. Math., 47, pp. 751-780]: When can we achieve the desirable uniform exponential decay property of the vibration energy for two coupled vibrating strings with a stabilizer or damper installed at the coupling point? An approach using abstract semigroups is shown. The paper proves the following: For the cases when the stabilizers are "symmetrically placed," the system's energy does not decay to zero with respect to time for some initial states. For all the other cases, if the proportion of the wave speeds on the two vibrating strings c_2/c_1 is a rational number, applying the formula of the growth order of a C_0 -semigroup on a Hilbert space, derived by Huang [2], proves that the semigroup satisfies the Spectrum-Determined Growth Assumption, so the uniform exponential decay property depends completely on the representative form of the ratio c_2/c_1 . Finally, the paper proves that if c_2/c_1 is an irrational number, then the energy does not decay uniformly exponentially.

Key words. wave equation, exponential decay, strong decay, C_0 -semigroup, spectrum-determined growth assumption

AMS(MOS) subject classifications. 93D15, 35B37, 35L05, 35P15

1. Introduction. Chen, Coleman, and West [1] studied the following coupled vibrating string system with a damper installed at the coupling point:

$$(1.1) \quad \begin{aligned} m_1 y_{tt} - T_1 y_{xx} &= 0, & x \in (0, 1), & \quad t > 0, \\ m_2 y_{tt} - T_2 y_{xx} &= 0, & x \in (1, 2), & \quad t > 0, \end{aligned}$$

where m_1, T_1, m_2, T_2 are positive constant numbers,

$$(1.2) \quad y(0, t) = 0 \quad \text{or} \quad y_x(0, t) = 0, \quad t > 0,$$

$$(1.3) \quad y(2, t) = 0 \quad \text{or} \quad y_x(2, t) = 0, \quad t > 0,$$

$$(1.4) \quad \begin{aligned} T_1 y_x(1^-, t) &= T_2 y_x(1^+, t) \quad \text{and} \\ y_t(1^-, t) - y_t(1^+, t) &= -K_1 T_1 y_x(1^-, t), \quad K_1 > 0, \end{aligned}$$

or

$$(1.4)' \quad \begin{aligned} y(1^-, t) &= y(1^+, t) \quad \text{and} \\ T_1 y_x(1^-, t) - T_2 y_x(1^+, t) &= -K_2 y_t(1, t), \quad K_2 > 0, \quad t > 0. \end{aligned}$$

Condition (1.4), or (1.4)', is prescribed at the coupling point, where a damper stabilizer is installed. In [1], the authors have also illustrated possible mechanical designs for those dampers satisfying the condition (1.4) or (1.4)'.

The system (1.1)-(1.4) or (1.4)' yields a total of six different combinations:

- (Case I) (1.1) + (1.2)₁ + (1.3)₁ + (1.4),
- (Case II) (1.1) + (1.2)₁ + (1.3)₂ + (1.4),
- (Case III) (1.1) + (1.2)₂ + (1.3)₂ + (1.4),
- (Case IV) (1.1) + (1.2)₁ + (1.3)₁ + (1.4)',
- (Case V) (1.1) + (1.2)₁ + (1.3)₂ + (1.4)',
- (Case VI) (1.1) + (1.2)₂ + (1.3)₂ + (1.4)'.

* Received by the editors June 3, 1987; accepted for publication (in revised form) November 30, 1987.

† Department of Basic Sciences, Southwestern Petroleum Institute, Nanchong, Sichuan, People's Republic of China.

The system's energy at time t is

$$(1.5) \quad E(t) = \int_0^1 \frac{1}{2} (m_1 y_t^2 + T_1 y_x^2) dx + \int_1^2 \frac{1}{2} (m_2 y_t^2 + T_2 y_x^2) dx.$$

We say that the energy $E(t)$ satisfies the uniform exponential decay property if there exist positive constant numbers μ and M such that

$$(1.6) \quad E(t) \leq M e^{-\mu t} E(0) \quad \text{for all } t > 0.$$

The energy $E(t)$ is said to decay strongly if

$$(1.7) \quad E(t) \rightarrow 0 \quad \text{as } t \rightarrow +\infty \quad \text{for all initial states } E(0) < +\infty.$$

In this paper we consider the following problem posed by Chen, Coleman, and West [1]: When can we achieve the desirable uniform exponential decay property (1.6)? Can we formulate some general sufficient conditions involving m_1 , T_1 , m_2 , T_2 , and K_1 , K_2 only?

The answers to the above depend on the wave speeds

$$c_1 \equiv \sqrt{T_1/m_1}, \quad c_2 \equiv \sqrt{T_2/m_2}.$$

Under the assumption $c_1 = c_2$ Chen, Coleman, and West [1] used the method of characteristics to prove that (Case II) and (Case V) have the property (1.6) and that for any of the other four cases the energy $E(t)$ does not decay strongly. They also discussed the case when c_2/c_1 is a rational number. However, for all cases except Case (I, $c_1/c_2 = 2$), no clear answers to the above problem were given in [1]. In the present paper we study this problem for general wavespeeds c_1 and c_2 , and solve it completely.

2. C_0 -semigroups for an equivalent hyperbolic system. For $x \in (0, 1)$ we set

$$\begin{aligned} w_1(x, t) &= \frac{1}{2}\sqrt{m_1}[-c_1 y_x(x, t) + y_t(x, t)], \\ w_2(x, t) &= \frac{1}{2}\sqrt{m_2}[-c_2 y_x(2-x, t) + y_t(2-x, t)], \\ w_3(x, t) &= \frac{1}{2}\sqrt{m_1}[c_1 y_x(x, t) + y_t(x, t)], \\ w_4(x, t) &= \frac{1}{2}\sqrt{m_2}[c_2 y_x(2-x, t) + y_t(2-x, t)], \\ w(x, t) &= (w_1(x, t), w_2(x, t), w_3(x, t), w_4(x, t))'. \end{aligned}$$

Then (1.1) becomes

$$(2.1) \quad \begin{aligned} \frac{\partial}{\partial t} w(x, t) &= \text{diag}(-c_1, -c_2, c_1, c_2) \frac{\partial}{\partial x} w(x, t) \\ &\equiv A \frac{\partial}{\partial x} w(x, t), \quad 0 < x < 1, \quad t > 0, \end{aligned}$$

an equivalent hyperbolic system. In (2.1), A represents the matrix with diagonal entries as indicated. After straightforward calculations, we get from (1.2)–(1.4)' the following boundary conditions:

$$(2.2) \quad \begin{aligned} B_0^{(j)} w(0, t) &= 0, \quad j = 1, 2, 3, \\ B_0^{(1)} &= \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} D \quad \text{for } (1.2)_1 + (1.3)_1, \\ B_0^{(2)} &= \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} D \quad \text{for } (1.2)_1 + (1.3)_2, \\ B_0^{(3)} &= \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} D \quad \text{for } (1.2)_2 + (1.3)_2, \end{aligned}$$

$$\begin{aligned}
 (2.3) \quad & B_1^{(j)} w(1, t) = 0, \quad j = 1, 2, \\
 & B_1^{(1)} = \begin{pmatrix} 1 - K_1 a_1 & -1 & 1 + K_1 a_1 & -1 \\ -a_1 & -a_2 & a_1 & a_2 \end{pmatrix} D \quad \text{for (1.4),} \\
 & B_1^{(2)} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ a_1 - K_2 & a_2 & -a_1 - K_2 & -a_2 \end{pmatrix} D \quad \text{for (1.4)',}
 \end{aligned}$$

where

$$a_1 = \frac{T_1}{c_1}, \quad a_2 = \frac{T_2}{c_2}, \quad D = \text{diag}(m_1^{-1/2}, m_2^{-1/2}, m_1^{-1/2}, m_2^{-1/2}).$$

From (1.5) we have

$$(2.4) \quad E(t) = \sum_{j=1}^4 \int_0^1 |w_j(x, t)|^2 dx = \|w(\cdot, t)\|_{L^2_{\mathcal{H}}([0,1]; \mathbb{C}^4)}^2.$$

Obviously, the underlying Hilbert space is $\mathcal{H} = L^2([0, 1]; \mathbb{C}^4)$. We define in \mathcal{H}

$$(2.5) \quad \mathcal{A} \equiv A \frac{d}{dx}, \quad \mathcal{D}(\mathcal{A}) = \{w \in H^1([0, 1]; \mathbb{C}^4) : B_0 w(0) = 0, \quad B_1 w(1) = 0\},$$

where “ B_0, B_1 ” is any of the combinations from (2.2) and (2.3), and the space $H^1([0, 1]; \mathbb{C}^4)$ is the Sobolev space of order 1. Thus we have the following theorem.

THEOREM 2.1. \mathcal{A} is dissipative and densely defined in \mathcal{H} .

Proof. Obviously $\mathcal{D}(\mathcal{A})$ is dense in \mathcal{H} . For $w \in \mathcal{D}(\mathcal{A})$ we have

$$\begin{aligned}
 \text{Re}(\mathcal{A}w, w)_{\mathcal{H}} &= \text{Re} \int_0^1 \left(A \frac{d}{dx} w, w \right)_{\mathbb{C}^4} dx \\
 &= \frac{1}{2} \int_0^1 \frac{d}{dx} (\overline{Aw'} \cdot w) dx = \frac{1}{2} (\overline{Aw(1)'} \cdot w(1) - \overline{Aw(0)'} \cdot w(0)).
 \end{aligned}$$

By (2.2) we can get $\overline{Aw(0)'} \cdot w(0) = 0$. From (2.3) we have

$$\begin{aligned}
 \overline{Aw(1)'} \cdot w(1) &= \text{Re} \begin{cases} [m_1^{-1/2}(w_1 + w_3) - m_2^{-1/2}(w_2 + w_4)] \sqrt{T_1}(\bar{w}_3 - \bar{w}_1)(1) & (\text{for } B_1^{(1)}) \\ [\sqrt{T_1}(w_3 - w_1) + \sqrt{T_2}(w_4 - w_2)] m_1^{-1/2}(\bar{w}_1 + \bar{w}_3)(1) & (\text{for } B_1^{(2)}) \end{cases} \\
 &= - \begin{cases} K_1 T_1 |w_1(1) - w_3(1)|^2 \\ K_2 m_1^{-1} |w_1(1) + w_3(1)|^2 \end{cases} \leq 0.
 \end{aligned}$$

THEOREM 2.2. Let $\sigma(\mathcal{A})$ ($\sigma_p(\mathcal{A})$) denote the (point) spectrum set of \mathcal{A} , $\rho(\mathcal{A})$ denote the resolvent set of \mathcal{A} . Then we have the following:

(a) $\sigma(\mathcal{A}) = \sigma_p(\mathcal{A}) = \{\lambda \in \mathbb{C} : \Delta(\lambda) \equiv \det D_\lambda = 0\}$, where

$$D_\lambda = \begin{pmatrix} B_0 \\ B_1 e^{\lambda A^{-1}} \end{pmatrix}.$$

(b) For $\lambda \in \rho(\mathcal{A})$, $f \in \mathcal{H}$ we have

$$\begin{aligned}
 (2.6) \quad & R(\lambda; \mathcal{A})f \equiv (\lambda - \mathcal{A})^{-1}f \\
 &= e^{\lambda x A^{-1}} \left[D_\lambda^{-1} \begin{pmatrix} 0 \\ B_1 \end{pmatrix} \int_0^1 e^{\lambda(1-\tau)A^{-1}} A^{-1} f(\tau) d\tau \right. \\
 &\quad \left. - \int_0^x e^{-\lambda \tau A^{-1}} A^{-1} f(\tau) d\tau \right].
 \end{aligned}$$

(c) $R(\lambda; A)$ is compact.

Proof. We solve the resolvent equation

$$(2.7) \quad (\lambda - \mathcal{A})w = f, \quad \lambda \in \mathbb{C}, \quad f \in \mathcal{H} \text{ is arbitrary, } w \in \mathcal{D}(\mathcal{A}).$$

In fact, (2.7)₁ yields componentwise

$$w(x) = e^{\lambda x A^{-1}}(w(0) - \int_0^x e^{-\lambda \tau A^{-1}} A^{-1} f(\tau) d\tau).$$

When we observe that

$$B_0 w(0) = 0 \quad \text{and} \quad B_1 w(1) = B_1 e^{\lambda A^{-1}} \left(w(0) - \int_0^1 e^{-\lambda \tau A^{-1}} A^{-1} f(\tau) d\tau \right) = 0,$$

that is,

$$D_\lambda w(0) = \begin{pmatrix} 0 \\ B_1 \end{pmatrix} \int_0^1 e^{\lambda(1-\tau)A^{-1}} A^{-1} f(\tau) d\tau,$$

(a) and (b) hold. From the familiar compact imbedding theorem, the compactness of $R(\lambda; \mathcal{A})$ follows from the fact that there exists a constant $C_\lambda > 0$ such that

$$\|R(\lambda; \mathcal{A})f\|_{H^1[(0, 1); \mathbb{C}^4]} \leq C_\lambda \|f\|_{\mathcal{H}}.$$

Combining Theorem 2.1 and Theorem 2.2, we have the following theorem.

THEOREM 2.3. \mathcal{A} generates a C_0 -semigroup of contractions $S(t)$ on \mathcal{H} (see [3, Lumer-Philips Theorem]). Moreover, for arbitrary initial condition $w(\cdot, 0) = f_0 \in \mathcal{H}$, the hyperbolic system (2.1) + (2.2) + (2.3) has a unique mild solution $w(\cdot, t) = S(t)f_0$ in \mathcal{H} . If $f_0 \in \mathcal{D}(\mathcal{A})$, $w(\cdot, t) = S(t)f_0$ is the strong solution of this system.

For the future, we calculate, respectively, the following:

(2.8)

$$\begin{aligned} \text{(Case I)} \quad m_1 m_2 \Delta_1(\lambda) &= m_1 m_2 \det \begin{pmatrix} B_0^{(1)} \\ B_1^{(1)} e^{\lambda A^{-1}} \end{pmatrix} \\ &= a_1(z_1 + z_1^{-1})(z_2 - z_2^{-1}) \\ &\quad + a_2(z_2 + z_2^{-1})[z_1 - z_1^{-1} + K_1 a_1(z_1 + z_1^{-1})] \\ &\equiv g_1(z_1, z_2); \\ \text{(Case II)} \quad m_1 m_2 \Delta_2(\lambda) &= m_1 m_2 \det \begin{pmatrix} B_0^{(2)} \\ B_1^{(1)} e^{\lambda A^{-1}} \end{pmatrix} \\ &= a_1(z_1 + z_1^{-1})(z_2 + z_2^{-1}) \\ &\quad + a_2(z_2 - z_2^{-1})[z_1 - z_1^{-1} + K_1 a_1(z_1 + z_1^{-1})] \\ &\equiv g_2(z_1, z_2); \\ \text{(Case III)} \quad m_1 m_2 \Delta_3(\lambda) &= m_1 m_2 \det \begin{pmatrix} B_0^{(3)} \\ B_1^{(1)} e^{\lambda A^{-1}} \end{pmatrix} \\ &= a_1(z_1 - z_1^{-1})(z_2 + z_2^{-1}) \\ &\quad + a_2(z_2 - z_2^{-1})[z_1 + z_1^{-1} + K_1 a_1(z_1 - z_1^{-1})] \\ &\equiv g_3(z_1, z_2); \\ \text{(Case IV)} \quad m_1 m_2 \Delta_4(\lambda) &= m_1 m_2 \det \begin{pmatrix} B_0^{(1)} \\ B_1^{(2)} e^{\lambda A^{-1}} \end{pmatrix} \\ &= (-z_2 + z_2^{-1})[a_1(z_1 + z_1^{-1}) + K_2(z_1 - z_1^{-1})] \\ &\quad - a_2(z_1 - z_1^{-1})(z_2 + z_2^{-1}) \\ &\equiv g_4(z_1, z_2); \end{aligned}$$

$$\begin{aligned}
(\text{Case V}) \quad m_1 m_2 \Delta_5(\lambda) &= m_1 m_2 \det \begin{pmatrix} B_0^{(2)} \\ B_1^{(2)} e^{\lambda A^{-1}} \end{pmatrix} \\
&= (-z_2 - z_2^{-1})[a_1(z_1 + z_1^{-1}) + K_2(z_1 - z_1^{-1})] \\
&\quad - a_2(z_1 - z_1^{-1})(z_2 - z_2^{-1}) \\
&\equiv g_5(z_1, z_2); \\
(\text{Case VI}) \quad m_1 m_2 \Delta_6(\lambda) &= m_1 m_2 \det \begin{pmatrix} B_0^{(3)} \\ B_1^{(2)} e^{\lambda A^{-1}} \end{pmatrix} \\
&= (-z_2 - z_2^{-1})[a_1(z_1 - z_1^{-1}) + K_2(z_1 + z_1^{-1})] \\
&\quad - a_2(z_1 + z_1^{-1})(z_2 - z_2^{-1}) \\
&\equiv g_6(z_1, z_2),
\end{aligned}$$

where $z_1 = e^{\lambda/c_1}$, $z_2 = e^{\lambda/c_2}$.

3. The spectrum-determined growth assumption. Let L generate a C_0 -semigroup e^{tL} on a Hilbert space. Then two related important data are

$$\sigma_0(L) = \sup \{ \operatorname{Re} \lambda : \lambda \in \sigma(L) \},$$

$$\omega_0(L) = \inf \{ \omega : \|e^{tL}\| \leq M e^{\omega t} \text{ for some } M \text{ and all } t > 0 \}.$$

Generally, we have that $\sigma_0(L) \leq \omega_0(L)$ and that e^{tL} is exponentially stable if and only if $\omega_0(L) < 0$ (cf. [4]). We say that the semigroup e^{tL} satisfies the Spectrum-Determined Growth Assumption if $\omega_0(L) = \sigma_0(L)$ (cf. [9]). Huang [2] shows in a counterexample that the identity $\omega_0(L) = \sigma_0(L)$ is false even if L has a compact resolvent. In the very same paper [2] Huang gives an extremely useful formula for $\omega_0(L)$:

$$(3.1) \quad \omega_0(L) = \inf \{ \delta : \delta > \sigma_0(L) \text{ and } \sup_{\operatorname{Re} \lambda \geq \delta} \|R(\lambda; L)\| < +\infty \}.$$

This formula was first used to prove that an open question posed by Pritchard and Zabczyk [4] has a completely affirmative answer (cf. [5]). The formula has also been applied to the study of exponential stability for several linear systems in Hilbert spaces (cf. [6], [7]).

Now we show our result in the present section.

THEOREM 3.1. *Let c_2/c_1 be a rational number. Then the semigroup $S(t)$ satisfies the Spectrum-Determined Growth Assumption.*

To prove Theorem 3.1 we need the following lemma.

LEMMA 3.1. *Let $c_2/c_1 = p/q$, where p, q are positive integers. Let $\varphi : \mathbb{R} \rightarrow \mathbb{C}^2$, $\varphi(\omega) = (e^{i\omega/c_1}, e^{i\omega/c_2})$. Then $\varphi(\mathbb{R})$ is a compact set in \mathbb{C}^2 .*

Proof. For $r \in \mathbb{R}$, we use $I(r)$ to denote the integer satisfying $0 \leq r - I(r) < 1$. Then we have

$$\begin{aligned}
e^{i\omega/c_1} &= \exp \left(ip \frac{\omega}{pc_1} \right) = \exp \left(ip \left[\frac{\omega}{pc_1} - 2\pi I \left(\frac{\omega}{2\pi pc_1} \right) \right] \right), \\
e^{i\omega/c_2} &= \exp \left(iq \frac{\omega}{pc_1} \right) = \exp \left(iq \left[\frac{\omega}{pc_1} - 2\pi I \left(\frac{\omega}{2\pi pc_1} \right) \right] \right).
\end{aligned}$$

Since $0 \leq \omega/pc_1 - 2\pi I(\omega/2\pi pc_1) < 2\pi$, we have that

$$(e^{i\omega_n/c_1}, e^{i\omega_n/c_2}) \xrightarrow{n} (y_1, y_2) \in \mathbb{C}^2$$

implies

$$(e^{i\omega_{n_k}/c_1}, e^{i\omega_{n_k}/c_2}) \xrightarrow{k} (e^{ipt_0}, e^{iqt_0}) = \varphi(pt_0c_1) = (y_1, y_2)$$

for some $\{n_k\}$ and t_0 satisfying $\omega_{n_k}/pc_1 - 2\pi I(\omega_{n_k}/2\pi pc_1) \xrightarrow{k} t_0$. This means $\varphi(\mathbb{R})$ is closed in \mathbb{C}^2 . Therefore $\varphi(\mathbb{R})$ is compact in \mathbb{C}^2 because $\varphi(\mathbb{R}) \subset S^1 \times S^1$, where $S^1 = \{\lambda \in \mathbb{C}: |\lambda| = 1\}$.

Proof of Theorem 3.1. From (3.1) and Theorem 2.3 we only need to prove that

$$(3.2) \quad \sup_{\delta \leq \operatorname{Re} \lambda \leq 1} \|R(\lambda; \mathcal{A})\| < +\infty \quad \text{for } \delta \in (\sigma_0(\mathcal{A}), 1).$$

In fact, there exists a constant number C_δ such that

$$(3.3) \quad \begin{aligned} \|e^{\lambda x A^{-1}}\|_{\mathcal{L}(\mathbb{C}^4)} &\leq C_\delta, \left\| \begin{pmatrix} 0 \\ B_1 \end{pmatrix} \right\|_{\mathcal{L}(\mathbb{C}^4)} \leq C_\delta, \\ \left\| \int_0^1 e^{\lambda(1-\tau)A^{-1}} A^{-1} f(\tau) d\tau \right\|_{\mathbb{C}^4} &\leq C_\delta \|f\|_{\mathcal{H}}, \\ \left\| \int_0^x e^{-\lambda \tau A^{-1}} A^{-1} f(\tau) d\tau \right\|_{\mathbb{C}^4} &\leq C_\delta \|f\|_{\mathcal{H}}, \end{aligned}$$

for $f \in \mathcal{H}$, $x \in [0, 1]$ and $\operatorname{Re} \lambda \in [\delta, 1]$. Furthermore, from (2.8) we have that for $\lambda = \tau + i\omega$, $\tau \in [\delta, 1]$, $\omega \in \mathbb{R}$,

$$|\Delta(\lambda)| = \left| \det \begin{pmatrix} B_0 \\ B_1 e^{\lambda A^{-1}} \end{pmatrix} \right| = G(\tau, \varphi(\omega)),$$

where $\varphi(\cdot)$ is the same as in Lemma 3.1 and $G(\cdot, \cdot)$ is some continuous function defined on $[\delta, 1] \times \varphi(\mathbb{R})$. Therefore, by Lemma 3.1 we obtain that there exists $(\tau_1, \omega_1) \in [\delta, 1] \times \mathbb{R}$ such that

$$\inf_{\delta \leq \operatorname{Re} \lambda \leq 1} |\Delta(\lambda)| = G(\tau_1, \varphi(\omega_1)) = |\Delta(\tau_1 + i\omega_1)|.$$

Using Theorem 2.2(a) we have $\Delta(\tau_1 + i\omega_1) \neq 0$. It follows that there exists a constant C'_δ such that

$$(3.4) \quad \|D_\lambda^{-1}\|_{\mathcal{L}(\mathbb{C}^4)} = |\Delta(\lambda)|^{-1} \|D_\lambda^*\|_{\mathcal{L}(\mathbb{C}^4)} \leq C'_\delta \quad \text{for } \operatorname{Re} \lambda \in [\delta, 1],$$

where D_λ^* is the adjoint matrix of D_λ .

Combining (2.6), (3.3), and (3.4), we see that (3.2) holds and the proof is complete.

4. Stability and instability results. A C_0 -semigroup e^{tL} on Banach space X is strongly asymptotically stable if and only if $\|e^{tL}f\| \rightarrow 0$ as $t \rightarrow +\infty$ for any f in X . If $\|e^{tL}\| \leq M$ for some M and all $t \geq 0$, the condition $\|e^{tL}f\| \rightarrow 0$ as $t \rightarrow +\infty$ for any f in $\mathcal{D}(L)$ implies the above. Concerning strong asymptotic stability, we present the following theorem of Huang [8].

THEOREM 4.1. *Let e^{tL} be a uniformly bounded C_0 -semigroup on a Banach space, and let $\operatorname{Re} \lambda < 0$ for all $\lambda \in \sigma(L)$. Then e^{tL} is strongly asymptotically stable. Conversely, let e^{tL} be strongly asymptotically stable. Then e^{tL} is uniformly bounded, $\operatorname{Re} \lambda \leq 0$ for all $\lambda \in \sigma(L)$, and the imaginary axis contains neither point nor residual spectrum of L .*

Since $\Delta_3(0) = \Delta_4(0) = 0$, from Theorem 2.2(a) and the above results we obtain the following theorem.

THEOREM 4.2. For (Case III) and (Case IV), the strong decay property (1.7) does not hold for any positive numbers m_1, m_2, T_1, T_2 , and K_1, K_2 .

It is well known that for (Case III) and (Case IV) the stabilizers are “symmetrically placed”— $T_1 y_x(0, t) = 0 = T_2 y_x(2, t)$, $T_1 y_x(1^-, t) = T_2 y_x(1^+, t)$; $y(0, t) = 0 = y(2, t)$, $y(1^-, t) = y(1^+, t)$.” So, symmetry should be avoided in the design of a point stabilizer (cf. also [1]).

THEOREM 4.3. Let $c_2/c_1 = p/q$ be a rational number, where p and q are relatively prime positive integers. Then the system's energy $E(t)$ does not decay strongly if the following condition holds, respectively:

$$\text{(Case I)} \quad p \equiv 1 \pmod{2} \quad \text{and} \quad q \equiv 1 \pmod{2},$$

$$\text{(Case II)} \quad p \equiv 1 \pmod{2} \quad \text{and} \quad q \equiv 0 \pmod{2},$$

$$\text{(Case V)} \quad p \equiv 0 \pmod{2} \quad \text{and} \quad q \equiv 1 \pmod{2},$$

$$\text{(Case VI)} \quad p \equiv 1 \pmod{2} \quad \text{and} \quad q \equiv 1 \pmod{2}.$$

Conversely, if the pair “ p, q ” is not of the form shown above for each respective case, then the energy $E(t)$ satisfies the uniform exponential decay property (1.6).

Proof. From (2.8) we have

$$(4.1) \quad \Delta(\lambda) = e^{-\lambda(1/c_1 + 1/c_2)} \cdot P(e^{\lambda/p c_1}),$$

where $P(\cdot)$ is some polynomial. Therefore, from Theorem 2.2(a) we get the distribution of the spectrum of \mathcal{A}

$$(4.2) \quad \sigma(\mathcal{A}) = \{pc_1[\ln |u| + i(2n\pi + \arg u)]: u \neq 0, P(u) = 0, n = 0, \pm 1, \pm 2, \dots\}.$$

Since the contraction semigroup $S(t)$ satisfies the Spectrum-Determined Growth Assumption, from (4.2) and (4.1) we have that $S(t)$ is exponentially stable if and only if

$$(4.3) \quad \Delta(i\omega) \neq 0 \quad \text{for all } \omega \in \mathbb{R}.$$

Solving the simple trigonometric equations from (2.8) we obtain

$$(4.4) \quad \begin{aligned} \Delta_1(i\omega) = 0 & \quad \text{has solutions iff} \quad \frac{c_2}{c_1} = \frac{4m \pm 1}{4n + 1}, \\ \Delta_2(i\omega) = 0 & \quad \text{has solutions iff} \quad \frac{c_1}{c_1} = \frac{4m \pm 1}{4n + 2}, \\ \Delta_5(i\omega) = 0 & \quad \text{has solutions iff} \quad \frac{c_2}{c_1} = \frac{4m \pm 2}{4n + 1}, \\ \Delta_6(i\omega) = 0 & \quad \text{has solutions iff} \quad \frac{c_2}{c_1} = \frac{4m \pm 1}{4n + 1}, \end{aligned}$$

where m, n are some integers. The proof is complete.

THEOREM 4.4. Let c_2/c_1 be an irrational number. Then for Cases I, II, V, and VI the strong decay property (1.7) holds, but the uniform exponential decay property (1.6) does not hold.

To prove Theorem 4.4 we first prove the following lemma.

LEMMA 4.1. Let c_2/c_1 be an irrational number. Then we have

$$(4.5) \quad \inf \{|\Delta(i\omega)|: \omega \in \mathbb{R}\} = 0,$$

where $\Delta(\cdot)$ is the same as in Theorem 2.2(a).

Proof. We let $\varphi(\cdot)$, $I(\cdot)$, S^1 be the same as in Lemma 3.1 and its proof.

Observing $g_1(i, i) = g_2(i, 1) = g_3(1, 1) = g_4(1, 1) = g_5(1, i) = g_6(i, i) = 0$ from (2.8), it is sufficient to prove that $\varphi(\mathbb{R})$ is dense in $S^1 \times S^1$. For every $t_2 \in [0, 2\pi]$ and $n \in \mathbb{Z}$, the set of all integers, we have

$$\varphi(c_2(t_2 + 2n\pi)) = (e^{idt_2} e^{i2\pi[nd - I(nd)]}, e^{it_2}),$$

where $d = c_2/c_1$ is an irrational number. So it is also sufficient to prove that $\{x(n) \equiv nd - I(nd) : n \in \mathbb{Z}\}$ is dense in $[0, 1]$. In fact, the map $x(\cdot)$ is one to one from \mathbb{Z} into $[0, 1]$. For $(a, b) \subset [0, 1]$, we set $X = \{x(1), x(2), \dots, x(N), x(N+1)\}$, where N is some positive integer satisfying $1/N < b - a$. Then there exist $x(n_1), x(n_2) \in X$ such that $0 < x(n_2) - x(n_1) < 1/N$, and so there exists $0 < K \in \mathbb{Z}$ such that $K(x(n_2) - x(n_1)) \in (a, b)$, that is, $K(n_2 - n_1)d - K[I(n_2d) - I(n_1d)] = x(K(n_2 - n_1)) \in (a, b) \subset (0, 1)$. The proof is complete.

Proof of Theorem 4.4. The strong decay property (1.7) follows from (4.4) and Theorem 4.1. Now we prove that the semigroup $S(t)$ is not exponentially stable for Cases I, II, V, and VI. In fact, we need to prove only

$$(4.6) \quad \sup \{\|R(i\omega; \mathcal{A})f\|_{\mathcal{H}} : \omega \in \mathbb{R}, f \in \mathcal{H}, \|f\|_{\mathcal{H}} = 1\} = +\infty,$$

for Cases I, II, V, and VI.

For Case I we set

$$D_\lambda^{-1} = \begin{pmatrix} B_0^{(1)} \\ B_1^{(1)} e^{\lambda A^{-1}} \end{pmatrix}^{-1} = \Delta_1(\lambda)^{-1} D_\lambda^* = \Delta_1(\lambda)^{-1} (d_{kl}^{(1)}(\lambda)), \quad k, l = 1, 2, 3, 4,$$

$$f_\lambda(\tau) = (0, 0, 0, e^{-\lambda(1-\tau)/c_2})',$$

$$v(x) = e^{\lambda x A^{-1}} D_\lambda^{-1} \begin{pmatrix} 0 \\ B_1^{(1)} \end{pmatrix} \int_0^1 e^{\lambda(1-\tau)A^{-1}} A^{-1} f_\lambda(\tau) d\tau = (v_1(x), \cdot, \cdot, \cdot)',$$

where $D_\lambda^* = (d_{kl}^{(1)}(\lambda))$ is the adjoint matrix of D_λ . Calculating, we have

$$\begin{pmatrix} 0 \\ B_1^{(1)} \end{pmatrix} \int_0^1 e^{\lambda(1-\tau)A^{-1}} A^{-1} f_\lambda(\tau) d\tau = (c_2 \sqrt{m_2})^{-1} (0, 0, -1, a_2)',$$

$$d_{13}^{(1)}(\lambda) = -\frac{a_2}{\sqrt{m_1 m_2}} (e^{\lambda/c_2} + e^{-\lambda/c_2}),$$

$$d_{14}^{(1)}(\lambda) = -\frac{1}{\sqrt{m_1 m_2}} (e^{\lambda/c_2} - e^{-\lambda/c_2}),$$

and so we have

$$v_1(x) = \frac{2a_2}{c_2 \sqrt{m_1 m_2 m_2}} \Delta_1(\lambda)^{-1} e^{-\lambda/c_2} e^{-\lambda x/c_1}.$$

Therefore, for $\lambda = i\omega$, $\omega \in \mathbb{R}$ we can obtain

$$(4.7) \quad \begin{aligned} & \|f_{i\omega}(\tau)\|_{\mathcal{H}} = 1, \\ & \|v(x)\|_{\mathcal{H}} \geq \left(\int_0^1 |v_1(x)|^2 dx \right)^{1/2} = \frac{2a_2}{c_2 \sqrt{m_1 m_2 m_2}} |\Delta_1(i\omega)|^{-1}, \\ & \|e^{i\omega x A^{-1}} \int_0^x e^{-i\omega \tau A^{-1}} A^{-1} f_{i\omega}(\tau) d\tau\|_{\mathcal{H}} \leq \max \left(\frac{1}{c_1}, \frac{1}{c_2} \right). \end{aligned}$$

Combining (2.6), (4.7), and (4.5), we have (4.6).

For Cases II, V, and VI the proof is similar. Here we give only the results of our calculations:

$$\begin{pmatrix} 0 \\ B_1^{(2)} \end{pmatrix} \int_0^1 e^{\lambda(1-\tau)A^{-1}} A^{-1} f_\lambda(\tau) d\tau = (c_2 \sqrt{m_2})^{-1} (0, 0, -1, -a_2)',$$

$$d_{13}^{(2)}(\lambda) = -d_{13}^{(5)}(\lambda) = d_{13}^{(6)}(\lambda) = -\frac{a_2}{\sqrt{m_1 m_2}} (e^{\lambda/c_2} - e^{-\lambda/c_2}),$$

$$d_{14}^{(2)}(\lambda) = -d_{14}^{(5)}(\lambda) = d_{14}^{(6)}(\lambda) = -\frac{1}{\sqrt{m_1 m_2}} (e^{\lambda/c_2} + e^{-\lambda/c_2}).$$

Remark 1. For Case $(c_1 = c_2)$ and Case I $(c_1/c_2 = 2)$, the above results are consistent with those in [1].

Remark 2. The original idea is from the answer to Question 190 in [11], according to which we prove that $x(n)$ is dense in $[0, 1]$ in Lemma 4.1. Then, it is known that it is also a special case of Example 2.1 in [10, p. 8].

Acknowledgments. The author thanks his graduate adviser Professor Fa-Lun Huang for his genial teachings and invaluable discussions. The author also thanks Professor Goong Chen for the interesting lectures given during his visit at Sichuan University in March 1986.

REFERENCES

- [1] G. CHEN, M. COLEMAN AND H. H. WEST, *Pointwise stabilization in the middle of the span for second order systems, nonuniform and uniform exponential decay of solutions*, SIAM J. Appl. Math., 47 (1987), pp. 751–780.
- [2] F. L. HUANG, *Characteristic conditions for exponential stability of linear dynamical systems in Hilbert spaces*, Ann. Differential Equations, 1 (1985), pp. 43–56.
- [3] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, Berlin, 1983, p.14.
- [4] A. J. PRITCHARD AND J. ZABCZYK, *Stability and stabilizability of infinite dimensional systems*, SIAM Rev., 23 (1981), pp. 25–52.
- [5] F. L. HUANG AND K. S. LIU, *A problem of exponential stability for linear dynamical systems in Hilbert spaces*, Kexue Tongbao English edition, 33 (1988), pp. 460–462; Chinese edition, 32 (1987), pp. 161–163.
- [6] G. CHEN, S. G. KRANTZ, D. W. MA, C. E. WAYNE, AND H. H. WEST, *The Euler–Bernoulli beam equation with boundary energy dissipation*, in Operator Methods for Optimal Control Problems, S. J. Lee, ed., Marcel Dekker, New York.
- [7] G. CHEN AND H. K. WANG, *Pointwise stabilization for coupled quasilinear and linear wave equations*, in Proc. Third International Conference on Control and Identification of Distributed Systems, Lecture Notes in Control and Inform. Sci. 75, Springer-Verlag, Berlin, New York, 1987, pp. 40–63.
- [8] F. L. HUANG, *Asymptotic stability theory for linear dynamical systems in Banach spaces*, Kexue Tongbao, English edition, Chinese, 10 (1983), pp. 584–586.
- [9] R. TRIGGIANI, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383–403.
- [10] L. KUIPERS AND H. NIEDERREITER, *Uniform Distribution of Sequences*, John Wiley, New York, 1974.
- [11] H. C. AUCH, *Theorems and Problems in Real Analysis*, USSR Education Press, 1965, Chinese Translation by Y. D. LI, Y. C. ZHANG, AND H. A. ZHANG, People Education Press, 1981.

A MIXED-STRATEGY MINIMAX THEOREM WITHOUT COMPACTNESS*

STEVE ALPERN† AND SHMUEL GAL‡

Abstract. Minimax theorems for infinite games generally require that both players choose their pure strategies from compact sets and have semicontinuity requirements in both variables. This paper proves the following theorem. Let X be a compact Hausdorff space and let (Y, A) be a measurable space. Let $f: X \times Y \rightarrow \mathbb{R}$ be a measurable function which is bounded below and lower semicontinuous on X for all fixed y in Y . Let M be any convex set of probability measures (mixed strategies) on (Y, A) . Then

$$\min_{\mu \in B(X)} \sup_{\rho \in M} \int \int f(x, y) d\mu d\rho = \sup_{\rho \in M} \min_{\mu \in B(X)} \int \int f(x, y) d\mu d\rho,$$

where $B(X)$ denotes the regular Borel probability measures on X .

Key words. minimax theorem, zero-sum two-person game, mixed strategy

AMS(MOS) subject classification. 90D05

1. Introduction. In this paper we establish a new mixed-strategy minimax theorem for a two-person zero-sum game given in the normal form $f: X \times Y \rightarrow \mathbb{R}$. This is interpreted in the usual way, so that if the minimizer picks a pure strategy x in X and the maximizer picks a pure strategy y in Y then the payoff (to the maximizer) is $f(x, y)$. We will assume that the pure strategy space X is a compact Hausdorff space, and that the mixed strategies available to the minimizer are the regular Borel probability measures on X , collectively denoted by $B(X)$. Our approach is asymmetric in that we do not assume that the maximizer's pure strategies are necessarily topologized. In this case it is well known that there is a certain arbitrariness about the appropriate definition of the maximizer's mixed strategies. We will initially take a general approach and merely assume that the maximizer may choose any mixed strategy in a given set M of probability measures on Y . We will assume that M is a convex set of measures on a common σ -algebra A , and for interpretive reasons we may assume that M contains all point masses. If f is a real (Borel $\times A$) measurable function on $X \times Y$ which is bounded below, we define the usual mixed extension of f , $F: B(X) \times M \rightarrow \mathbb{R}$, by the integral

$$F(\mu, \rho) = \int \int f d(\mu \times \rho) = \int \int f(x, y) d\mu(x) d\rho(y) = \int \int f(x, y) d\rho(y) d\mu(x),$$

where the equivalence of the various forms follows from Fubini's Theorem. Using a pure strategy minimax theorem of Kneser [K], we prove the following mixed-strategy result.

THEOREM 1. *Let X be a compact Hausdorff space and let (Y, A) be a measurable space. Let $f: X \times Y \rightarrow \mathbb{R}$ be a measurable function which is bounded below and lower semicontinuous in x for all fixed y . Let M be any convex set of probability measures on (Y, A) . Then*

$$\min_{\mu \in B(X)} \sup_{\rho \in M} F(\mu, \rho) = \sup_{\rho \in M} \min_{\mu \in B(X)} F(\mu, \rho).$$

By specializing the mixed strategies M available to the maximizer, we obtain corollaries to Theorem 1 which extend known results. For example, if we take Y to

* Received by the editors December 16, 1987; accepted for publication (in revised form) January 19, 1988.

† Department of Mathematics, London School of Economics, University of London, London WC2A 2AE, U.K.

‡ IBM Israel Scientific Center, Technion City, Haifa 32000, Israel.

be a topological space with A its Borel algebra and $M = B(Y)$, we obtain the following generalization of Glicksberg's Theorem [G1] (without compactness of Y or lower semicontinuity in y).

COROLLARY 1. *Let X and Y be topological spaces, with X compact Hausdorff. Let $f: X \times Y \rightarrow \mathbb{R}$ be a measurable function which is bounded below and lower semicontinuous in x for all fixed y . Then*

$$\min_{\mu \in B(X)} \sup_{\rho \in B(Y)} F(\mu, \rho) = \sup_{\rho \in B(Y)} \min_{\mu \in B(X)} F(\mu, \rho).$$

If we take the σ -algebra of Theorem 1 to be the power set of Y and M to be all convex combinations of point masses, then we obtain the following generalization of the theorem of Peck and Dulmage [PD] (with their assumption of continuity reduced to semicontinuity). The lower bound for f (required for Fubini's Theorem) is no longer needed.

COROLLARY 2. *Let X be a compact Hausdorff space and Y be any set. Then for any real function f on $X \times Y$ which is lower semicontinuous on X for every fixed y in Y , we have*

$$\min_{\mu \in B(X)} \sup_{\rho \in M} F(\mu, \rho) = \sup_{\rho \in M} \min_{\mu \in B(X)} F(\mu, \rho),$$

where M is the set of finite mixtures $\rho = \sum_{1 \leq k \leq n} \rho_k y_k$ with $\sum_{1 \leq k \leq n} \rho_k = 1$, $\rho_k \geq 0$ and y_k in Y , $k = 1, \dots, n$, and

$$F(\mu, \rho) = \sum_{1 \leq k \leq n} \rho_k \int_X f(x, y_k) d\mu(x).$$

In [G1, Appendix 1], Theorem 1 was established for a class of payoff functions given by the "time required for capture" for certain search games. The existence of a more general result implicit in [G1, Appendix 1] was observed in [A], where it was shown how this result could be used to establish a conjectured minimax for a search game of Baston and Bostock [BB].

The present paper is organized as follows. In § 2 we give some background material on the weak star (weak*) topology on $B(X)$ and on lower semicontinuous (l.s.c.) functions. These will be combined with Kneser's Theorem in § 3 to give a proof of Theorem 1. In § 4 we will consider some examples.

2. The weak star topology and semicontinuous functions. Let $C(X)$ be the Banach space of continuous real functions on a compact Hausdorff space X , and let $C^*(X)$ denote its conjugate space (see [R]). The weak* topology on $C^*(X)$ is given by the sequential convergence $g_i \rightarrow g$ in $C^*(X)$ if and only if $g_i(h) \rightarrow g(h)$ for all h in $C(X)$. The Riesz Representation Theorem identifies the Borel probability measures $B(X)$ with a closed convex subset of the unit sphere in $C^*(X)$ by the formula $\mu(h) = \int h(x) d\mu(x)$. Since the unit ball of $C^*(X)$ is compact in the weak* topology (Aloaglu Theorem), so is $B(X)$. Recall that a real function ψ on a topological space Z is lower semicontinuous if $z_i \rightarrow z$ implies $\liminf \psi(z_i) \geq \psi(z)$, or equivalently, if $\{z: \psi(z) > r\}$ is open for all real r .

LEMMA 1. *Suppose that for every y in a measure space (Y, A, ρ) the map $\phi_y: X \rightarrow \mathbb{R}$ is lower semicontinuous on the topological space X , and that for all x in X , $\phi_y(x)$ is measurable with respect to (Y, A, ρ) . Then the mean $\phi_\rho: X \rightarrow \mathbb{R}$ defined by $\phi_\rho = \int \phi_y(x) d\rho(y)$ is also lower semicontinuous.*

Proof. Assume that $x_i \rightarrow x$ in X . Then

$$\begin{aligned}\lim \phi_\rho(x_i) &= \lim \int \phi_y(x_i) d\rho(y) \\ &\cong \int \lim \phi_y(x_i) d\rho(y) \quad \text{by Fatou's Lemma} \\ &\cong \int \phi_y(x) d\rho(y) \quad \text{since } \phi_y \text{ is l.s.c.} \\ &= \phi_\rho(x).\end{aligned}$$

LEMMA 2. If ϕ is a lower semicontinuous positive function on a compact Hausdorff space X , then the function ϕ' defined by $\phi'(\mu) = \mu(\phi) = \int \phi(x) d\mu$ is lower semicontinuous on $B(X)$ with respect to the weak* topology.

Proof. Assume that $\mu_i \rightarrow \mu$ in the weak* topology on $B(X)$. Fix any real r and define $V = V_r = \{x: \phi(x) > r\}$. Since ϕ is lower semicontinuous, the set V must be open. Given any $\varepsilon > 0$, the regularity of μ ensures that we can find a closed subset D of V , with $\mu(V - D) < \varepsilon$. By Urysohn's Lemma there exists a function g in $C(X)$ such that g equals 1 on D , zero on the complement of V , and is between zero and 1 on all of X . It follows that $\int g d\mu \geq \mu(V) - \varepsilon$. But the weak* convergence of the μ_i to μ implies that $\mu_i(V) \geq \int g d\mu_i \rightarrow \int g d\mu \geq \mu(V) - \varepsilon$. Since ε was arbitrary, we have $\lim \mu_i(V_r) \geq \mu(V_r)$. Since ϕ is positive we may compute its expected value by the formula $\int_0^\infty \mu(V_r) dr = \int_0^\infty \phi(x) d\mu$ [G1, p. 13]. Combining the above results, we obtain

$$\begin{aligned}\lim \phi'(\mu_i) &= \lim \int_0^\infty \phi(x) d\mu_i \\ &= \lim \int_0^\infty \mu_i(V_r) dr \\ &\geq \int_0^\infty \lim \mu_i(V_r) dr \quad \text{by Fatou's Lemma} \\ &\geq \int_0^\infty \mu(V_r) dr \\ &= \int \phi(x) d\mu \\ &= \phi'(\mu).\end{aligned}$$

3. Proof of Theorem 1. Our mixed-strategy minimax theorem, Theorem 1, relies mainly on the following pure strategy minimax theorem of Kneser [K]. Kneser's Theorem has been generalized by Fan [F], Sion [S], Peck and Dulmage [PD], and Tserkelson [T], but the original result is the simplest and is sufficient for our needs.

KNESER'S THEOREM. *Let K and L be convex subsets of linear spaces and let $F: K \times L \rightarrow R$ be linear. Suppose that K is compact for some topology for which $F(\mu, \rho)$ is lower semicontinuous in μ for all fixed ρ in L . Then*

$$\min_{\mu \in K} \sup_{\rho \in L} F(\mu, \rho) = \sup_{\rho \in L} \min_{\mu \in K} F(\mu, \rho).$$

Proof of Theorem 1. Without loss of generality we may assume that f is positive. Let $F: K \times L \rightarrow R$ be the mixed extension of $f: X \times Y \rightarrow R$, with $K = B(X)$ and $L = M$. Take as the topology on $B(X)$ the weak* topology it inherits as a subset of $C^*(X)$. As noted above, $B(X)$ is compact, so all we need to apply Knese's Theorem is the lower semicontinuity of the functions $\phi'_\rho(\mu) = F(\mu, \rho)$, ρ in M , with respect to the weak* topology. According to the hypotheses of Theorem 1 the positive functions $\phi_y(x)$, y in Y , defined by $\phi_y(x) = f(x, y)$, are lower semicontinuous in x . Hence for any ρ in M , Lemma 1 ensures that their mean ϕ_ρ , defined by $\phi_\rho(x) = \int f(x, y) d\rho(y)$, is lower semicontinuous on the compact Hausdorff space X . It now follows from Lemma 2 that the conjugate map ϕ'_ρ on $B(X)$ is also lower semicontinuous.

4. Examples. Theorem 1 was originally established by Gal to prove that the "Princess and Monster Game" has a value. In that game the Princess and Monster, respectively, pick continuous paths $y(t)$ and $x(t)$ for t in $[0, \infty)$, in a compact path-connected search space S . The Monster has a speed restriction $|x(t_1) - x(t_2)| \leq v|t_1 - t_2|$. The payoff $f(x, y)$ to the maximizing Princess is the first time T for which $|x(t) - y(t)| \leq r$, a given capture radius. The demonstration that the payoff function f and the relevant path spaces X and Y satisfy the hypotheses of Theorem 1 can be found in [G1, Appendix 1]. Theorem 1 can also be used to show that the following high-low search game of Baston and Bostock [BB] has a value: The maximizing Hider picks a point y in the open interval $Y = (0, 1)$. Without knowing y the Searcher makes a sequence of guesses g_1, g_2, \dots . Each guess is made with the knowledge of whether the previous guesses were less than, equal to, or greater than y . The payoff is the sum of the errors $|g_t - y|$, from $t = 1$ to ∞ . The set of pure search strategies X (that is, the rules for picking the guesses g_t) can be identified with the (compact) Hilbert cube and the payoff can be shown to be lower semicontinuous [A].

In this section we give a simpler application of Theorem 1, which is related to the well-known sequential search methods for approximating the maximum of a unimodal function [Ki], [G2], [GM]. In those problems there is an unknown function h , defined on $[0, 1]$, drawn from the set of continuous unimodal functions Y (this means that for some number m , h is strictly increasing on $[0, m]$ and strictly decreasing on $(m, 1]$). The searcher picks successive points x_1, x_2, \dots at which to evaluate the function h , always waiting until the previous point has been evaluated before picking the next one. These points must all be at least a distance e apart, where e is a given "resolution." Depending on the ordering of the evaluations of the previous points, the searcher can infer a new "interval of uncertainty" for the point m where h is maximized. For example, if $x_1 < x_2$, then the uncertainty interval is, respectively, $(x_1, 1]$, $[0, x_2]$, or (x_1, x_2) , depending on whether $h(x_1) < h(x_2)$, $h(x_1) > h(x_2)$, or $h(x_1) = h(x_2)$. Kiefer [Ki] showed that when the payoff is the length of the uncertainty interval after a specified number N of guesses, then a "Fibonacci search" is the minimax pure search strategy.

Consider now the simple case where the searcher is given only $N = 2$ searches (points to evaluate), treated as a game against Nature. That is, the maximizing Nature picks a unimodal function $y = y(t)$ in Y , and the minimizing searcher picks two search points x_1 and x_2 . Since no relevant information about the maximum of y is obtained

from the single evaluation $y(x_1)$, the second guess does not have to be considered as a function of this evaluation. By symmetry considerations we may also assume $x_1 < x_2$. Thus the pure strategy set of the searcher is the set $X = \{(x_1, x_2): 0 \leq x_1 \leq x_1 + e \leq x_2\}$. The length $f(x, y)$ of the uncertainty interval corresponding to a function y and search strategy $x = (x_1, x_2)$ is

$$f(x, y) = \begin{cases} 1 - x_1 & \text{if } y(x_1) < y(x_2), \\ x_2 & \text{if } y(x_1) > y(x_2), \\ x_2 - x_1 & \text{if } y(x_1) = y(x_2). \end{cases}$$

Clearly X is compact. For any fixed continuous function y in Y , f is continuous on the two open sets where y is greater at x_2 or greater at x_1 . Since the value of f on the set where the two evaluations are equal is $x_2 - x_1$, which is less than or equal to the minimum of $1 - x_1$ and x_2 , it follows that f is lower semicontinuous on X for any fixed y . Hence Theorem 1 establishes that the game has a value and the searcher has an optimal mixed strategy. Unlike the two examples given in the first paragraph, the optimal strategies for this game can be computed. The value is $\frac{1}{2} + e/2$ and we leave the determination of the optimal strategies to the reader (the searcher has an optimal strategy which is pure).

Another application of Theorem 1 is to the problem of getting safely, by a continuous trajectory y , from a given point A to a given point B , while avoiding k "mines" x_1, \dots, x_k placed by an opponent in a compact set D which separates A from B . The trajectory y is constrained to lie within some set S in R^n , which contains A , B , and D . With possible further restrictions, the set of allowable trajectories is called Y . The other pure strategy set X is the compact product of k copies of D . Given an allowable trajectory y in Y and mine placements $x = (x_1, \dots, x_k)$ in $X = D^k$, the payoff $f(x, y)$ is $+1$ if the maximizer y successfully reaches B by always keeping a distance greater than a given danger radius r from any of the points x_i . Otherwise, $f(x, y)$ is -1 . The payoff function is lower semicontinuous in x for any fixed trajectory y , so Theorem 1 applies regardless of any further restrictions (turning radius, speed) that may be placed on trajectories. See [G0] for details.

REFERENCES

- [A] S. ALPERN, *Search for point in interval, with high-low feedback*, Math. Proc. Cambridge Philos. Soc., 98 (1985), pp. 569-578.
- [BB] V. J. BASTON AND F. A. BOSTOCK, *A high-low search game on the unit interval*, Math. Proc. Cambridge Philos. Soc., 97 (1985), pp. 345-348.
- [F] K. FAN, *Minimax theorems*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 42-47.
- [G0] S. GAL, *Problems associated with naval mining*, Israel J. Tech., 6 (1968), pp. 316-321.
- [G1] ———, *Search Games*, Math. Sci. Engrg., 149 (1980).
- [G2] ———, *Multidimensional minimax search for a maximum*, SIAM J. Appl. Math., 23 (1972), pp. 513-526.
- [GM] S. GAL AND C. A. MICCHELLI, *Optimal sequential and nonsequential procedures for evaluating a functional*, Appl. Anal., 10 (1980), pp. 105-120.
- [G1] I. L. GLICKSBERG, *Minimax theorems for upper and lower semicontinuous payoffs*, Rand Corporation, Research Memo. RM-478, 1950, 4pp.
- [Ki] J. KIEFER, *Sequential minimax search for a maximum*, Proc. Amer. Math. Soc., 4 (1953), pp. 502-505.
- [K] H. KNESER, *Sur un théorème fondamental de la théorie des jeux*, C. R. Acad. Sci. Paris Ser. A, 234 (1952), pp. 2418-2420.
- [PD] J. E. L. PECK AND A. L. DULMADGE, *Games on a compact set*, Canad. J. Math., 9 (1957), pp. 450-458.
- [R] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.
- [S] M. SION, *On general minimax theorems*, Pacific J. Math., 8 (1958), pp. 171-176.
- [T] F. TERKELSON, *Some minimax theorems*, Math. Scand., 31 (1972), pp. 405-413.

CANONICAL FORMS FOR A CLASS OF DISTRIBUTED PARAMETER CONTROL SYSTEMS*

RICHARD REBARBER†

Abstract. A class of linear distributed parameter control systems with scalar control is considered in which the uncontrolled system is governed by a holomorphic semigroup. A canonical form is constructed for these systems, and is used to help construct feedback controls solving an eigenvalue specification problem. Explicit formulas for the feedback element and the closed-loop eigenvectors are obtained in terms of the original eigenvectors and the basis which is biorthogonal to the eigenvectors. This will be applied to an eigenvalue specification problem for a structurally damped beam. These feedback controls are given by infinite series, which cannot be computed in practice, so the canonical form is used to analyze the effect of truncated series on the distributed system.

The canonical form is a functional equation with an associated canonical state space equation that is equivalent to it. The development in this paper is primarily done in the frequency domain, and hinges on an infinite-dimensional generalization of the characteristic polynomial, which is a meromorphic function.

Key words. distributed parameter systems, canonical form, spectral determination, feedback control, frequency domain, holomorphic semigroups, elastic beam, structural damping, control spillover

AMS(MOS) subject classifications. 93C20, 93B60, 93B55, 93B10

1. Introduction. In this paper we discuss the construction and use of a control canonical form for a class of distributed parameter systems. The canonical form is a functional equation with an associated equivalent canonical state space equation. The canonical form is used by mapping the original state space equation to the canonical state space equation, and analyzing this system by using the functional equation. This makes it easier to determine the evolution of the controlled system. We use the control canonical form to help construct feedback controls solving an eigenvalue specification problem. Although the feedback control is given by an infinite series, it can lead to a usable control scheme by truncating the series to get a finite-dimensional control. We illustrate this numerically in the case of a structurally damped vibrating beam. We also use the canonical form to analyze the effect of the finite-dimensional control on the distributed system. This analysis is in the same spirit as the analysis of the effect of control spillover into higher-order modes (Balas [1]).

The systems we consider are of the form

$$(1.1) \quad \dot{z}(t) = Az(t) + bu(t), \quad z(0) = z_0,$$

where $z(t) \in H$, a Hilbert space, $u(t)$ is a scalar control, and b is a one-dimensional admissible input element (see Definition 1.1 and [5]). We assume that A has a Riesz basis of eigenvectors $\{\varphi_k\}_{k \in I}$, i.e., every $z \in H$ can be written as $\sum_{k \in I} z_k \varphi_k$, and there exists $m, M > 0$ such that

$$(1.2) \quad m \left\| \sum_{k \in I} z_k \varphi_k \right\|^2 \leq \sum_{k \in I} |z_k|^2 \leq M \left\| \sum_{k \in I} z_k \varphi_k \right\|^2,$$

where $\|\cdot\|$ is the norm on H . Let $\{\lambda_k\}_{k \in I}$ be the associated eigenvalues.

Let the dual of H be represented by Hilbert space H' (which can, of course, be chosen to be H), and let $\langle \cdot, \cdot \rangle$ be an inner product on $H' \otimes H$. Let $\{\psi_k\}_{k \in I}$ be the Riesz

* Received by the editors October 28, 1985, accepted for publication (in revised form) November 25, 1987. This research was supported in part by Air Force Office of Scientific Research grant AFOSR-86-0079.

† Department of Mathematics, University of Nebraska, Lincoln, Nebraska 68588.

basis of H' that is biorthogonal to $\{\varphi_k\}_{k \in I}$, i.e., $\{\psi_k, \varphi_j\} = \delta_{j,k}$, where $\delta_{j,k}$ is the Kronecker delta.

In this paper we are interested in the case where the eigenvalues lie between two curves of the form

$$(1.3) \quad \Gamma_a := \{a \pm ix e^{\pm i\sigma} \mid x \in (0, \infty)\},$$

where $\sigma \in [0, \pi/2)$. If $\sigma \in (0, \pi/2)$, A generates a semigroup that is holomorphic on

$$(1.4) \quad \tilde{\Omega} := \{z \in \mathbb{C} \mid |\arg(z)| < \sigma\}$$

and strongly continuous on $\bar{\Omega}$. The semigroup generated by A is given by

$$(1.5) \quad S(t) \left(\sum_{k \in I} x_k \varphi_k \right) = \sum_{k \in I} x_k e^{\lambda_k t} \varphi_k.$$

We now give an outline of this paper. In § 1.1 we discuss a model for a structurally damped vibrating beam with a class of boundary values, based on a general model for linear elastic systems by Chen and Russell [2]. We discuss a specific example with one end hinged, the other sliding, and a control force that approximates a single-point actuator in the center of the beam. We then state the main eigenvalue specification results of the paper, and apply them to the specific example. In § 1.2, motivation for the canonical form is given, using the familiar finite-dimensional canonical form as a point of reference. We briefly present an infinite-dimensional example (Russell [11]), which is more straightforward than ours, but has many of the same features. In § 2, the development of the canonical form is carried out in detail. We first define the canonical state space equation, then the canonical functional equation, and finally show their equivalence. In § 3, we prove the eigenvalue specification results stated in § 1.1. We also analyze the effects of the related finite-dimensional control scheme on the distributed system.

1.1. Examples. Systems of this sort arise from models for structural damping of vibrating systems. Structural damping occurs when the rates of damping of the normal modes of vibration are proportional to the oscillation frequencies of the modes. The following model for structural damping in linear elastic systems is given in Chen and Russell [2]. Suppose that a conservative elastic system is represented by

$$\ddot{w} + Bw = 0,$$

where $w \in X$, a Hilbert space, and B is a positive, self-adjoint operator on X with domain $\mathcal{D}(B)$. Then the equation

$$(1.6) \quad \ddot{w} + 2\gamma B^{1/2} \dot{w} + Bw = 0, \quad 0 < \gamma < 1$$

displays structural damping, since the eigenvalues of the associated first-order operator

$$(1.7) \quad A = \begin{bmatrix} 0 & I \\ -B & -2\gamma B^{1/2} \end{bmatrix}$$

lie on Γ_0 (given by (1.3)) for some $\sigma \in [0, \pi/2)$, and the associated eigenvectors form a Riesz basis of the appropriate space.

Structurally damped beam. The Euler-Bernoulli equation

$$\ddot{w}(x, t) + (EI/p) \dot{w}'''(x, t) = f(x)u(t),$$

where the dot denotes differentiation with respect to time $t \in \bar{\Omega}$ and the prime denotes differentiation with respect to the position $x \in [0, L]$, is an approximate mathematical

model for a long slender beam, where $w(x, t)$ is the lateral deflection of the beam, $f(x)u(t)$ is the external distributed transverse force and $f \in L^2[0, L]$. We assume that the flexural rigidity EI and the density ρ of the beam are constant. For our purposes, the easiest boundary conditions to deal with are of the following form:

$$(1.8) \quad \begin{aligned} \alpha_0 w(0, t) + \beta_0 w'(0, t) &= 0, & \alpha_1 w(L, t) + \beta_1 w'(L, t) &= 0, & \alpha_0^2 + \beta_0^2 &\neq 0, \\ \alpha_0 w''(0, t) + \beta_0 w'''(0, t) &= 0, & \alpha_1 w''(L, t) + \beta_1 w'''(L, t) &= 0, & \alpha_1^2 + \beta_1^2 &\neq 0. \end{aligned}$$

For instance, when $\alpha_1 = 0$ the right end is sliding, and when $\beta_1 = 0$, the right end is hinged.

Without loss of generality, we will assume that $L = 1$. Let

$$X = L^2[0, 1],$$

$$B: X \rightarrow X: w \rightarrow (EI/\rho)w'''' ,$$

$$\begin{aligned} \mathcal{D}(B) = \{w \in H^4[0, 1] \mid \alpha_0 w(0) + \beta_0 w'(0) &= 0, \alpha_1 w(1) + \beta_1 w'(1) = 0, \\ \alpha_0 w''(0) + \beta_0 w'''(0) &= 0; \alpha_1 w''(1) + \beta_1 w'''(1) = 0\}. \end{aligned}$$

Then B has a set of eigenvectors $\{\Phi_k\}_{k \in \mathbb{Z}^+}$ that form a Riesz basis for X , and associated real nonnegative eigenvalues $\{\sigma_k\}_{k \in \mathbb{Z}^+}$, which are $O(k^4)$. With this choice of $\mathcal{D}(B)$, $B^{1/2}$ is given by the differential operator $B^{1/2}w = -(EI/\rho)^{1/2}w''$ and

$$\mathcal{D}(B^{1/2}) = \{w \in H^2[0, 1] \mid \alpha_0 w(0) + \beta_0 w'(0) = 0, \alpha_1 w(1) + \beta_1 w'(1) = 0\}.$$

For other boundary conditions, $B^{1/2}$ is not a differential operator.

Using (1.6) to model structural damping in the beam leads to

$$(1.9) \quad \ddot{w}(x, t) - 2\gamma(EI/\rho)^{1/2}\dot{w}'' + (EI/\rho)w''''(x, t) = f(x)u(t),$$

with the boundary conditions (1.7) and initial conditions

$$(1.10) \quad w(x, 0) = w_0, \quad \dot{w}(x, 0) = w_1$$

where $w_0 \in \mathcal{D}(B^{1/2})$ and $w_1 \in X$. Numerical tests in [17] imply that (1.9) might be a model for a structurally damped beam with other boundary conditions as well. The damping term $-2\gamma(EI/\rho)^{1/2}\dot{w}''$ can be interpreted as a lateral force on the beam negatively proportional to the rate of bending [14].

To put this in a state space setting, let

$$Y := \mathcal{D}(B^{1/2}), \quad H := Y \oplus X,$$

where H has the norm $\|[z_1, z_2]^T\|^2 := \|z_1''\|_{L^2[0,1]}^2 + \|z_2\|_{L^2[0,1]}^2$. Let

$$(1.11) \quad z(t) = [w(t), \dot{w}(t)]^T, \quad z_0 = [w_0, w_1]^T, \quad b = [0, f(\cdot)]^T,$$

so z_0 and b are in H , and (1.8), (1.9), and (1.10) can be written in the form (1.1), where A is given by (1.7). A has domain $\mathcal{D}(A) = \mathcal{D}(B) \oplus Y$, and A generates a holomorphic semigroup on H . To see this, it is sufficient to note that A has eigenvectors

$$(1.12) \quad \varphi_{\pm k} := \frac{1}{\sqrt{2}} \begin{bmatrix} \sigma_k^{-1/2} \Phi_k \\ e^{\pm i\eta} \Phi_k \end{bmatrix}, \quad e^{\pm i\eta} = (-\gamma \pm \sqrt{1 - \gamma^2}), \quad k \in \mathbb{Z}^+$$

with associated eigenvalues $\lambda_{\pm k} = (\sigma_{\pm k})^{1/2} e^{\pm i\eta}$, which lie on Γ_0 given by (1.3) with $\sigma = \eta - \pi/2$. These eigenvectors are easily seen to form a Riesz basis for H . For the remainder of this section, let I be the index set $\mathbb{Z}^+ \cup \mathbb{Z}^-$, so A generates the semigroup given by (1.5).

For our purposes, it is convenient to think of the dual of H as being

$$H' := Y' \oplus X,$$

where

$$Y' := \mathcal{D}(B^{1/2})' = \left\{ \sum_{k \in I} x_k \Phi_k \mid \sum_{k \in I} |x_k^2 / \sigma_k| < \infty \right\} = \mathcal{D}(B^{-1/2}).$$

The inner product for $[z_1, z_2]^T \in H'$ and $[y_1, y_2]^T \in H$ is

$$(1.13) \quad \langle [z_1, z_2]^T, [y_1, y_2]^T \rangle := \int_0^1 (B^{-1/2} z_1(x) B^{1/2} y_1(x) + z_2(x) y_2(x)) dx.$$

It is easy to see that

$$(1.14) \quad \psi_{\pm k} := \frac{1}{\sqrt{2} \sin \eta} \begin{bmatrix} -(\pm i \sigma_k^{1/2} e^{-(\pm i \eta)} \Phi_k) \\ \pm i \Phi_k \end{bmatrix}, \quad k \in \mathbb{Z}^+$$

is a Riesz basis for H' , and that $\langle \varphi_k, \psi_\ell \rangle = \delta_{k, \ell}$, so $\{\psi_k\}_{k \in I}$ is the dual basis to $\{\varphi_k\}_{k \in I}$.

When $z_0 \in \mathcal{D}(A)$, $S(t)z_0$ is in H for all $t \in \bar{\Omega}$ and is the classical solution of

$$(1.15) \quad \dot{z}(t) = Az(t), \quad z(0) = z_0.$$

If z_0 is in H , then $z(t) = S(t)z_0$ is in H and is the solution of (1.15) in a generalized sense, but (1.15) is no longer an equation in H . A can be extended to

$$(1.16) \quad \hat{A}: H \rightarrow \mathcal{H}: \langle \eta, \hat{A}z \rangle_* := \langle A'\eta, z \rangle \quad \forall \eta \in \mathcal{D}(A'),$$

where A' is the adjoint of A when the inner product (1.13) is used, and $\mathcal{H} := \mathcal{D}(A)'$.

In this context $\mathcal{D}(A')$ is the domain of A' in H' , so $\mathcal{D}(A') = X \oplus Y$, and

$$\mathcal{H} = X \oplus Y' = X \oplus \mathcal{D}(B^{-1/2}).$$

Therefore (1.15) can be interpreted as an equation in \mathcal{H} , and $S(t)z_0$ is the solution of this equation.

If b is in \mathcal{H} , it can be written as $\hat{A}\tilde{b}$ for some \tilde{b} in H , so b can be expanded as

$$(1.17) \quad b = \sum_{k \in I} b_k \varphi_k \quad \text{where } b_k = \langle \psi_k, b \rangle_* = \lambda_k \langle \psi_k, \tilde{b} \rangle.$$

We can now give the definition of an admissible input element that we will be using in this paper. This is a slight modification of the definition in [5].

DEFINITION 1.1. If b is an element of \mathcal{H} , then b is admissible on $\ell_\theta = \{x e^{i\theta} \mid x > 0\}$ if the family of operators $B(t): L^2[\ell_\theta] \rightarrow H$ defined by

$$\langle B(t)u, y \rangle = \int_{[0, t]} \langle b, S(t-s)'y \rangle u(s) ds \quad \text{for every } y \in \mathcal{D}(A')$$

(where $[0, t]$ is the straight line segment from zero to t) is a strongly continuous family of bounded operators for all $t \in \ell_\theta$. b is an "admissible input element" if b is admissible on ℓ_θ for $\theta \in [-\sigma, \sigma]$.

It is shown in [5] that if b is an admissible input element, then $z(t) = S(t)z_0 + B(t)u$ is a generalized solution of (1.1) for $t \in \bar{\Omega}$, where (1.1) is an equation in \mathcal{H} and $z(t)$ is in H . If $z_0 = 0$ and b is given by (1.17), then

$$z(t) = \sum_{k \in I} \left(b_k \int_{[0, t]} e^{\lambda_k(t-s)} u(s) ds \right) \varphi_k$$

and b is an admissible input element if the coefficients of φ_k are square summable whenever $u \in L^2[\ell_\theta]$ for $\theta \in [-\sigma, \sigma]$. It is easily seen from standard perturbation results (e.g., [6]) that bounded input elements are also admissible.

We can frequently write the control system in the form (1.1) with an admissible input element when the control input is not distributed across the spatial extent of the beam, but is concentrated at a finite number of interior or boundary points (e.g., [5], [10]). Sometimes we can even solve an eigenvalue specification problem if the input element is not even admissible [10].

We now consider the following specific control system, which describes a structurally damped beam with a hinged left end and a sliding right end:

$$(1.18) \quad \begin{aligned} \ddot{w}(x, t) - \dot{w}''(x, t) + w''''(x, t) &= f(x)u(t), \\ w(0) = w'(1) = w''(0) = w'''(1) &= 0. \end{aligned}$$

We take $f(x)$ to approximate a single point actuator at $x = .5$, i.e., $f(x) = 1/(2\varepsilon)$ for $|x| < \varepsilon$, and $f(x) = 0$ for $|x| \geq \varepsilon$. For this system $\Phi_k(x) = \sqrt{2} \sin((-\pi/2 + \pi k)x)$, $\sigma = \pi/6$, and $\eta = 2\pi/3$, so the eigenvalues of A are

$$(1.19) \quad \lambda_{\pm k} = (-\pi/2 + \pi k)^2 e^{\pm i2\pi/3},$$

and the eigenvectors $\{\varphi_k\}_{k \in I}$ and the dual vectors $\{\psi_k\}_{k \in I}$ are given by (cf. (1.12), (1.14))

$$(1.20) \quad \varphi_{\pm k}(x) := \begin{bmatrix} (1/(-\pi/2 + \pi k))^2 \sin((-\pi/2 + \pi k)x) \\ e^{\pm i2\pi/3} \sin((-\pi/2 + \pi k)s) \end{bmatrix},$$

$$(1.21) \quad \psi_{\pm k}(x) := (2\sqrt{3}) \begin{bmatrix} -(\pm i(-\pi/2 + \pi k)) e^{-(\pm i2\pi/3)} \sin((-\pi/2 + \pi k)x) \\ \pm i \sin((-\pi/2 + \pi k)x) \end{bmatrix},$$

for $k \in Z^+$.

The input element coefficients (1.17) are in this case

$$(1.22) \quad \begin{aligned} b_{\pm k} = \langle b, \psi_{\pm k} \rangle &= (2/\sqrt{3}) \int_0^1 \pm i \sin((-\pi/2 + \pi k)s) f(s) ds \\ &= \pm i\sqrt{2/3}(-1)^{\sigma(k)} \sin(\varepsilon(-\pi/2 + \pi k))/(\varepsilon(-\pi/2 + \pi k)), \end{aligned}$$

where $\sigma(k) = 0$ if k is 1 mod 4 or 2 mod 4, and $\sigma(k) = 1$ if k is 3 mod 4 or 4 mod 4. Note that as ε goes to zero, this goes to

$$(1.23) \quad b_{\pm k} = \sqrt{2/3} \pm i(-1)^{\sigma(k)},$$

which is what we would get if f were the Dirac delta distribution at $x = .5$.

We now consider the eigenvalue specification problem for the system (1.1), state our results (which are proved in § 3), and apply them to the specific problem.

Eigenvalue specification problem. Given a set of complex numbers $\{\alpha_k\}_{k \in I}$, can a feedback control $h^*: H \rightarrow \mathbb{R}$ be found such that the closed-loop operator $A + bh^*$ has eigenvalues $\{\alpha_k\}_{k \in I}$.

If the feedback element h^* is bounded, it can be represented by an element $h \in H'$ via $h^*z = \langle h, z \rangle$; and any element h of H' has an associated bounded functional h^* . In [15], Sun considers systems that include those we are considering in this paper. He proves as a necessary and sufficient condition that there exists a bounded functional $h^*: H \rightarrow \mathbb{R}$ such that $A + bh^*$ has eigenvalues at $\{\alpha_k\}_{k \in I}$ is

$$(1.24) \quad \sum_{k \in I} |(\alpha_k - \lambda_k)/b_k|^2 < \infty.$$

This answers the existence question completely, at least as long as we only consider bounded input and feedback elements. Unfortunately, this rules out many interesting

cases, including boundary control. In [10] it is shown that (1.24) is sufficient when b is only admissible for the systems under consideration in this paper. Another drawback of Sun's result is that the eigenvectors of the closed-loop operator are not explicitly demonstrated, although they can be found by solving an infinite-dimensional linear system of equations.

The restriction to bounded feedback elements means that we cannot move the eigenvalues uniformly away from the original values. In this paper we will use the following definition of an admissible feedback element, which depends on b .

DEFINITION 1.2. Suppose $h^*: H \rightarrow \mathbb{R}$ is linear and possibly unbounded with domain $\mathcal{D}(h^*)$, and b is admissible. For all $z \in H$ we can compute $\hat{A}z + bh^*z$ as an element of \mathcal{H} . Let

$$(1.25) \quad \mathcal{D}(A + bh^*) := \{z \in \mathcal{D}(h^*) \mid \hat{A}z + bh^*z \in H\}$$

and for $z \in \mathcal{D}(A + bh^*)$,

$$(1.26) \quad (A + bh^*)z := \hat{A}z + bh^*z.$$

h^* is admissible if $A + bh^*$ generates a holomorphic semigroup.

We will address the following problem: Find a linear functional h^* such that $A + bh^*$ has eigenvalues at $\{\alpha_k\}_{k \in I}$, compute the resulting feedback control $u(t)$ in terms of the initial state, and find the closed-loop eigenvectors $\{\chi_k\}_{k \in I}$. This is solved by the following theorem, which is proved in § 3. The solution depends on the construction of meromorphic functions called *cardinal functions*, which are an infinite-dimensional generalization of characteristic polynomials. Cardinal functions are defined in § 2.2, and a class of cardinal functions are constructed in [8]. We will give a cardinal function for the specific control system described above. The use of the cardinal function is central to the development of the canonical form for (1.1).

THEOREM 1.3. Assume $\{\lambda_k\}_{k \in I}$ is the zero set of a cardinal function p . Suppose $\{\alpha_k\}_{k \in I}$ is the zero set of a cardinal function q with the same poles as p . Let

$$(1.27) \quad \chi_k = (1/p'(\lambda_k)b_k) \sum_{j \in I} [p(\alpha_k)b_j/(\alpha_k - \lambda_j)]\varphi_j,$$

$$(1.28) \quad h_j^\alpha = b_j p'(\lambda_j) \sum_{k \in I} q(\lambda_k)/[q'(\alpha_j)(\lambda_k - \alpha_j)b_k p'(\lambda_k)]\psi_k,$$

$$(1.29) \quad h^* = \sum_{j \in I} p(\alpha_j)h_j^\alpha/b_j p'(\lambda_j).$$

Suppose either one of the two following conditions holds:

(1) b satisfies

$$(1.30) \quad m < b_k < M \quad \text{for all } k \in I.$$

(2) $\{\alpha_k\}_{k \in I}$ satisfies (1.24), and b is a bounded input element, so that $\{b_k\}_{k \in I} \in \ell_2$.

Then $A + bh^*$ has eigenvalues at $\{\alpha_k\}_{k \in I}$ and a Riesz basis of eigenvectors $\{\chi_k\}_{k \in I}$. Furthermore, $\{h_j^\alpha\}_{j \in I}$ is a Riesz basis for H' and $\langle \chi_k, h_j^\alpha \rangle = \delta_{j,k}$. If the initial state z_0 is given by $\sum_{k \in I} z_k \varphi_k$, the control $u(t) = h^*z(t)$ is given by

$$(1.31) \quad u(t) = \sum_{j \in I} \sum_{k \in I} [z_k p(\alpha_j) q(\lambda_k)/q'(\alpha_j)(\lambda_k - \alpha_j)b_k p'(\lambda_k)] e^{\alpha_j t}.$$

If condition (1) holds, $A + bh^*$ is interpreted as in (1.25) and (1.26). If condition (2) holds, the feedback element h^* is bounded. \square

Remark 1. If $\{\lambda_k\}_{k \in I}$ is the zero set of any of the cardinal functions constructed in [8], then there is a cardinal function q with the same poles, and zeros at $\{\alpha_k\}_{k \in I}$, if (1.24) holds along with conditions 1 or 2 in Theorem 1.3.

Remark 2. When $\{\lambda_k\}_{k \in I}$ is the eigenvalue set associated with the structurally damped beam model (1.8)–(1.10), there is a cardinal function p with zeros at $\{\lambda_k\}_{k \in I}$.

Remark 3. Other eigenvalue specification results of this sort can be found in [10].

We can apply these results to the specific beam example (1.18). In this case λ_k is given by (1.19) and b_k is given by (1.22). If we let $\mu_k = \lambda_k - 10$, it is shown in [8] that

$$(1.32) \quad p(\lambda) = \prod_{k \in I} (\lambda - \lambda_k) / (\lambda - \mu_k)$$

is a cardinal function. It is a consequence of results in [8] that if

$$(1.33) \quad \alpha_k = \lambda_k - \beta_k \quad \text{where } \{\beta_k\}_{k \in I} \in \ell_\infty,$$

then

$$(1.34) \quad q(\lambda) = \prod_{k \in I} (\lambda - \alpha_k) / (\lambda - \mu_k)$$

is a cardinal function. In this example, we will realize the closed-loop eigenvalues

$$\alpha_{\pm k} = \lambda_{\pm k} - 8/k^2.$$

Then $\{\alpha_k\}_{k \in I}$ satisfies (1.24) when b_k is given by (1.22).

It is possible to approximate $p(\alpha_j)$, $q(\lambda_k)$, $p'(\lambda_k)$, and $q'(\alpha_j)$ numerically, which we did for j and $k = \pm 1, 2, \dots, 9$. Letting

$$p_l(\lambda) = \prod_{k=-l}^l (\lambda - \lambda_k) / (\lambda - \mu_k),$$

we approximate $p(\alpha_j)$ by $p_l(\alpha_j)$, where l was chosen as that $|(p_l(\alpha_j)/p_{l-1}(\alpha_j)) - 1| < .005$. Then the feedback element is given by $h^* = \sum_{j \in I} c_j \psi_j + c_j \psi_{-j}$, where ψ_j is given by (1.21) and $\{c_j\}_{j=1}^9$ is (by (1.28) and (1.29))

$$\begin{aligned} &\{28.98 - 15.84i, 2.10 - 2.03i, -0.46 + 1.41i, -0.17 + 0.97i, 0.08 \\ &\quad -0.75i, 0.05 - 0.65i, -1.67 + 38.24i, -1.87 + 56.15i, 9.15 - 93.5i\}. \end{aligned}$$

If h^* is written as $[h_1(s), h_2(s)]$, then the control is given by

$$u(t) = \langle h^*, z(t) \rangle (\text{cf. (1.13)}) = \int_0^1 \tilde{h}_1(x) w''(x, t) + h_2(x) w(x, t) dx,$$

where $w(x, t)$ solves (1.18) and

$$\begin{aligned} \tilde{h}_1(x) &= B^{1/2} h_1(x) = 10^{2*} \sum_{j>0} a_j \sin((\pi k - \pi/2)x), \\ h_2(x) &= 10^{2*} \sum_{j>0} b_j \sin((\pi k - \pi/2)x), \end{aligned}$$

with the coefficients given approximately by

$$\begin{aligned} \{a_j\}_{j=1}^9 &= \{-.7625, -.0654, .0256, .0145, -.0102, -.0084, .4749, .6857, -1.2627\}, \\ \{b_j\}_{j=1}^9 &= \{-.3658, -.0468, .0326, .0224, -.0173, -.0149, .8831, 1.2968, -2.1593\}. \end{aligned}$$

If we know the initial states $w_0 \in \mathcal{D}(B^{1/2})$ and $w_1 \in X$, we can compute the control $u(t)$ using (1.31). As an example, we take

$$w_0(x) = (\cos(\pi x) - 1), \quad w_1(x) = x.$$

Expanding these gives

$$\begin{aligned} w_0(x) &= \sum_{k>0} \sin((\pi k - \pi/2)x) / [\pi(.5 + k)(-1.5 + k)(-.5 + k)], \\ w_1(x) &= \sum_{k>0} \sin((\pi k - \pi/2)x) (-1)^{k+1} / (\pi k - \pi/2)^2. \end{aligned}$$

Then we can write $z_0 = [w_0, w_1]$ as $\sum_{k>1} z_k \varphi_k + z_k \varphi_k$, where φ_k is given by (1.20) and $\{z_k\}_{k=1}^9$ is approximately $\{-1.481 + .5241i, 2.6657 + 1.5023i, 1.0578 + .6240i, .6911 + .3923i, .5193 + .3039i, .4177 + .2384i, .3500 + .2041i, .3016 + .1726i, .2650 + .1542i, .2365 + .1356i\}$. In this case (1.31) gives

$$u(t) = \sum_{j>0} (f_j e^{\alpha_j t} + \bar{f}_j e^{\bar{\alpha}_j t}) = \sum_{j>0} \operatorname{Re} (f_j e^{\alpha_j t}),$$

$$\{f_j\}_{j=1}^9 = \{-1.4810 - .5241i, 2.6657 + 1.5023i, 1.0578 + .6240i, .6911 + .3923i, .5193 + .3039i, .4177 + .2384i, .3500 + .2041i, .3016 + .1726i, .2650 + .1542i\}.$$

If we use $f(x) = \delta(x - .5)$, where δ is the Dirac delta function and the external force is a point accuator at $x = .5$, then we can go through the same procedure, but the closed-loop eigenvalues will not have to satisfy (1.24), since the input element coefficients (1.23) satisfy (1.30). In this case we can let α_k be the zero set of any cardinal function with poles at $\{\mu_k\}_{k \in I}$. For instance, we can realize any closed-loop eigenvectors $\{\alpha_k\}_{k \in I}$ such that $\alpha_k = \lambda_k - \beta_k$, where $\{\beta_k\}_{k \in I}$ is a bounded sequence, since $q(\lambda)$ given by (1.34) is still a cardinal function in this case.

Since it is not possible to compute the infinite series for $u(t)$, an obvious question arises regarding the properties of the control

$$(1.35) \quad u_{n,m}(t) = \sum_{j \in I_n} \sum_{k \in I_m} [z_k p(\alpha_j) q(\lambda_k)] / [q'(\alpha_j)(\lambda_k - \alpha_j) b_k p'(\lambda_k)] e^{\alpha_j t},$$

where $I_n := \{\pm k \mid k = 1, 2, \dots, n\}$. We are interested in the effect of using $u_{n,m}(t)$ as the control in (1.1); in particular, does the control add energy to the higher-order modes of the system? As an illustration of the usefulness of the canonical form, we will analyze the effect of this control scheme, and obtain the following result.

THEOREM 1.4. *Suppose $z(t)$ solves (1.1) with $u = u_{n,m}$, and suppose (1.1) satisfies either of the conditions in Theorem 1.3. Then there exists M , independent of $z_0 \in H$ and $t \in \tilde{\Omega}$ such that*

$$\|z(t)\| \leq MC(t) \|z_0\| \quad \text{where } C(t) = \sup \{|e^{\lambda_k t}|, |e^{\alpha_j t}|\}_{j \in I_n, k \in I}.$$

Furthermore, suppose we fix ε . Then, for a given z_0 , there exists M and N large enough that for $n \geq N$ and $m \geq M$, and for $t \in \tilde{\Omega}$,

$$\|z(t)\| \leq M_1 C_1(t) \|z_0\| + \varepsilon C_2(t),$$

for some M_1 independent of t and z_0 , and $C_1(t) = \sup \{|e^{\alpha_j t}|\}_{j \in I_n}$, $C_2(t) = \sup \{|e^{\lambda_k t}|\}_{k \in I}$. \square

To illustrate this, consider the specific example (1.18) with the damping term $\dot{w}''(x, t)$ removed, so that σ is zero instead of $\pi/6$ and everything else is the same. Then the uncontrolled system is conservative. Suppose our goal is to increase the stability of the system. If we assume that $f(x) = \delta(x - .5)$, we can construct the control given by (1.35) using $\alpha_k = \lambda_k - 8$, since condition (1) of Theorem 1.3 holds. Theorem 1.4 says that, given ε , we can find n and m , depending on z_0 , large enough so that for $t \in [0, \infty)$,

$$\|w''(\cdot, t)\|_{L^2[0,1]}^2 + \|w(\cdot, t)\|_{L^2[0,1]}^2 \leq C e^{-8t} (\|w_0\|^2 + \|w_1\|^2) + \varepsilon$$

for some C independent of t and z_0 .

This approach to the eigenvalue specification problem retains the distributive nature of the system until the last step, where (1.35) is computed. This is somewhat different from the more typical method of using a reduced-order model and then analyzing the effect of the control spillover (e.g., [1]). However, Theorem 1.4 is in the

spirit of the analysis of control spillover in [1], in the sense that we analyze the effect of a finite-dimensional control scheme on all the modes of an infinite-dimensional system.

1.2. Control canonical forms. We consider a control canonical form to be a scalar functional equation and an associated canonical state space equation. These two equations are equivalent in the sense that we can recover the solution of one from the solution of the other. To use the canonical form, we map the original system to the canonical state space equation, and then use the scalar equation to analyze the properties of the solution. Canonical forms relevant to this discussion have been used to study control in distributed parameter systems by Russell [11], [13], Ho [4], Clarke and Williamson [3], and Teglas [16]. In [4], [11], [13], [16] the cardinal functions are entire, while in this paper we allow meromorphic cardinal functions, which allows us to construct canonical forms for a larger class of distributed parameter systems. This involves a frequency domain (i.e., Laplace transform) approach which proves to be rather general. To motivate this work, we will first discuss the familiar finite-dimensional canonical form to get an idea of the machinery that we need, and then discuss an infinite-dimensional example done by Russell [11] which will indicate the approach we will take.

Finite-dimensional canonical form. Let the characteristic polynomial of the system be normalized so that $p(z) = a_1 + a_2z + a_3z^2 + \cdots + a_nz^{n-1} + z^n$. When the controllability matrix $(A^{n-1}b, A^{n-2}b, \dots, Ab, b)$ is nonsingular, there exists a nonsingular matrix T such that $x = Ty$ brings the original system to the canonical state space equation. If $y = [y_1, y_2, \dots, y_n]^T$, then

$$(1.36) \quad \begin{aligned} y_i(t) &= y_{i+1}(t), & i = 1, 2, \dots, n-1, \\ \dot{y}_n(t) &= -(a_1y_1(t) + \cdots + a_{n-1}y_{n-1}(t) + a_ny_n(t)) + u(t). \end{aligned}$$

This can be written as

$$(1.37) \quad p(D)y_1(t) = u(t).$$

We can study this equation using the Laplace transform. Let $(\mathcal{L}y)(z)$ be the Laplace transform of y . It is well known that

$$(1.38) \quad \mathcal{L}(p(D)y(t)) = p(z)(\mathcal{L}y)(z) + f(z),$$

where f is an $n-1$ -degree polynomial determined by p and the initial conditions.

We will rewrite this slightly to indicate the way we will generalize these notions. For $y \in C^{n-1}[0, T]$, define the functional $\langle \eta, y \rangle := p(D)y(s)|_{s=0}$. Then (1.37) becomes

$$(1.39) \quad (\eta * y)(t) := \langle \eta, y(t + \cdot) \rangle = u(t).$$

Note that the Fourier transform of η , $\mathcal{F}\eta$, is

$$(1.40) \quad \mathcal{F}\eta(z) := \langle \eta, e^{z\cdot} \rangle = p(z).$$

Our goal will be to develop the following:

- (1) A function $p(z)$ for distributed parameter systems that parallels the role of the characteristic polynomial in finite-dimensional systems.
- (2) A functional η , satisfying (1.40), for which (1.39) makes sense and can be solved for a class of controls u .
- (3) A Laplace transform theory, for which there is an analogue to (1.38).
- (4) A state space equation that is equivalent to the functional equation.
- (5) A way of recovering the solution of the state space equation from the solution of the functional equation, analogous to (1.36).

An infinite-dimensional example. The most straightforward example of an infinite-dimensional canonical form of this type is the hyperbolic case discussed by Russell in [11], [12].

We assume that the eigenvalues are of the form $\{2\pi ji + \varepsilon_j + \alpha\}_{j \in \mathbb{Z}}$, where $\{\varepsilon_j\}_{j \in \mathbb{Z}}$ is a square summable sequence and $\alpha \in \mathbb{C}$. It is well known that $\{e^{\lambda_j}\}_{j \in \mathbb{Z}}$ is a Riesz basis for $L^2[-1, 1]$. This space will serve as the canonical state space, and the canonical state space equation in $L^2[-1, 1]$ is

$$\dot{x}(t) = Dx(t) + \hat{b}u(t).$$

Here D is a “differentiation” operator in the sense that $D(\sum_{k \in \mathbb{Z}} x_k e^{\lambda_k s}) = \sum_{k \in \mathbb{Z}} x_k \lambda_k e^{\lambda_k s}$. The domain of D , $\mathcal{D}(D)$, is the subspace of $H^1[-1, 1]$ that contains all elements of the form $\sum_{k \in \mathbb{Z}} x_k e^{\lambda_k}$ for which $\{x_k \lambda_k\}_{k \in \mathbb{Z}}$ is square summable. The canonical input element \hat{b} does not belong to $L^2[-1, 1]$, but it is an admissible input element.

The canonical functional equation is of the form (1.39), where the “generating functional” η is an element of the dual space of $H^1[-1, 1]$. $\mathcal{D}(D)$ has the alternate characterization as all x in $H^1[-1, 1]$ such that $\langle \eta, x \rangle = 0$. For these systems (1.39) is a zero-order neutral equation. The convolution operator can also be defined in the frequency domain by using the functions p , which is related to η by (1.40). To solve this functional equation, we can use the two-sided Laplace transform theory developed in Russell [12]. It can be shown that the solution of the canonical state space equation is $x(t) = y(t + \cdot)$ for all real t , where y solves $\eta * y = u$.

In general, it is not possible to find a useful state space that has $\{e^{\lambda_k}\}_{k \in I}$ as a Riesz basis. In the general approach, the canonical state space E will be a space of equivalence classes of a space X , where the equivalence class containing x is denoted by $\{x\}$. E will have $\{\{e^{\lambda_k}\}\}_{k \in I}$ as a Riesz basis. We can put the canonical form above into this framework as follows. Let $X^a := \{x \mid e^{-\rho|\cdot|}x \in L^2(-\infty, \infty), \text{ for all } \rho > a\}$. We then write $x_1 \sim x_2$ if $x_1 = x_2$ in $L^2[-1, 1]$. In this way we can think of the canonical state space as being the space of all equivalence classes of X^a .

More generally, we start with spaces X and Y , and a bilinear form $\langle \cdot, \cdot \rangle$ defined on $Y \otimes X$. Then we say that $x_1 \sim x_2$ if $\langle y, x_1 \rangle = \langle y, x_2 \rangle$ for all $y \in Y$. In the above hyperbolic case, X is X^a , Y is $L^2[-1, 1]$, and $\langle y, x \rangle$ is $\int_{-1}^1 x(t)y(t) dt$. This approach, which is introduced in Russell [11], is the one we will use.

2. Canonical form.

2.1. A Laplace transform. The canonical state space and functional equation are constructed in the frequency domain, so we need to use a Laplace transform relevant to holomorphic semigroups, described in Rebarber [7]. We summarize the needed results here, starting with some definitions.

DEFINITION 2.1. Let a be real and $\theta \in [0, \pi/2)$, and let

$$\Gamma_{a,\theta} := \Gamma_{a,\theta}^1 \cup \Gamma_{a,\theta}^2 := \{a + ix e^{i\theta} \mid x \in [0, \infty)\} \cup \{a - ix e^{-i\theta} \mid x \in [0, \infty)\},$$

oriented counterclockwise; we write $\lambda > \Gamma_{a,\theta}$ when λ is to the right of $\Gamma_{a,\theta}$ in the complex plane, and $\lambda < \Gamma_{a,\theta}$ when λ is to the left of $\Gamma_{a,\theta}$. Most of the time we think of σ as being fixed, so we let

$$\Gamma_a := \Gamma_{a,\sigma}.$$

We also let

$$\ell_\theta := \{x e^{i\theta} \mid x > 0\}, \quad \Omega_{a,\theta} := \{\lambda \mid \lambda > \Gamma_{a,\theta}\}, \quad \Omega_a := \Omega_{a,\sigma}.$$

DEFINITION 2.2. X^a is defined as the space of all functions f that are holomorphic in $\tilde{\Omega}$, have L^2 boundary values on ℓ_σ and $\ell_{-\sigma}$, and satisfy, for some $C = C(\rho, f) > 0$,

$$(2.1) \quad \|e^{-\rho} f\|_{L^2[\ell_\sigma]} \leq C \quad \text{for all } \theta \in [-\sigma, \sigma] \text{ and } \rho > a.$$

X^a will be the domain of the Laplace transform.

DEFINITION 2.3. Ψ^a is defined as the space of all functions ψ that are holomorphic in Ω_a , such that for all $\rho > a$ and $\theta \in (-\sigma, \sigma)$ there exists $m(\rho, \psi)$, $M(\rho, \theta, \psi)$, and $\varepsilon(\rho, \theta, \psi)$, all positive, such that

$$(2.2) \quad \|\psi\|_{L^2[\Gamma_r]} \leq m \quad \text{for } r > \rho,$$

$$(2.3) \quad |\psi(\lambda)| \leq m \quad \text{for } \lambda \in \Omega_\rho,$$

$$(2.4) \quad |\psi(\lambda)| \leq M|\lambda|^{-\varepsilon} \quad \text{for } \lambda \in \Omega_{\rho, \theta}.$$

Ψ^a will be the range of the Laplace transform. We can put norms on these spaces. For $f \in X^a$ and $\rho > a$, let

$$\|f\|_\rho := \left(\int_{\ell_\sigma \cup \ell_{-\sigma}} |e^{-\rho t} f(t)|^2 |dt| \right)^{1/2}.$$

For $\psi \in \Psi^a$ and $\rho > a$, let

$$\|\psi\|_\rho = \left(\int_{\Gamma_\rho} |\psi(\lambda)|^2 |d\lambda| \right)^{1/2}.$$

We are now ready to define the Laplace transform and its inverse.

DEFINITION 2.4. Let

$$S_{a, \theta} := \{a + \xi e^{-i\theta} \mid \operatorname{Re}(\xi) > 0\}.$$

For $\lambda \in S_{a, \theta}$, with $\theta \in [-\sigma, \sigma]$, and $f \in X^a$, let

$$\mathcal{L}_\theta f(\lambda) := \int_{\ell_\theta} e^{-\lambda t} f(t) dt.$$

It is shown in [7] that $\{\mathcal{L}_\theta f\}_{|\theta| \leq \sigma}$ is a set of analytic continuations, so we can define $\mathcal{L}f(\lambda)$ for any $\lambda \in \Omega_a$, which is just $\bigcup_{|\theta| \leq \sigma} S_{a, \theta}$, by

$$\mathcal{L}f(\lambda) := \mathcal{L}_\theta f(\lambda) \quad \text{for } \lambda \in S_{a, \theta}.$$

DEFINITION 2.5. For $\rho > a$, $t \in \tilde{\Omega}$, and $\psi \in \Psi^a$, let

$$\mathcal{K}\psi(\lambda) := \left(\frac{1}{2\pi i} \right) \int_{\Gamma_\rho} e^{\lambda t} \psi(\lambda) d\lambda.$$

It is shown in [7] that $\mathcal{K}\psi$ is independent of ρ .

THEOREM 2.6 [7]. Let $\rho > a$. If X^a has norm $\|f\|_\rho$ and Ψ^a has norm $\|\psi\|_\rho$, then \mathcal{L} is an isomorphism between X^a and Ψ^a , and $\mathcal{L}^{-1} = \mathcal{K}$.

Theorem 2.6 will allow us to translate results between the frequency domain and the time domain. In the remainder of the paper, we will freely do this.

2.2. Canonical state space. The canonical state space will be a space of equivalence classes.

DEFINITION 2.7. Let $M^{a,b}$ be the set of all functions in Ψ^a that are holomorphic in $\{\lambda > \Gamma_a\} \cup \{\lambda < \Gamma_b\}$ and for which, for all $\rho < b$ and $\theta \in [\sigma, \pi/2]$, there exists $m(\rho, \psi)$, $M(\rho, \theta, \psi)$, and $\varepsilon(\rho, \theta, \psi)$, all positive, such that

$$\begin{aligned}\|\psi\|_{L^2[\Gamma, \cdot]} &\leq m \quad \text{for all } r < \rho, \\ |\psi(\lambda)| &\leq m \quad \text{for } \lambda < \Gamma_\rho, \\ |\psi(\lambda)| &\leq M|\lambda|^{-\varepsilon} \quad \text{for } \lambda < \Gamma_{\rho, \theta}.\end{aligned}$$

Let $X^{a,b} := \mathcal{L}^{-1}M^{a,b}$.

We will choose a and b above so that $\{\lambda_k\}_{k \in I}$ lies between Γ_a and Γ_b .

A *cardinal function* is our infinite-dimensional generalization of the characteristic polynomial. From the discussion following the description of the finite-dimensional canonical form in § 1.2 we can determine some properties we would like a cardinal function to have. The functional equation associated with the canonical form will be of the form (1.39), and we would like the generalization of (1.38) to be

$$\mathcal{L}(\eta * y)(\lambda) = p(\lambda)\mathcal{L}y(\lambda) + \varphi(\lambda),$$

where $\varphi \in M^{c,d}$ for some c and d , and φ depends linearly on the initial condition. If $z(t, \cdot)$ solves (1.1), we want to be able to recover $z(t, \cdot)$ from the solution y of the functional equation, so the Laplace transforms of the solutions of $\dot{z}(t) = Az(t)$ and $\eta * y = 0$ should have the same poles. Since

$$\mathcal{L}y(\lambda) = \varphi(\lambda)/p(\lambda) \quad \text{and} \quad \mathcal{L}z(\lambda, \cdot) = \sum_{k \in I} (x_k/(\lambda - \lambda_k))\varphi_k,$$

p should have zeros at $\{\lambda_k\}_{k \in I}$. We would like the solution to the functional equation to be in $X^{a,b}$ whenever u is in $X^{a,b}$. Therefore, we want the map

$$P: M^{a,b} \rightarrow M^{a,b}: \psi \rightarrow \psi/p$$

to be bounded. We want $\eta * y$ to be in X^a whenever y is in $X^{a,b}$, so we also want the map

$$P^{-1}: M^{a,b} \rightarrow M^{a,b}: \psi \rightarrow \psi p$$

to be bounded. This implies that p should be bounded below for $\lambda > \Gamma_a$ and $\lambda < \Gamma_b$, and bounded above for $\lambda > \Gamma_a$. It is easy to see that, in general, there is no entire function with this property with zeros at the sets $\{\lambda_k\}_{k \in I}$ we are considering (see Definition 2.8, condition (6) below). Hence we have to consider functions with poles to the left of Γ_b . Since we want y to be in $M^{a,b}$ whenever u is zero, we want φ to have the same poles as p . Other restrictions on p arise from more careful consideration of the map P . A precise definition of a cardinal function follows.

DEFINITION 2.8. Let p be a meromorphic function with zero set $\{\lambda_k\}_{k \in I}$ and pole set $\{\mu_k\}_{k \in I}$. Assume that all of the λ_k 's and μ_k 's are distinct, and that $I = Z^+ \cup Z^-$. Then p is a cardinal function if

(1) There exist real constants d , c , and m_1 , with $d < c$, such that $|p(\lambda)| \leq m_1$ for all $\lambda \in \{\lambda > \Gamma_c\} \cup \{\lambda < \Gamma_d\}$.

(2) There exist real constants a , b , and m_2 , with $a > b > c$, such that $|p(\lambda)| \geq m_2 > 0$ for all $\lambda \in \{\lambda > \Gamma_a\} \cup \{\lambda < \Gamma_b\}$.

(3) There exists m_3 such that $|p'(\lambda_k)|, |p'(\mu_k)/(p(\mu_k))^2| \geq m_3 > 0$ for all $k \in I$.

(4) There exist paths $\{\Lambda_j\}_{j \in Z}$ between Γ_a and Γ_b such that

(a) There exists m_4 , independent of j , such that $|p(\lambda)| \geq m_4 > 0$ for $\lambda \in \Lambda_j$;

(b) $\{\text{length}(\Lambda_j)\}_{j \in Z}$ is bounded;

(c) $(\inf_{\lambda \in \Lambda_j} \text{Im}(\lambda)) \rightarrow_{j \rightarrow \infty} \infty$ and $(\sup_{\lambda \in \Lambda_j} \text{Im}(\lambda)) \rightarrow_{j \rightarrow -\infty} -\infty$.

- (5) There are paths $\{\tilde{\Lambda}_j\}_{j \in \mathbb{Z}}$ between Γ_c and Γ_d such that
- (a) There exists $m_5 > 0$, independent of j , such that $|p(\lambda)| \leq m_5$ for $\lambda \in \tilde{\Lambda}_j$;
 - (b) $\{\text{length}(\tilde{\Lambda}_j)\}_{j \in \mathbb{Z}}$ is bounded;
 - (c) $(\inf_{\lambda \in \tilde{\Lambda}_j} \text{Im}(\lambda)) \rightarrow_{j \rightarrow \infty} \infty$ and $(\sup_{\lambda \in \tilde{\Lambda}_j} \text{Im}(\lambda)) \rightarrow_{j \rightarrow -\infty} -\infty$.
- (6) $\inf \{\text{Im}(e^{-i\sigma}(\lambda_k - \lambda_{k-1})), \text{Im}(e^{i\sigma}(\lambda_{-k} - \lambda_{-k+1}))\}_{k \geq 1} =: \delta > 0$, $\text{Im}(\lambda_k) \geq 0$ for $k > 0$, and $\text{Im}(\lambda_k) \leq 0$ for $k < 0$.
- (7) $\inf \{\text{Im}(e^{-i\sigma}(\mu_k - \mu_{k-1})), \text{Im}(e^{i\sigma}(\mu_{-k} - \mu_{-k+1}))\}_{k \geq 1} =: \tilde{\delta} > 0$, $\text{Im}(\mu_k) \geq 0$ for $k > 0$, and $\text{Im}(\mu_k) \leq 0$ for $k < 0$.

Remark 1. Conditions (1) and (2) imply that $\{\lambda_k\}_{k \in I}$ lies between Γ_a and Γ_b , and $\{\mu_k\}_{k \in I}$ lies between Γ_c and Γ_d . The spacing conditions (6) and (7) are essential for much of our subsequent work. Two classes of cardinal functions are constructed in [8]. In one class the zeros and poles are spaced asymptotically linearly, and in the other class they grow like k^α for $\alpha > 1$.

Remark 2. It is a consequence of conditions (1) and (2), proved in [10], that there exists M_3 such that $|p'(\lambda_k)|, |p'(\mu_k)/(p(\mu_k))^2| \leq M_3$ for all $k \in I$.

For ease of notation, let

$$(2.5) \quad e_k := \exp(\lambda_k \cdot), \quad f_k := \exp(\mu_k \cdot), \quad \tau_k := 1/(\cdot - \lambda_k), \quad \nu_k := 1/(\cdot - \mu_k).$$

We will need some results from [7] regarding spanning properties of $\{e_k\}_{k \in I}$ in X^a .

DEFINITION 2.9. Let $\{\alpha_k\}_{k \in I}$ be a set of complex numbers between Γ_a and Γ_b . Let

$$X\{\alpha_k\} := \left\{ \sum_{k \in I} x_k \exp(\alpha_k \cdot) \mid \sum_{k \in I} |x_k|^2 < \infty \right\},$$

$$\Psi\{\alpha_k\} := \left\{ \sum_{k \in I} x_k (\cdot - \alpha_k)^{-1} \mid \sum_{k \in I} |x_k|^2 < \infty \right\}$$

and define the operator

$$P: \Psi\{\lambda_k\} \rightarrow \Psi\{\mu_k\}: \psi \rightarrow p\psi.$$

THEOREM 2.10 [7]. *Let $\{\alpha_k\}_{k \in I}$ be the zero set or the pole set of a cardinal function and let $r > a$. Then $X\{\alpha_k\}$ is a Hilbert space with norm $\|\cdot\|_r$, and $\{\exp(\alpha_k \cdot)\}_{k \in I}$ is a Riesz basis for $X\{\alpha_k\}$. $\Psi\{\alpha_k\}$ is a Hilbert space with norm $\|\cdot\|_r$, and $\{(\cdot - \alpha_k)^{-1}\}_{k \in I}$ is a Riesz basis for $\Psi\{\alpha_k\}$.*

Furthermore, P is a Hilbert space isomorphism between $\Psi\{\lambda_k\}$ and $\Psi\{\mu_k\}$.

Let

$$p_k(\lambda) := p(\lambda)/p'(\lambda_k)(\lambda - \lambda_k).$$

Since $\{\tau_k\}_{k \in I}$ is a Riesz basis for $\Psi\{\lambda_k\}$ and P is an isomorphism, $\{P\tau_k\}_{k \in I}$ is a Riesz basis for $\Psi\{\mu_k\}$. Using condition (3) in Definition 2.8 and Remark 2 following Definition 2.8, $\{p_k\}_{k \in I}$ is a Riesz basis for $\Psi\{\mu_k\}$. If we define

$$g_k := \mathcal{L}^{-1} p_k,$$

then $\{g_k\}_{k \in I}$ is a Riesz basis for $X\{\mu_k\}$.

We are now ready to define our equivalence classes in the way suggested at the end of § 1.2. Let

$$X := X^{a,b}, \quad Y := X\{\mu_k\}, \quad M := M^{a,b},$$

$$\Phi := \Psi\{\mu_k\}, \quad X_1 := X\{\lambda_k\}, \quad \Psi_1 := \Psi\{\lambda_k\}.$$

We now define a bilinear form $\langle \cdot, \cdot \rangle$ on $M \otimes \Phi$. Using this the dual of Φ will be represented as Ψ_1 , and the dual of Ψ_1 as Φ . Let $-\Gamma_\rho$ denote Γ_ρ oriented in the opposite direction. For $\varphi \in \Phi$, $\psi \in M$, and $\rho \in (c, b)$,

$$\langle \varphi, \psi \rangle := \left(\frac{1}{2\pi i} \right) \int_{-\Gamma_\rho} \varphi(\lambda) \psi(\lambda) d\lambda,$$

which is easily seen to be independent of $\rho \in (c, b)$. For $f \in X$ and $g \in Y$, define

$$\langle g, f \rangle := \langle \mathcal{L}g, \mathcal{L}f \rangle.$$

We can define equivalence relations on M and X : for $\psi, \chi \in M$, write $\psi \sim \chi$ if $\langle \varphi, \psi \rangle = \langle \varphi, \chi \rangle$ for every $\varphi \in \Phi$. For $f, g \in X$, write $f \sim g$ if $\langle h, f \rangle = \langle h, g \rangle$ for every $h \in Y$. Let E be the set of equivalence classes of X with respect to this equivalence relation. We denote an element of E containing x by $\{x\}$. Let \mathcal{E} be the equivalence classes of M . For $y \in Y$ and $\{x\} \in E$, let $\langle y, \{x\} \rangle := \langle y, x \rangle$. For $\varphi \in \Phi$ and $\{\psi\} \in \mathcal{E}$, let $\langle \varphi, \{\psi\} \rangle := \langle \varphi, \psi \rangle$.

We now show that $\{p_k\}_{k \in I}$ is the Riesz basis of Φ which is dual to $\{\tau_k\}_{k \in I}$, which is a Riesz basis of Ψ_1 :

$$\langle p_k, \tau_j \rangle = \left(\frac{1}{2\pi i} \right) \int_{-\Gamma_r} \{ p(\lambda) / [p'(\lambda_k)(\lambda - \lambda_j)(\lambda - \lambda_k)] \} d\lambda, \quad c < r < b.$$

Because of the conditions on the cardinal function p , this is just the sum of the residues of the integrand to the left of Γ_r . Hence

$$\langle p_k, \tau_j \rangle = \delta_{j,k},$$

and so

$$(2.6) \quad \langle g_k, e_j \rangle = \delta_{j,k}.$$

For $\psi \in M$, define

$$\|\psi\|_\rho^* := \sup_{\varphi \in \Phi} \frac{|\langle \varphi, \psi \rangle|}{\|\varphi\|_\rho}$$

and $\|x\|_\rho^* := \|\mathcal{L}x\|_\rho^*$. It is clear that E is a Hilbert space with the norm $\|\{x\}\|_\rho^* := \|x\|_\rho^*$, because Φ is a Hilbert space. We would like to replace this norm with a more convenient one. For this we need the three following lemmas. The first follows immediately from the Dominated Convergence Theorem.

LEMMA 2.11. *Let $\psi = \sum_{k \in I} x_k \tau_k \in \Psi_1$ and $\varphi = \sum_{k \in I} y_k p_k \in \Phi$. Then*

$$\langle \varphi, \psi \rangle = \sum_{k \in I} x_k y_k. \quad \square$$

The next two lemmas show that E is isomorphic to X_1 .

LEMMA 2.12. *For any $x \in X$,*

$$x \sim x_1 := \sum_{k \in I} \langle g_k, x \rangle e_k.$$

Proof. By Lemma 2.11 and (2.6), if $y \in Y$ and $y = \sum_{k \in I} y_k g_k$,

$$\langle y, x_1 \rangle = \sum_{k \in I} y_k \langle g_k, x \rangle = \langle y, x \rangle. \quad \square$$

Thus we can represent every $\{x\} \in E$ as $\{x_1\}$, where $x_1 \in X_1$.

LEMMA 2.13. *For every $\psi \in \Psi_1$ and $\rho > c$, there exists $m_\rho, M_\rho > 0$ such that*

$$m_\rho \|\psi\|_\rho \leq \|\psi\|_\rho^* \leq M_\rho \|\psi\|_\rho.$$

Proof. Pick $r \in (c, b)$. By Hölder's inequality,

$$\|\psi\|_\rho^* \leq \|\psi\|_r \sup_{\varphi \in \Phi} \{\|\varphi\|_r / \|\varphi\|_\rho\}.$$

Using properties of elements in Ψ^c found in [7], we can show that $\|\psi\|_r \leq 0(1)\|\psi\|_\rho$, which shows the first inequality in the lemma. Let $\tilde{\varphi} = \sum_{k \in I} \tilde{x}_k p_k$. Then, using Theorem 2.10, there exists c_ρ and C_ρ such that

$$\|\psi\|_\rho^* \geq \langle \tilde{\varphi}, \psi \rangle / \|\tilde{\varphi}\|_\rho \geq c_\rho \left(\sum_{k \in I} |x_k|^2 \right)^{1/2} \geq C_\rho \|\psi\|_\rho;$$

this finishes the proof of the lemma. \square

We see from these lemmas that the map

$$T: X_1 \rightarrow E: x_1 \rightarrow \{x_1\}$$

is a Hilbert space isomorphism. The dual of E can be represented by Y , and $\{\{e_k\}\}_{k \in I}$ is a Riesz basis for E , with dual basis $\{g_k\}_{k \in I}$. Lemma 2.13 and Theorem 2.10 show that we can put the following more convenient norm on E . Let $x_1 = \sum_{k \in I} x_k e_k$, and define

$$\|\{x_1\}\| := \left(\sum_{k \in I} |x_k|^2 \right)^{1/2}.$$

2.3. The canonical state space equation. The semigroup generator for the canonical form will be a “differentiation” operator D on E , which is defined on an appropriate subspace of “differentiable” elements of E . The eigenvectors of D will be $\{\{e_k\}\}_{k \in I}$. Let

$$\tilde{X}^a := \{f \in X^a \mid f \text{ is absolutely continuous on } \tilde{\Omega} \text{ and } f' \in X^a\},$$

and let $\tilde{\Psi}^a := \mathcal{L}\tilde{X}^a$.

Let $\alpha \notin \{\lambda_k\}$ and $\Gamma_a > \alpha > \Gamma_b$. If $f \in \tilde{X}^a$, then, if we differentiate by parts, $\mathcal{L}f$ is of the form

$$(2.7) \quad \psi(\lambda) = (x + \tilde{\psi}(\lambda)) / (\lambda - \alpha), \quad x \in \mathbb{C}, \quad \tilde{\psi} \in \tilde{\Psi}^a,$$

where $x = f(0)$ and $\tilde{\psi} = \mathcal{L}(-\alpha f + f')$. Let

$$\tilde{M} := \{(x_0 + \tilde{\psi}) / (\cdot - \alpha) \mid x_0 \in \mathbb{C}, \tilde{\psi} \in M\},$$

and let $\tilde{X} := \mathcal{L}^{-1}\tilde{M}$. Let $\tilde{\mathcal{E}}$ be the set of all equivalence classes of \tilde{M} with respect to the relation \sim , and let \tilde{E} be the set of equivalence classes of \tilde{X} .

LEMMA 2.14. *Let $\psi_1(\lambda) = (x_1 + \tilde{\psi}_1(\lambda)) / (\lambda - \alpha)$ and $\psi_2(\lambda) = (x_2 + \tilde{\psi}_2(\lambda)) / (\lambda - \alpha)$, where $\tilde{\psi}_1, \tilde{\psi}_2 \in \tilde{M}$. If $\psi_1 \sim \psi_2$, then $x_1 = x_2$ and $\tilde{\psi}_1 \sim \tilde{\psi}_2$.*

Proof. Suppose that $\psi / (\cdot - \alpha) \sim c / (\cdot - \alpha)$ for some $\tilde{\psi} \in M$ and $c \in \mathbb{C}$. Then $\langle \nu_k, \psi / (\cdot - \alpha) \rangle = \langle \nu_k, c / (\cdot - \alpha) \rangle$ for all $k \in I$, which means that $c / (\mu_k - \alpha) = \tilde{\psi}(\mu_k) / (\mu_k - \alpha)$. However, if c is not zero then this is impossible, since it is shown in [7] that $\sum_{k \in I} |\tilde{\psi}(\mu_k)|^2 < \infty$.

Since $\psi_1 \sim \psi_2$, $(\tilde{\psi}_1 - \tilde{\psi}_2) / (\cdot - \alpha) \sim (x_1 - x_2) / (\cdot - \alpha)$. The above observations imply that $x_1 - x_2 = 0$. Therefore $(\tilde{\psi}_1(\mu_k) - \tilde{\psi}_2(\mu_k)) / (\mu_k - \alpha) = 0$ for all $k \in I$. This implies that $\langle \nu_k, \tilde{\psi}_1 \rangle = \langle \nu_k, \tilde{\psi}_2 \rangle$ for all $k \in I$. Since $\{\nu_k\}_{k \in I}$ is a Riesz basis for Φ , $\tilde{\psi}_1 \sim \tilde{\psi}_2$. \square

COROLLARY 2.15. *If $f_1 \sim f_2$, and $f_1, f_2 \in \tilde{X}$, then $f'_1 \sim f'_2$.*

Proof. By (2.7) and Lemma 2.14, we see that $\mathcal{L}(-\alpha f_1 + f'_1) \sim \mathcal{L}(-\alpha f_2 + f'_2)$, so $f'_1 \sim f'_2$. \square

This corollary allows us to define differentiation on \tilde{E} as $D\{f\} := \{f'\}$. If we restrict D to

$$(2.8) \quad \mathcal{D}(D) = \left\{ \sum_{k \in I} x_k \{e_k\} \mid \sum_{k \in I} |x_k \lambda_k|^2 < \infty \right\},$$

then D is the infinitesimal generator of the strongly continuous semigroup

$$S(t) \left(\sum_{k \in I} x_k \{e_k\} \right) = \sum_{k \in I} x_k e^{\lambda_k t} \{e_k\}.$$

It is of interest to us to characterize how the semigroup behaves in the frequency domain, since the equivalence of the canonical state space equation and the canonical functional equation is shown in the frequency domain. Let $r > a$ and $\rho < b$, and let $\Gamma_{r,\rho} := \Gamma_r \cup -\Gamma_\rho$. For $f \in L^2[\Gamma_{r,\rho}]$, define

$$(T_{r,\rho}f)(\lambda) := \left(\frac{1}{2\pi i} \right) \int_{\Gamma_{r,\rho}} [f(\zeta)/(\lambda - \zeta)] d\zeta,$$

for $\lambda > \Gamma_r$ and $\lambda < \Gamma_\rho$. We will restrict attention to how $T_{r,\rho}$ behaves on functions in

$$\mathcal{G} := \{e^{-t}\psi \mid \psi \in \Psi_1, t \in \tilde{\Omega}\} \cup \{q\psi \mid q \text{ is a cardinal function and } \psi \in M\}.$$

For such functions it is easy to see that $(T_{r,\rho}f)(\lambda)$ is independent of r and ρ , as long as z is to the left of both Γ_r and Γ_ρ or to the right of both Γ_r and Γ_ρ . Thus, we can unambiguously define $Tf(\lambda) = (T_{r,\rho}f)(\lambda)$ for any $\lambda > \Gamma_a$ or $\lambda < \Gamma_b$ by the appropriate choice of $r > a$ and $\rho < b$. Next we show T is a projection into M .

LEMMA 2.16. $Tf \in M$ whenever $f \in \mathcal{G}$.

Proof. For $r > a, \rho < b$ and $\lambda > \Gamma_r$ or $\lambda < \Gamma_b$.

$$\begin{aligned} Tf(\lambda) &= (1/2\pi i) \left(\int_{\Gamma_r^+} + \int_{\Gamma_r^2} + \int_{-\Gamma_\rho^1} + \int_{-\Gamma_\rho^2} \right) [f(\zeta)/(\lambda - \zeta)] d\zeta \\ (2.9) \quad &:= F_1(\lambda) + F_2(\lambda) + F_3(\lambda) + F_4(\lambda) \end{aligned}$$

(cf. Definition 2.1). We can use the Carleson measure theorem [5], as in [7], to show that there exists $C(r, \rho)$ such that $\|F_i\|_s \leq C$ for $i = 1, 2, 3$, or 4 and $s > r$ or $s < \rho$ whenever $f \in L^2[\Gamma_r]$. Hence $\|Tf\|_s$ is bounded for $s > r$ or $s < \rho$. Using Hölder's inequality and (2.9), we see that

$$|F_1(\lambda)| \leq (1/\pi) \|f\|_r^2 \left\{ \int_0^\infty |dt/(\lambda - i e^{i\sigma} t - \rho)|^2 \right\}^{1/2}.$$

The term in brackets is a uniformly bounded function of λ in $\{\lambda < \Gamma_\eta\} \cup \{\lambda > \Gamma_s\}$ whenever $s > r$ and $\eta < \rho$. For $\lambda < \Gamma_{\eta,\theta}$ or $\lambda > \Gamma_{s,\omega}$, where $\theta \in (\sigma, \pi/2)$ and $\omega \in (0, \sigma)$, the term in brackets is $O(|\lambda|^{-1})$. Similar estimates can be made for F_2, F_3 , and F_4 , showing that $Tf \in M$. \square

We can use this projection to characterize the action of the semigroup in the frequency domain.

THEOREM 2.17. Let $\psi(\lambda, t) := T(e^{-t}\psi)(\lambda)$ for $\psi \in M$. If $\mathcal{L}^{-1}\psi = y$, then $\mathcal{L}^{-1}\psi(\lambda, t) = y(t + \cdot)$.

Proof. Let $s > r > a, \rho < b$, and $\tau \in \tilde{\Omega}$. Then

$$\begin{aligned} \mathcal{L}^{-1}(\psi(\cdot, t))(\tau) &= \left(-\frac{1}{4\pi^2} \right) \int_{\Gamma_s} e^{\lambda\tau} \int_{\Gamma_{r,\rho}} [e^{\zeta t}\psi(\zeta)/(\lambda - \zeta)] d\zeta d\lambda \\ &= \left(-\frac{1}{4\pi^2} \right) \int_{\Gamma_{r,\rho}} e^{\zeta t}\psi(\zeta) \int_{\Gamma_s} [e^{\lambda\tau}/(\lambda - \zeta)] d\lambda d\zeta \\ &= \left(\frac{1}{2\pi i} \right) \int_{\Gamma_{r,\rho}} e^{\zeta\tau}\psi(\zeta) e^{\zeta t} d\zeta. \end{aligned}$$

By the properties of the space Ψ^a found in [7], we can show that $(1/2\pi i) \int_{\Gamma_\rho} e^{\xi(t+\tau)} \psi(\xi) d\xi$ is equal to the sum of the residues of its integrand to the left of Γ_ρ , which is zero. Hence

$$\mathcal{L}^{-1}(\psi(\cdot, t))(\tau) = (1/2\pi i) \int_{\Gamma_\rho} e^{\xi(t+\tau)} \psi(\xi) d\xi = y(t+\tau). \quad \square$$

If $x \in X_1$, $S(t)\{x(t)\} = \{x(t+\cdot)\}$, so this theorem shows that

$$\{\mathcal{L}S(t)\{x\}\} = \{T(e^{\lambda \cdot}(\mathcal{L}x))\}.$$

This will be helpful when we show the equivalence of the canonical state space equation and the canonical functional equation.

The canonical input element will be

$$\hat{b} := \sum_{k \in I} (1/p'(\lambda_k)) \{e_k\}.$$

LEMMA 2.18. \hat{b} is an admissible input element, as defined in Definition 1.1.

Proof. The coefficients $(1/p'(\lambda_k))$ are bounded because of condition (3) in Definition 2.8. Using this, and the restrictions on $\{\lambda_k\}_{k \in I}$ given by condition (6) in Definition 2.8, we can use the Carleson Measure Theorem as in [5] to show that \hat{b} is an admissible input element. \square

The canonical state space equation is then

$$(2.10) \quad \{\dot{x}(t)\} = D\{x(t)\} + \hat{b}u(t), \quad \{x(0)\} = \{x_0\}.$$

2.4. The canonical functional equation. We will start this section by defining a class of functionals on \tilde{X} that are represented by cardinal functions in the frequency domain. Fix $\alpha \notin \{\lambda_k\}_{k \in I}$ between Γ_b and Γ_a , and let C be a small circle around α , not surrounding any λ_k , oriented counterclockwise. Then, for $\rho \in (c, b)$ and $\psi(\lambda) = (x + \tilde{\psi}(\lambda))/(\lambda - \alpha) \in \tilde{M}$, define

$$(2.11) \quad (p, \psi) := \left(\frac{1}{2\pi i} \right) \left\{ \int_{-\Gamma_\rho} [p(\lambda) \tilde{\psi}(\lambda)/(\lambda - \alpha)] d\lambda + \int_C [p(\lambda)x/(\lambda - \alpha)] d\lambda \right\}.$$

It is easy to see that (p, ψ) is independent of $\rho \in (c, b)$. In the remainder of this section assume that $\rho \in (c, b)$.

LEMMA 2.19. If ψ and $\chi \in \tilde{M}$ and $\psi \sim \chi$, then $(p, \psi) = (p, \chi)$.

Proof. Let $\psi(\lambda) = (x + \tilde{\psi}(\lambda))/(\lambda - \alpha)$ and $\chi(\lambda) = (y + \tilde{\chi}(\lambda))/(\lambda - \alpha)$. By Lemma 2.14, $x = y$ and $\tilde{\psi} \sim \tilde{\chi}$. Let $\varphi(\lambda) = (p(\lambda) - p(\alpha))/(\lambda - \alpha)$. It is readily seen that $\varphi \in \Phi$. Hence

$$\begin{aligned} (p, \psi) - (p, \chi) &= (1/2\pi i) \int_{-\Gamma_\rho} [\varphi(\lambda) + p(\alpha)/(\lambda - \alpha)] (\tilde{\psi}(\lambda) - \tilde{\chi}(\lambda)) d\lambda \\ &= \int_{-\Gamma_\rho} [p(\alpha)/(\lambda - \alpha)] (\tilde{\psi}(\lambda) - \tilde{\chi}(\lambda)) d\lambda, \end{aligned}$$

by the equivalence of $\tilde{\psi}$ and $\tilde{\chi}$. Because of the decay properties of $\tilde{\psi}$ and $\tilde{\chi}$, the last integral is equal to the sum of the residues of the integrand to the right of Γ_ρ . Since the integrand has no poles to the right of Γ_ρ , this is zero, proving the lemma. \square

We can now define $(p, \{\psi\}) := (p, \psi)$ unambiguously for $\{\psi\} \in \tilde{\mathcal{E}}$. For $x \in \tilde{X}$ and $\{x\} \in \tilde{E}$ let

$$(2.12) \quad (\eta, x) := (p, \mathcal{L}x), \quad (\eta, \{x\}) := (\eta, x).$$

The cardinal function p is seen to be the Fourier transform of η in the sense that $\mathcal{F}\eta(\lambda) := (\eta, e^{\lambda \cdot}) = p(\lambda)$ whenever $\Gamma_b < \lambda < \Gamma_a$: This is because $(\mathcal{L}(e^{\lambda \cdot}))(\zeta) = (\zeta - \lambda)^{-1} = (1 + \tilde{\psi}(\zeta))/(\zeta - \alpha)$, where $\tilde{\psi}(\zeta) = (\lambda - \alpha)/(\zeta - \lambda)$, so

$$(\eta, e^{\lambda \cdot}) = (1/2\pi i) \left\{ \int_{-\Gamma_p} [p(\zeta)\tilde{\psi}(\zeta)/(\zeta - \alpha)] d\zeta + \int_C [p(\zeta)/(\zeta - \alpha)] d\zeta \right\},$$

which we find by calculating residues to be $p(\lambda) - p(\alpha) + p(\alpha) = p(\lambda)$. If q is a cardinal function, $\mathcal{F}^{-1}q$ is defined as the functional μ such that $(\mu, x) := (q, \mathcal{L}x)$ for $x \in \tilde{X}$.

In analogy to the two-sided transform theory developed in Russell [12], we can characterize $\mathcal{D}(D)$ in terms of the “generating functional” η .

THEOREM 2.20. $\mathcal{D}(D) = \{\{x\} \in \tilde{E} \mid (\eta, \{x\}) = 0\}$.

Because this result is not used in the rest of this paper, and because the proof is quite lengthy, we will refer the reader to the author’s thesis [9] for the proof.

Remark. If the poles $\{\mu_k\}_{k \in I}$ are fixed, then the space E is fixed. For every cardinal function with poles at $\{\mu_k\}_{k \in I}$ there is an associated differentiation operator D with $\mathcal{D}(D)$ given by (2.8). For this cardinal function there is also a generating functional defined by (2.12). Theorem 2.20 shows that the associated operator and functional are very closely related.

An interesting fact that comes from the proof of Theorem 2.20 is that every $\{x\} \in \tilde{E}$ can be expanded as

$$(2.13) \quad \{x\} = y_0\{e^{\alpha \cdot}\} + \sum_{k \in I} y_k\{e_k\},$$

for some sequence $\{y_k\}_{k \in I}$ such that $\sum_{k \in I} |y_k \lambda_k|^2 < \infty$. Hence

$$\langle \eta, \{x\} \rangle = y_0 p(\alpha).$$

Equation (2.13) is a generalization of the fact that a Riesz basis of exponentials of $L^2[-1, 1]$ can be made into a Riesz basis of $H^1[-1, 1]$ by appending another exponential [12].

Our next step is to define a convolution related to a cardinal function p . In § 2.3, we defined a projection T into M on a class of functions \mathcal{G} that includes functions of the form $q\psi$, where q is a cardinal function and $\psi \in M$. We will use this projection to define our convolution. Let

$$p * \psi := T(p\psi),$$

and for $y \in X$, let

$$\eta * y = \mathcal{L}^{-1}(p * \mathcal{L}y).$$

We can show that this corresponds to a usual notion of convolution when y is differentiable, i.e., when $y \in \tilde{X}$.

LEMMA 2.21. *If $y \in \tilde{X}$ and $t \in \tilde{\Omega}$, then $(\eta * y)(t) = (\eta, y(t + \cdot))$.*

This is another result that will not be used in the sequel and has a somewhat lengthy proof, so the reader will again be referred to the author’s thesis [9] for the proof.

The canonical functional equation will be

$$(2.14) \quad (\eta * y)(t) = u(t), \quad y \sim x_0,$$

which we wish to solve for x_0 , u , and y in X . The result in Lemma (2.21) could be used to help solve this when $x_0 \in \tilde{X}$, but we get more information by solving (2.14) in the frequency domain.

We need to develop some machinery to help solve this functional equation in the frequency domain before we can show the equivalence of the canonical functional equation and the canonical state space equation. This next result is an infinite-dimensional generalization of (1.38).

THEOREM 2.22. For $\mathcal{L}y = \psi \in M$, $\mathcal{L}(\eta * y) = p\psi + \chi$, where $\chi \in \Phi$.

Proof. For $\psi \in M$, $\rho \in (c, b)$, and $\lambda \in \Omega_a$, let

$$R(p\psi)(\lambda) := (1/2\pi i) \int_{\Gamma_\rho} [\psi(\zeta)p(\zeta)/(\lambda - \zeta)] d\zeta.$$

It is easy to see that this is independent of ρ as long as $\lambda > \Gamma_\rho$, which defines $R(p\psi)$ unambiguously for $\lambda \in \Omega_c$. Using the properties of $\psi \in M$ and of the cardinal function p , we see that $R(p\psi)(\lambda)$ is equal to the sum of the residues of $\psi p/(\lambda - \cdot)$ to the left of Γ_c . Thus

$$R(p\psi)(\lambda) = \sum_{k \in I} [\psi(\mu_k)p^2(\mu_k)/p'(\mu_k)]v_k.$$

Using condition (3) in Definition 2.8 and results in [7], we see that $R(p\psi) \in \Phi$.

Let $r > a$ and $c < \rho < b$. Since $p\psi \in \Psi^a$, it is easy to show that

$$(1/2\pi i) \int_{\Gamma_r} [\psi(\zeta)p(\zeta)/(\lambda - \zeta)] d\zeta = \psi(\lambda)p(\lambda) \quad \text{for } \lambda > \Gamma_r.$$

Then, using $\eta^*y = \mathcal{L}^{-1}(T(p\psi))$

$$\begin{aligned} (2.15) \quad \mathcal{L}(\eta * x)(\lambda) &= \int_{\Gamma_r} [\psi(\zeta)p(\zeta)/(\lambda - \zeta)] d\zeta - \int_{\Gamma_\rho} [\psi(\zeta)p(\zeta)/(\lambda - \zeta)] d\zeta \\ &= p\psi(\lambda) - R(p\psi)(\lambda). \end{aligned}$$

Since the three functions above are all holomorphic in any region not intersecting $\{\Gamma_b < \lambda < \Gamma_a\} \cup \{\mu_k\}_{k \in I}$, we can extend (2.15) to any such region. Since $R(p\psi) \in \Phi$, (2.15) proves the theorem. \square

Theorem 2.22 will help us in solving the canonical functional equation as follows: if $\varphi = -R(p\psi)$, $\eta * y = u$ implies that $p\psi + \varphi = \mathcal{L}u$, which means that

$$(2.16) \quad \psi = \chi + (\mathcal{L}u/p), \quad \chi \in \Psi_1,$$

where $\chi = \varphi/p \in \Psi_1$ by Theorem 2.10.

LEMMA 2.23. Any ψ of the form (2.16) solves $p * \psi = \mathcal{L}u$.

Proof. Let $\varphi = p\chi \in \Phi$. If $\lambda > \Gamma_r$ or $\lambda < \Gamma_\rho$, then

$$(p * \psi)(\lambda) = (1/2\pi i) \int_{\Gamma_{r,\rho}} [\mathcal{L}u(\zeta)/(\lambda - \zeta)] d\zeta + (1/2\pi i) \int_{\Gamma_{r,\rho}} [\varphi(\zeta)/(\lambda - \zeta)] d\zeta.$$

The second integral can be seen to be the sum of the residues of its integrand between Γ_r and Γ_ρ , which is zero. The first integral is the sum of the residues of its integrand in the region $\{\lambda > \Gamma_r\} \cup \{\lambda < \Gamma_\rho\}$, which yields $\mathcal{L}u(\lambda)$, proving the lemma. \square

We see that to solve (2.14), we let $y = \mathcal{L}^{-1}\psi$, where ψ is given by (2.16), and we choose $\chi \in \Psi_1$ to satisfy the “initial condition” $y \sim x_0$. This is very easy to do, because of the following result.

LEMMA 2.24. For $\psi \in M$, $\psi/p \sim 0$.

Proof. $\langle p_k, \psi/p \rangle = (1/2\pi i) \int_{-\Gamma_\rho} [\psi(\lambda)/p'(\lambda_k)(\lambda - \lambda_k)] d\lambda$. It is easily seen that this is the sum of the residues of the integrand to the left of Γ_ρ , which is zero. Since this is true for every $k \in I$, and $\{p_k\}_{k \in I}$ is a Riesz basis for Φ , $\psi/p \sim 0$. \square

Thus, if we want $(\chi + \mathcal{L}u/p) \sim \mathcal{L}x_0$, we simply choose $\chi \sim \mathcal{L}x_0$, so

$$(2.17) \quad \chi = \sum_{k \in I} \langle p_k, y_0 \rangle \tau_k.$$

To summarize, we get the following result.

THEOREM 2.25. *The solution of the canonical functional equation (2.14) is given by $y = \mathcal{L}^{-1}\psi$, where $\psi = \chi + \mathcal{L}u/p$ and χ is given by (2.17).*

2.5. Equivalence of state space and functional equations. We are now ready to show the equivalence of the canonical state space equation and the canonical functional equation, using the frequency domain structure developed above.

THEOREM 2.26. *If $\{x(t)\}$ solves (2.10), and y solves (2.14), then $\{x(t)\} = \{y(t + \cdot)\}$.*

Proof. Let $x_0 = \sum_{k \in I} x_k \{e_k\}$. By the variation of parameters formula,

$$(2.18) \quad \{x(t)\} = \sum_{k \in I} \{e_k\} \left(x_k e^{\lambda_k t} + (1/p'(\lambda_k)) \int_{[0, t]} e^{\lambda_k(t-s)} u(s) ds \right).$$

Let $r > a$ and $\rho < b$. From Theorem 2.25, $\mathcal{L}y = (\mathcal{L}u/p) + \chi = (\mathcal{L}u/p) + \sum_{k \in I} x_k \tau_k$. By Theorem 2.17,

$$\begin{aligned} \mathcal{L}(y(t + \cdot)) &= (1/2\pi i) \int_{\Gamma_{r,\rho}} [e^{\xi t} \mathcal{L}y(\zeta)/(\lambda - \zeta)] d\zeta \\ &= (1/2\pi i) \int_{\Gamma_{r,\rho}} [e^{\xi t} \chi(\zeta)/(\lambda - \zeta)] d\zeta \\ &\quad + (1/2\pi i) \int_{\Gamma_{r,\rho}} [e^{\xi t} \mathcal{L}u(\zeta)/p(\zeta)(\lambda - \zeta)] d\zeta \\ &=: \chi(\lambda, t) + \varphi(\lambda, t). \end{aligned}$$

$\chi(\lambda, t)$ is easily seen to be the sum of the residues of the integrand between Γ_r and Γ_ρ , which is $\sum_{k \in I} x_k e^{\lambda_k t} \tau_k$. $\varphi(\lambda, t) \sim \sum_{k \in I} y_k(t) \tau_k$, where

$$(2.19) \quad y_k(t) = \langle p_k, \varphi(\cdot, t) \rangle.$$

Hence $\{y(t + \cdot)\} = \sum_{k \in I} (x_k e^{\lambda_k t} + y_k(t)) \{e_k\}$. Comparing this with (2.18), we see that the theorem is true if the following claim is true.

CLAIM 2.27. $y_k(t) = (1/p'(\lambda_k)) \int_{[0, t]} e^{\lambda_k(t-s)} u(s) ds$.

Proof of Claim 2.27. Let $\sigma < \rho$. Then, from (2.19),

$$\begin{aligned} y_k(t) &= \left(-\frac{1}{4\pi^2} \right) \int_{-\Gamma_\sigma} [p(\lambda)/p'(\lambda_k)(\lambda - \lambda_k)] \int_{\Gamma_{r,\rho}} [e^{\xi t} \mathcal{L}u(\zeta)/p(\zeta)(\lambda - \zeta)] d\zeta d\lambda \\ &= \left(\frac{1}{2\pi i p'(\lambda_k)} \right) \int_{\Gamma_{r,\rho}} [e^{\xi t} \mathcal{L}u(\zeta)/(\zeta - \lambda_k)] d\zeta. \end{aligned}$$

For $\lambda \in S_\theta$ (cf. Definition 2.4),

$$\begin{aligned} (\mathcal{L}y_k(t))(\lambda) &= \left(\frac{1}{2\pi i p'(\lambda_k)} \right) \int_{\ell_\theta} e^{-\lambda t} \int_{\Gamma_{r,\rho}} [\mathcal{L}u(\zeta) e^{\xi t}/(\zeta - \lambda_k)] d\zeta dt \\ (2.20) \quad &= \left(\frac{1}{2\pi i p'(\lambda_k)} \right) \int_{\Gamma_{r,\rho}} [\mathcal{L}u(\zeta)/(\zeta - \lambda_k)(\lambda - \zeta)] d\zeta \\ &= \mathcal{L}u(\lambda)/[p'(\lambda_k)(\lambda - \lambda_k)]. \end{aligned}$$

On the other hand, using a result in [7],

$$(2.21) \quad \mathcal{L} \left[(1/p'(\lambda_k)) \int_{[0,t]} e^{\lambda_k(t-s)} u(s) ds \right] = \mathcal{L}u(\lambda) / [p'(\lambda_k)(\lambda - \lambda_k)].$$

Comparing (2.20) with (2.21), we see that Claim 2.27 is true, which completes the theorem. \square

Hence, the functional equation is equivalent to the state space equation. Furthermore, it gives us information about the evolution of the solution of the state space equation which we did not previously have. In the next section, we apply these results to the solution of an eigenvalue specification problem, and use the canonical form to analyze the effect of a finite-dimensional control scheme on the distributed parameter system.

3. Application to control.

3.1. Application to spectral determination. In this section we will consider the eigenvalue specification problem discussed in § 1.

In [11], Russell uses a canonical form for a class of hyperbolic systems to show that Sun's condition (1.24) is sufficient for solving the eigenvalue specification problem. We will do the same for the systems under consideration in this paper. We will also consider a case where the input element is not bounded but is admissible. One of the benefits of a canonical form approach is that we obtain formulas for the feedback element and the closed-loop eigenvectors and dual basis.

We will first prove the result for the canonical state space equation.

THEOREM 3.1. *If $\{\alpha_k\}_{k \in I}$ is the zero set of a cardinal function q with poles at $\{\mu_k\}_{k \in I}$, the pole set of p , then there exists a feedback element d^* such that $D + \hat{b}d^*$ (interpreted as in (1.25) and (1.26)) has eigenvalues at $\{\alpha_k\}_{k \in I}$ and eigenvectors $\{\{e^{\alpha_k}\}\}_{k \in I}$ which form a Riesz basis for E . If*

$$(3.1) \quad \sum_{k \in I} |\alpha_k - \lambda_k|^2 < \infty,$$

then $d^*x = \langle d, x \rangle$ for some $d \in Y$.

Proof. Let $q_k(\lambda) := q(\lambda) / [q'(\alpha_k)(\lambda - \alpha_k)]$, and $g_k^\alpha := \mathcal{L}^{-1}q_k$. Since q is a cardinal function, we can go through the same procedure with q as we did with p . Hence $\{\{e^{\alpha_k}\}\}_{k \in I}$ is a Riesz basis for E , $\{((\cdot - \alpha_k)^{-1})\}_{k \in I}$ is a Riesz basis for \mathcal{E} , $\{q_k\}_{k \in I}$ is a Riesz basis for Φ , and $\{g_k^\alpha\}_{k \in I}$ is a Riesz basis for Y . $\{g_k^\alpha\}_{k \in I}$ is dual to $\{\{e^{\alpha_k}\}\}_{k \in I}$, and $\{q_k\}_{k \in I}$ is dual to $\{((\cdot - \alpha_k)^{-1})\}_{k \in I}$. Let $\mu = \mathcal{F}^{-1}q$.

Let $u(t) = (\eta * y)(t)$, where y solves

$$(3.2) \quad \mu * y = 0, \quad y \sim x_0$$

and let $\{x(t)\} := \{y(t + \cdot)\}$. By Theorem 2.26, $\{x(t)\}$ solves

$$(3.3) \quad \{\dot{x}(t)\} = \tilde{D}\{x(t)\}, \quad \{x(0)\} = \{x_0\},$$

where \tilde{D} is the differentiation operator on

$$\mathcal{D}(\tilde{D}) = \left\{ \sum_{k \in I} x_k \{e^{\alpha_k}\} \mid \sum_{k \in I} |x_k \alpha_k|^2 < \infty \right\}.$$

y also satisfies (2.14), so $\{x(t)\}$ also satisfies

$$(3.4) \quad \{\dot{x}(t)\} = D\{x(t)\} + \hat{b}u(t), \quad \{x(0)\} = \{x_0\}.$$

We now show that $u(t)$ is of the form $d^*\{x(t)\}$.

Since y solves (3.3), $y(t) = \sum_{k \in I} \langle g_k^\alpha, x_0 \rangle e^{\alpha_k t}$. Hence

$$(3.5) \quad u(t) = \sum_{k \in I} \langle g_k^\alpha, x_0 \rangle p(\alpha_k) e^{\alpha_k t},$$

$$(3.6) \quad \{x(t)\} = \sum_{k \in I} \langle g_k^\alpha, x_0 \rangle p(\alpha_k) e^{\alpha_k t} \{e^{\alpha_k t}\}.$$

Note that $\{p(\alpha_k)\}_{k \in I}$ is a bounded sequence, by condition (1) of Definition 2.8. Let d^* be the input element represented by

$$(3.7) \quad d := \sum_{k \in I} p(\alpha_k) g_k^\alpha$$

in the sense that, for $\{x\} \in E$,

$$d^*\{x\} := \sum_{k \in I} p(\alpha_k) \langle g_k^\alpha, \{x\} \rangle$$

and let $\mathcal{D}(d^*) = \{\sum_{k \in I} x_k \{e^{\alpha_k t}\} \mid \sum_{k \in I} |p(\alpha_k) x_k| < \infty\}$. Then (3.6) and (3.5) imply that

$$u(t) = d^*\{x(t)\},$$

so (3.4) becomes

$$(3.8) \quad \{\dot{x}(t)\} = (D + \hat{b}d^*)\{x(t)\}, \quad \{x(0)\} = \{x_0\}.$$

Using (3.3) and (3.8), it is easy to show that $D + \hat{b}d^*$ has eigenvalues $\{\alpha_k\}_{k \in I}$ and associated eigenvectors $\{e^{\alpha_k t}\}_{k \in I}$, which form a Riesz basis for E , and that $D + \hat{b}d^*$ is in fact \tilde{D} . If $\sum_{k \in I} |p(\alpha_k)|^2 < \infty$, then d is an element of Y . Therefore the theorem is true once we prove the following result.

LEMMA 3.2. *If $\{\alpha_k\}_{k \in I}$ satisfies (3.1), then $\sum_{k \in I} |p(\alpha_k)|^2 < \infty$.*

Proof. Let $r_1 > 0$ be such that every λ_k is at least a distance of r_1 from Γ_c , and let $0 < r_2 < r_1$. Let C_k be a circle of radius r_1 centered at λ_k . For k large enough, α_k belongs to a circle of radius r_2 centered at λ_k . For such k ,

$$\begin{aligned} |p(\alpha_k) - p(\lambda_k)| &= \left(\frac{1}{2\pi} \right) \left| \int_{C_k} [p(\zeta)/(\alpha_k - \zeta)] d\zeta - \int_{C_k} [p(\zeta)/(\lambda_k - \zeta)] d\zeta \right| \\ &\leq \left(\frac{1}{2\pi} \right) |\lambda_k - \alpha_k| \left\{ \int_{C_k} |p(\zeta)/(\lambda_k - \zeta)(\alpha_k - \zeta)| |d\zeta| \right\}. \end{aligned}$$

Since $|\alpha_k - \zeta| > r_1 - r_2$ and $|\lambda_k - \zeta| = r_1$ whenever $\zeta \in C_k$, and p is uniformly bounded on C_k , we see that the term in brackets is uniformly bounded. Therefore,

$$(3.9) \quad |p(\alpha_k)| \leq 0(1) |\lambda_k - \alpha_k|.$$

When we compare this with (3.1), the proofs of Lemma 3.2 and Theorem 3.1 are completed. \square

We are now in a position to prove Theorem 1.3.

Proof of Theorem 1.3. To translate the result to the original state space equation (1.1), we need a mapping between H and E , which takes (2.10) to (1.1). Let

$$T_0: E \rightarrow H: \{e_k\} \rightarrow \varphi_k,$$

$$T_1: E \rightarrow E: \{e_k\} \rightarrow \{e_k\} p'(\lambda_k) b_k$$

and let $T = T_0 T_1$, so $T^{-1} A T = D$.

First suppose condition (1) of Definition 2.8 holds, so T is an isomorphism between E and H . T can be extended to an isomorphism between $\mathcal{H} = \mathcal{D}(A')'$ and $\mathcal{D}(D')'$ and $b = T\hat{b}$, where b and \hat{b} are admissible input elements. The transformation $\{x(t)\} = T^{-1}z(t)$ then takes the original state space equation (1.1) to the canonical state space equation (2.10).

Let $h^* = \sum_{k \in I} p(\alpha_k)(T^{-1})^* g_k^\alpha = (T^{-1})^* d$, where d^* is given by (3.7). Then

$$(3.10) \quad T^{-1}(A + bh^*)T = D + \hat{b}d^*.$$

This implies that $A + bh^*$ has the same eigenvalues as $D + \hat{b}d^*$, and that $\{T\{e^{\alpha_k}\}\}_{k \in I}$ is a Riesz basis for H .

Now suppose condition (2) of Definition 2.8 holds. T is not an isomorphism, but we would still like to claim that $A + bh^*$ has the same eigenvalues as $D + \hat{b}d^*$. We follow an argument given by Russell in [11].

The element d given by (3.7) is an element of Y , because (1.24) implies (3.1), so it can be written as $d = \sum_{k \in I} d_k g_k$, where $\sum_{k \in I} |d_k|^2 < \infty$. We need more information about the coefficients $\{d_k\}_{k \in I}$.

LEMMA 3.3. *If $x = \sum_{k \in I} x_k g_k = \sum_{k \in I} y_k g_k^\alpha$, then there exists $M(x)$ such that $|x_k - y_k| < M|\lambda_k - \alpha_k|$ for all $k \in I$.*

Proof. For $\rho \in (c, b)$,

$$\begin{aligned} |x_k - y_k| &= |\langle x, e_k \rangle - \langle x, e^{\alpha_k} \rangle| \\ &= (1/2\pi) \left| \int_{-\Gamma_\rho} \mathcal{L}x(\lambda) ((\lambda - \lambda_k)^{-1} - (\lambda - \alpha_k)^{-1}) d\lambda \right| \\ &\leq (1/2\pi) \|\mathcal{L}x\|_\rho |\lambda_k - \alpha_k| \left\{ \int_{\Gamma_\rho} |1/(\lambda - \lambda_k)(\lambda - \alpha_k)|^2 d\lambda \right\}^{1/2}. \end{aligned}$$

The term in brackets is uniformly bounded for all $k \in I$, so the lemma is true. \square

From (1.24) and (3.9) we see that $\sum_{k \in I} |p(\alpha_k)/b_k|^2 < \infty$, so Lemma 3.3 implies that $\sum_{k \in I} |d_k/b_k|^2 < \infty$. Since

$$h = (T^{-1})^* d = \sum_{k \in I} (d_k/p'(\lambda_k)b_k)\psi_k,$$

this implies that $h \in H$.

Let

$$T_3: E \rightarrow E: \{e^{\alpha_k}\} \rightarrow \{e^{\alpha_k}\} p'(\lambda_k)b_k,$$

so $T_3^*: Y \rightarrow Y: g_k^\alpha \rightarrow g_k^\alpha p'(\lambda_k)b_k$. Using Lemma 3.3 and (3.9), we can use an argument found in [11] to show that $T_1 T_3^{-1}$ is an isomorphism of E . Since T_0 is an isomorphism, $T_4 := T_0 T_1 T_3^{-1}$ is an isomorphism between E and H . It is easy to see that $T_3^{-1}(D + \hat{b}d^*)T_3 = D + \hat{b}d^*$. Combining this with (3.10), we see $T_4^{-1}(A + bh^*)T_4 = D + \hat{b}d^*$. Therefore $A + bh^*$ has eigenvalues at $\{\alpha_k\}_{k \in I}$ and a Riesz basis of eigenvectors $\{\chi_k\}_{k \in I}$, where $\chi_k = T_4\{e^{\alpha_k}\}$. The dual basis to $\{\chi_k\}_{k \in I}$ is given by $\{h_k^\alpha\}_{k \in I}$, where $h_k^\alpha = (T_4^{-1})^* g_k^\alpha$. This conclusion is also true if condition (1) of Definition 2.8, instead of condition (2), holds, since T_4 is clearly an isomorphism in this case.

We can write χ_k in terms of the original eigenvectors $\{\varphi_k\}_{k \in I}$, and h_k^α and h in terms of the original dual basis $\{\psi_k\}_{k \in I}$. Recall that $\{e^{\alpha_k}\} = \sum_{j \in I} \langle g_j, e^{\alpha_k} \rangle \{e_j\}$.

For $\rho \in (c, b)$,

$$\begin{aligned} \langle g_j, e^{\alpha_k} \rangle &= \langle p_j, (\cdot - \alpha_k)^{-1} \rangle \\ &= (2\pi i)^{-1} \int_{-\Gamma_\rho} p(\lambda) / [p'(\lambda_j)(\lambda - \lambda_j)(\lambda - \alpha_k)] d\lambda \\ &= p(\alpha_k) / p'(\lambda_j)(\alpha_k - \lambda_j). \end{aligned}$$

Hence

$$\chi_k = (1/p'(\lambda_k)b_k) \sum_{j \in I} [p(\alpha_k)b_j/(\alpha_k - \lambda_j)]\varphi_j.$$

Furthermore, $g_j^\alpha = \sum_{k \in I} \langle g_j^\alpha, e_k \rangle g_k$. Since

$$\begin{aligned}\langle g_j^\alpha, e_k \rangle &= \left(\frac{1}{2\pi i} \right) \int_{-\Gamma_p} q(\lambda) / [q'(\alpha_j)(\lambda - \alpha_j)(\lambda - \lambda_k)] d\lambda \\ &= q(\lambda_k) / q'(\alpha_j)(\lambda_k - \alpha_j), \\ h_j^\alpha &= b_j p'(\lambda_j) \sum_{k \in I} q(\lambda_k) / [q'(\alpha_j)(\lambda_k - \alpha_j) b_k p'(\lambda_k)] \psi_k.\end{aligned}$$

The feedback element is

$$h^* = \sum_{j \in I} p(\alpha_j) h_j^\alpha / b_j p'(\lambda_j) = \sum_{j \in I} p(\alpha_j) \sum_{k \in I} q(\lambda_k) / [q'(\alpha_j)(\lambda_k - \alpha_j) b_k p'(\lambda_k)] \psi_k.$$

From (3.5) we can see that

$$u(t) = \sum_{j \in I} \sum_{k \in I} [x_k p(\alpha_j) q(\lambda_k) / [q'(\alpha_j)(\lambda_k - \alpha_j)]] e^{\alpha_j t}.$$

Since $z_0 = \sum_{k \in I} z_k \varphi_k$ and $x_0 = T^{-1} z_0$, we see that $x_k = z_k / (b_k p'(\lambda_k))$, so $u(t)$ can be written as in (1.31). This completes the proof of the lemma. \square

3.2. Analysis of truncated control using the canonical form. In this section we will use the canonical form to study the effect of the truncated control (1.35).

Proof of Theorem 1.4. First assume that b satisfies (1.30), so T is an isomorphism. Assume $z(t)$ satisfies (1.1) so $\{x(t)\} = T^{-1} z(t)$ solves (2.10). By the results in § 2, $\{x(t)\} = \{y(t + \cdot)\}$, where y solves (2.14). Using the control $u_{n,m}$, we can solve for $\mathcal{L}y$ by using Theorem 2.25, so $\mathcal{L}y = \chi + \mathcal{L}u_{n,m}/p$, where $\chi = \sum_{k \in I} x_k / (\lambda - \lambda_k)$.

To simplify $\mathcal{L}u_{n,m}/p$, we first compute its residues, which are at $\{\lambda_k\}_{k \in I}$ and $\{\alpha_j\}_{j \in I_n}$:

$$\text{Res}(\alpha_i) = \lim_{\lambda \rightarrow \alpha_i} \sum_{k \in I_m} \sum_{j \in I_n} p(\alpha_j) q(\lambda_k) x_k (\lambda - \alpha_i) / [q'(\alpha_j)(\lambda_k - \alpha_j)(\lambda - \alpha_j) p(\lambda)],$$

which is equal to zero if $|i| \geq n$, and is equal to

$$(3.11) \quad \sum_{k \in I_m} q(\lambda_k) x_k / [q'(\alpha_i)(\lambda_k - \alpha_i)] = \sum_{k \in I_m} q_i(\lambda_k) x_k =: y_i$$

if $i \in I_n$. It is easily seen that $\langle q_i, e_k \rangle = q_i(\lambda_k)$, so

$$\sum_{i \in I_n} y_i \{e^{\alpha_i}\} = \sum_{i \in I_n} \sum_{k \in I_m} \langle q_i, e_k \rangle x_k \{e^{\alpha_i}\} = \sum_{i \in I_n} \left\langle q_i, \sum_{k \in I_m} x_k e_k \right\rangle \{e^{\alpha_i}\} = \sum_{k \in I_m} x_k \{e_k\}.$$

Hence there exists M such that

$$(3.12) \quad \sum_{i \in I_n} |y_i|^2 \leq M \sum_{k \in I_m} |x_k|^2.$$

To compute the residue of $\mathcal{L}u_{n,m}/p$ at λ_i , we see that

$$\begin{aligned}\text{Res}(\lambda_i) &= \sum_{j \in I_n} \sum_{k \in I_m} p(\alpha_j) q(\lambda_k) x_k / [q'(\alpha_j) p'(\lambda_i)(\lambda_k - \alpha_j)(\lambda_i - \alpha_j)] \\ &= - \sum_{j \in I_n} \sum_{k \in I_m} p_i(\alpha_j) q_j(\lambda_k) x_k = - \sum_{j \in I_n} \sum_{k \in I_m} \langle p_i, e^{\alpha_j} \rangle \langle q_j, e^{\lambda_k} \rangle x_k \\ &= \left\langle p_i, \sum_{j \in I_n} \langle q_j, x_0^m \rangle e^{\alpha_j} \right\rangle,\end{aligned}$$

where $x_0^m = \sum_{k \in I_m} x_k e^{\lambda_k}$. Then

$$\begin{aligned}(3.13) \quad x_i + \text{Res}(\lambda_i) &= \langle p_i, x_0 \rangle + \text{Res}(\lambda_i) \\ &= \left\langle p_i, \sum_{j \in I} \langle q_j, x_0 \rangle e^{\alpha_j} \right\rangle - \left\langle p_i, \sum_{j \in I_n} \langle q_j, x_0^m \rangle e^{\alpha_j} \right\rangle \\ &= \left\langle p_i, \sum_{j \in I} \langle q_j, x_0 - x_0^m \rangle e^{\alpha_j} \right\rangle + c_i,\end{aligned}$$

where $c_i = \langle p_i, \sum_{|j|>n} \langle q_j, x_0^m \rangle e^{\alpha_j} \rangle$. The first term on the right side is just the i th coefficient of $x_0 - x_0^m$, which is x_i if $|i| > m$ and zero if $i \in I_m$. Hence $\text{Res}(\lambda_i) = c_i$ if $|i| > m$, and $\text{Res}(\lambda_i) = c_i - x_i$ if $i \in I_m$. We can note that $\sum_{k \in I} c_k \{e_k\} = \sum_{|j|>n} \langle q_j, x_0^m \rangle \{e^{\alpha_j}\}$, so

$$(3.14) \quad \sum_{k \in I} |c_k|^2 \leq M \sum_{|j|>n} |\langle q_j, x_0^m \rangle|^2.$$

Putting the computation of these residues together, we get

$$\mathcal{L}y(\lambda) = \chi(\lambda) + \mathcal{L}u/p(\lambda) = \sum_{k \in I_n} y_k/(\lambda - \alpha_k) + \sum_{|k|>m} x_k/(\lambda - \lambda_k) + \sum_{k \in I} c_k/(\lambda - \lambda_k).$$

Then

$$\{x(t)\} = \sum_{k \in I_n} y_k e^{\alpha_k t} \{e^{\alpha_k}\} + \sum_{|k|>m} x_k e^{\lambda_k t} e_k + \sum_{k \in I} c_k e^{\lambda_k t} e_k,$$

so

$$(3.15) \quad z(t) = \sum_{k \in I_n} y_k p'(\lambda_k) b_k e^{\alpha_k t} \chi_k + \sum_{|k|>m} x_k p'(\lambda_k) b_k e^{\lambda_k t} \varphi_k + \sum_{k \in I} c_k p'(\lambda_k) b_k e^{\lambda_k t} \varphi_k.$$

Since $\{\varphi_k\}_{k \in I}$ and $\{\psi_k\}_{k \in I}$ are both Riesz basis of H , there exists M independent of $z_0 \in H$ and $t \in \tilde{\Omega}$ such that

$$(3.16) \quad \begin{aligned} \|z(t)\|^2 &\leq \sup_{k \in I_n} |e^{\alpha_k t}|^2 \left(\sum_{k \in I_n} |y_k p'(\lambda_k) b_k|^2 \right) + \sup_{|k|>m} |e^{\lambda_k t}|^2 \left(\sum_{|k|>m} |z_k|^2 \right) \\ &\quad + \sup_{k \in I} |e^{\lambda_k t}|^2 \left(\sum_{k \in I} |c_k p'(\lambda_k) b_k|^2 \right). \end{aligned}$$

The first part of the theorem clearly follows from (3.16), (3.12), and (3.14). Now pick $\varepsilon > 0$. Since $\{z_k\}_{k \in I}$ is in ℓ_2 , the second term in parenthesis on the right side of (3.16) can be made less than $\varepsilon/2$ for large enough m . The third term in parentheses is bounded by (3.14), which can be made less than $\varepsilon/2$ for large enough n . This completes the proof of the theorem when (1.30) is satisfied.

Now suppose b is a bounded input element and (1.24) holds. Referring to (3.11), $\sum_{j \in I_n} |y_j p'(\lambda_j) b_j|^2 = \sum_{j \in I_n} |\sum_{k \in I_m} x_k q_j(\lambda_k) p'(\lambda_j) b_j|^2$. Since $T_3 T_1^{-1}$ is an isomorphism, and $T_3 T_1^{-1} (\sum_{k \in I_m} x_k p'(\lambda_k) b_k \{e_k\}) = \sum_{j \in I} (\sum_{k \in I_m} x_k q_j(\lambda_k) p'(\lambda_j) b_j) \{e^{\alpha_j}\}$, we see that there exists M such that

$$(3.17) \quad \sum_{j \in I_n} |y_j p'(\lambda_j) b_j|^2 \leq M \sum_{k \in I_m} |x_k p'(\lambda_k) b_k|^2 = M \sum_{k \in I_m} |z_k|^2.$$

It can be shown that $\sum_{j \in I} c_j p'(\lambda_j) b_j \{e_j\} = T_1 T_3^{-1} (\sum_{|j|>n} \langle q_j, T_3 x_0^m \rangle \{e^{\alpha_j}\})$, so there exists M such that

$$(3.18) \quad \sum_{j \in I} |c_j p'(\lambda_j) b_j|^2 \leq M \sum_{|j|>n} |\langle q_j, T_3 x_0^m \rangle|^2.$$

We can note that $T_3 x_0^m = T_3 T_1^{-1} T_0^{-1} (\sum_{k \in I_m} z_k \varphi_k)$, and $T_3 T_1^{-1} T_0^{-1}$ is an isomorphism, so there exists C independent of m and n such that

$$(3.19) \quad \sum_{|j|>n} |\langle q_j, T_3 x_0^m \rangle|^2 \leq C.$$

Referring to (3.16), we use (3.17)–(3.19) to prove the theorem in the same way when b is a bounded input element and (1.24) holds. \square

REFERENCES

- [1] M. BALAS, *Modal control of certain flexible dynamic systems*, SIAM J. Control Optim., 16 (1978), pp. 450-462.
- [2] G. CHEN AND D. L. RUSSELL, *A mathematical model for linear elastic systems with structural damping*, Quart. Appl. Math., 34 (1982), pp. 433-454.
- [3] B. M. N. CLARKE AND D. WILLIAMSON, *Control canonical forms and eigenvalue assignment by feedback for a class of linear hyperbolic systems*, SIAM J. Control Optim., 19 (1981), pp. 711-729.
- [4] L. F. HO, *Controllability and spectral assignability of a class of hyperbolic control systems with retarded control canonical forms*, Ph.D. thesis, University of Wisconsin, Madison, WI, 1981.
- [5] L. F. HO AND D. L. RUSSELL, *Admissible input elements for systems in Hilbert space and a Carleson measure criterion*, SIAM J. Control Optim., 21 (1983), pp. 614-639.
- [6] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, 1976.
- [7] R. REBARBER, *A Laplace transform relevant to holomorphic semigroups*, Proc. Roy. Soc. Edinburgh Sect. A, 105 (1987), pp. 243-258.
- [8] ———, *A class of meromorphic functions used for spectral determination*, to appear.
- [9] ———, *Control of holomorphic semigroups generated by a class of spectral operators*, Ph.D. thesis, University of Wisconsin, Madison, WI, 1984.
- [10] ———, *Spectral assignability for distributed parameter systems with unbounded scalar control*, SIAM J. Control Optim., 27 (1989), to appear.
- [11] D. L. RUSSELL, *Closed loop eigenvalue specification for infinite dimensional systems: Augmented and deficient hyperbolic cases*, Tech. Summary Report 2021, Math. Research Center, University of Wisconsin, Madison, WI, 1979.
- [12] ———, *Uniform bases of exponentials, neutral groups, and a transform theory for $H^m[a, b]$* , Tech. Summary Report 2149, Math. Research Center, University of Wisconsin, Madison, WI, 1980.
- [13] ———, *Functional equations as control canonical forms for distributed parameter control systems and a state space theory for certain differential equations of infinite order*, in Volterra and Functional Differential Equations, K. B. Hannsgen, T. L. Herdman, H. W. Stech, R. L. Wheeler, eds., Marcel Dekker, New York, 1974.
- [14] ———, *Mathematical models for the elastic beam and their control-theoretic implications*, Semigroup Theory and Applications, H. Brezis, M. G. Crandall, and S. F. Kappel, eds., Longman, New York, 1985.
- [15] S.-H. SUN, *On spectrum distribution of completely controllable linear systems* (L. F. Ho, trans.) SIAM J. Control Optim., 19 (1981), pp. 730-743.
- [16] R. TEGLAS, *On the control canonical structure of a class of scalar input systems*, SIAM J. Control Optim., 22 (1984), pp. 552-569.
- [17] S.-F. C. YU, *Computing the eigenvalues for a damped Euler-Bernoulli beam equation with the Chebyshev spectral method*, M.Sc. thesis, Pennsylvania State University, University Park, PA, 1985.

OPTIMALITY CONDITIONS FOR QUASI-DIFFERENTIABLE PROGRAMS WITH APPLICATION TO TWO-LEVEL OPTIMIZATION*

YO ISHIZUKA†

Abstract. As an extension of the quasi-differentiability of Demyanov et al. (V. F. Demyanov and A. M. Rubinov, *Soviet Math. Dokl.*, 21 (1980), pp. 14–17; V. F. Demyanov and N. L. Polyakova, U.S.S.R. Comput. Math. and Math. Phys., 20 (1981), pp. 34–43), this paper introduces a new class of quasi-differentiable functions whose directional derivatives are given by a positively homogeneous function. Optimality conditions for quasi-differentiable programs are obtained by use of the generalized Farkas theorem. As an application an optimality condition is derived for an optimization problem with two-level structure.

Key words. optimality conditions, directional derivatives, quasi-differentiable programs, the generalized Farkas theorem, two-level optimization

AMS(MOS) subject classifications. 90C30, 90C31

1. Introduction. The notion of quasi-differentiability was introduced by Pshenichnyi [7]. A quasi-differentiable function in the sense of Pshenichnyi is a function whose directional derivative $f'(x; s)$ is represented by

$$(1.1) \quad f'(x; s) = \max_{\gamma \in \Gamma} \langle \gamma, s \rangle,$$

where Γ is a compact set of R^n and $\langle \cdot, \cdot \rangle$ denotes the inner product in R^n .

In optimization problems, however, we often have to deal with functions not in the class of Pshenichnyi. For example, even if a function f is quasi-differentiable, $-f$ is not necessarily so. Thus, a composite function of quasi-differentiable functions may not be quasi-differentiable.

So, Demyanov and Ruvinov [1] and Demyanov and Polyakova [2] have extended Pshenichnyi's quasi-differentiability to functions with directional derivatives such as

$$f'(x; s) = \max_{\gamma \in \Gamma} \langle \gamma, s \rangle + \min_{\lambda \in \Lambda} \langle \lambda, s \rangle,$$

where Γ and Λ are compact sets, and they have investigated quasi-differential calculus and optimality conditions for quasi-differentiable optimization problems. Recently, Shapiro [9] has refined these optimality conditions.

Meanwhile, it has been shown that the directional derivative of an optimal-value function f —one of the most typical and important nondifferentiable functions appearing in optimization problems—is represented by (see [3], [6])

$$f'(x; s) = \max_{a \in A} \min_{\gamma \in \Gamma(a)} \langle \gamma, s \rangle,$$

where $\Gamma(\cdot)$ is a set-valued map defined on a set A . Thus, an optimal-value function does not belong in general to the class presented in [1], [2]. To deal with optimization problems including such functions, we need to consider the class of functions that admit directional derivatives with a form such as

$$(1.2) \quad f'(x; s) = \max_{a \in A} \min_{\gamma \in \Gamma(a)} \langle \gamma, s \rangle + \min_{b \in B} \max_{\lambda \in \Lambda(b)} \langle \lambda, s \rangle.$$

* Received by the editors April 6, 1987; accepted for publication (in revised form) February 5, 1988.

† Department of Mechanical Engineering, Faculty of Science and Technology, Sophia University, Tokyo, Japan.

The aim of this paper is to derive optimality conditions for the problem with objective and constraint functions that are in such a class. To achieve that, we also introduce the generalized Farkas theorem for a system of inequalities of positively homogeneous functions such as (1.2). As an application, we consider an optimization problem with two-level structure.

In § 2 we mention notation and preliminary results including the generalized Farkas theorem. In § 3, as a new class of quasi-differentiable functions, we introduce the class of directionally differentiable functions mentioned above, and derive optimality conditions for the quasi-differentiable optimization problem. In § 4 we formulate a two-level optimization problem [10], [11], and show its optimality condition.

2. Preliminaries. In the sequel, $\text{co } D$ and $\text{cl } D$ stand for convex hull and closure of $D \subset R^n$, respectively.

For a set-valued map $\Gamma: A \rightarrow 2^{R^n}$ on a set A into the power set of R^n , we denote by $\prod_{a \in A} \Gamma(a)$ the Cartesian product of the family $\mathcal{D} \equiv \{\Gamma(a) \subset R^n \mid a \in A\}$, i.e., $\prod_{a \in A} \Gamma(a)$ is the set of all selection functions of \mathcal{D} :

$$\prod_{a \in A} \Gamma(a) \equiv \{\gamma: A \rightarrow R^n \mid \gamma(a) \in \Gamma(a) \forall a \in A\}.$$

We shall also make use of the following notation.

$$\begin{aligned} \Gamma(A) &\equiv \bigcup_{a \in A} \Gamma(a), & \gamma(A) &\equiv \bigcup_{a \in A} \gamma(a), \\ \sigma[\mathcal{D}] &\equiv \left\{ \gamma(A) \subset R^n \mid \gamma(\cdot) \in \prod_{a \in A} \Gamma(a) \right\}. \end{aligned}$$

We note that $\prod_{a \in A} \Gamma(a)$ and $\sigma[\mathcal{D}]$ are not empty under the Axiom of Choice.

We denote by $\delta^*(\cdot \mid D)$ the supporting function of a set $D \subset R^n$, and by $\delta_*(s \mid D)$ we denote $-\delta^*(-s \mid D)$, i.e.,

$$\delta^*(s \mid D) \equiv \sup_{d \in D} \langle d, s \rangle, \quad \delta_*(s \mid D) \equiv \inf_{d \in D} \langle d, s \rangle.$$

The directional derivative of a function $f: R^n \rightarrow R^1$ at $x \in R^n$ in the direction $s \in R^n$ is defined as

$$f'(x; s) \equiv \lim_{\alpha \downarrow 0} \alpha^{-1} \{f(x + \alpha s) - f(x)\}.$$

We say that f is directionally differentiable at x if $f'(x; s)$ exists for any $s \in R^n$.

PROPOSITION 2.1. *Let D be a nonempty set of R^n . Then*

$$\delta^*(s \mid D) \geq 0 \quad \forall s \in R^n \Leftrightarrow 0 \in \text{cl co } D.$$

Proof. This is an immediate consequence of Theorems 13.1 and 32.2 of [8]. \square

PROPOSITION 2.2. *Let $\Gamma: A \rightarrow 2^{R^n}$ be a set-valued map such that $\Gamma(a)$ is nonempty and compact at every $a \in A$. Then, for any $s \in R^n$, we have*

$$(2.1) \quad \sup_{a \in A} \delta_*(s \mid \Gamma(a)) = \min_{\gamma(\cdot) \in \prod_{a \in A} \Gamma(a)} \delta^*(s \mid \gamma(A)).$$

Proof. Let $s \in R^n$ be arbitrarily fixed. Define

$$\alpha = \sup_{a \in A} \delta_*(s \mid \Gamma(a)), \quad \beta = \inf_{\gamma(\cdot) \in \prod_{a \in A} \Gamma(a)} \delta^*(s \mid \gamma(A)).$$

Let us consider a function $\tilde{\gamma}(\cdot) \in \prod_{a \in A} \Gamma(a)$ such that

$$\langle \tilde{\gamma}(a), s \rangle = \delta_*(s \mid \Gamma(a)) \quad \forall a \in A.$$

Then we have

$$(2.2) \quad \begin{aligned} \alpha &= \sup_{a \in A} \langle \bar{\gamma}(a), s \rangle = \delta^*(s | \bar{\gamma}(A)) \\ &\geq \beta. \end{aligned}$$

Now, if we suppose that $\alpha > \beta$, then there exists $\bar{\gamma}(\cdot) \in \prod_{a \in A} \Gamma(a)$ such that

$$\begin{aligned} \alpha &> \delta^*(s | \bar{\gamma}(A)) = \sup_{a \in A} \langle \bar{\gamma}(a), s \rangle \\ &\geq \sup_{a \in A} \delta_*(s | \Gamma(a)) \\ &= \alpha. \end{aligned}$$

This is a contradiction. Thus we have $\alpha = \beta$. Furthermore, from (2.2) there exists $\bar{\gamma}(\cdot) \in \prod_{a \in A} \Gamma(a)$ satisfying $\alpha = \delta^*(s | \bar{\gamma}(A)) = \beta$. Hence (2.1) holds. \square

Let us consider the following positively homogeneous functions h_i defined on R^n :

$$(2.3) \quad h_i(s) \equiv \sup_{a_i \in A_i} \delta_*(s | \Gamma_i(a_i)) + \inf_{b_i \in B_i} \delta^*(s | \Lambda_i(b_i)), \quad i \in \{0\} \cup I,$$

where $\Gamma_i: A_i \rightarrow 2^{R^n}$, $\Lambda_i: B_i \rightarrow 2^{R^n}$, and I is a finite index set. In order to derive optimality conditions for quasi-differentiable optimization problems, we generalize the ordinary Farkas theorem to make it valid for a system of inequalities described by positively homogeneous functions (2.3). The results obtained here are extensions of those in [5], in which simpler functions such as

$$h_i(s) = \sup_{a_i \in A_i} \delta_*(s | \Gamma_i(a_i))$$

are considered.

For functions h_i , let us assume the following.

Assumption 2.1. $\Gamma_i(a_i)$ and $\Lambda_i(b_i)$ are nonempty and compact sets of R^n at any $a_i \in A_i$ and $b_i \in B_i$, $i \in \{0\} \cup I$.

We make an assumption that corresponds to a constraint qualification for quasi-differentiable optimization problems (see Assumption 3.2 in § 3).

Assumption 2.2. For any $\gamma_i(\cdot) \in \prod_{a_i \in A_i} \Gamma_i(a_i)$ and $b_i \in B_i$, $i \in I$, it holds that

$$0 \notin \text{cl co} \bigcup_{i \in I} \{\gamma_i(A_i) + \Lambda_i(b_i)\}.$$

Moreover, we shall need the following assumption.

Assumption 2.3. For any $s \in R^n$, the infimums $\inf_{b_i \in B_i} \delta^*(s | \Lambda_i(b_i))$, $i \in I$, are attained, i.e.,

$$(2.4) \quad \inf_{b_i \in B_i} \delta^*(s | \Lambda_i(b_i)) = \min_{b_i \in B_i} \delta^*(s | \Lambda_i(b_i)), \quad i \in I.$$

The generalized Farkas theorem is stated as follows.

LEMMA 2.1. *Let Assumptions 2.1 and 2.2 be satisfied. If there is no $s \in R^n$ with*

$$(2.5) \quad h_0(s) < 0, \quad h_i(s) \leq 0, \quad i \in I,$$

then, for arbitrarily fixed $\gamma_i(\cdot) \in \prod_{a_i \in A_i} \Gamma_i(a_i)$ and $b_i \in B_i$, $i \in \{0\} \cup I$, there exist scalars μ_i , $i \in I$, satisfying

$$(2.6) \quad 0 \in \text{cl co} \left[\gamma_0(A_0) + \Lambda_0(b_0) + \sum_{i \in I} \mu_i \{\gamma_i(A_i) + \Lambda_i(b_i)\} \right], \quad \mu_i \geq 0, \quad i \in I,$$

and furthermore, the converse is true provided that Assumption 2.3 is satisfied.

Proof. Let us choose arbitrarily fixed $\bar{\gamma}_i(\cdot) \in \prod_{a_i \in A_i} \Gamma_i(a_i)$ and $\bar{b}_i \in B_i$, $i \in \{0\} \cup I$, and define a set

$$Z \equiv \bigcup_{\substack{\mu_i \geq 0, \\ i \in I}} \text{cl co} \left[\bar{\gamma}_0(A_0) + \Lambda_0(\bar{b}_0) + \sum_{i \in I} \mu_i \{ \bar{\gamma}_i(A_i) + \Lambda_i(\bar{b}_i) \} \right].$$

Suppose, to show a contradiction, that there is no $s \in R^n$ satisfying (2.5), and that $0 \notin Z$. Under Assumption 2.2, it can be shown that Z is closed and convex (see Proposition A in the Appendix of [5]). Then we can apply the separation theorem to find $\bar{s} \in R^n$ and α satisfying $\langle z, \bar{s} \rangle < \alpha < 0$ for all $z \in Z$. Hence

(2.7)

$$\langle \bar{\gamma}_0(a_0) + \lambda_0, \bar{s} \rangle + \sum_{i \in I} \mu_i \langle \gamma_i(a_i) + \lambda_i, \bar{s} \rangle < \alpha < 0 \quad \forall a_i \in A_i \quad \forall \lambda_i \in \Lambda_i(\bar{b}_i), \quad i \in \{0\} \cup I.$$

Letting $\mu_i = 0$, $i \in I$, we have

$$\begin{aligned} 0 &> \sup_{a_0 \in A_0} \langle \bar{\gamma}_0(a_0), \bar{s} \rangle + \max_{\lambda_0 \in \Lambda_0(\bar{b}_0)} \langle \lambda_0, \bar{s} \rangle \\ &\geq \sup_{a_0 \in A_0} \delta_*(\bar{s} | \Gamma_0(a_0)) + \inf_{b_0 \in B_0} \delta^*(\bar{s} | \Lambda_0(b_0)) \\ &= h_0(\bar{s}). \end{aligned}$$

Meanwhile, if we suppose that there exist $j \in I$, $\bar{a}_j \in A_j$, and $\bar{\lambda}_j \in \Lambda_j(\bar{b}_j)$ such that $\langle \bar{\gamma}_j(\bar{a}_j) + \bar{\lambda}_j, \bar{s} \rangle > 0$, then, by letting μ_j be large enough, we can show a contradiction in (2.7). Therefore

$$\langle \bar{\gamma}_i(a_i) + \lambda_i, \bar{s} \rangle \leq 0 \quad \forall a_i \in A_i \quad \forall \lambda_i \in \Lambda_i(\bar{b}_i), \quad i \in I,$$

and hence we have

$$\begin{aligned} 0 &\geq \sup_{a_i \in A_i} \langle \bar{\gamma}_i(a_i), \bar{s} \rangle + \max_{\lambda_i \in \Lambda_i(\bar{b}_i)} \langle \lambda_i, \bar{s} \rangle \\ &\geq h_i(\bar{s}), \quad i \in I. \end{aligned}$$

This, together with $h_0(\bar{s}) < 0$, is a contradiction. Therefore it must hold that $0 \in Z$. Now, let us assume that (2.5) holds, and that, for arbitrarily fixed $\gamma_i(\cdot) \in \prod_{a_i \in A_i} \Gamma_i(a_i)$ and $b_i \in B_i$, $i \in \{0\} \cup I$, there exist μ_i , $i \in I$, satisfying (2.6). Suppose that there exists $\bar{s} \in R^n$ satisfying (2.5). Then, from Proposition 2.2 and Assumption 2.3, we can find $\bar{\gamma}_i(\cdot) \in \prod_{a_i \in A_i} \Gamma_i(a_i)$ and $\bar{b}_i \in B_i$, $i \in \{0\} \cup I$ such that

$$\begin{aligned} (2.8) \quad &\delta^*(\bar{s} | \bar{\gamma}_0(A_0)) + \delta^*(\bar{s} | \Lambda_0(\bar{b}_0)) < 0, \\ &\delta^*(\bar{s} | \bar{\gamma}_i(A_i)) + \delta^*(\bar{s} | \Lambda_i(\bar{b}_i)) \leq 0, \quad i \in I. \end{aligned}$$

For these $\bar{\gamma}_i(\cdot)$ and \bar{b}_i , there exist scalars $\bar{\mu}_i \geq 0$, $i \in I$ such that

$$0 \in \text{cl co} \left[\bar{\gamma}_0(A_0) + \Lambda_0(\bar{b}_0) + \sum_{i \in I} \bar{\mu}_i \{ \bar{\gamma}_i(A_i) + \Lambda_i(\bar{b}_i) \} \right].$$

From Proposition 2.1, this is equivalent to the following condition:

$$\delta^* \left(s \mid \bar{\gamma}_0(A_0) + \Lambda_0(\bar{b}_0) + \sum_{i \in I} \bar{\mu}_i \{ \bar{\gamma}_i(A_i) + \Lambda_i(\bar{b}_i) \} \right) \geq 0 \quad \forall s \in R^n,$$

which contradicts (2.8). \square

Finally, we present the following proposition.

PROPOSITION 2.3. *Define the sets*

$$S \equiv \{s \in R^n \mid h_i(s) \leq 0, i \in I\},$$

$$S^0 \equiv \{s \in R^n \mid h_i(s) < 0, i \in I\}.$$

Under Assumptions 2.1–2.3, if $h_i, i \in I$, are continuous on R^n , then it holds that $\text{cl } S^0 = S$.

Proof. From Proposition 2.1, Assumption 2.2 holds if and only if, for every fixed $\gamma_i(\cdot) \in \prod_{a_i \in A_i} \Gamma_i(a_i)$ and $b_i \in B_i, i \in I$, there exists $s \in R^n$ such that $\delta^*(s \mid \gamma_i(A_i) + \Lambda_i(b_i)) < 0, i \in I$. Since $\delta^*(s \mid \gamma_i(A_i) + \Lambda_i(b_i)) \geq h_i(s), S^0$ is not empty. Since $S^0 \subset S$ and S is closed (because h_i are continuous), it is obvious that $\text{cl } S^0 \subset S$. Thus we have only to show $S \subset \text{cl } S^0$. For arbitrarily fixed $\bar{s} \in S$, consider functions $\bar{\gamma}_i(\cdot) \in \prod_{a_i \in A_i} \Gamma_i(a_i)$ and elements $\bar{b}_i \in B_i, i \in I$ such that

$$\begin{aligned} \langle \bar{\gamma}_i(a_i), \bar{s} \rangle &= \delta_*(\bar{s} \mid \Gamma_i(a_i)) \quad \forall a_i \in A_i, \\ \delta^*(\bar{s} \mid \Lambda_i(\bar{b}_i)) &= \inf_{b_i \in B_i} \delta^*(\bar{s} \mid \Lambda_i(b_i)), \quad i \in I. \end{aligned}$$

Note that the compactness of $\Gamma_i(a_i)$ and Assumption 2.3 ensure the existence of such $\bar{\gamma}_i(\cdot)$ and \bar{b}_i . Then we have

$$0 \geq h_i(\bar{s}) = \delta^*(\bar{s} \mid \bar{\gamma}_i(A_i) + \Lambda_i(\bar{b}_i)),$$

and, for these $\bar{\gamma}_i(\cdot)$ and \bar{b}_i , there exists $s^0 \in R^n$ satisfying $\delta^*(s^0 \mid \bar{\gamma}_i(A_i) + \Lambda_i(\bar{b}_i)) < 0, i \in I$. Denoting $\bar{s} + \theta s^0$ by $s(\theta)$, we have

$$\begin{aligned} h_i(s(\theta)) &= \sup_{a_i \in A_i} \delta_*(s(\theta) \mid \Gamma_i(a_i)) + \inf_{b_i \in B_i} \delta^*(s(\theta) \mid \Lambda_i(b_i)) \\ &\leq \delta^*(s(\theta) \mid \bar{\gamma}_i(A_i)) + \delta^*(s(\theta) \mid \Lambda_i(\bar{b}_i)) \\ &\leq \delta^*(\bar{s} \mid \bar{\gamma}_i(A_i)) + \theta \delta^*(s^0 \mid \bar{\gamma}_i(A_i)) \\ &\quad + \delta^*(\bar{s} \mid \Lambda_i(\bar{b}_i)) + \theta \delta^*(s^0 \mid \Lambda_i(\bar{b}_i)) \\ &= \delta^*(\bar{s} \mid \bar{\gamma}_i(A_i) + \Lambda_i(\bar{b}_i)) + \theta \delta^*(s^0 \mid \bar{\gamma}_i(A_i) + \Lambda_i(\bar{b}_i)) \\ &< 0 \quad \forall \theta > 0. \end{aligned}$$

This shows $s(\theta) \in S^0$ for all $\theta > 0$. Since for every $\varepsilon > 0$ there exists $\theta > 0$ such that $s(\theta) \in B(\bar{s}; \varepsilon)$, we have $S^0 \cap B(\bar{s}; \varepsilon) \neq \emptyset$, where $B(\bar{s}; \varepsilon) = \{s \in R^n \mid \|s - \bar{s}\| < \varepsilon\}$. Therefore $\bar{s} \in \text{cl } S^0$, and then $S \subset \text{cl } S^0$. \square

3. Quasi-differentiable optimization problem. As mentioned in § 1, we extend the quasi-differentiability in the sense of Demyanov et al. to a more general case as follows.

DEFINITION 3.1. Let $f: R^n \rightarrow R^1$ be directionally differentiable at $x \in R^n$, and let $f'(x; \cdot)$ be continuous on R^n . We say f is quasi-differentiable at x if there exists a pair of families $(\mathcal{D}_*f(x), \mathcal{D}^*f(x))$ such that

$$\begin{aligned} \mathcal{D}_*f(x) &= \{D_*f(x; a) \subset R^n \mid a \in A(x)\}, \\ \mathcal{D}^*f(x) &= \{D^*f(x; b) \subset R^n \mid b \in B(x)\}, \\ f'(x; s) &= \max_{D_* \in \mathcal{D}_*f(x)} \delta_*(s \mid D_*) + \min_{D^* \in \mathcal{D}^*f(x)} \delta^*(s \mid D^*) \quad \forall s \in R^n, \end{aligned}$$

where $A(x)$ and $B(x)$ are index sets, $D_*f(x; \cdot)$ and $D^*f(x; \cdot)$ are set-valued maps such that $D_*f(x; a), a \in A(x)$, and $D^*f(x; b), b \in B(x)$, are nonempty and compact

sets. Since such a pair is not uniquely determined, we denote by $\mathcal{D}f(x)$ the set of those pairs of families that generate the same directional derivative $f'(x; s)$.

The following proposition is obvious.

PROPOSITION 3.1. *Let f_i , $i = 1, 2$, be quasi-differentiable at $\bar{x} \in R^n$, and*

$$(\mathcal{D}_* f_i(\bar{x}), \mathcal{D}^* f_i(\bar{x})) \in \mathcal{D}f_i(\bar{x}), \quad i = 1, 2;$$

then $f_1 + f_2$ and αf_i (α is a scalar) are quasi-differentiable at \bar{x} , and we have¹

$$(\mathcal{D}_* f_1(\bar{x}) + \mathcal{D}_* f_2(\bar{x}), \mathcal{D}^* f_1(\bar{x}) + \mathcal{D}^* f_2(\bar{x})) \in \mathcal{D}(f_1 + f_2)(\bar{x}),$$

$$(\alpha \mathcal{D}_* f_i(\bar{x}), \alpha \mathcal{D}^* f_i(\bar{x})) \in \mathcal{D}(\alpha f_i)(\bar{x}) \quad \text{if } \alpha \geq 0,$$

$$(\alpha \mathcal{D}^* f_i(\bar{x}), \alpha \mathcal{D}_* f_i(\bar{x})) \in \mathcal{D}(\alpha f_i)(\bar{x}) \quad \text{if } \alpha < 0.$$

More generally, the following proposition holds.

PROPOSITION 3.2. *Let f_j , $j = 1, \dots, N$ be quasi-differentiable at $\bar{x} \in R^n$, $f(x) \equiv (f_1(x), \dots, f_N(x))$, and $F: R^N \rightarrow R^1$. If F is continuously differentiable, and if*

$$(\mathcal{D}_* f_j(\bar{x}), \mathcal{D}^* f_j(\bar{x})) \in \mathcal{D}f_j(\bar{x}), \quad j = 1, \dots, N,$$

then the composite function $F \circ f$ is quasi-differentiable at \bar{x} , and it holds that

$$(\mathcal{F}_*, \mathcal{F}^*) \in \mathcal{D}(F \circ f)(\bar{x}),$$

where

$$\mathcal{F}_* \equiv \sum_{j \in J^+} \alpha_j \mathcal{D}_* f_j(\bar{x}) + \sum_{j \in J^-} \alpha_j \mathcal{D}^* f_j(\bar{x}),$$

$$\mathcal{F}^* \equiv \sum_{j \in J^+} \alpha_j \mathcal{D}^* f_j(\bar{x}) + \sum_{j \in J^-} \alpha_j \mathcal{D}_* f_j(\bar{x}),$$

$$\alpha_j \equiv \nabla_{f_j} F(f(\bar{x})), \quad j = 1, \dots, N,$$

$$J^+ \equiv \{j = 1, \dots, N \mid \alpha_j \geq 0\},$$

$$J^- \equiv \{j = 1, \dots, N \mid \alpha_j < 0\}.$$

Proof. Since

$$\begin{aligned} (F \circ f)'(\bar{x}; s) &= \sum_{j=1}^N \nabla_{f_j} F(f(\bar{x})) f'_j(\bar{x}; s) \\ &= \left(\sum_{j=1}^N \alpha_j f_j \right)'(\bar{x}; s), \end{aligned}$$

we have

$$\mathcal{D}(F \circ f)(\bar{x}) = \mathcal{D} \left(\sum_{j=1}^N \alpha_j f_j \right) (\bar{x}).$$

Then the conclusion follows from Proposition 3.1. \square

Now, let us consider the following minimization problem with inequality constraints:

$$(MP) \quad \min_x f_0(x) \quad \text{subject to } f_i(x) \leq 0, \quad i = 1, \dots, l.$$

¹ For families $\mathcal{D}_1, \mathcal{D}_2$ of sets in R^n , we define $\mathcal{D}_1 + \mathcal{D}_2 \equiv \{D_1 + D_2 \subset R^n \mid D_i \in \mathcal{D}_i, i = 1, 2\}$ and $\alpha \mathcal{D}_i \equiv \{\alpha D_i \subset R^n \mid D_i \in \mathcal{D}_i\}$.

At a point $\bar{x} \in R^n$, we define the following sets:

$$\begin{aligned} I &\equiv \{i = 1, \dots, l \mid f_i(\bar{x}) = 0\}, \\ S &\equiv \{s \in R^n \mid f'_i(\bar{x}; s) \leq 0, i \in I\}, \\ S^0 &\equiv \{s \in R^n \mid f'_i(\bar{x}; s) < 0, i \in I\}. \end{aligned}$$

Then we can easily obtain the following optimality condition described by the directional derivatives.

PROPOSITION 3.3. *Assume that functions $f_i, i = 0, 1, \dots, l$ are directionally differentiable at \bar{x} , $f'_i(\bar{x}; \cdot)$ are continuous on R^n , and that the constraint qualification $\text{cl } S^0 = S$ is satisfied. If \bar{x} is an optimal solution of (MP) then there is no $s \in R^n$ with*

$$f'_0(\bar{x}; s) < 0, \quad f'_i(\bar{x}; s) \leq 0, \quad i \in I.$$

Let us assume the following:

Assumption 3.1. Functions $f_i, i = 0, 1, \dots, l$ are quasi-differentiable at \bar{x} .

Assumption 3.2 (Constraint qualification). Given $(\mathcal{D}_* f_i(\bar{x}), \mathcal{D}^* f_i(\bar{x})) \in \mathcal{D}f(\bar{x})$, we have

$$0 \notin \text{cl co} \bigcup_{i \in I} \{D_{*i} + D_i^*\} \quad \forall D_{*i} \in \sigma[\mathcal{D}_* f_i(\bar{x})] \quad \forall D_i^* \in \mathcal{D}^* f_i(\bar{x}), \quad i \in I.$$

We are now in a position to show the optimality condition for the quasi-differentiable optimization problem.

THEOREM 3.1. *Let Assumptions 3.1 and 3.2 be satisfied. If \bar{x} is an optimal solution of (MP) then, for arbitrarily fixed $D_{*i} \in \sigma[\mathcal{D}_* f_i(\bar{x})]$ and $D_i^* \in \mathcal{D}^* f_i(\bar{x}), i = 0, 1, \dots, l$, there exist scalars $\mu_i, i = 1, \dots, l$ satisfying*

$$\begin{aligned} (3.1) \quad & 0 \in \text{cl co} \left\{ D_{*0} + D_0^* + \sum_{i=1}^l \mu_i (D_{*i} + D_i^*) \right\}, \\ & \mu_i f_i(\bar{x}) = 0, \quad \mu_i \geq 0, \quad i = 1, \dots, l. \end{aligned}$$

Proof. From the quasi-differentiability of f_i and Propositions 2.3 and 3.3, there is no $s \in R^n$ satisfying

$$\begin{aligned} & \max_{D_{*0} \in \mathcal{D}_* f_0(\bar{x})} \delta_*(s \mid D_{*0}) + \min_{D_0^* \in \mathcal{D}^* f_0(\bar{x})} \delta^*(s \mid D_0^*) < 0, \\ & \max_{D_{*i} \in \mathcal{D}_* f_i(\bar{x})} \delta_*(s \mid D_{*i}) + \min_{D_i^* \in \mathcal{D}^* f_i(\bar{x})} \delta^*(s \mid D_i^*) \leq 0, \quad i \in I. \end{aligned}$$

It follows from Assumption 3.2 and Lemma 2.1 that, for arbitrarily fixed $D_{*i} \in \sigma[\mathcal{D}_* f_i(\bar{x})]$ and $D_i^* \in \mathcal{D}^* f_i(\bar{x}), i \in I$, there exist scalars $\mu_i \geq 0, i \in I$, such that

$$0 \in \text{cl co} \left\{ D_{*0} + D_0^* + \sum_{i \in I} \mu_i (D_{*i} + D_i^*) \right\}.$$

Therefore, letting $\mu_i = 0, i \notin I$, we have (3.1). \square

Example 3.1. Consider the following unconstrained minimization problem:

$$(3.2) \quad \min_{x \in R^2} f(x) \equiv (\sqrt{|x_1|} + \sqrt{|x_2|})^2$$

whose minimum solution is $\bar{x} = 0$. Clearly, f is continuous and positively homogeneous on R^2 . Therefore f is directionally differentiable at $\bar{x} = 0$, and it holds that

$$f'(\bar{x}; s) = f(s) = (\sqrt{|s_1|} + \sqrt{|s_2|})^2 \quad \forall s \in R^2.$$

Since $f'(\bar{x}; s)$ (i.e., $f(s)$) is *not* locally Lipschitz near every point in $\{s \in \mathbb{R}^2 \mid s_1 s_2 = 0\} \setminus \{0\}$ (see Fig. 3.1), there does not exist a quasi-differential of f in the sense of Demyanov et al. at $\bar{x} = 0$. But an easy calculation shows that $f'(\bar{x}; s)$ is represented as

$$\begin{aligned} f'(\bar{x}; s) = f(s) &= \min_{b \in B} \left\{ \frac{1}{b} |s_1| + \frac{1}{1-b} |s_2| \right\} \\ &= \min_{b \in B} \max_{\lambda \in \Lambda(b)} \langle \lambda, s \rangle, \end{aligned}$$

where

$$\begin{aligned} B &\equiv \{b \in \mathbb{R}^1 \mid 0 < b < 1\}, \\ \Lambda(b) &\equiv \left\{ \left(-\frac{1}{b}, -\frac{1}{1-b} \right), \left(-\frac{1}{b}, \frac{1}{1-b} \right), \left(\frac{1}{b}, -\frac{1}{1-b} \right), \left(\frac{1}{b}, \frac{1}{1-b} \right) \right\}. \end{aligned}$$

Thus f is quasi-differentiable at $\bar{x} = 0$, and we have

$$(\mathcal{D}_* f(\bar{x}), \mathcal{D}^* f(\bar{x})) \in \mathcal{D}f(\bar{x}),$$

where

$$\mathcal{D}_* f(\bar{x}) \equiv \{\{0\}\}, \quad \mathcal{D}^* f(\bar{x}) \equiv \{\Lambda(b) \subset \mathbb{R}^2 \mid b \in B\}.$$

Noting that $\sigma[\mathcal{D}_* f(\bar{x})] = \{\{0\}\}$ and $0 \in \text{co } \Lambda(b)$ for every $b \in B$, we can confirm that the following necessary condition given in Theorem 3.1 for problem (3.2):

$$0 \in \text{cl co } \{D_* + D^*\} \quad \forall (D_*, D^*) \in (\sigma[\mathcal{D}_* f(\bar{x})], \mathcal{D}^* f(\bar{x}))$$

holds at $\bar{x} = 0$.

4. Two-level optimization problem. In this section we consider a minimization problem whose objective and constraint functions include optimal-value functions.

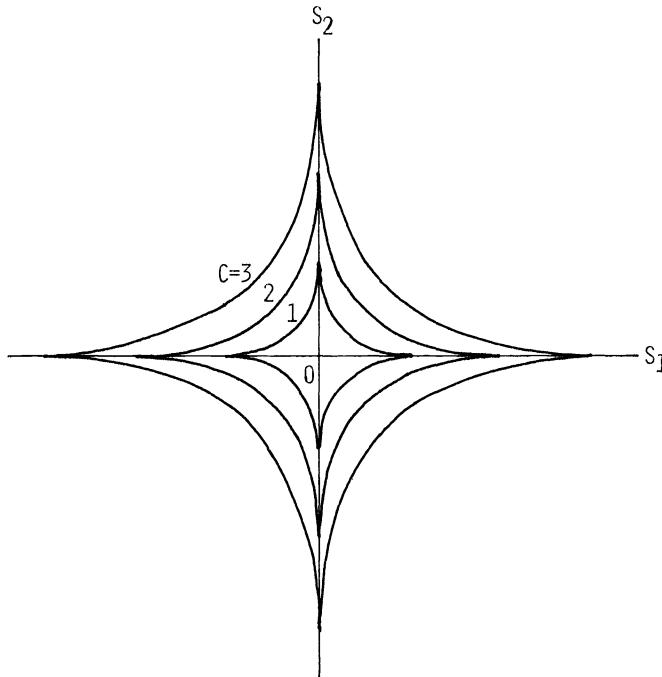


FIG. 3.1. Directional derivative of f in Example 3.1. $f'(\bar{x}; s) = c$ ($c = 1, 2, 3$, $\bar{x} = 0$).

At the lower level, N subsystems choose the optimal decision vectors $y_j \in R^{m_j}$, $j = 1, \dots, N$, minimizing their performance indexes f_j under given parameter $x \in R^n$ from the center. While, at the upper level, the center decides parameter values to be assigned to the subsystems so as to optimize the central objective, which appraises the values of subsystems' performances. Such a two-level optimization problem (TOP) is formulated as follows [10], [11]:

$$\begin{aligned} \text{(TOP)} \quad & \min_x F_0(x, v(x)) \quad \text{subject to } F_i(x, v(x)) \leq 0, \quad i = 1, \dots, l, \\ & v_j(x) = \min_{y_j} f_j(x, y_j) \quad \text{subject to } g_{ji}(x, y_j) \leq 0, \quad i = 1, \dots, p_j, \quad j = 1, \dots, N, \end{aligned}$$

where $v(x) = (v_1(x), \dots, v_N(x))$. Shimizu and Ishizuka [10] derived optimality conditions for (TOP) by assuming the linear independence constraint qualification for lower-level problems. Under the linear independence constraint qualification, optimal-value functions v_j are quasi-differentiable in the sense of Pshenichnyi; therefore functions $\tilde{F}_i(x) \equiv F_i(x, v(x))$ belong to the class of Demyanov et al. provided $F_i \in C^1$. In this section we shall assume the Slater constraint qualification and convexity to the lower-level problems under which \tilde{F}_i are no longer quasi-differentiable in the sense of Demyanov et al.

Let us define the following:

$$\begin{aligned} L_j(x, y_j, u_j) &\equiv f_j(x, y_j) + \sum_{i=1}^{p_j} u_{ji} g_{ji}(x, y_j), \\ S_j(x) &\equiv \{y_j \in R^{m_j} \mid g_{ji}(x, y_j) \leq 0, \quad i = 1, \dots, p_j\}, \\ P_j(x) &\equiv \{y_j \in S_j(x) \mid f_j(x, y_j) = v_j(x)\}, \\ K_j(x, y_j) &\equiv \{u_j^* \in R^{p_j} \mid \nabla_{y_j} L_j(x, y_j, u_j^*) = 0, u_{ji}^* g_{ji}(x, y_j) = 0, u_{ji}^* \geq 0, \quad i = 1, \dots, p_j\}, \\ K_j(x) &\equiv K_j(x, P_j(x)), \\ I &\equiv \{i = 1, \dots, l \mid F_i(x, v(x)) = 0\}, \end{aligned}$$

and, for some neighborhood $N(\bar{x})$ of \bar{x} , let us assume the following.

Assumption 4.1. f_j and g_j are continuously differentiable on $N(\bar{x}) \times R^{m_j}$.

Assumption 4.2. $S_j(\bar{x})$ is not empty, and $S_j(\cdot)$ is uniformly bounded at \bar{x} , i.e., there is a neighborhood $N'(\bar{x})$ of \bar{x} such that $S_j(N'(\bar{x}))$ is bounded, $j = 1, \dots, N$.

Assumption 4.3. There exist points $y_j^0 \in R^{m_j}$, $j = 1, \dots, N$, satisfying $g_j(\bar{x}, y_j^0) < 0$, $j = 1, \dots, N$.

Assumption 4.4. Functions $f_j(\bar{x}, \cdot)$ and $g_j(\bar{x}, \cdot)$ are convex in $y_j \in R^{m_j}$, $j = 1, \dots, N$.

Assumption 4.5. F_i , $i = 0, 1, \dots, l$ are continuously differentiable on $N(\bar{x}) \times R^N$.

PROPOSITION 4.1. Under Assumptions 4.1–4.3, $P_j(\bar{x})$ and $K_j(\bar{x})$ are nonempty and compact sets, $j = 1, \dots, N$.

Proof. See Theorem 3.7 in [4]. \square

PROPOSITION 4.2. Under Assumptions 4.1–4.4, v_j , $j = 1, \dots, N$, are quasi-differentiable at \bar{x} , $v'_j(\bar{x}; \cdot)$ are continuous on R^n , and we have

$$v'_j(\bar{x}; s) = \min_{y_j^* \in P_j(\bar{x})} \max_{u_j^* \in K_j(\bar{x})} \langle \nabla_x L_j(x, y_j^*, u_j^*), s \rangle \quad \forall s \in R^n.$$

Proof. This is an immediate consequence of Theorem 12 in [3]. \square

Let us define the following:

$$\begin{aligned} J_i^- &\equiv \{j = 1, \dots, N \mid \nabla_{v_j} F_i(\bar{x}, v(\bar{x})) < 0\}, \\ J_i^+ &\equiv \{j = 1, \dots, N \mid \nabla_{v_j} F_i(\bar{x}, v(\bar{x})) \geq 0\}, \\ Z_{ji}(\bar{x}, y_j, u_j) &\equiv \nabla_{v_j} F_i \cdot \nabla_x L_j(\bar{x}, y_j, u_j), \end{aligned}$$

where $\nabla_{v_j} F_i = \nabla_{v_j} F_i(\bar{x}, v(\bar{x}))$.

PROPOSITION 4.3. Under Assumptions 4.1–4.5, $\tilde{F}_i(x) \equiv F_i(x, v(x))$, $i = 0, 1, \dots, l$, are quasi-differentiable at \bar{x} , and we have

$$(\mathcal{D}_* \tilde{F}_i(\bar{x}), \mathcal{D}^* \tilde{F}_i(\bar{x})) \in \mathcal{D} \tilde{F}_i(\bar{x})$$

where

$$\begin{aligned} \mathcal{D}_* \tilde{F}_i(\bar{x}) &= \left\{ \sum_{j \in J^-} Z_{ji}(\bar{x}, y_j^*, K_j(\bar{x})) \mid y_j^* \in P_j(\bar{x}), j \in J^- \right\}, \\ \mathcal{D}^* \tilde{F}_i(\bar{x}) &= \left\{ \nabla_x F_i + \sum_{j \in J^+} Z_{ji}(\bar{x}, y_j^*, K_j(\bar{x})) \mid y_j^* \in P_j(\bar{x}), j \in J^+ \right\}, \end{aligned}$$

where $\nabla_x F_i \equiv \nabla_x F_i(\bar{x}, v(\bar{x}))$.

Proof. For functions $h_k : R^n \rightarrow R^1$ such that $h_k(x) = x_k$, $k = 1, \dots, n$, it holds that

$$(\{\{0\}\}, \{\{e_k\}\}) \in \mathcal{D} h_k(x) \quad \forall x \in R^n, \quad k = 1, \dots, n,$$

where $e_k \in R^n$ and $e_{ki} = 0$, $i \neq k$, $e_{kk} = 1$. From Proposition 4.2, we have

$$(\mathcal{D}_* v_j(\bar{x}), \mathcal{D}^* v_j(\bar{x})) \in \mathcal{D} v_j(\bar{x}), \quad j = 1, \dots, N,$$

where

$$\begin{aligned} \mathcal{D}_* v_j(\bar{x}) &\equiv \{\{0\}\}, \\ \mathcal{D}^* v_j(\bar{x}) &\equiv \{\nabla_x L_j(\bar{x}, y_j^*, K_j(\bar{x})) \mid y_j^* \in P_j(\bar{x})\}. \end{aligned}$$

Since $\tilde{F}(x) = F(h(x), v(x))$, this proposition follows from Proposition 3.2. \square

Since

$$\sigma[\mathcal{D}_* \tilde{F}_i(\bar{x})] = \left\{ \sum_{j \in J^-} \bigcup_{y_j^* \in P_j(\bar{x})} Z_{ji}(\bar{x}, y_j^*, u_{ji}^*(y_j^*)) \mid u_{ji}^* : P_j(\bar{x}) \rightarrow K_j(\bar{x}), j \in J^- \right\},$$

we obtain the following optimality condition for (TOP).

Assumption 4.6. For any functions $u_{ji}^* : P_j(\bar{x}) \rightarrow K_j(\bar{x})$, $j \in J^-$, and points $y_{ji}^* \in P_j(\bar{x})$, $j \in J^+$, $i \in I$, it holds that

$$0 \notin \text{cl co} \bigcup_{i \in I} \left\{ \nabla_x F_i + \sum_{j \in J^-} \bigcup_{y_j^* \in P_j(\bar{x})} Z_{ji}(\bar{x}, y_j^*, u_{ji}^*(y_j^*)) + \sum_{j \in J^+} Z_{ji}(\bar{x}, y_{ji}^*, K_j(\bar{x})) \right\}.$$

THEOREM 4.1. Let Assumptions 4.1–4.6 be satisfied. If \bar{x} is an optimal solution of (TOP) then, for arbitrarily fixed functions $u_{ji}^* : P_j(\bar{x}) \rightarrow K_j(\bar{x})$, $j \in J^-$, and points $y_{ji}^* \in P_j(\bar{x})$, $j \in J^+$, $i = 0, 1, \dots, l$, there exist scalars μ_i , $i = 1, \dots, l$, satisfying

$$\begin{aligned} 0 \in \text{cl co} & \left[\nabla_x F_0 + \sum_{j \in J_0^-} \bigcup_{y_j^* \in P_j(\bar{x})} Z_{j0}(\bar{x}, y_j^*, u_{j0}^*(y_j^*)) \right. \\ & + \sum_{j \in J_0^+} Z_{j0}(\bar{x}, y_{j0}^*, K_j(\bar{x})) \\ & + \sum_{i=1}^l \mu_i \left\{ \nabla_x F_i + \sum_{j \in J_i^-} \bigcup_{y_j^* \in P_j(\bar{x})} Z_{ji}(\bar{x}, y_j^*, u_{ji}^*(y_j^*)) \right. \\ & \quad \left. \left. + \sum_{j \in J_i^+} Z_{ji}(\bar{x}, y_{ji}^*, K_j(\bar{x})) \right\} \right], \end{aligned}$$

$$\mu_i F_i(\bar{x}, v(\bar{x})) = 0, \quad \mu_i \geq 0, \quad i = 1, \dots, l.$$

Proof. This theorem follows from Theorem 3.1 and Proposition 4.3. \square

The following corollary is an immediate consequence of Theorem 4.1.

COROLLARY 4.1. *Under the same assumptions as in Theorem 4.1, if \bar{x} is an optimal solution of (TOP) then, for arbitrarily fixed points $u_{ji}^* \in K_j(\bar{x})$, $j \in J_i^-$, and $y_{ji}^* \in P_j(\bar{x})$, $j \in J_i^+$, $i = 0, \dots, l$, there exist scalars μ_i , $i = 1, \dots, l$ satisfying*

$$0 \in \text{co} \left[\nabla_x F_0 + \sum_{j \in J_0^-} Z_{j0}(\bar{x}, P_j(\bar{x}), u_{j0}^*) + \sum_{j \in J_0^+} Z_{j0}(\bar{x}, y_{j0}^*, K_j(\bar{x})) \right. \\ \left. + \sum_{i=1}^l \mu_i \left\{ \nabla_x F_i + \sum_{j \in J_i^-} Z_{ji}(\bar{x}, P_j(\bar{x}), u_{ji}^*) + \sum_{j \in J_i^+} Z_{ji}(\bar{x}, y_{ji}^*, K_j(\bar{x})) \right\} \right], \\ \mu_i F_i(\bar{x}, v(\bar{x})) = 0, \quad \mu_i \geq 0, \quad i = 1, \dots, l.$$

REFERENCES

- [1] V. F. DEMYANOV AND A. M. RUVINOV, *On quasidifferentiable functionals*, Soviet Math. Dokl., 21 (1980), pp. 14-17.
- [2] V. F. DEMYANOV AND L. N. POLYAKOVA, *Minimization of a quasi-differentiable function in a quasi-differentiable set*, U.S.S.R. Comput. Math. and Math. Phys., 20 (1981), pp. 34-43.
- [3] A. V. Fiacco, *Optimal value continuity and differential stability bounds under the Mangasarian-Fromovitz constraint qualification*, in Mathematical Programming with Data Perturbations II, A. V. Fiacco, ed., Marcel Dekker, New York, 1983, pp. 65-90.
- [4] J. GAUVIN AND F. DUBEAU, *Differentiable properties of the marginal function in mathematical programming*, Math. Programming Stud., 19 (1982), pp. 101-109.
- [5] Y. ISHIZUKA, *Farkas' theorem of nonconvex type and its application to a min-max problem*, J. Optim. Theory Appl., 57 (1988), pp. 341-354.
- [6] R. JANIN, *Directional derivative of the marginal function in nonlinear programming*, Math. Programming Stud., 21 (1984), pp. 110-126.
- [7] B. N. PSHENICHNYI, *Necessary Conditions for an Extremum*, Marcel Dekker, New York, 1971.
- [8] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [9] A. SHAPIRO, *On optimality conditions in quasidifferentiable optimization*, SIAM J. Control Optim., 22 (1984), pp. 611-617.
- [10] K. SHIMIZU AND Y. ISHIZUKA, *Optimality conditions and algorithms for parameter design problems with two-level structure*, IEEE Trans. Automat. Control., AC-30 (1985), pp. 986-993.
- [11] T. TANINO AND T. OGAWA, *An algorithm for solving two-level convex optimization problems*, Internat. J. Systems Sci., 15 (1984), pp. 163-174.

A DIGITAL PI-CONTROLLER FOR DISTRIBUTED PARAMETER SYSTEMS*

TOSHIHIRO KOBAYASHI

Abstract. In this paper a discrete-time multivariable tuning PI-controller (proportional controller plus integral controller) is designed for a stable distributed parameter system with an unbounded input and output operator. First, the systems with discrete-time controls and observations are analyzed. Next, it is shown that the integral feedback matrix has the same form as that in the case of continuous-time controls, and it can be determined by steady-state information only. Moreover, it is shown that continuous-time set-point tracking is realized as the result of discrete-time set-point tracking. The theory is applied to the system governed by a second-order evolution equation with a first-order damping term. It is shown that the integral feedback matrix is independent of the damping operator. Last, the theory is applied to a heat equation and an Euler-Bernoulli beam equation.

Key words. distributed parameter system, flexible structure, digital PI-controller, unbounded control and observation

AMS(MOS) subject classifications. 35B37, 35L15, 73K05, 93C20, 93C55, 93D15

1. Introduction. From the practical point of view, the purpose of control is not always to find an optimal control, but to find, on the basis of measurements, a control that stabilizes and regulates the system so that it behaves in a robust way against both external and internal perturbations. In addition, the controller should be tuned with minimum information about the process, for example, measuring only step responses.

A solution to this control problem in finite-dimensional theory has been given by Davison, who introduced multivariable robust PI-controller (proportional controller plus integral controller) theory. The tuning of the analogue controller for an unknown system has been partly solved in [2], where an algorithm is presented to tune the controller's integral part. The tuning of the proportional part seems still to be an open problem and some preliminary results have been reported in [8], [10], and [11]. In [10] and [11], digital robust PI-controller theories also have been investigated. For infinite-dimensional stable systems, Pohjolainen has given a solution of the continuous-time control problem [9].

In this paper the theory of a digital multivariable PI-controller is presented for an exponentially stable distributed parameter system with boundary or pointwise controls and observations. The theory is applied to flexible structures, that is, systems governed by second-order evolution equations with damping terms.

2. System description and problem formulation. Let us consider the system described by a first-order evolution equation on a complex Banach space X :

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + Ew, & x(0) &= x_0 \in X, \\ (1) \quad y(t) &= Cx(t), & t > 0, \\ e(t) &= y(t) - y_d, \end{aligned}$$

where $x(t)$ is the system state, $u(t) \in E^p$ is the control input, and $u(t)$ is assumed to be Hölder continuous, where E^p is a p -dimensional complex Euclidean space. The error $e(t)$ is the difference between the constant reference signal y_d and the output $y(t) \in E^r$, and $w \in E^q$ is an unknown constant disturbance. The operator A is the generator of an exponentially stable semigroup $U(t)$ on X such that $\|U(t)\| \leq M e^{-\omega t}$

* Received by the editors March 3, 1986; accepted for publication (in revised form) November 25, 1987.

† Department of Control Engineering, Kyushu Institute of Technology, Tobata, Kitakyushu 804, Japan.

for constants $M \geq 1$, $\omega > 0$. For the present we assume that B is a bounded linear operator from E^p to X (we denote $B \in L(E^p, X)$) and $E \in L(E^q, X)$, $C \in L(X, E')$. When the operator A generates an analytic semigroup, we can take the operator C such that $D(C) \supset D(A)$ and C is A -bounded [9].

A multivariable PI-controller for (1) has the form

$$(2) \quad u(t) = Pe(t) + Kz(t), \quad \dot{z}(t) = e(t),$$

where the constant matrices P and K are the gains of the proportional part and the integral part of the controller, respectively. The basic role of the integral part is to change the system steady state, so that as soon as the new steady state is achieved, the regulation $y(t) \rightarrow y_d$ will also occur. The purpose of the proportional part is to speed up the system response, if possible.

Another useful feature is robustness, defined by Davison as follows.

DEFINITION OF ROBUSTNESS. Given the system (1), suppose that there exists a multivariable PI-controller (2) that regulates the system (1). Let the plant parameters A , B , and C be perturbed so that the closed-loop system remains stable. If the regulation still occurs in the presence of such perturbations, the controller (2) is said to be a robust multivariable PI-controller.

Without exact knowledge of the parameters A , B , C , and E , for the system (1) we consider the following problem.

Problem 1. It is desired to find a robust multivariable PI-controller (2) for the system (1) so that output regulation will occur for all constant reference signals y_d and all constant disturbances w , and for all bounded linear operators E such as $E \in L(E^q, X)$.

The basic solution for Problem 1 has been obtained by Pohjolainen [9] in the case where A generates an exponentially stable analytic semigroup. He has constructed a robust multivariable integral-controller such that

$$(3) \quad u(t) = \varepsilon Kz(t),$$

where ε is a positive tuning parameter and the integral-feedback matrix K is given by

$$(4) \quad \begin{aligned} K &= (CA^{-1}B)^* \\ &= (CA^{-1}B)^T[(CA^{-1}B)(CA^{-1}B)^T]^{-1}. \end{aligned}$$

He also shows that there exists $\varepsilon^* > 0$ so that the closed-loop system is exponentially stable if ε in (3) satisfies $0 < \varepsilon < \varepsilon^*$.

Remark 1. $CA^{-1}B = -G(0)$, where $G(s)$ is the transfer function matrix of the system (1). The value $G(0)$ is experimentally determined on the basis of step responses [2], [9].

3. Analysis of discrete-time control systems. For the system (1), assume that the control $u(t)$ are constant vectors on each sampling period $[kT, (k+1)T)$, that is,

$$(5) \quad u(t) = u_k, \quad kT \leq t < (k+1)T, \quad k = 0, 1, \dots$$

where T is a sampling period. Then the system (1) has the unique mild solution such that $x \in C([0, t_1]; X)$ for any finite time $t_1[1]$. Since $U(t)$ is a strongly continuous semigroup on X , the system state $x_k = x(kT)$ are given at each sampling times as follows:

$$(6) \quad x_{k+1} = U(T)x_k + \int_0^T U(T-t)(Bu_k + Ew) dt, \quad k = 0, 1, \dots$$

Thus we have the discrete-time system on X :

$$(7) \quad \begin{aligned} x_{k+1} &= U(T)x_k + B_U u_k + w_U, & x_0 &\in X, \\ y_k &= Cx_k, \\ e_k &= y_k - y_d, & z_{k+1} - z_k &= e_k, \end{aligned}$$

where $B_U u_k = \int_0^T U(T-t) B u_k dt$, $w_U = \int_0^T U(T-t) E w dt$, $y_k = y(kT)$.

Since A is exponentially stable, $U(T)$ is also stable, that is, the spectral radius $r(U(T))$ of $U(T)$ is strictly less than 1 and the condition

$$(8) \quad \|U(T)^k\| = \|U(kT)\| \leq M(\exp(-\omega T))^k$$

holds.

We shall investigate the system state x_k and the operator B_U . Since for $B \in L(E^p, X)$ and $u \in E^p$,

$$(9) \quad \begin{aligned} B_U u &= \int_0^T U(T-t) B u dt \\ &= \int_0^T U(T-t) A A^{-1} B u dt \\ &= \int_0^T A U(T-t) A^{-1} B u dt \\ &= - \int_0^T \frac{d}{dt} U(T-t) A^{-1} B u dt \\ &= [U(T) - I] A^{-1} B u, \end{aligned}$$

we have $B_U \in L(E^p, D(A))$ and we also have $w_U \in D(A)$. Here we assume the existence of a Banach space $Y (\supset D(A))$, dense in X , such that [1, Chap. 8]

$$(A1) \quad U(t)x \in Y \quad \text{for } x \in X, \quad t > 0.$$

Then $x_{k+1} \in Y$, $k=0, 1, \dots$. Thus we can take the output operator C such that $C \in L(Y, E^r)$ and may treat boundary or pointwise outputs.

In order to investigate B_U when $B: E^p \rightarrow X$ is unbounded, we assume the existence of a Banach space W , with X dense in W , such that

$$(A2) \quad \begin{aligned} (a) \quad & W \supset R(B) \supset X, \\ (b) \quad & B \in L(E^p, W), \\ (c) \quad & U(t) \in L(W, X), \quad t > 0, \\ (d) \quad & \|U(t)w\|_X \leq g(t)\|w\|_W \quad \text{for all } w \in W \quad \text{with } g \in L^1(0, 1)[1]. \end{aligned}$$

Then

$$\begin{aligned} \|B_U u\|_X &= \left\| \int_0^T U(T-t) B u dt \right\|_X \\ &\leq \int_0^T g(T-t) \|B\|_{L(E^p, W)} \|u\|_{E^p} dt \\ &\leq \|g\|_{L^1(0, T)} \|B\|_{L(E^p, W)} \|u\|_{E^p}. \end{aligned}$$

Thus we obtain that $B_U \in L(E^p, X)$. We can apply the same discussion to the operator E . From this we can also treat boundary or pointwise inputs and disturbances. Even if the input operator B is unbounded for the continuous-time case, B_U may be bounded for the discrete-time case.

Next consider the case where $U(t)$ is an analytic semigroup. Since from (9) we have $B_U \in L(E^p, D(A))$ and $w_U \in D(A)$ for $B \in L(E^p, X)$ and $E \in L(E', X)$, it follows from (7) that $x_{k+1} \in D(A)$, $k=0, 1, \dots$. Thus we can take the operator C such that CA^{-1} is bounded on X . We can treat boundary or pointwise outputs.

Moreover, in order to obtain a more precise estimation, we shall use the fractional powers $(-A)^\gamma$, $0 < \gamma < 1$, which are well defined, since A generates an exponentially stable analytic semigroup [13]. For this we assume the existence of an operator G such that for $u \in E^p$, $Gu \in D((-A)^{1-\gamma})$. The existence of such an operator G is investigated in the cases of boundary and pointwise inputs [3], [14]. Then G is the Green mapping of a related boundary value problem, and

$$\begin{aligned} B_U u &= \int_0^T (-A)U(T-t)Gu \, dt \\ (10) \quad &= \int_0^T (-A)^\gamma U(T-t)(-A)^{1-\gamma}Gu \, dt \end{aligned}$$

holds. From this we have

$$\begin{aligned} B_U u &= \int_0^T (-A)^{\gamma-1}(-A)U(T-t)(-A)^{1-\gamma}Gu \, dt \\ (11) \quad &= (-A)^{\gamma-1}[I - U(T)](-A)^{1-\gamma}Gu \end{aligned}$$

since $(-A)^{\gamma-1} \in L(X)$ and $U(t)(-A)^{1-\gamma}Gu \in D(-A)$ for $0 < \gamma < 1$. It follows from (11) that $B_U \in L(E^p, D((-A)^{1-\gamma}))$. Moreover, if $w_U \in D((-A)^{1-\gamma})$, we have that $x_{k+1} \in D((-A)^{1-\gamma})$, $k=0, 1, \dots$ and we can take the operator C such that $C \in L(D((-A)^{1-\gamma}), E')$. Thus we can treat boundary or pointwise inputs and outputs.

Last, consider a more special class of the operator A . Let V and H be complex Hilbert spaces such that V , H , and V' (the dual space of V) satisfy the inclusion relation $V \subset H \subset V'$ with each space dense in the following with continuous injection. Let A be in $L(V, V')$, satisfying the coercive condition, that is,

$$(12) \quad \operatorname{Re}(-Ax, x) \geq \alpha \|x\|_V^2, \quad \alpha > 0, \quad x \in V.$$

Then the operator A generates an exponentially stable analytic semigroup $U(t)$ on H and V' , respectively [12].

In this case, for $B \in L(E^p, V')$, $E \in L(E^q, V')$ and $x_0 \in H$, there exists a unique solution x of the system (1) with discrete-time inputs (5) and $x \in L^2(0, t_1; V) \cap C([0, t_1]; H)$ for any finite t_1 [7]. We can treat boundary or pointwise inputs and disturbances, since we can take B and E such that $B \in L(E^p, V')$ and $E \in L(E^q, V')$ [7].

Moreover we shall investigate the system state x_k . Define

$$(13) \quad v_k = A^{-1}(Bu_k + Ew)$$

and then $v_k \in V$ for $Bu_k + Ew \in V'$ [12]. Now $x_v(t) = x(t) + v_k$ satisfies

$$\begin{aligned} \frac{dx_v(t)}{dt} &= \frac{dx(t)}{dt} \\ (14) \quad &= Ax(t) + Bu_k + Ew \\ &= Ax(t) + Av_k \\ &= Ax_v(t) \end{aligned}$$

on the interval $(kT, (k+1)T)$. Hence we have

$$x_v((k+1)T) = U(T)(x_k + v_k).$$

Since $x_k + v_k \in H$, $x_v((k+1)T) \in D(A) \subset V$. As v_k is also in V , $x_{k+1} = x_v((k+1)T) - v_k$ is in V . From this we also obtain that $U(T) \in L(H, V)$, $B_U \in L(E^p, V)$ and $w_U \in V$ for $B \in L(E^p, V')$ and $E \in L(E^q, V')$. Thus we can take the operator C whose domain is V . Therefore we can also treat boundary or pointwise outputs [7].

Remark 2. If $-A$ is a self-adjoint, positive-definite operator on X with compact resolvent, then there exist a sequence $\{\lambda_n, \phi_n; n=1, 2, \dots\}$ of eigenvalues and the corresponding orthonormal eigenvectors such that

$$0 \leq \lambda_1 \leq \lambda_2 \leq \dots$$

and

$$-A\phi_n = \lambda_n\phi_n, \quad n=1, 2, \dots$$

In this case we obtain the fractional powers of $-A$:

$$(-A)^\gamma x = \sum_{n=1}^{\infty} (\lambda_n)^\gamma (x, \phi_n) \phi_n, \quad x \in D((-A)^\gamma),$$

$$D((-A)^\gamma) = \left\{ x \left| \sum_{n=1}^{\infty} (\lambda_n)^{2\gamma} |(x, \phi_n)|^2 < \infty \right. \right\}$$

for real γ .

For example, we can consider Y in (A1) and $D((-A)^{1/2})$ as V , and W in (A2) as V' .

4. Solution of Problem 1 by discrete-time controls. In this section we shall solve Problem 1 using only discrete-time input-output data.

First we consider the following problem for the discrete-time system (7).

PROBLEM 2. Find a robust multivariable PI-controller for the discrete-time system (7) so that output regulation occurs, that is, $y_k \rightarrow y_d$ as $k \rightarrow \infty$ for all constant reference signals y_d and for all constant disturbances w .

We obtain from (7)

$$(15) \quad \begin{pmatrix} e_k \\ x_{k+1} - x_k \end{pmatrix} = \begin{pmatrix} I & C \\ 0 & U(T) \end{pmatrix} \begin{pmatrix} e_{k-1} \\ x_k - x_{k-1} \end{pmatrix} + \begin{pmatrix} 0 \\ B_U \end{pmatrix} (u_k - u_{k-1}).$$

For this system we consider a linear feedback control law

$$(16) \quad u_k = \alpha \varepsilon K e_k + \varepsilon K z_k,$$

that is,

$$(17) \quad \begin{aligned} u_k - u_{k-1} &= \alpha \varepsilon K (e_k - e_{k-1}) + \varepsilon K (z_k - z_{k-1}) \\ &= \alpha \varepsilon K C (x_k - x_{k-1}) + \varepsilon K e_{k-1}, \end{aligned}$$

where K is the $p \times r$ integral-feedback matrix, which is also the proportional one, and α, ε are positive tuning parameters. The resultant closed-loop system becomes

$$(18) \quad \begin{aligned} \begin{pmatrix} e_k \\ x_{k+1} - x_k \end{pmatrix} &= \begin{pmatrix} I & C \\ 0 & U(T) \end{pmatrix} + \varepsilon \begin{pmatrix} 0 & 0 \\ B_U K & \alpha B_U K C \end{pmatrix} \begin{pmatrix} e_{k-1} \\ x_k - x_{k-1} \end{pmatrix} \\ &= (U_0 + \varepsilon B_K) \begin{pmatrix} e_{k-1} \\ x_k - x_{k-1} \end{pmatrix} \end{aligned}$$

on the extended state space $X_e = E^r \times X$.

Now the following theorem gives the solution to Problem 2.

THEOREM 1. Assume that $\text{rank}[C(I - U(T))^{-1}B_U] = r$. Then there is a positive number ε^* , dependent of K , and a regulating control of the form (16) for all $p \times r$ matrices K satisfying $\sigma\{C[I - U(T)]^{-1}B_U K\} \subset C^-$ and for all $\varepsilon \in (0, \varepsilon^*)$. A good selection of K is

$$(19) \quad K = -[C(I - U(T))^{-1}B_U]^* \text{diag}\{p_1, \dots, p_r\}, \quad p_i > 0.$$

Proof. To prove the theorem, we proceed as in the proof of Theorem 3.4 in [9].

The stability of the closed-loop system (18) can be seen from the behavior of the spectrum of U_0 on the unit circle, because the rest of the system remains in the unit circle and stable for sufficiently small values of ε .

Since the spectrum of U_0 is separated into two parts such that $\sigma(U_0) = \sigma(U(T)) \cup \{1\}$, we can take a positively-oriented closed curve Γ , surrounding an open set containing the point $z = 1$ and enclosing $\sigma(U(T))$ in its exterior. Then the spectrum decomposition assumption holds for the perturbed operator $U_0 + \varepsilon B_K$ as far as

$$(20) \quad 0 \leq \varepsilon < \min_{\lambda \in \Gamma} \{\|B_K\|^{-1} \|R(\lambda; U(T))\|^{-1}\},$$

where $R(\lambda; U(T)) = [\lambda I - U(T)]^{-1}$ is the resolvent. The projection, corresponding to the decomposition, is given by

$$(21) \quad P(\varepsilon) = \frac{1}{2\pi i} \int_{\Gamma} R(\lambda; U_0 + \varepsilon B_K) d\lambda.$$

For the values of ε satisfying (20) and $\lambda \in \Gamma$, the resolvent can be expressed as a converging Neumann series

$$(22) \quad \begin{aligned} R(\lambda; U_0 + \varepsilon B_K) &= R(\lambda; U_0)[I + \varepsilon B_K R(\lambda; U_0)]^{-1} \\ &= R(\lambda; U_0) \sum_{n=0}^{\infty} \varepsilon^n [B_K R(\lambda; U_0)]^n. \end{aligned}$$

Since the series converges uniformly for $\lambda \in \Gamma$ and the terms depend continuously on λ , we can integrate it termwise to have the corresponding series for $P(\varepsilon)$. This gives

$$(23) \quad \begin{aligned} P(\varepsilon) &= \frac{1}{2\pi i} \int_{\Gamma} R(\lambda; U_0) \sum_{n=0}^{\infty} \varepsilon^n [B_K R(\lambda; U_0)]^n d\lambda \\ &= \sum_{n=0}^{\infty} \varepsilon^n \frac{1}{2\pi i} \int_{\Gamma} R(\lambda; U_0) [B_K R(\lambda; U_0)]^n d\lambda \\ &= \sum_{n=0}^{\infty} \varepsilon^n P^{(n)}, \end{aligned}$$

where $P^{(n)} = 1/2\pi i \int_{\Gamma} R(\lambda; U_0) [B_K R(\lambda; U_0)]^n d\lambda \in L(X_e)$.

The series is convergent for ε satisfying (20) so that $P(\varepsilon)$ is analytic near $\varepsilon = 0$. The range of $P(\varepsilon)$ is isomorphic with the eigenspace $P(0)X_e$ of U_0 for the eigenvalues at the point $z = 1$ [4]. Perturbations of the eigenvalues inside the closed curve Γ may now be computed as the nonzero eigenvalues of a finite-dimensional operator $P(\varepsilon) - (U_0 + \varepsilon B_K - I)P(\varepsilon)$ [4]. We can represent this operator as follows:

$$\begin{aligned} P(\varepsilon)(U_0 + \varepsilon B_K - I)P(\varepsilon) &= (U_0 + \varepsilon B_K - I)P(\varepsilon) \\ &= \frac{1}{2\pi i} \int_{\Gamma} (U_0 + \varepsilon B_K - I)R(\lambda; U_0 + \varepsilon B_K) d\lambda \\ &= \frac{1}{2\pi i} \int_{\Gamma} (\lambda - 1)R(\lambda; U_0 + \varepsilon B_K) d\lambda \end{aligned}$$

because $(U_0 + \varepsilon B_K - I)R(\lambda; U_0 + \varepsilon B_K) = I + (\lambda - 1)R(\lambda; U_0 + \varepsilon B_K)$ and the corresponding integral of I on Γ is zero by the Cauchy integral theorem. If we apply (22) as in the previous formula, we shall have a uniformly convergent power series with continuous elements for all ε satisfying (20):

$$(24) \quad (U_0 + \varepsilon B_K - I)P(\varepsilon) = \sum_{n=0}^{\infty} \varepsilon^n U^{(n)},$$

where $U^{(n)} = 1/2\pi i \int_{\Gamma} (\lambda - 1)R(\lambda; U_0)[B_K R(\lambda; U_0)]^n d\lambda \in L(X_e)$ for all n . Let us compute the first term in the series:

$$(25) \quad \begin{aligned} U^{(0)} &= \frac{1}{2\pi i} \int_{\Gamma} (\lambda - 1)R(\lambda; U_0) d\lambda \\ &= \frac{1}{2\pi i} \int_{\Gamma} (\lambda - 1) \begin{bmatrix} \frac{1}{\lambda - 1} & \frac{1}{\lambda - 1} CR(\lambda; U(T)) \\ 0 & R(\lambda; U(T)) \end{bmatrix} d\lambda \\ &= \begin{bmatrix} \frac{1}{2\pi i} \int_{\Gamma} I d\lambda & \frac{1}{2\pi i} \int_{\Gamma} CR(\lambda; U(T)) d\lambda \\ 0 & \frac{1}{2\pi i} \int_{\Gamma} (\lambda - 1)R(\lambda; U(T)) d\lambda \end{bmatrix} \\ &= 0. \end{aligned}$$

The integrals are zero because all the operators are analytic inside Γ and continuous in Γ .

The second term becomes

$$(26) \quad \begin{aligned} U^{(1)} &= \frac{1}{2\pi i} \int_{\Gamma} (\lambda - 1)R(\lambda; U_0)B_K R(\lambda; U_0) d\lambda \\ &= \frac{1}{2\pi i} \int_{\Gamma} \begin{bmatrix} \frac{1}{\lambda - 1} CR(\lambda; U(T))B_U K & \frac{1}{\lambda - 1} CR(\lambda; U(T))B_U KCR(\lambda; U(T)) \\ R(\lambda; U(T))B_U K & R(\lambda; U(T))B_U KCR(\lambda; U(T)) \end{bmatrix} d\lambda \\ &= \begin{bmatrix} C[I - U(T)]^{-1}B_U K & C[I - U(T)]^{-1}B_U K C[I - U(T)]^{-1} \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

In this term the integrals of analytic operators disappear and the remaining integrals are computed by using Cauchy integral formula. Thus

$$(27) \quad (U_0 + \varepsilon B_K - I)P(\varepsilon) = \sum_{n=1}^{\infty} \varepsilon^n U^{(n)}.$$

The perturbation of the eigenvalues for small ε is dominated by the first term $\varepsilon U^{(1)}$. The eigenvalues of $U^{(1)}$ in the invariant subspace $R(P(0))$ are those of $C[I - U(T)]^{-1}B_U K$ by (26). Thus the perturbed eigenvalues of $(U_0 + \varepsilon B_K)P(\varepsilon)$ are approximately $\lambda'_i \approx 1 + \varepsilon \lambda_i$, where $\lambda_i \in \sigma(C[I - U(T)]^{-1}B_U K)$ for sufficiently small values of ε . If we select K such that $\sigma(C[I - U(T)]^{-1}B_U K) \subset C^-$, which is permissible because $\text{rank}(C[I - U(T)]^{-1}B_U) = r$, the eigenvalues of $U_0 + \varepsilon B_K$ will go into the unit circle. If we take

$$K = -C\{[I - U(T)]^{-1}B_U\}^* \text{diag}\{p_1, \dots, p_r\}, \quad p_i > 0 \ (i = 1, \dots, r),$$

then $\lambda'_i \approx 1 - \varepsilon p_i$ ($i = 1, \dots, r$) and they will depart with the derivatives $\partial \lambda'_i(0)/\partial \varepsilon = -p_i$. This implies that we can design the derivatives by choosing the parameters p_1, \dots, p_r . In this sense the K given by (19) is a good selection.

Since all the eigenvalues of $U_0 + \varepsilon B_K$ are in the unit circle, the feedback system (18) becomes stable. Thus we have proved the theorem.

We can show that the condition $\text{rank} \{C[I - U(T)]^{-1} B_U\} = r$ is a necessary condition for the existence of a stabilizing control of the form (17).

The system (15) is stabilizable with bounded state feedback if and only if the modes corresponding to the eigenvalues at the point $z = 1$ are controllable. This is possible if and only if

$$(28) \quad \text{rank} \left[P(U_0 - I)P, P \begin{pmatrix} 0 \\ B_U \end{pmatrix} \right] = r,$$

where P is the projector corresponding to the part of the system at the point $z = 1$. The projector is given by

$$P = \frac{1}{2\pi i} \int_{\Gamma} R(\lambda; U_0) d\lambda = \begin{pmatrix} I & C[I - U(T)]^{-1} \\ 0 & 0 \end{pmatrix},$$

where Γ is the circle surrounding the point $z = 1$. Thus we have

$$\begin{aligned} \text{rank} \left(P(u_0 - I)P, P \begin{pmatrix} 0 \\ B_U \end{pmatrix} \right) &= \text{rank} \left(0, \begin{pmatrix} C[I - U(T)]^{-1} B_U \\ 0 \end{pmatrix} \right) \\ &= \text{rank} \{C[I - U(T)]^{-1} B_U\} = r, \end{aligned}$$

which proves the necessity.

Remark 3. In (19), $C[I - U(T)]^{-1} B_U = \hat{G}(1)$, where $\hat{G}(z)$ is the pulse transfer function of the system (7).

Since A generates an exponentially stable semigroup, there exists the inverse $A^{-1} \in L(X, D(A))$. For $x \in X$

$$\begin{aligned} \int_0^t U(t)x dt &= \int_0^t U(t)AA^{-1}x dt \\ &= [U(t) - I]A^{-1}x \end{aligned}$$

holds. Thus we have

$$\begin{aligned} C[I - U(T)]^{-1} B_U u &= C[I - U(T)]^{-1} \int_0^T U(t)Bu dt \\ &= C[I - U(T)]^{-1} [U(T) - I]A^{-1}Bu = -CA^{-1}Bu, \end{aligned}$$

if $CA^{-1}B$ is well defined. Here, for B_U given by (10), we must replace $-CA^{-1}B$ by CG , but the results are same. When we select K as (19), we obtain

$$(29) \quad \begin{aligned} K &= -[C(I - U(T))^{-1} B_U]^* \text{diag} \{p_1, \dots, p_r\} \\ &= (CA^{-1}B)^* \text{diag} \{p_1, \dots, p_r\}, \end{aligned}$$

which is the same as the feedback matrix K in (4) for the continuous-time case.

Basic knowledge for tuning the I -controller is given by the matrix $CA^{-1}B$. This matrix can be measured from the process with the aid of the following algorithm.

i
↓

ALGORITHM 1. Select the input u_k such that $u_k = u_i = [0, \dots, 0, 1, 0, \dots, 0]^T \in E^p$ for all k . Measure the output y_{ik} . Iterate for $i = 1, \dots, p$.

In this case we have

$$\begin{aligned}
 y_{ik} &= Cx_k = CU^k(T)x_0 + C \sum_{j=0}^{k-1} U^j(T)B_U u_i \\
 &= CU^k(T)x_0 + C[I - U^k(T)][I - U(T)]^{-1}B_U u_i \\
 (30) \quad &= CU^k(T)x_0 + C[I - U^k(T)][I - U(T)]^{-1}[U(T) - I]A^{-1}Bu_i \\
 &= CU^k(T)x_0 + C[U^k(T)A^{-1}Bu_i - A^{-1}Bu_i] \\
 &\rightarrow -CA^{-1}Bu_i
 \end{aligned}$$

as $k \rightarrow \infty$, since (8) holds. On the limit the measured vector y_{ik} will be the i th column of the matrix $-CA^{-1}B$. Thus if we select the inputs as $u_k \equiv u_i$, $i = 1, \dots, p$ and measure the corresponding steady-state outputs y_i , we can determine the matrix $-CA^{-1}B = [y_1, y_2, \dots, y_p]$. This implies that the integral feedback matrix K can be determined only by steady-state information.

Next let us investigate the response between sampling periods. For $0 \leq m \leq 1$

$$\begin{aligned}
 x_{k,m} &= x(kT + mT) \\
 (31) \quad &= U(mT)x_k + \int_0^{mT} U(mT-t)(Bu_k + Ew) dt \\
 &= U(mT)x_k + \int_0^{mT} U(t)(Bu_k + Ew) dt.
 \end{aligned}$$

Putting $x^* = \lim_{k \rightarrow \infty} x_k$, $u^* = \lim_{k \rightarrow \infty} u_k$, we have from (7)

$$\begin{aligned}
 (32) \quad x^* &= U(T)x^* + B_U u^* + w_U \\
 &= U(T)x^* + \int_0^T U(t)(Bu^* + Ew) dt,
 \end{aligned}$$

$$(33) \quad y^* = Cx^* = y_d.$$

From (32) we obtain

$$[U(T) - I]x^* + [U(T) - I]A^{-1}(Bu^* + Ew) = 0.$$

Since $\text{Null}[U(T) - I] = \{0\}$, we have

$$(34) \quad x^* + A^{-1}(Bu^* + Ew) = 0.$$

Using this, we obtain from (31)

$$\begin{aligned}
 (35) \quad \lim_{k \rightarrow \infty} x_{k,m} &= U(mT)x^* + \int_0^{mT} U(t)(Bu^* + Ew) dt \\
 &= U(mT)x^* + [U(mT) - I]A^{-1}(Bu^* + Ew) \\
 &= x^*
 \end{aligned}$$

and

$$(36) \quad \lim_{k \rightarrow \infty} y(kT + mT) = \lim_{k \rightarrow \infty} Cx(kT + mT) = Cx^* = y_d,$$

which shows continuous-time output regulation. This is also shown in the case in which B_U and w_U are given, such as (10).

The continuous-time closed-loop system is stable, if the corresponding discrete-time one (18) is stable [5].

We have obtained the following theorem, which gives the solution of Problem 1.

THEOREM 2. *For Problem 1, there is a positive number ε^* , dependent on K , and a robust discrete-time stabilizing control of the form (16) for all $p \times r$ matrices K satisfying $\sigma(CA^{-1}BK) \subset C^+$ and for all ε ($0 < \varepsilon < \varepsilon^*$), if and only if $\text{rank}(CA^{-1}B) = r$. A good selection of K is*

$$(37) \quad K = (CA^{-1}B)^* \text{diag}\{p_1, \dots, p_r\}, \quad p_i > 0.$$

5. Flexible structures. In this section the theory is applied to flexible structures, that is, the generalized oscillating systems with damping.

Consider the following oscillating system on a Hilbert space H :

$$(38) \quad \ddot{z} + 2\zeta\Lambda\dot{z} = Az + Bu + Ew, \quad y = Cz,$$

where $-A$ and Λ are self-adjoint operators with

$$(39) \quad \begin{aligned} (-Az, z) &\geq a\|z\|^2, \quad a > 0, \quad z \in D(A), \\ (\Lambda\dot{z}, \dot{z}) &\geq b\|\dot{z}\|^2, \quad \dot{z} \in D(\Lambda) \supset D(A) \end{aligned}$$

for some real number b . Moreover $B: E^p \rightarrow H$, $C: E^q \rightarrow H$, $E: H \rightarrow E^r$, and ζ is a positive constant.

Second-order systems of the type (38) are useful in applications. Since $\sqrt{-A}$ is well defined and $D(\sqrt{-A})$ is a Hilbert space with the graph norm [13], we can write the system (38) as a first-order system on a Hilbert space $X = D(\sqrt{-A}) \times H$:

$$(40) \quad \begin{aligned} \frac{d}{dt} \begin{pmatrix} z \\ \dot{z} \end{pmatrix} &= \begin{pmatrix} 0 & I \\ A & -2\zeta\Lambda \end{pmatrix} \begin{pmatrix} z \\ \dot{z} \end{pmatrix} + \begin{pmatrix} 0 \\ B \end{pmatrix} u + \begin{pmatrix} 0 \\ E \end{pmatrix} w, \\ y &= [C \quad 0] \begin{pmatrix} z \\ \dot{z} \end{pmatrix}. \end{aligned}$$

By Theorem 2.14 in [1] it is seen that \bar{A}

$$\bar{A} = \begin{pmatrix} 0 & I \\ A & -2\zeta\Lambda \end{pmatrix}, \quad D(\bar{A}) = D(A) \times D(\Lambda)$$

generates a strongly continuous semigroup $\bar{U}(t)$ on X , and for $b > 0$, $\bar{U}(t)$ is exponentially stable. Thus when $\Lambda = I$, $\Lambda = -\sqrt{-A}$, and $\Lambda = -A$, the operators \bar{A} generates exponentially stable semigroups $\bar{U}(t)$. Here we considered A to be a fourth-order differential operator.

Moreover, we can show that for $\Lambda = -\sqrt{-A}$ and $\Lambda = -A$, the operators \bar{A} generate analytic semigroups. These Λ have the bounded inverses Λ^{-1} . Introducing a new variable v such as $v = \dot{z} + 2\zeta\Lambda z$, we can reduce (38) to the first-order equation on $H \times H$:

$$(41) \quad \frac{d}{dt} \begin{pmatrix} v \\ \dot{z} \end{pmatrix} = \begin{pmatrix} \frac{1}{2\zeta} A\Lambda^{-1} & -\frac{1}{2\zeta} A\Lambda^{-1} \\ \frac{1}{2\zeta} A\Lambda^{-1} & -2\zeta\Lambda - \frac{1}{2\zeta} A\Lambda^{-1} \end{pmatrix} \begin{pmatrix} v \\ \dot{z} \end{pmatrix} + \begin{pmatrix} B \\ B \end{pmatrix} u + \begin{pmatrix} E \\ E \end{pmatrix} w.$$

The operator \tilde{A} ,

$$\tilde{A} = \begin{pmatrix} \frac{1}{2\zeta} A\Lambda^{-1} & -\frac{1}{2\zeta} A\Lambda^{-1} \\ \frac{1}{2\zeta} A\Lambda^{-1} & -2\zeta\Lambda - \frac{1}{2\zeta} A\Lambda^{-1} \end{pmatrix}, \quad D(\tilde{A}) = D(A\Lambda^{-1}) \times D(\Lambda)$$

can be viewed as

$$\tilde{A} = \begin{pmatrix} \frac{1}{2\zeta} A\Lambda^{-1} & 0 \\ \frac{1}{2\zeta} A\Lambda^{-1} & -2\zeta\Lambda \end{pmatrix} + \begin{pmatrix} 0 & -\frac{1}{2\zeta} A\Lambda^{-1} \\ 0 & -\frac{1}{2\zeta} A\Lambda^{-1} \end{pmatrix} = \tilde{A}_0 + \tilde{A}_p.$$

Since for $\Lambda = -\sqrt{-A}$ and $\Lambda = -A$, $A\Lambda^{-1}$ and $-\Lambda$ are generators of analytic semigroups, \tilde{A}_0 generates an analytic semigroup on $H \times H$ [6]. Moreover the operator $A\Lambda^{-1}$ is relatively bounded with respect to Λ . Thus the operator \tilde{A} generates an analytic semigroup on $H \times H$ [6]. Then the Cauchy problem for the system

$$(42) \quad \ddot{z} + 2\zeta\Lambda\dot{z} = Az$$

has a unique solution for $z_0 \in D(A)$ and $\dot{z}_0 \in D(\Lambda)$. Furthermore for $z_0 \in D(\Lambda)$ and $\dot{z}_0 \in H$, there exists a unique analytic weak solution in a sector that contains the positive real axis [6]. This implies that the operator \tilde{A} also generates analytic semigroups $\bar{U}(t)$ on X for $\Lambda = -\sqrt{-A}$ and $\Lambda = -A$, respectively.

Remark 4. When the operator $-A$ has a compact resolvent, we can easily seek $\sigma(\bar{A})$, the spectrum of \bar{A} . In the cases of $\Lambda = -\sqrt{-A}$ and $\Lambda = -A$, $\sigma(\bar{A})$ are contained in sectors $S(\omega, \beta) = \{\lambda \in \mathbb{C} \mid |\arg(\lambda - \beta)| < \pi/2 + \omega, \beta \neq \lambda\}$, where $\pi/2 > \omega > 0$, $\beta \in \mathbb{R}$. This implies that $\bar{U}(t)$ is an analytic semigroup.

Next, for the system (40), we consider the feedback matrix K . If the operator equation

$$\begin{pmatrix} 0 & I \\ A & -2\zeta\Lambda \end{pmatrix} \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} = \begin{pmatrix} 0 \\ B \end{pmatrix}$$

has a solution such that $G_1 = A^{-1}B$, $G_2 = 0$, we have

$$[C \ 0] \begin{pmatrix} 0 & I \\ A & -2\zeta\Lambda \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ B \end{pmatrix} = CA^{-1}B.$$

This is independent of the damping term $2\zeta\Lambda\dot{z}$ of (38) and, moreover, is independent of the form of the damping operator Λ . The feedback matrix K is still

$$(43) \quad K = (CA^{-1}B)^* \text{diag}\{p_1, \dots, p_r\}$$

for the second-order system (38). Here we can also use the corresponding Green mapping G in place of $-A^{-1}B$.

Remark 5. In the case where \bar{A} generates an exponentially stable analytic semigroup, although \bar{A} is not self-adjoint as in Remark 2, fractional powers of $-\bar{A}$ are expressed in terms of eigenvalues λ_n , eigenvectors ϕ_n , and the corresponding eigenvectors ψ_n of $-\bar{A}^*$, such that $(\phi_n, \psi_m)_X = \delta_{nm}$,

$$(-\bar{A})^\gamma x = \sum_{n=1}^{\infty} \lambda_n^\gamma (x, \psi_n) \phi_n, \quad x \in D((-\bar{A})^\gamma),$$

$$D((-\bar{A})^\gamma) = \left\{ x \mid \sum_{n=1}^{\infty} |\lambda_n|^{2\gamma} |(x, \psi_n)|^2 < \infty \right\}$$

for all real $-1 < \gamma < 1$.

6. Examples. In this section, we shall give simple examples to illustrate our theory.

Example 6.1. Let us consider the following heating system:

$$\begin{aligned}
 x_t(t, \eta) &= x_{\eta\eta}(t, \eta) + \delta(\eta - 0.1)[u_1(t) + w] + \delta(\eta - 0.9)u_2(t), \quad 0 < \eta < 1, \\
 x(t, 0) &= x(t, 1) = 0, \\
 y_k &= \begin{pmatrix} x(kT, 0.3) \\ x(kT, 0.7) \end{pmatrix}, \quad k = 0, 1, \dots,
 \end{aligned}
 \tag{44}$$

where w is an unknown constant. For this system we take $H = L^2(0, 1)$ and $Au = \Delta u$; Δ is the Laplacian with the Dirichlet conditions at $x = 0$ and $x = 1$.

For the pointwise controls u_1 and u_2 , the operator B is $B = [\delta(\eta - 0.1), \delta(\eta - 0.9)]$. Thus we can take $V = H_0^1(0, 1) = \{v \in L^2(0, 1) \text{ such that } dv/d\eta \in L^2(0, 1) \text{ and } v(0) = v(1) = 0\}$. Then $V' = H^{-1}(0, 1)$ [7], and $B \in L(E^2, H^{-1}(0, 1))$. The constant disturbance is $w\delta(\eta - 0.1) \in V'$.

In this case we have

$$(-Av, v) = - \int_0^1 \frac{d^2 v(\eta)}{d\eta^2} v(\eta) d\eta \cong \alpha \left\{ \int_0^1 \left[\frac{dv(\eta)}{d\eta} \right]^2 + \int_0^1 v^2(\eta) d\eta \right\},$$

where $\alpha = \pi^2/(1 + \pi^2)$, since $\int_0^1 v^2(\eta) d\eta \leq 1/\pi^2 \int_0^1 v_\eta^2(\eta) d\eta$. Thus $A \in L(V, V')$ and A generates an exponentially stable analytic semigroup $U(t)$ such that $\|U(t)\| \leq e^{-\pi^2 t}$. Moreover, since $x(kT, \cdot) \in V = H_0^1(0, 1) \subset C(0, 1)$ from Sobolev's Embedding Theorem [13], the pointwise observation y_k has a meaning and $CA^{-1}B$ is well defined.

Let us determine the matrix $CA^{-1}B$. Putting $A^{-1}B = [g(\eta), h(\eta)]$, we have

$$\begin{aligned}
 g_{\eta\eta}(\eta) &= \delta(\eta - 0.1), & g(0) &= g(1) = 0, \\
 h_{\eta\eta}(\eta) &= \delta(\eta - 0.9), & h(0) &= h(1) = 0.
 \end{aligned}
 \tag{45}$$

Solve (45). We can determine $g(\eta)$ from the conditions

$$\begin{aligned}
 g(\eta) &= \begin{cases} -a\eta, & 0 \leq \eta \leq 0.1, \\ b(\eta - 1), & 0.1 \leq \eta \leq 1, \end{cases} \\
 g(\eta)|_{0.1-} &= g(\eta)|_{0.1+}, & g_\eta(\eta)|_{0.1+} - g_\eta(\eta)|_{0.1-} &= 1.
 \end{aligned}$$

The solution $g(\eta)$ of (45) becomes

$$g(\eta) = \begin{cases} -0.9\eta, & 0 \leq \eta \leq 0.1, \\ 0.1(\eta - 1), & 0.1 \leq \eta \leq 1. \end{cases}$$

Similarly, we have

$$h(\eta) = \begin{cases} -0.1\eta, & 0 \leq \eta \leq 0.9, \\ 0.9(\eta - 1), & 0.9 \leq \eta \leq 1. \end{cases}$$

Thus we obtain the matrix $CA^{-1}B$ as

$$CA^{-1}B = \begin{pmatrix} g(0.3) & h(0.3) \\ g(0.7) & h(0.7) \end{pmatrix} = \begin{pmatrix} -0.07 & -0.03 \\ -0.03 & -0.07 \end{pmatrix}.$$

A suitable selection for the feedback matrix K is

$$\begin{aligned}
 K &= (CA^{-1}B)^{-1} \text{diag} \{p_1, p_2\} \\
 &= \begin{pmatrix} -17.5 & 7.5 \\ 7.5 & -17.5 \end{pmatrix} \begin{pmatrix} p_1 & 0 \\ 0 & p_2 \end{pmatrix}, \quad p_1, p_2 > 0.
 \end{aligned}$$

From Remark 2 it follows that for $\gamma > \frac{1}{4}$, $B_U \in L(E^2, D((-A)^{1-\gamma}))$ and $C \in L(D((-A)^{1-\gamma}), E^2)$.

Thus, using the feedback matrix K determined above, we can design a discrete-time PI-controller for the system (44) with unbounded input and output operators. From a practical point of view, the matrix $(CA^{-1}B)^{-1}$ must be experimentally determined by Algorithm 1.

Example 6.2. Let us consider the transverse vibrations of a simply supported Euler-Bernoulli beam described by the following partial differential equation:

$$(46) \quad \begin{aligned} \ddot{z}(t, \eta) - 2\zeta \dot{z}_{\eta\eta}(t, \eta) &= -z_{\eta\eta\eta\eta}(t, \eta) + \delta(\eta - 0.1)[u_1(t) + w] + \delta(\eta - 0.9)u_2(t), \\ z(t, 0) = z(t, 1) = z_{\eta\eta}(t, 0) = z_{\eta\eta}(t, 1) &= 0, \end{aligned} \quad 0 < \eta < 1,$$

where a damping constant ζ is $0 < \zeta < 1$.

The basic space H for $z(t, \cdot)$ is $L^2(0, 1)$, and the operator A is defined as

$$(47) \quad \begin{aligned} Av &= -v_{\eta\eta\eta\eta}, \\ D(A) &= \{v \in H^4(0, 1): v(0) = v(1) = v_{\eta\eta}(0) = v_{\eta\eta}(1) = 0\}. \end{aligned}$$

To describe the system in first-order form, we take the space X to be $H_0^2(0, 1) \times L^2(0, 1)$, where $H_0^2(0, 1) = \{v \in H^2(0, 1): v(0) = v(1) = 0\}$ and $\sqrt{-A}v = v_{\eta\eta}$ with $D(\sqrt{-A}) = H_0^2(0, 1)$. Then the operator \bar{A} ,

$$\bar{A} = \begin{pmatrix} 0 & I \\ A & -2\zeta\sqrt{-A} \end{pmatrix}, \quad D(\bar{A}) = D(A) \times D(\sqrt{-A})$$

generates an exponentially stable analytic semigroup $\bar{U}(t)$ on X . It is easily seen that \bar{A} has eigenvalues

$$\lambda_n = (n\pi)^2(-\zeta + i\sqrt{1-\zeta^2}), \quad \bar{\lambda}_n, \quad n = 1, 2, \dots$$

and eigenvectors

$$\phi_n = \frac{1}{\lambda_n} \begin{pmatrix} \sin(n\pi\eta) \\ \lambda_n \sin(n\pi\eta) \end{pmatrix}, \quad \bar{\phi}_n = \frac{1}{\bar{\lambda}_n} \begin{pmatrix} \sin(n\pi\eta) \\ \bar{\lambda}_n \sin(n\pi\eta) \end{pmatrix},$$

and A^* has the same eigenvalues and eigenvectors:

$$\psi_n = \frac{1}{\lambda_n} \begin{pmatrix} \sin(n\pi\eta) \\ -\lambda_n \sin(n\pi\eta) \end{pmatrix}, \quad \bar{\psi}_n = \frac{1}{\bar{\lambda}_n} \begin{pmatrix} \sin(n\pi\eta) \\ -\bar{\lambda}_n \sin(n\pi\eta) \end{pmatrix}.$$

Note from Remark 5 that the input operator for the first-order system corresponding to (40) is in $L(E^2, D((- \bar{A})^{3/4-\varepsilon}))$ for any small $\varepsilon > 0$ in the discrete-time case, although the input operator is unbounded for the continuous-time case.

Our observations are the slopes at the ends of the beam:

$$(48) \quad y_k = \begin{pmatrix} z_\eta(kT, 0) \\ -z_\eta(kT, 1) \end{pmatrix} = Cz(kT, \cdot).$$

This operator C is bounded on $X = H_0^2(0, 1) \times L^2(0, 1)$, since $H^2(0, 1) \subset C^1(0, 1)$ by Sobolev's Embedding Theorem.

Let us determine the matrix CG . Setting the Green mapping $G = [g(\eta), h(\eta)]$, we have

$$(49) \quad \begin{aligned} -\frac{d^4 g(\eta)}{d\eta^4} &= \delta(\eta - 0.1), & g(0) = g(1) = g''(0) = g''(1) &= 0, \\ -\frac{d^4 h(\eta)}{d\eta^4} &= \delta(\eta - 0.9), & h(0) = h(1) = h''(0) = h''(1) &= 0, \end{aligned}$$

since $B = [\delta(\eta - 0.1), \delta(\eta - 0.9)]$. We can obtain the solutions

$$g(\eta) = \begin{cases} 0.9\eta^3 - 0.171\eta, & 0 \leq \eta \leq 0.1, \\ -0.1(\eta - 1)^3 + 0.099(\eta - 1), & 0.1 \leq \eta \leq 1, \end{cases}$$

$$h(\eta) = \begin{cases} 0.1\eta^3 - 0.099\eta, & 0 \leq \eta \leq 0.9, \\ -0.9(\eta - 1)^3 + 0.171(\eta - 1), & 0.9 \leq \eta \leq 1. \end{cases}$$

Thus we have the matrix CG :

$$CG = \begin{pmatrix} -g_\eta(0) & -h_\eta(0) \\ g_\eta(1) & h_\eta(1) \end{pmatrix} = \begin{pmatrix} 0.171 & 0.099 \\ 0.099 & 0.171 \end{pmatrix}.$$

A suitable feedback matrix K for the system (46) is

$$K = \begin{pmatrix} -0.171 & -0.099 \\ -0.099 & -0.171 \end{pmatrix}^{-1} \begin{pmatrix} p_1 & 0 \\ 0 & p_2 \end{pmatrix}, \quad p_1, p_2 > 0.$$

The numerical results for (46) are shown in Figs. 1 and 2 and Table 1.

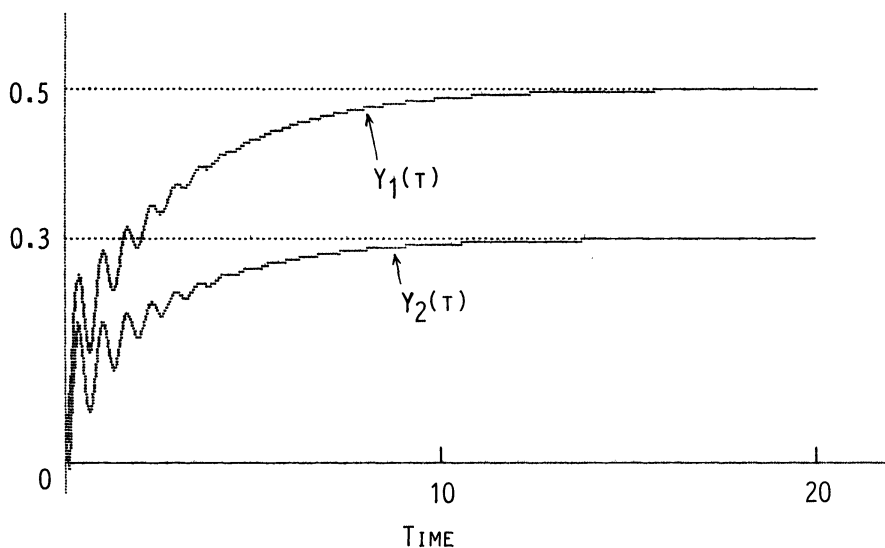


FIG. 1. Response of the output.

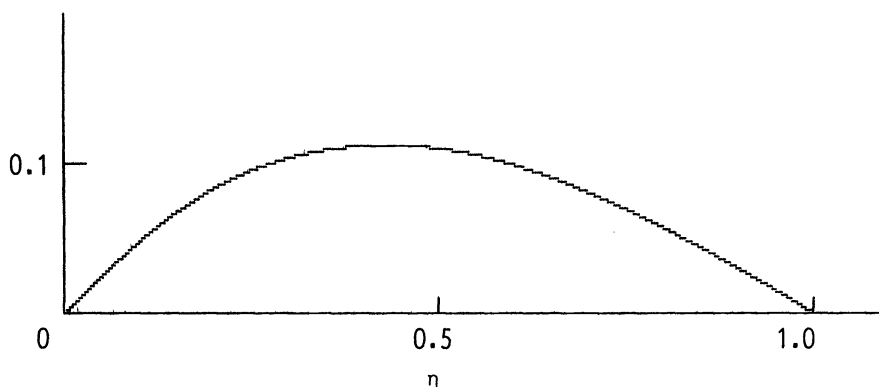


FIG. 2. The steady state $z_s(\eta)$.

TABLE 1
First eight eigenvalues of the open-loop system and the closed-loop system.

Open-loop system	Closed-loop system
1	0.99683
1	0.99674
$0.98541 \pm j0.97081 \times 10^{-1}$	$0.98690 \pm j0.96969 \times 10^{-1}$
$0.88808 \pm j0.36796$	$0.88940 \pm j0.36784$
$0.58030 \pm j0.70745$	$0.58045 \pm j0.70743$

The values of the parameters are

$$\zeta = 0.1, \quad w = 5.0, \quad z(0, \eta) = \dot{z}(0, \eta) = 0,$$

$$y_d = \begin{pmatrix} 0.5 \\ 0.3 \end{pmatrix}, \quad T = 0.01,$$

$$\varepsilon = 0.02, \quad \alpha = 0, \quad p_1 = p_2 = 1.$$

In Fig. 1 we give the response of the outputs $y_1(t) = z_\eta(t, 0)$, $y_2(t) = -z_\eta(t, 1)$. In Fig. 2 we give the steady state $z_s(\eta) = \lim_{t \rightarrow \infty} z(t, \eta)$. In Table 1, the first eight eigenvalues of U_0 (the open-loop system) and $U_0 + \varepsilon B_K$ (the corresponding closed-loop system), respectively, are given.

These numerical results show the efficiency of the theory. In this numerical simulation we cannot find the effectiveness of tuning the proportional parameter α .

We can also consider the boundary controls

$$z_{\eta\eta}(t, 0) = u_1(t) + w, \quad z_{\eta\eta}(t, 1) = u_2(t).$$

For this case the input operator is also bounded in the discrete-time case.

7. Conclusions. In this paper we investigated a digital multivariable PI-controller that regulates the outputs for an exponentially stable distributed parameter system with an unbounded input and an unbounded output operator. First we analyzed the systems with discrete-time controls and observations. We showed that the input operator B_U may be bounded in the case of discrete-time controls, even if the input operator B is unbounded in the case of continuous-time controls.

Next we showed that the integral feedback matrix K given by (19) has the same form as that given by (4) in the case of continuous-time controls. Since $CA^{-1}B = -\hat{G}(1)$ (the value at $z = 1$ of the pulse transfer function matrix $\hat{G}(z)$ of the system), we can determine K experimentally only by steady-state information, without exact knowledge of the system parameters A , B , C , and E . Moreover, we showed that continuous-time set-point tracking can be realized as the result of discrete-time set-point tracking.

We also applied the theory to the generalized oscillating system described by a second-order evolution equation with a first-order damping term. We showed that the integral feedback matrix K given by (37) is independent of the damping term. Thus we can determine K independent of the form of the damping operator Λ .

To illustrate the theory we applied it to a heating system and a vibration beam system. Moreover, we gave the numerical results for the beam example and demonstrated the efficiency of the theory.

Last, we discussed the robustness of the proposed controller. Suppose that the operator A is perturbed so that it remains a generator of an exponentially stable semigroup $U'(t)$ on X , and the operators B and C are perturbed so that they remain

bounded operators such as $B' \in L(E^p, X)$ and $C' \in L(X, E')$, respectively. When the operator A' generates an analytic semigroup, suppose that the operator C is perturbed so that it remains an A' -bounded operator C' such that $D(C') \supset D(A')$. Then we can easily see that the proposed controller (37) will regulate the outputs and stabilize the closed-loop system, if $\sigma(C'(A')^{-1}B'K) \subset C^+$ and $\text{rank}[C'(A')^{-1}B'] = r$. In this case the proposed controller has the robustness so that asymptotic regulation, in the presence of disturbance, occurs independently of perturbations in the plant's parameters of the system.

REFERENCES

- [1] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sciences 3, Springer-Verlag, Berlin, New York, 1978.
- [2] E. J. DAVISON, *The robust control of a servomechanism problem for linear time-invariant multivariable systems*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 25-34.
- [3] A. ICHIKAWA, *A semigroup model for parameter equations with boundary and pointwise noise*, in Stochastic Space-Time Models and Limit Theorems, L. Arnold and P. Kotelenetz, eds., D. Reidel, Boston, pp. 81-94.
- [4] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.
- [5] T. KOBAYASHI, *Discrete-time servomechanism design of parabolic distributed-parameter systems*, Internat. J. Control, 41 (1985), pp. 845-864.
- [6] S. G. KREIN, *Linear Differential Equations in Banach Space*, American Mathematical Society, Providence, RI, 1971.
- [7] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1971.
- [8] J. PENTTINEN AND H. N. KOIVO, *Multivariable tuning regulators for unknown systems*, Automatica, 16 (1980), pp. 393-398.
- [9] S. A. POHJOLAINEN, *Robust multivariable PI-controller for infinite dimensional systems*, IEEE Trans. Automat. Control (1982), pp. 17-30.
- [10] B. PORTER, *Design of tunable set-point tracking controllers for linear multivariable plants*, Internat. J. Control, 35 (1982), pp. 1107-1115.
- [11] B. PORTER AND A. BRADSHAW, *Design of error-actuated controllers for multivariable plants with unknown dynamics and unmeasurable outputs*, Internat. J. Control, 37 (1983), pp. 1-16.
- [12] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.
- [13] K. YOSHIDA, *Functional Analysis*, Springer-Verlag, Berlin, 1978.
- [14] J. ZABCZYK, *On decomposition of generators*, SIAM J. Control Optim., 16 (1978), pp. 523-534.

PSEUDO-RATIONAL INPUT/OUTPUT MAPS AND THEIR REALIZATIONS: A FRACTIONAL REPRESENTATION APPROACH TO INFINITE-DIMENSIONAL SYSTEMS*

YUTAKA YAMAMOTO†

Abstract. This paper studies matrix fractional representation for impulse responses of a certain class of infinite-dimensional linear systems which contains, in particular, delay-differential systems. Such impulse responses are called *pseudo-rational* in this paper. This fractional representation is effectively used to derive concrete function space models from the abstract shift realizations. Given a fractional representation $Q^{-1} * P$, a standard observable realization, analogous to Fuhrmann realizations for finite-dimensional systems, is associated to it. A new notion of coprimeness called *approximate left coprimeness* is introduced, and it is shown that the standard observable realization associated to the representation $Q^{-1} * P$ is canonical if and only if Q and P are approximately left coprime. Some examples are discussed to illustrate the relationships among various coprimeness concepts that have appeared in the literature.

Key words. pseudo-rational impulse responses, delay-differential systems, matrix fractional representation, shift realization, canonical realization, approximate left coprimeness

AMS(MOS) subject classifications. 93B15, 93C05, 93C20

1. Introduction. Consider an impulse response matrix $A(t)$ whose Laplace transform is not necessarily rational. For each input $u(t)$, this impulse response induces the output

$$(1.1) \quad y(t) = \int_0^t A(t-\tau)u(\tau) d\tau$$

under the hypothesis that the initial state is zero. Following the general framework given in [6], [18], let us assume that the present time is zero and that we observe outputs *after* time zero corresponding to inputs applied *before* time zero. In this case, (1.1) takes the form

$$(1.2) \quad y(t) = \int_{-\infty}^0 A(t-\tau)u(\tau) d\tau.$$

This correspondence $f_A: u(\cdot) \mapsto y(\cdot)$ is called the *input/output map associated with the impulse response matrix* A . One of the central observations in abstract realization theory [6] is that we can obtain the canonical state space by separating the time-axis as above and considering the “sections” of the input/output pair at zero, that is, taking the equivalence classes obtained by identifying the inputs that induce the same outputs (see [6] for details).

To put it more precisely, let Ω be the space of inputs having compact support in $(-\infty, 0]$ (that is, each input should be applied only for a finite time period). Let Γ denote the space of outputs observed on $[0, \infty)$. (Their precise definitions will be given

* Received by the editors January 9, 1984; accepted for publication February 1, 1988. A preliminary version of this work was presented at the Joint U.S.-Japan Seminar on Algebraic System Theory, Gainesville, FL, April 7-12, 1983. This work was supported by the Japan Society for the Promotion of Science. Part of this research was done while the author was at the Center for Mathematical System Theory, University of Florida, Gainesville, Florida.

† Department of Applied Systems Science, Faculty of Engineering, Kyoto University, Kyoto 606, Japan.

in § 2.) To obtain the canonical realization according to the above program, we need only to take either the quotient space $\Omega/\ker f_A$ or, equivalently, the subspace $\text{im } f_A$ in Γ as the state space, and consider the following commutative diagram:

$$\begin{array}{ccc} \Omega & \xrightarrow{f_A} & \Gamma \\ p \searrow & & \nearrow j \\ & \Omega/\ker f_A \xrightarrow{f_A} \text{im } f_A & \end{array}$$

Here p and j denote the canonical projection and embedding, respectively. The free-state transition is naturally induced from the left shift semigroup either in Ω or in Γ , and the suitable input/state and state/output correspondences are induced from p and j . Let us call a realization obtained as above a *shift realization*.

This realization framework was first introduced for discrete-time linear systems of Kalman (see, e.g., [6]), and has been used as a basis for a number of algorithms (e.g., [6], [13]). It also gives rise to a concrete state space representation, once a representation (e.g., a fractional representation for the transfer matrix) for f_A is given and the finite-dimensionality of $\Omega/\ker f_A$ is guaranteed [5], [6].

Analogous approaches for continuous-time systems have been pursued by a number of authors [4], [7], [9], [18]. However, although most natural from the abstract point of view, a concrete representation for shift realizations, naturally associated with the theory of differential equations, has not been obtained when the finite-dimensionality is not guaranteed.

The present paper attacks this problem by introducing a fractional representation for impulse response matrices. After fixing the basic realization framework, we shall introduce in § 2 a particular type of fractional representations for impulse response matrices. Roughly speaking, they consist of *Schwartz distributions having compact support contained in $(-\infty, 0]$* . We say that an impulse response matrix A is *pseudo-rational* if it can be written as $A = Q^{-1} *|_{[0, \infty)}$ for matrices with such entries. The precise definition will be introduced in § 2. It is seen that this class contains the class of delay-differential systems, retarded or neutral. We then associate to such a representation an observable realization (denoted Σ^Q) similar to the Fuhrmann observable realization [5] in the finite-dimensional case. It turns out that this observable realization is relatively easy to compute, and brings out a familiar function space model, such as an M_2 model for delay-differential systems. A computational procedure for obtaining a concrete realization is illustrated in Example 3.8.

The question of finding the canonical realization then turns out to be that of determining when this realization Σ^Q is canonical. Our main result is the following: *The system Σ^Q is canonical if and only if the pair of matrices Q and P are approximately left coprime* (the meaning will be clarified below). This result clarifies the relationship between function space reachability and coprimeness conditions in the class of pseudo-rational impulse responses. In § 5 we shall also discuss various coprimeness properties, using examples to illustrate their relationships to canonicity and nonequivalence of various coprimeness notions.

1.1. Notation and convention. Throughout the paper k denotes a fixed field, either R or C . Every function or distribution is assumed to be k -valued. Every vector space is over k . For a vector space V (or a ring, algebra being appropriate to the context), V^p is its p -fold product, and $V^{p \times m}$ denotes the set of $p \times m$ matrices having entries in V . If V is a topological space, V^p and $V^{p \times m}$ carry the standard product topology. For

a vector x , x^T denotes its transpose. As usual, $\hat{x}(s)$ denotes the (two-sided) Laplace transform of x .

The following spaces of distributions will be extensively used in the sequel. As usual, $\mathcal{D}(\mathbb{R})$ denotes the space of C^∞ -functions on $(-\infty, \infty)$ with compact support; its dual space $\mathcal{D}'(\mathbb{R})$ is the space of all distributions on $(-\infty, \infty)$. We set $\mathcal{E}'(\mathbb{R}^-)$ as the space of distributions having compact support in $(-\infty, 0]$; the Dirac distribution δ at the origin, its derivative δ' , the Dirac distribution δ_{-a} ($a > 0$) at point $-a$, etc., are examples of elements in $\mathcal{E}'(\mathbb{R}^-)$. Let $\mathcal{D}'_+(R)$ (often abbreviated as \mathcal{D}'_+) denote the space of distributions having support bounded on the left. Clearly $\mathcal{E}'(\mathbb{R}^-)$ is a subspace of \mathcal{D}'_+ . For $\alpha \in \mathcal{D}'_+$, $l(\alpha)$ denotes the greatest lower bound of its support, i.e., $l(\alpha) := \inf\{t \in \text{supp } \alpha\}$ ([9]). (For more details see, e.g., [9], [19].) In the sequel, we shall treat \mathcal{D}'_+ as a convolution algebra. We now introduce a space of distributions on the half-line $[0, \infty)$. Let $\mathcal{D}[0, \infty)$ be the space of infinitely differentiable functions on $(-\infty, \infty)$ with compact support contained in $[0, \infty)$; its topology is defined in the standard way, as in [15]. The dual space of $\mathcal{D}[0, \infty)$ is denoted by $\mathcal{D}'[0, \infty)$.

The truncation operator $\pi: \psi \rightarrow \psi|_{[0, \infty)}$ is then generalized to the space \mathcal{D}'_+ of distributions as follows:

$$(1.3) \quad \langle \pi\alpha, \psi \rangle := \langle \alpha, \psi \rangle, \quad \alpha \in \mathcal{D}'_+, \quad \psi \in \mathcal{D}[0, \infty).$$

This gives a well-defined element $\pi\alpha \in \mathcal{D}'[0, \infty)$, because we may regard ψ as an element of $\mathcal{D}(\mathbb{R})$. It is easy to see that π is continuous with respect to the strong dual topologies of \mathcal{D}'_+ and $\mathcal{D}'[0, \infty)$ [14, Chap. 4].

2. Preliminaries: input/output maps and pseudo-rationality. Before introducing the notion of pseudo-rational impulse responses, let us first review the realization framework developed in [18]. Let $C_0[0, \infty)$ be the space of continuous functions on $(-\infty, \infty)$ with compact support in $[0, \infty)$. The dual space of $C_0[0, \infty)$ is denoted by $M[0, \infty)$, and regarded as a space of (Radon) measures on $[0, \infty)$. Let A be a $p \times m$ matrix whose entries a_{ij} belong to $M[0, \infty)$. Our (*constant, linear, and continuous-time*) *input/output map* f_A associated with the *impulse response* A is specified as

$$(2.1) \quad f_A(\omega) := \pi(A * \omega)$$

where ω is an L^2 -input with bounded support contained in $(-\infty, 0]$ π is the truncation $\pi\psi := \psi|_{[0, \infty)}$, and $*$ denotes convolution. (Note that (2.1) is precisely the same as (1.2).)

Let us specify the input and output function spaces precisely. Let $\Omega := \bigcup_{n=1}^{\infty} L^2[-n, 0]$. Under suitable identifications, Ω is regarded as consisting of L^2 -functions having compact support contained in $(-\infty, 0]$. Its m -fold product Ω^m is our *space of m -inputs*. Now let Γ be the space $L^2_{\text{loc}}[0, \infty)$ of all locally square-integrable functions on $[0, \infty)$. Its p -fold product Γ^p is our *space of p -outputs*. Γ^p is a Fréchet space with respect to the following seminorms:

$$\|\psi\|_{[0, a]} := \left\{ \int_0^a \|\psi\|^2 dt \right\}^{1/2}, \quad a > 0.$$

Our input/output map f_A then gives a continuous linear correspondence: $\Omega^m \rightarrow \Gamma^p$, which commutes with left shift operators $\sigma_t, \tilde{\sigma}_t$ defined as follows [18]:

$$(2.2) \quad \begin{aligned} (\sigma_t \omega)(s) &:= \begin{cases} \omega(s+t), & s \leq t, \\ 0, & s > t, \end{cases} & \omega \in \Omega^m, \\ (\tilde{\sigma}_t \gamma)(s) &:= \gamma(s+t), & \gamma \in \Gamma^p. \end{aligned}$$

It is well known that the realization of f_A is completely specified by the following commutative diagram [6], [13], [18]:

$$\begin{array}{ccc} \Omega^m & \xrightarrow{f_A} & \Gamma^p \\ & \searrow g & \nearrow h \\ & X & \end{array}$$

Here X is the state space equipped with a semigroup $\Phi(t)$ governing the free-state transition, g gives the correspondence: inputs given on $(-\infty, 0] \mapsto$ states at time zero, and h gives the correspondence: states at time 0 \mapsto output functions on $[0, \infty)$. Since there is a one-to-one correspondence between realizations and commutative diagrams as above, we shall speak about system properties in terms of the triple (X, g, h) [18].

DEFINITIONS 2.3. The system (X, g, h) is said to be *quasi-reachable* if g has dense image; *topologically observable* if h is an injective topological isomorphism into Γ^p ; *canonical* if it is both quasi-reachable and topologically observable; and *topologically observable in bounded time* $T > 0$ if $h_T: X \rightarrow h(X)|_{[0, T]}$ is continuously invertible, where $h(X)|_{[0, T]}$ is equipped with the subspace topology induced from $(L^2[0, T])^p$.

The following theorem is known for the existence and uniqueness of canonical realizations [18].

THEOREM 2.4. Let f_A be an input/output map. Its canonical realization is unique up to isomorphism and is given by the triple $(\overline{\text{im } f_A}, f_A, j)$, where $\overline{\text{im } f_A}$ is the closure of the image of f_A in Γ^p , $j: \overline{\text{im } f_A} \rightarrow \Gamma^p$ is the canonical inclusion, and $\overline{\text{im } f_A}$ is equipped with the left shift operators as the semigroup for free-state transition.

Although this canonical realization is certainly topologically observable, it is not necessarily topologically observable in bounded time. This means that a long time period, depending on each state, may be required to determine whether the initial state is close to zero. This is undesirable, and the notion of topological observability in bounded time is introduced to avoid this difficulty. However, not all input/output maps admit a realization that is topologically observable in bounded time, and this is one reason why it is difficult to obtain a concrete representation for the canonical state space $\overline{\text{im } f_A}$. The notion of T -bounded (or simply *bounded*) input/output maps (or impulse responses) is therefore introduced; this requires that the canonical realization be topologically observable in bounded time for some $T > 0$ [20]. Then the canonical state space $\overline{\text{im } f_A}$ is topologically isomorphic to $\overline{\text{im } f_A}|_{[0, T]}$, which is determined by the free output data on the finite interval $[0, T]$. Hence the canonical realization is expected to have a much simpler structure in this case. In order to give a condition for T -boundedness, we now introduce the concept of pseudo-rationality as follows.

DEFINITION 2.5. Let A be an impulse response matrix. A (or its associated input/output map f_A) is said to be *pseudo-rational* if A can be written in the form $A = \pi(Q^{-1} * P)$ for some $p \times p$ and $p \times m$ matrices Q and P with entries in $\mathcal{E}'(R^-)$ such that

$$(\det Q)^{-1} \text{ exists in } \mathcal{D}'_+, \text{ supp } Q^{-1} \subset [0, \infty), \text{ and } \text{ord}(\det Q)^{-1} = -\text{ord}(\det Q),$$

where $\text{ord } \alpha$ denotes the order of distribution α [15].

Remark 2.6. In many cases in which Q^{-1} and A are functions, we can write simply $A = Q^{-1} * P$, without the truncation π .

Example 2.7. As is well known [8], the impulse response of a delay-differential system (retarded or neutral) with constant coefficients can be expressed as the fraction of polynomials in distributions of type δ' , δ_{-a_i} ($a_i > 0$). Therefore, this class of impulse responses are pseudo-rational. It is also easy to see that the impulse responses of systems containing distributed delays are pseudo-rational. Another example is an impulse response that has bounded support.

The notion of pseudo-rationality is related to T -boundedness as follows.

THEOREM 2.8. *Let $A = \pi(Q^{-1} * P)$ be a pseudo-rational impulse response matrix. Then A is T -bounded for any T greater than $-l(\det Q)$.*

For the proof, see [20]. This theorem means that, for a pseudo-rational impulse response, we can take, as a canonical state space, $\overline{\text{im } f_A}|_{[0, T]}$, which is a Hilbert space, rather than $\overline{\text{im } f_A}$, which is primarily only a Fréchet space. Therefore, it is natural to expect that computation of the canonical realization is easier for a pseudo-rational impulse response. Indeed, the following result holds.

THEOREM 2.9. *Let $A = \pi(Q^{-1} * P)$ be pseudo-rational. Then the canonical state space $\overline{\text{im } f_A}$ is a closed subspace of*

$$X^Q := \{\gamma \in \Gamma^p; \pi(Q * \gamma) = 0\}$$

where $\pi(Q * \gamma) = 0$ must be understood in the sense of distributions.

To make the discussion on X^Q precise, we need to define the convolution of $\alpha \in \mathcal{E}'(R^-)$ and $\beta \in \mathcal{D}'[0, \infty)$. This is given in the Appendix.

To prove Theorem 2.9, we need the following lemma.

LEMMA 2.10. *X^Q is a closed subspace of Γ^p .*

Proof. Suppose that $x_n \in X^Q$ and $x_n \rightarrow x$ in Γ^p . Then $\pi(Q * x_n) \rightarrow \pi(Q * x)$ in $\mathcal{D}'[0, \infty)$ by the continuity of π and by the separate continuity of convolution [15]. Thus $\pi(Q * x) = 0$, and hence x belongs to X^Q . \square

Proof of Theorem 2.9. Take any $\omega \in \Omega^m$. Then we have

$$\begin{aligned} \pi(Q * f_A(\omega)) &= \pi(Q * \pi(A * \omega)) = \pi(Q * \pi(\pi(Q^{-1} * P) * \omega)) \\ &= \pi(Q * \pi(Q^{-1} * P * \omega)) \quad (\text{by Lemma A3}) \\ &= \pi(Q * Q^{-1} * P * \omega) \quad (\text{by Lemma A3}) \\ &= \pi(P * \omega) = 0 \quad (\text{by Lemma A2}) \end{aligned}$$

because $\text{supp } P * \omega \subset (-\infty, 0]$. Therefore, $\text{im } f_A \subset X^Q$, which readily implies $\overline{\text{im } f_A} \subset X^Q$ in view of Lemma 2.10. \square

3. Topologically observable realization Σ^Q . Using the space X^Q introduced in the preceding section, we give a topologically observable realization of the impulse response matrix A , and exploit some of the basic properties of such realizations.

THEOREM 3.1. *Let $A = \pi(Q^{-1} * P) \in \Gamma^{p \times m}$ be pseudo-rational. Then the system Σ^Q given by the following formulas is a topologically observable realization of A :*

X^Q is the state space.

$\tilde{\sigma}_t$ (restricted to X^Q) is the semigroup of the input-free state evolution.

$F := d/dt$ (is the infinitesimal generator of $\tilde{\sigma}_t$).

$D(F) := (H_{\text{loc}}^1[0, \infty))^p \cap X^Q$, $(H_{\text{loc}}^1[0, \infty))$ is the subspace of $L_{\text{loc}}^2[0, \infty)$ consisting of those elements having a derivative belonging to $L_{\text{loc}}^2[0, \infty)$.

$G := k^m \rightarrow X^Q : u \mapsto A(\cdot)u$.

$Hx := x(0)$ for $x \in (C[0, \infty))^p \cap X^Q$.

The system equation is

$$(3.2) \quad \frac{dx}{dt} = Fx + Gu(t),$$

$$(3.3) \quad y = Hx.$$

Proof. Since X^Q is closed under $\tilde{\sigma}_t$, $\tilde{\sigma}_t$ constitutes a semigroup in X^Q . That Σ^Q is a well-defined system can be shown as in [18]. The system is clearly topologically observable. Thus we need only to prove that Σ^Q is a realization of A .

For sufficiently smooth u , we have

$$x(t) = \tilde{\sigma}_t x(0) + \int_0^t \tilde{\sigma}_{t-\tau} Gu(\tau) d\tau.$$

If an input $u \in \Omega^m$ is applied under the assumption that the initial state is zero before the application of the input, then we have

$$x(0)(\eta) = \int_{-\infty}^0 \tilde{\sigma}_{-\tau} A(\eta) u(\tau) d\tau.$$

Note here that $x(0)$ is a function of $\eta \geq 0$ as an element of X^Q . The zero initial-state response of this system is then given by

$$\begin{aligned} y(t) &= Hx(t) = H\tilde{\sigma}_t x(0) \\ &= \int_{-\infty}^0 \tilde{\sigma}_t \tilde{\sigma}_{-\tau} A(\eta) u(\tau) d\tau|_{\eta=0} \\ &= \int_{-\infty}^0 A(\eta + t - \tau) u(\tau) d\tau|_{\eta=0} \\ &= \int_{-\infty}^0 A(t - \tau) u(\tau) d\tau = \pi(A * u). \end{aligned}$$

This shows that the impulse response of this system is A , and hence Σ^Q is a realization of A . \square

Remark 3.4. The topologically observable realization Σ^Q can be defined even when $A \notin \Gamma^{p \times m}$ [18].

The following proposition is obtained by a minor modification of Theorem 2.8.

PROPOSITION 3.5. *Let Q be a $p \times p$ matrix satisfying the conditions of pseudorationality. Let $\pi_T: \Gamma^p \rightarrow (L^2[0, T])^p: \psi \mapsto \psi|_{[0, T]}$ be the truncation mapping. Then for any $T > -l(\det Q)$,*

$$(3.6) \quad \pi_T|_{X^Q}: X^Q \rightarrow \pi_T(X^Q)$$

is a topological isomorphism, where the space $\pi_T(X^Q)$ is endowed with the subspace topology induced from $(L^2[0, T])^p$.

This shows that the system Σ^Q is topologically observable in bounded time T . In particular, the state space X^Q is isomorphic to a Hilbert space.

We then have the following corollary.

COROLLARY 3.7. (i) *The semigroup $\tilde{\sigma}_t$ in X^Q has at most exponential growth.*

(ii) *Every $x(\tau) \in X^Q$ is Laplace transformable as a function of τ .*

Proof. The proof of (i) is immediate from the standard semigroup theory, since X^Q is isomorphic to a Hilbert space.

(ii) Take any $T > 0$ satisfying the condition of Proposition 3.5. We have

$$\|\tilde{\sigma}_t x\|_{[0, T]} \leq C \exp(\beta t) \|x\|_{[0, T]}.$$

This readily implies that the Laplace transform of $x(\tau)$ is absolutely convergent for $\operatorname{Re} s > \beta$. \square

We close this section by indicating how to compute the realization Σ^Q in terms of an example. (For a more general discussion, see [21].)

Example 3.8. Consider the following transfer function

$$(3.9) \quad W(s) := \left[\frac{s+1}{s(sz-1)}, \frac{1}{s} \right]^T, \quad z := e^s,$$

which admits the fractional representation

$$(3.10) \quad \dot{W} = \begin{pmatrix} sz-1 & -1 \\ 0 & s \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} =: \hat{Q}(s)^{-1} \hat{P}(s).$$

Let us exhibit Σ^Q . First notice that a smooth pair $(\gamma_1, \gamma_2)^T \in \Gamma^2$ belongs to X^Q if and only if

$$(3.11) \quad \gamma_1'(t+1) - \gamma_1(t) - \gamma_2(t) = 0, \quad \gamma_2'(t) = 0 \quad \text{for } t \geq 0.$$

This implies that $\gamma_2(t) = \gamma_2(0) = \text{constant}$, and

$$(3.12) \quad \gamma_1(t) = \gamma_1(1) + \int_1^t \{\gamma_1(\tau-1) + \gamma_2(\tau-1)\} d\tau, \quad 1 \leq t \leq 2,$$

and similarly for $t \geq 2$. Then the values of $\gamma_1(t)$ and $\gamma_2(t)$ are completely determined once $\gamma_1(1)$, $\gamma_2(0)$ and $\gamma_1|_{[0,1]}$ are known. Taking the closure of all such $(\gamma_1, \gamma_2)^T$ in Γ^2 , we see that

$$(3.13) \quad X^Q \simeq R^2 \times L^2[0, 1].$$

Note that $l(\det Q) = -1$, and hence X^Q must be completely determined by the function values on $[0, T]$ for any $T > 1$. The isomorphism (3.13) reflects this fact. Let us now indicate how to obtain the differential equation description in the space $R^2 \times L^2[0, 1]$. This procedure is not as trivial as it may appear, since $R^2 \times L^2[0, 1]$ is not endowed with a natural left shift semigroup, and we must induce the transition semigroup of X^Q to $R^2 \times L^2[0, 1]$ via (3.13). This can be done by a repeated application of (3.12). To compute the infinitesimal generator F of the free transition in $R^2 \times L^2[0, 1]$, we must shift $\gamma_1(t)$ by ε , compute the differences

$$\{\tilde{\sigma}_\varepsilon \gamma_1 - \gamma_1\}/\varepsilon, \quad \{(\tilde{\sigma}_\varepsilon \gamma_1)(1) - (\gamma_1)(1)\}/\varepsilon,$$

and then take the limits as $\varepsilon \rightarrow 0$ (see Fig. 1).

Transforming $D(F)$ defined in Theorem 3.1 via (3.13), we see that the domain of F is given by

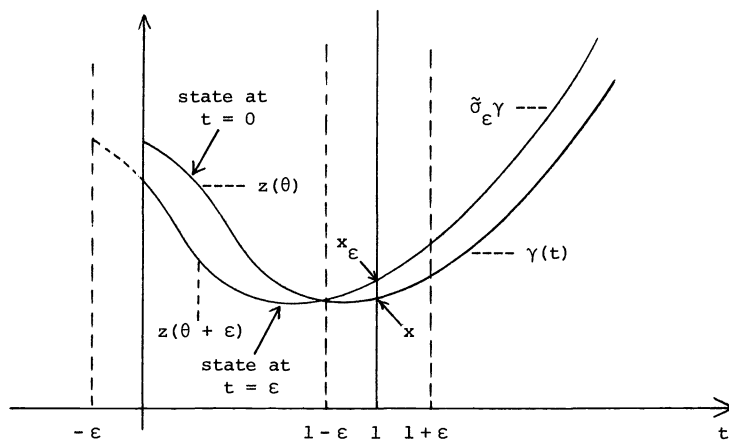
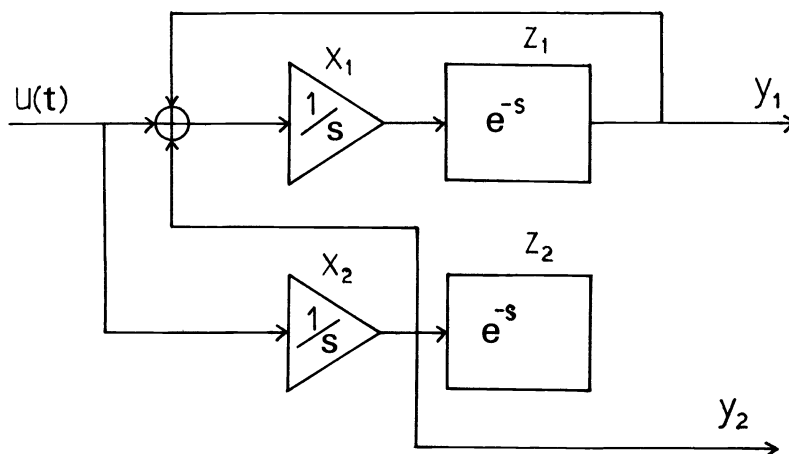
$$D(F) = \{(x_1, x_2, z) \in R^2 \times L^2[0, 1]; z \in H^1[0, 1], x_1 = z(1)\}$$

where x_1, x_2, z denote $\gamma_1(1), \gamma_2(0)$, and $\gamma_1|_{[0,1]}$, respectively.

This leads us to the following functional differential equation description of the observable realization Σ^Q :

$$(3.14) \quad \frac{d}{dt} \begin{bmatrix} x_1 \\ x_2 \\ z \end{bmatrix} = \begin{bmatrix} x_2 + z(0) \\ 0 \\ (\partial/\partial\theta)z(\theta) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} u, \quad y = [z(0), x_2]^T.$$

As we shall see in § 5, this system is canonical. The standard M_2 -realization, which requires a delay part also for x_2 , hence is redundant and noncanonical. This situation is clearly shown in Fig. 2.

FIG. 1. State transition in $R \times L^2[0, 1]$.FIG. 2. Redundant M_2 -realization.

4. Reachability and coprimeness. In the previous section, we have introduced a topologically observable realization Σ^Q for a pseudo-rational impulse response $A = \pi(Q^{-1} * P)$. The realization is, by construction, always *topologically observable in bounded time*, and hence its state space is topologically isomorphic to Hilbert space. Furthermore, it contains the canonical realization as a closed subsystem by Theorem 2.9. Our next objective is to give a condition for Σ^Q to be quasi-reachable (and hence canonical). We shall prove that Σ^Q is *quasi-reachable if and only if* Q and P are *approximately left coprime* (the meaning of which will be clarified below).

DEFINITIONS 4.1. Let Q and P be $p \times p$ and $p \times m$ matrices with entries in $\mathcal{E}'(R^-)$, respectively. The pair (Q, P) is said to be *approximately left coprime* if there exist matrix sequences R_n and S_n with $\mathcal{E}'(R^-)$ -entries such that

$$(4.2) \quad Q * R_n + P * S_n \rightarrow \delta I_p \quad \text{in } \mathcal{E}'(R^-).$$

(Q, P) is said to be *spectrally left coprime* if $[\hat{Q}(\lambda), \hat{P}(\lambda)]$ has full rank for every $\lambda \in C$. Let L be a subalgebra of $\mathcal{E}'(R^-)$. (Q, P) is said to be *left coprime over L* (or *L -left coprime*) if we have the following:

- (a) The entries of Q and P belong to L ;
- (b) There exist matrices R and S over L such that $Q * R + P * S = \delta I_p$.

It is simply said to be *left coprime* when $L = \mathcal{E}'(R^-)$.

A typical example of L is the polynomial ring $k[\delta', \delta_{a_1}, \dots, \delta_{a_r}]$ ($a_i < 0$). This corresponds to the ring-theoretic approach to delay-differential systems. Approximate left coprimeness implies spectral left coprimeness, but not the converse (Example (5.10)).

Let us start by proving the following lemma.

LEMMA 4.3. *Let ψ be an element in $(\mathcal{D}'_+)^p$ such that $\pi(Q * \pi\psi) = 0$. Then there exists a sequence $\psi_n \in X^Q$ such that $\psi_n \rightarrow \pi\psi$. Furthermore, such ψ_n can be taken to be C^∞ on $[0, \infty)$.*

Proof. Let ρ_n be a sequence of C^∞ -functions with compact support such that (i) $\text{supp } \rho_n \subset (-\infty, 0]$; (ii) $\rho_n \rightarrow \delta$ in $\mathcal{D}'(R)$ [15]. Then $\phi_n := \rho_n * \psi$ converges to ψ , and hence $\psi_n := \pi\phi_n \rightarrow \pi\psi$ [15]. Clearly ψ_n belongs to $C^\infty[0, \infty)$. Then,

$$\begin{aligned} \pi(Q * \psi_n) &= \pi(Q * \pi\phi_n) = \pi(Q * \phi_n) \quad (\text{by Lemma A.3}) \\ &= \pi(Q * \rho_n * \psi) = \pi(\rho_n * Q * \psi) \\ &= \pi(\rho_n * \pi(Q * \psi)) \quad (\text{by Lemma A.3}) \\ &= 0, \end{aligned}$$

so that $\psi_n \in X^Q$. \square

We can now state and prove the main result of this section.

THEOREM 4.4. *The system Σ^Q is quasi-reachable (i.e., $\overline{\text{im } f_A} = X^Q$, and is hence canonical) if and only if the pair (Q, P) is approximately left coprime in the fractional representation $A = \pi(Q^{-1} * P)$.*

Proof. (Necessity.) Let us first consider the case in which Q^{-1} belongs to $\Gamma^{p \times p}$. Let Q_i^{-1} denote the i th column of Q^{-1} . We have $\pi(Q * \pi Q_i^{-1}) = \pi(Q * Q_i^{-1}) = \pi(e_i \delta) = 0$ (e_i = the i th unit vector of k^p), so that πQ_i^{-1} belongs to X^Q for every i . Since Σ^Q is quasi-reachable, there exists a sequence $s_n^i \in \Omega^m$ such that $f_A(s_n^i) = \pi(Q^{-1} * P * s_n^i)$ converges to πQ_i^{-1} . Putting $S_n := (s_n^1, \dots, s_n^m)$, we have

$$(4.5) \quad \pi(Q^{-1} * P * S_n) \rightarrow \pi Q^{-1}.$$

Since $\pi(Q^{-1} * P * S_n) \in \Gamma^{p \times p}$, we can extend it to $(-\infty, \infty)$ by setting it to be zero on $(-\infty, 0)$. Denote the extension by Ψ_n . Also, since $Q^{-1} \in \Gamma^{p \times p}$, it is zero on $(-\infty, 0)$. Then we have $Q^{-1} = \lim_{n \rightarrow \infty} \Psi_n$. Define $R_n := \Psi_n - Q^{-1} * P * S_n$. Since $\Psi_n = Q^{-1} * P * S_n$ on $[0, \infty)$, $\text{supp } R_n$ is contained in $(-\infty, 0]$. Furthermore, $\text{supp } R_n$ is compact because $\text{supp } Q^{-1} * P * S_n$ is bounded on the left. Therefore, each element of R_n belongs to $\mathcal{E}'(R^-)$. Thus, we have

$$Q^{-1} = \lim_{n \rightarrow \infty} \Psi_n = \lim_{n \rightarrow \infty} (Q^{-1} * P * S_n + R_n).$$

Multiply Q from the left to both sides to obtain

$$\delta I_p = \lim_{n \rightarrow \infty} (Q * R_n + P * S_n).$$

When Q^{-1} does not belong to $\Gamma^{p \times p}$, we need only to approximate it by a sequence in X^Q by Lemma 4.3 and then apply the above argument. This proves the necessity.

(Sufficiency.) Conversely, suppose that (Q, P) is approximately left coprime, i.e., $Q * R_n + P * S_n \rightarrow \delta I_p$ for suitable R_n and S_n . Take any $x \in X^Q$ with C^∞ -entries. Then there exists an extension $\tilde{x} \in (C^\infty(-\infty, \infty))^m$ such that (i) $\pi \tilde{x} = x$, and (ii) $\text{supp } \tilde{x}$ is bounded on the left. Now put $\omega_n := S_n * Q * \tilde{x}$. Since $\tilde{x} \in (C^\infty(-\infty, \infty))^m$, ω_n belongs to $(C^\infty(-\infty, \infty))^m$ and its support is bounded on the left. Moreover,

$$\begin{aligned} \pi \omega_n &= \pi(S_n * Q * \tilde{x}) = \pi(S_n * \pi(Q * \tilde{x})) \quad (\text{by Lemma (A.3)}) \\ &= \pi(S_n * \pi(Q * \pi \tilde{x})) = \pi(S_n * \pi(Q * x)) = 0. \end{aligned}$$

Thus by Lemma A2, ω_n belongs to Ω^m . Then we have

$$\begin{aligned} Q^{-1} * P * \omega_n + R_n * Q * \tilde{x} &= Q^{-1} * P * S_n * Q * \tilde{x} + R_n * Q * \tilde{x} \\ &= Q^{-1} * (P * S_n + Q * R_n) * Q * \tilde{x} \rightarrow \tilde{x}. \end{aligned}$$

Therefore,

$$\begin{aligned} x = \pi \tilde{x} &= \lim_{n \rightarrow \infty} \{ \pi(Q^{-1} * P * \omega_n) + \pi(R_n * Q * \tilde{x}) \} \\ &= \lim_{n \rightarrow \infty} \{ \pi(Q^{-1} * P * \omega_n) + \pi(R_n * \pi(Q * \pi \tilde{x})) \} \\ &= \lim_{n \rightarrow \infty} \pi(Q^{-1} * P * \omega_n) = \lim_{n \rightarrow \infty} f_A(\omega_n). \end{aligned}$$

This shows that $x \in \overline{\text{im } f_A}$.

It remains only to prove that $X^Q \cap (C^\infty[0, \infty))^p$ is dense in X^Q . According to Lemma 4.3, for every x in X^Q , there exists a sequence x_n in $X^Q \cap (C^\infty[0, \infty))^p$ such that x_n converges to x in the sense of distributions. Furthermore, this sequence converges to x also with respect to the topology of Γ^p [15, 6.4]. Thus $X^Q \cap (C^\infty[0, \infty))^p$ is dense in X^Q . \square

Since the canonical state space $\overline{\text{im } f_A}$ should not depend on a particular representation $A = \pi(Q^{-1} * P)$, we may pose the question: “When are two spaces X^{Q_1} and X^{Q_2} equal?” The following theorem answers this question.

THEOREM 4.6. *Let Q_1 and Q_2 be matrices with entries in $\mathcal{E}'(R^-)$ that are invertible over \mathcal{D}'_+ . Then $X^{Q_1} \subset X^{Q_2}$ if and only if $Q_2 = D * Q_1$ for some $D \in (\mathcal{E}'(R^-))^{p \times p}$. In particular, $X^{Q_1} = X^{Q_2}$ if and only if $Q_2 = D * Q_1$ for some unimodular matrix D (i.e., $\det D = d\delta$, where d is a nonzero constant).*

Proof. Suppose $X^{Q_1} \subset X^{Q_2}$. Using Lemma 4.3, approximate the columns of πQ_1^{-1} by elements in $X^{Q_1} \cap (C^\infty[0, \infty))^p$, so that we have $\Psi_n \rightarrow \pi Q_1^{-1}$, $\Psi_n \in X^{Q_1}$. It follows that $\pi(Q_2 * \Psi_n) = 0$, and hence $\pi(Q_2 * Q_1^{-1}) = 0$ by the continuity of π and convolution. Therefore, $D := Q_2 * Q_1^{-1} \in (\mathcal{E}'(R^-))^{p \times p}$ by Lemma A2. Thus $Q_2 = D * Q_1$.

Conversely, suppose $Q_2 = D * Q_1$, and take any $x \in X^{Q_1}$. Then

$$\begin{aligned} \pi(Q_2 * x) &= \pi(D * Q_1 * x) = \pi(D * \pi(Q_1 * x)) \quad (\text{by Lemma A3}) \\ &= 0. \end{aligned}$$

Hence x belongs to X^{Q_2} .

Therefore, $X^{Q_1} = X^{Q_2}$ if and only if $Q_2 = D * Q_1$ and $Q_1 = C * Q_2$ for some C , $D \in (\mathcal{E}'(R^-))^{p \times p}$. This implies that $C * D = D * C = \delta I_p$, that is, C and D are invertible over $\mathcal{E}'(R^-)$. Hence $\det C$ and $\det D$ are invertible over $\mathcal{E}'(R^-)$. Here we quote from Schwartz [15, 6.10] the fact that the support of a fundamental solution of $(\det D) * v = \delta$ is not bounded unless $\det D$ is a constant multiple of a Dirac point mass δ_a . Since $\det D$ belongs to $\mathcal{E}'(R^-)$, a is not positive. But if a were negative, the inverse $\delta_a^{-1} = \delta_{-a}$ would not belong to $\mathcal{E}'(R^-)$, so a must be zero. Thus $\det D = d\delta$. Conversely, if D is thus, $X^{Q_1} = X^{Q_2}$ clearly follows. \square

5. Some examples. In this section we discuss various coprimeness conditions in terms of delay-differential systems.

The following example shows that left coprimeness over $\mathcal{E}'(R^-)$ does not necessarily imply left coprimeness over a subalgebra L even if Q and P have entries in L .

Example 5.1. Let us take the same example as Example 3.8:

$$(5.2) \quad W = \begin{pmatrix} sz - 1 & -1 \\ 0 & s \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} =: \hat{Q}(s)^{-1} \hat{P}(s), \quad z = e^s.$$

Let us first see that \hat{Q} and \hat{P} are not left coprime as polynomial matrices over $R[s, z]$. Indeed, observe that

$$\text{rank} [\hat{Q}(s_1, z_1), \hat{P}(s_1, z_1)] = 1 < 2$$

for $s_1 = -1$, $z_1 = -1$, which would be impossible if the pair were left coprime over $R[s, z]$. Now note that $z_1 = -1$ is impossible if we return to the original interpretation $z = e^s$, so that this choice (s_1, z_1) is meaningless when we consider (Q, P) as a pair over $\mathcal{E}'(R^-)$.

However, the pair (Q, P) is left coprime over $\mathcal{E}'(R^-)$. To show this, let us solve the following equation over $\mathcal{E}'(R^-)$ (written in terms of Laplace transforms):

$$(5.3) \quad \begin{pmatrix} se^s - 1 & -1 \\ 0 & s \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} [u \quad v] = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

A solution is given by

$$(5.4) \quad \begin{aligned} c &:= \frac{(se^s - 1)a - 1}{s + 1}, \\ d &:= \frac{(se^s - 1)b + 1}{s + 1}, \\ u &:= \frac{-s(se^s - 1)a + s}{s + 1} = -sc, \\ v &:= \frac{1 - s(se^s - 1)b}{s + 1} = 1 - sd \end{aligned}$$

where a and b are constants satisfying

$$(5.5) \quad ((se^s - 1)a - 1)|_{s=-1} = 0, \quad ((se^s - 1)b + 1)|_{s=-1} = 0.$$

In view of (5.5), c , d , u , and v are entire functions of s . According to the Paley-Wiener Theorem and Proposition A6 (see the Appendix), a , b , c , d , u , and v are Laplace transforms of elements of $\mathcal{E}'(R^-)$. Therefore,

$$\hat{R} := \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad \hat{S} := [u \quad v]$$

give rise to a solution to $Q * R + P * S = \delta I$, and (Q, P) is left coprime over $\mathcal{E}'(R^-)$.

Hence the realization Σ^Q (and hence system (3.14)) is canonical by Theorem 4.4.

Example 5.6. In Manitius [11, Ex. 5.1], the following example is discussed:

$$W(s) := [1/se^{h_1 s}, 1/se^{h_2 s}]^T \quad (0 < h_1 < h_2).$$

Though the standard M_2 -space to be associated with this transfer matrix is $R^2 \times (L^2[0, h_2])^2$, it is clearly redundant. It is asserted in [11] that the right state space is $R^2 \times L^2[0, h_1] \times L^2[0, h_2]$. This space can easily be obtained by computing X^Q for the obvious choice of a factorization $W(s) = \hat{Q}(s)^{-1} \hat{P}(s)$.

In the above examples, the canonical systems are still not very far from the M_2 -models. The next example is more striking in that it is realizable by an M_2 -model but its canonical realization is not expressible as a finite linear combination of delays and integrators.

Example 5.7. Let $W(s)$ be the following transfer function:

$$(5.8) \quad W(s) = \frac{e^s - 1}{se^s} \left(= \frac{1 - e^{-s}}{s}, \text{ i.e., zero-order hold element} \right).$$

The impulse response $A(t)$ is

$$A(t) = \begin{cases} 1, & 0 \leq t \leq 1, \\ 0, & t > 1. \end{cases}$$

Let $q := \delta'_{-1}$ and $p := \delta_{-1} - \delta$. Clearly $A = q^{-1} * p$ and A is realizable by an M_2 -model as in Fig. 3. But q and p are not approximately coprime since $\hat{q}(s)$ and $\hat{p}(s)$ have a common zero $s = 0$. The correct (coprime) fractional representation is

$$W(s) = \{(e^s - 1)/s\}/e^s,$$

which yields a new fractional representation $A = \delta_{-1} * a$, where

$$a(t) = L^{-1}[(e^s - 1)/s] = \begin{cases} 1, & -1 \leq t \leq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The pair (δ_{-1}, a) is coprime over $\mathcal{E}'(R^-)$ since $e^s \cdot 1 + ((e^s - 1)/s) \cdot (-s) = 1$. The canonical realization Σ^{q_1} induced by this factorization is given by the following functional differential equation:

$$(5.9) \quad \begin{aligned} \frac{d}{dt} x_t(\theta) &= \frac{\partial}{\partial \theta} x_t(\theta) + u(t), \\ y(t) &= x_t(0), \quad x_t(\cdot) \in X^{q_1} = L^2[0, 1]. \end{aligned}$$

The model (5.9) contains a distributed input term, and does not fall into the scope of M_2 -models. Observe also that this canonical realization is clearly internally stable with decay rate faster than any exponentials, whereas the M_2 -realization corresponding to Fig. 3 is not even BIBO stable (observation due to E. W. Kamen [22], [23]).

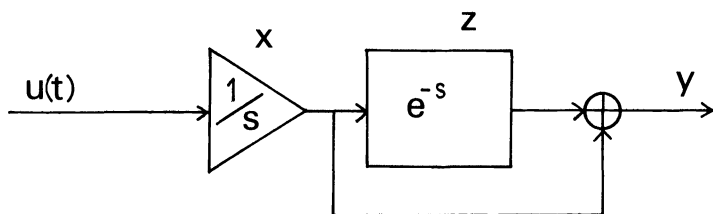


FIG. 3. Noncanonical realization for $(1 - e^{-s})/s$.

In the next example, we show that spectral left coprimeness does not imply approximate left coprimeness.

Example 5.10. Consider the transfer function matrix given by

$$W(s) = \begin{pmatrix} se^s & -1 \\ e^s & se^s \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} =: \hat{Q}^{-1} \hat{P}.$$

The pair (Q, P) is easily seen to be spectrally left coprime. But it is not approximately left coprime. To see this, let $x := 0$ and $y :=$ an arbitrary nonzero function in $L^2[0, 1]$. Then we have

$$[x, y] * Q = [\delta_{-1} * y, \delta'_{-1} * y], \quad [x, y] * P = 0.$$

Suppose that there exist sequences R_n and S_n of matrices over $\mathcal{E}'(R^-)$ such that $Q * R_n + P * S_n \rightarrow \delta I$. Since $\text{supp}(\delta_{-1} * y)$ and $\text{supp}(\delta'_{-1} * y)$ are both contained in $(-\infty, 0]$, it follows that

$$\begin{aligned} [x, y] &= [\pi x, \pi y] \\ &= \lim_{n \rightarrow \infty} \pi\{[x, y] * (Q * R_n + P * S_n)\} \\ &= \lim_{n \rightarrow \infty} \pi\{[\delta_{-1} * y, \delta'_{-1} * y] * R_n\} = 0. \end{aligned}$$

This is absurd, so (Q, P) is not approximately left coprime.

This example (and the realization Σ^Q) corresponds to the following retarded delay-differential equation:

$$(5.11) \quad \frac{d}{dt} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(t-1) \\ x_2(t-1) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t).$$

According to a necessary and sufficient condition for M_2 -quasi-reachability by Manitius [12], this system is not quasi-reachable. This also shows that (Q, P) is not approximately left coprime in view of Theorem 4.4.

Let us now examine this example from the viewpoint of systems over rings. Rewrite (5.11) as follows:

$$(5.12) \quad \frac{d}{dt} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 & \sigma \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t) =: F(\sigma)x + G(\sigma)u,$$

where σ is the formal unit-time delay operator: $(\sigma x)(t) := x(t-1)$. This system is *reachable* in the sense of systems over rings, because $[G(\sigma), F(\sigma)G(\sigma)]$ generates $(R[\sigma])^2$. This indicates that there is a good deal of discrepancy between these two approaches to delay-differential systems. Observe also that matrices $sI - F(\sigma)$ and $G(s)$ even satisfy the following Bezout identity over $R[s, \sigma]$:

$$\begin{bmatrix} s & -\sigma \\ 1 & s \end{bmatrix} \begin{bmatrix} -s & 1-s \\ 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} [1 + s^2 + \sigma, s^2 - s + \sigma] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus left coprimeness (Bezout property) over $R[s, \sigma]$ does not necessarily imply left coprimeness over $\mathcal{E}'(R^-)$. The converse is also false as the following easy example shows.

Example 5.13. Let

$$W(s) = \frac{1}{se^s - 1} = \frac{\sigma}{s - \sigma}, \quad \sigma := e^{-s}.$$

While $(se^s - 1, 1)$ is clearly coprime, $(\sigma, s - \sigma)$ is not coprime over $R[s, \sigma]$.

6. Concluding remarks. In this paper we have given a concrete function space representation of shift realizations for the class of pseudo-rational input/output maps, by using the matrix fractional representation over $\mathcal{E}'(R^-)$. Kamen [9] seems to have been the first to consider the fractional representation over $\mathcal{E}'(R^-)$. Under a coprimeness condition, he successfully derived a characterization of the kernel of an input/output map. However, since his input and output spaces are spaces of distributions, his result does not lead to the familiar function space models as presented here. His result was later extended to the multivariable delay-differential systems by Denham and Yamashita [2], but without an explicit coprimeness condition—this is somewhat mysterious in view of the results in § 4 and in [9].

A number of coprimeness notions have appeared in the literature: Callier and Desoer [1], Desoer and Vidyasagar [3], Vidyasagar, Schneider, and Francis [17], and others have studied various coprimeness notions over the ring of stable transfer functions for the study of input/output stability and feedback stabilizability. Neither our framework nor theirs properly contains the other. While their approaches are focused on the input/output analysis and synthesis, ours is more appropriate for realization. Khargonekar [10] studied left coprimeness over a commutative ring and derived its relationship to Fuhrmann realizations over a commutative ring. The difference between the present approach based on the convolution algebra $\mathcal{E}'(R^-)$ and the polynomial ring approach is as discussed in § 5.

Appendix. We give a precise definition of the convolution of $\alpha \in \mathcal{E}'(R^-)$ and $\beta \in \mathcal{D}'[0, \infty)$.

It is easy to see that the inclusion mapping $j: \mathcal{D}[0, \infty) \hookrightarrow \mathcal{D}(R)$ is not only an injection but also a topological homomorphism. Hence its adjoint, namely π , is clearly surjective by the Hahn-Banach Theorem. Now take any $\alpha \in \mathcal{E}'(R^-)$ and $\beta \in \mathcal{D}'[0, \infty)$. then there exists an extension $\beta_1 \in \mathcal{D}'(R)$ (actually it can be taken in $\mathcal{D}'_+(R)$) such that $\pi(\beta_1) = \beta$. We define the convolution $\pi(\alpha * \beta) \in \mathcal{D}'[0, \infty)$ by

$$(A1) \quad \pi(\alpha * \beta) := (\alpha * \beta_1).$$

To show that this formula is well defined, let us prove the following lemma.

LEMMA A2. $\pi\alpha = 0$ if and only if $\text{supp } \alpha \subset (-\infty, 0]$.

Proof. Suppose $\pi\alpha = 0$. This means $\langle \alpha, \psi \rangle = 0$ for every ψ in $\mathcal{D}[0, \infty)$, i.e., $\text{supp } \alpha \subset (-\infty, 0]$. Conversely, suppose that $\text{supp } \alpha$ is contained in $(-\infty, 0]$. By definition, $\langle \alpha, \psi \rangle = 0$ for every $\psi \in \mathcal{D}(R)$ with support contained in $(0, \infty)$. By continuity of distributions, $\langle \alpha, \psi \rangle = 0$ for every $\psi \in \mathcal{D}[0, \infty)$, that is, $\pi\alpha = 0$. \square

Now suppose that $\pi(\beta_1) = \pi(\beta_2)$. We need to show that $\pi(\alpha * \beta_1) = \pi(\alpha * \beta_2)$. Since $\pi(\beta_1 - \beta_2) = 0$, $\text{supp } (\beta_1 - \beta_2)$ is contained in $(-\infty, 0]$, and hence so is $\text{supp } (\alpha * (\beta_1 - \beta_2))$. Now the lemma above implies that $\pi(\alpha * (\beta_1 - \beta_2)) = 0$, and hence $\pi(\alpha * \beta_1) = \pi(\alpha * \beta_2)$. Thus (A1) is well defined.

In particular, we have the following lemma.

LEMMA A3. $\pi(\alpha * \pi\beta) = \pi(\alpha * \beta)$, $\alpha \in \mathcal{E}'(R^-)$, $\beta \in \mathcal{D}'_+$.

Proof. The proof is obvious since β itself is an extension of $\pi\beta$. \square

Paley-Wiener Theorem for Laplace transforms of distributions. The following version of the Paley-Wiener-Schwartz Theorem for Laplace transforms can be obtained via a minor modification of the one used for Fourier transforms [15], [16], and is often useful for checking if a given impulse response is pseudo-rational.

THEOREM A4. A distribution α has compact support contained in $[-a, 0]$ if and only if its (two-sided) Laplace transform $\hat{\alpha}(s)$ is an entire function satisfying the following estimate for some $C > 0$ and an integer $m \geq 0$:

$$(A.5) \quad \begin{aligned} |\hat{\alpha}(s)| &\leq C(1+|s|)^m \exp(a \cdot \text{Re } s) \quad \text{for } \text{Re } s \geq 0, \\ &\leq C(1+|s|)^m \quad \text{for } \text{Re } s < 0. \end{aligned}$$

The proof is obtained from the standard form for Fourier transforms [15], [16] by replacing the Fourier variable ξ by is and then shifting the result by multiplying the transform by $\exp(-as/2)$ so that the resulting distribution has support contained in $[-a/2, a/2]$.

An important consequence of this theorem is that it is closed under a finitely many-zero cancellation. Denote by $\text{PWS}(a, m, C)$ the class of entire functions satisfying estimate (A5). We have the following proposition.

PROPOSITION A6. *Let $\alpha \in \text{PWS}(a, m, C_1)$. Suppose that $\alpha(\lambda) = 0$. Then $\beta := \alpha/(s - \lambda)$ belongs to $\text{PWS}(a, m, C_2)$ for some C_2 .*

Proof. Clearly β is an entire function. If $|s - \lambda| \geq 1$, we have

$$\begin{aligned} |\beta(s)| &\leq |\alpha(s)| \leq C(1 + |s|)^m \exp(a \cdot \operatorname{Re} s) \quad \text{for } \operatorname{Re} s \geq 0, \\ &\leq C(1 + |s|)^m \quad \text{for } \operatorname{Re} s < 0. \end{aligned}$$

On the other hand, if $|s - \lambda| \leq 1$,

$$|\beta(s)| \leq M,$$

because β is continuous on the disc $|s - \lambda| \leq 1$. Putting $C_2 := \max\{C_1, M\}$ completes the proof. \square

Acknowledgments. The author thanks the Japan Society for the Promotion of Science for financial support and the Center for Mathematical System Theory, University of Florida, Gainesville, Florida, for financial support during his stay there from April 14–20, 1983.

REFERENCES

- [1] F. M. CALLIER AND C. A. DESOER, *An algebra of transfer functions for distributed linear time-invariant systems*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 651–662.
- [2] M. J. DENHAM AND K. YAMASHITA, *On the controllability of delay-differential systems*, Internat. J. Control, 30 (1979), pp. 813–822.
- [3] C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.
- [4] P. DEWILDE, *Input-output description of roomy systems*, SIAM J. Control Optim. 14 (1976), pp. 712–736.
- [5] P. A. FUHRMANN, *Algebraic system theory: an analyst's point of view*, J. Franklin Inst., 301 (1976), pp. 521–540.
- [6] R. E. KALMAN, P. L. FALB, AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.
- [7] R. E. KALMAN AND M. L. J. HAUTUS, *Realization of continuous-time linear dynamical systems: rigorous theory in the style of Schwartz*, in Ordinary Differential Equations, NRL–Mathematics Research Center Conference, Academic Press, New York, 1972, pp. 151–164.
- [8] E. W. KAMEN, *On an algebraic theory of systems defined by convolution operators*, Math. Systems Theory, 9 (1975), pp. 57–74.
- [9] ———, Personal communication.
- [10] ———, *Module structure of infinite-dimensional systems with applications to controllability*, SIAM J. Control Optim., 14 (1976), pp. 389–408.
- [11] E. W. KAMEN, P. P. KHARGONEKAR, AND A. TANNENBAUM, *Proper stable Bezout factorizations and feedback control of linear time-delay systems*, Internat. J. Control, 43 (1986), pp. 837–857.
- [12] P. P. KHARGONEKAR, *On matrix fraction representations for linear systems over commutative rings*, SIAM J. Control Optim., 20 (1982), pp. 172–197.
- [13] A. MANITIUS, *Completeness and F-completeness of eigenfunctions associated with retarded functional differential equations*, J. Differential Equations, 35 (1980), pp. 1–29.
- [14] ———, *Necessary and sufficient conditions of approximate controllability for general linear retarded systems*, SIAM J. Control Optim., 16 (1981), pp. 516–532.
- [15] J. RISSANEN, *Recursive identification of linear systems*, SIAM J. Control, 9 (1971), pp. 420–430.
- [16] H. H. SCHAEFER, *Topological Vector Spaces*, Springer-Verlag, New York, Berlin, 1971.
- [17] L. SCHWARTZ, *Théorie des distributions*, Second edition, Hermann, Paris, 1966.
- [18] F. TREVES, *Topological Vector Spaces, Distributions and Kernels*, Academic Press, New York, 1967.
- [19] M. VIDYASAGAR, H. SCHNEIDER, AND B. A. FRANCIS, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 880–894.
- [20a] Y. YAMAMOTO, *Realization theory of infinite-dimensional linear systems, Part I*, Math. Systems Theory, 15 (1981), pp. 55–77.
- [20b] ———, *Realization theory of infinite-dimensional linear systems, Part II*, Math. Systems Theory, 15 (1982), pp. 169–190.

- [21] ———, *Module structure of constant linear systems and its applications to controllability*, J. Math. Anal. Appl., 83 (1981), pp. 411–437.
- [22] ———, *A note on linear input/output maps of bounded-type*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 733–734.
- [23] Y. YAMAMOTO AND S. UESHIMA, *A new model for neutral delay-differential systems*, Internat. J. Control, 43 (1985), pp. 465–472.

FEEDBACK CONTROLLED DIFFERENTIAL INCLUSIONS AND STABILIZATION OF UNCERTAIN DYNAMICAL SYSTEMS*

D. P. GOODALL† AND E. P. RYAN‡

Abstract. Deterministic design of stabilizing feedback controls for a class of uncertain dynamical systems is considered. Adopting a problem formulation based on differential inclusions (to encompass all possible realizations of uncertainty), the authors describe a class of generalized feedback controls (with practical analogues in the form of discontinuous feedbacks) that guarantees global uniform ultimate boundedness. Moreover, under a matched uncertainty hypothesis, this class of controls guarantees global uniform asymptotic stability of the zero state and ultimate attainment of prescribed model behaviour.

Key words. uncertain dynamical systems, differential inclusions, stability, nonlinear feedback

AMS(MOS) subject classification. 93

1. Introduction. Consider a controlled dynamical system of the form

$$(1) \quad \dot{x}(t) = H(t, x(t), u(t)), \quad x(t) \in \mathbb{R}^n, \quad u(t) \in \mathbb{R}^m.$$

If (1) corresponds to a mathematical model of a “real world” process then, almost surely, some approximation, imprecision, or uncertainty will have been introduced during the modeling procedure; the function H is, at best, a “reasonable” representation of the true controlled vector field. If (1) is deemed to be a satisfactory model, then the theories of classical and controlled differential equations may provide the appropriate framework for analysis and control design. Note, however, that if control synthesis is an objective, then *discontinuous* feedback ($u(t) = D(t, x(t))$) is a natural candidate in many problems of stabilization and optimization, in which case the associated differential equation ($\dot{x}(t) = H_D(t, x(t))$, $H_D(t, x) := H(t, x, D(t, x))$), modeling the feedback controlled system, fails to satisfy the requisite hypotheses of the classical theory. Returning to the basic modeling problem, in many cases the determination of an acceptably accurate model of form (1) is impossible, i.e., uncertainty may be an intrinsic feature. For example, it may only be possible to determine a model structure that is specified up to a collection of parameters, the values of which are unknown (but possibly restricted to known sets); furthermore, realistic processes are frequently subject to extraneous disturbances (again possibly with known bounds). In cases for which such “randomness” in the model admits a statistical characterization, stochastic control theory may be appropriate. On the other hand, if structural properties and bounds relating to the uncertainty are known, then a deterministic treatment may be feasible.

Deterministic control of uncertain dynamical systems has been the focus of much research in recent years (see, e.g., [2], [3], [5]–[10], [13], and references therein) and is the subject of this paper. In particular, the difficulties, alluded to above, associated

* Received by the editors July 22, 1987; accepted for publication (in revised form) February 15, 1988. A preliminary version of this paper appeared in the Proceedings of the 25th Annual IEEE Conference on Decision and Control, Athens, Greece, 1986.

† Department of Computing, Electronic, and Mathematical Studies, GLOSCAT, Park Campus, Cheltenham GL50 2RR, United Kingdom. Present address: Department of Mathematics, Coventry Polytechnic, Coventry CV1 5FB, United Kingdom.

‡ School of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom. The research of this author was partly supported by United Kingdom Science and Engineering Research Council grant GR/D/60638.

with modeling uncertainty and discontinuous feedback are addressed. The approach is in the spirit of [13] but with a fundamental distinction: in [13], functional properties of the uncertain system are assumed that ensure, for any (bounded measurable) control $u(\cdot)$ and any admissible realization of uncertainty, the classical (Carathéodory) concept of solution of the underlying differential equation is adequate; in the present paper, the uncertain system is more generally defined via a differential inclusion, the right-hand side of which takes the form of a known multifunction, or set-valued map, $(t, x, u) \mapsto \mathcal{H}(t, x, u)$, so that $\mathcal{H}(t, x(t), u(t))$ is the set of all candidate velocities $\dot{x}(t)$ of the uncertain system at time t . Within this general class of systems, attention will be restricted to a subclass (essentially comprised of perturbed linear systems) for which $\mathcal{H}(t, x, u) = Ax + Bu + \mathcal{G}(t, x, u)$, where \mathcal{G} is a compact-valued multifunction with a specific structure made precise in § 3.

2. Preliminaries. Before describing the class of uncertain systems to be considered and the associated problem of stabilization by feedback to be studied, some notation and general properties of multifunctions are introduced [1], [4].

Let Y_1 and Y_2 be Hausdorff topological spaces. A multifunction $\mathcal{Y}: Y_1 \rightrightarrows Y_2$, $y \mapsto \mathcal{Y}(y) \subset Y_2$, is a mapping of Y_1 into the subsets of Y_2 . \mathcal{Y} (with nonempty values) is said to be *upper semicontinuous* at $y_1 \in Y_1$ if, for each open set $\mathcal{N}_2 \supset \mathcal{Y}(y_1)$, there exists a neighbourhood \mathcal{N}_1 of y_1 such that $\mathcal{Y}(\mathcal{N}_1) \subset \mathcal{N}_2$. \mathcal{Y} is said to be *upper semicontinuous* (u.s.c.) if it is u.s.c. at each $y_1 \in Y_1$.

If, as in the present paper, Y_1 and Y_2 are real Banach spaces and $\mathcal{Y}: Y_1 \rightrightarrows Y_2$ has *compact* values, then the above definition of u.s.c. of \mathcal{Y} at $y_1 \in Y_1$ is equivalent to the following: for each $\varepsilon > 0$, there exists $\delta > 0$ such that¹ $\mathcal{Y}(y) \subset \mathcal{Y}(y_1) + \varepsilon \mathcal{B}_{Y_2}$, for all $y \in y_1 + \delta \mathcal{B}_{Y_1}$, where \mathcal{B}_Z denotes the open unit ball in Banach space Z (with closure $\overline{\mathcal{B}}_Z$).

The following properties will be invoked later.

PROPOSITION 1. *Let $K \subset Y_1$ be compact. If $\mathcal{Y}: Y_1 \rightrightarrows Y_2$ is upper semicontinuous with compact values, then $\mathcal{Y}(K) \subset Y_2$ is compact.*

PROPOSITION 2. *Let $\mathcal{Y}_1: Y_1 \rightrightarrows Y_2$ and $\mathcal{Y}_2: Y_2 \rightrightarrows Y_3$ have nonempty values. Define the composition $\mathcal{Y}_2 \circ \mathcal{Y}_1: Y_1 \rightrightarrows Y_3$ by $y \mapsto (\mathcal{Y}_2 \circ \mathcal{Y}_1)(y) = \mathcal{Y}_2(\mathcal{Y}_1(y)) := \bigcup_{v \in \mathcal{Y}_1(y)} \mathcal{Y}_2(v)$. If \mathcal{Y}_1 and \mathcal{Y}_2 are upper semicontinuous, then $\mathcal{Y}_2 \circ \mathcal{Y}_1$ is upper semicontinuous.*

PROPOSITION 3. *If $f: Y_1 \rightarrow \mathbb{R}$ is continuous and $\mathcal{Y}: Y_1 \rightrightarrows Y_2$ is upper semicontinuous with compact values, then $f\mathcal{Y}: Y_1 \rightrightarrows Y_2$, $y \mapsto f(y)\mathcal{Y}(y)$, is upper semicontinuous with compact values.*

Here, multifunctions on $\mathbb{R} \times X \times U$, $\mathbb{R} \times X$, or U will be considered, with $X := \mathbb{R}^n$ and $U := \mathbb{R}^m$ ($1 \leq m \leq n$) throughout. The Euclidean inner product (on X or U as appropriate) and the induced norm are denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively. For a linear map L , $\|L\| = [\max \sigma(L^T L)]^{1/2}$, where σ denotes spectrum and no distinction is made between a map L and its matrix representation.

For a compact subset $K \neq \emptyset$ of X or U , $\xi(K) := \max \{\|v\|: v \in K\}$ and $\xi(\emptyset) := 0$. For subspace $S \subset X$, Π_S denotes the orthogonal projector onto S . Finally, for $x \in X$ and $S \subset X$, $\langle x, S \rangle := \{\langle x, s \rangle: s \in S\} \subset \mathbb{R}$.

3. The system. The system to be considered is of the form

$$(2) \quad \dot{x}(t) \in \mathcal{H}(t, x(t), u(t))$$

where $\mathcal{H}: \mathbb{R} \times X \times U \rightrightarrows X$ is a known multifunction with nonempty values. For a given function (control) $u: I \rightarrow U$, defined on some interval $I \subset \mathbb{R}$, $x: I \rightarrow X$ is deemed a

¹ For $y \in Y$ and $S_1, S_2 \subset Y$, $y + S_1 := \{y + s_1: s_1 \in S_1\}$, $S_1 + S_2 := \{s_1 + s_2: s_1 \in S_1; s_2 \in S_2\}$.

solution or trajectory of (2) if it is absolutely continuous and satisfies (2) almost everywhere on I . To ensure existence of solutions, conditions of upper semicontinuity, and convexity and compactness of its values, will be imposed on \mathcal{H} ; moreover, to ensure feedback stabilizability (in a sense to be defined), the allowable class of multifunctions must be further restricted. The requisite hypotheses are contained in assumptions (A1) (below) and (A2) (to be introduced later).

- (A1) There exists a controllable pair $(A, B) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times m}$, with B of full rank $m \leq n$, such that the multifunction $(t, x, u) \mapsto \mathcal{G}(t, x, u) := \mathcal{H}(t, x, u) - Ax - Bu$ has the following property. There exist multifunctions $\mathcal{G}_r: \mathbb{R} \times X \rightrightarrows \ker B^T$, $\mathcal{G}_m: \mathbb{R} \times X \rightrightarrows U$ and a continuous function $\beta: \mathbb{R} \rightarrow [0, \kappa_0]$, $\kappa_0 < 1$, such that we have the following:
- (i) $\Pi_{\ker B^T} \mathcal{G}(t, x, u) = \mathcal{G}_r(t, x)$ for all (t, x, u) ;
 - (ii) $\Pi_{\text{im } B} \mathcal{G}(t, x, u) = B[\mathcal{G}_m(t, x) + \beta(t)\mathcal{G}_c(u)]$ for all (t, x, u) , where $\mathcal{G}_c: u \mapsto \|u\| \mathcal{B}_U$;
 - (iii) \mathcal{G}_r and \mathcal{G}_m are upper semicontinuous with convex and compact values.

Remarks. (i) In the terminology of [2], [3], [13], the multifunctions \mathcal{G}_m and $\beta\mathcal{G}_c$ may be interpreted as modeling the *matched uncertainty* in the system, while \mathcal{G}_r models *residual uncertainty*.

(ii) A simple example of a class of matched uncertain systems that can be embedded into the more general class considered here is typified by the following:

$$\dot{x}(t) = Ax(t) + B[u(t) + g(t, x(t), u(t))]$$

where g is an unknown function with

$$\|g(t, x, u)\| \leq \psi(t, x) + \beta(t)\|u\| \quad \text{for all } (t, x, u)$$

and where the bounding functions $\psi: \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^+ := [0, \infty)$ and $\beta: \mathbb{R} \rightarrow [0, \kappa_0]$, $\kappa_0 < 1$, are known and continuous.

The fundamental problem to be studied is that of stabilization by feedback, viz. determine a (time-dependent) feedback strategy $(t, x) \mapsto \mathcal{F}(t, x)$ such that, loosely speaking, all trajectories of (2), with $u(t) \in \mathcal{F}(t, x(t))$, exhibit “stable” behaviour. Here, the class of admissible feedbacks is taken to be the class of upper semicontinuous multifunctions $\mathcal{F}: \mathbb{R} \times X \rightrightarrows U$ with nonempty, convex, and compact values, henceforth referred to as *generalized feedbacks*.

Remark. In the present paper, the proposed generalized feedbacks are singleton-valued except on a set $\Gamma_{\mathcal{F}}$ of Lebesgue measure zero in $\mathbb{R} \times X$. The set $\Gamma_{\mathcal{F}}$ may be identified as a switching surface of control discontinuities.

4. Stabilization by generalized feedback: problem formulation. The basic stabilization problem may now be stated as follows. Determine a generalized feedback \mathcal{F} such that the feedback-controlled differential inclusion

$$(3a) \quad \dot{x}(t) \in \mathcal{H}_{\mathcal{F}}(t, x(t)),$$

$$(3b) \quad \mathcal{H}_{\mathcal{F}}(t, x) := Ax + \mathcal{G}_r(t, x) + B\mathcal{F}(t, x) + B[\mathcal{G}_m(t, x) + \beta(t)(\mathcal{G}_c \circ \mathcal{F})(t, x)]$$

exhibits a compact set (positively invariant and containing the origin) that is globally uniformly asymptotically stable (g.u.a.s.) in the following sense.

DEFINITION 1. Compact $S \subset X$ is a g.u.a.s. set for system (3) if the following hold:

(i) *Existence and continuation of solutions.* For each $(t_0, x^0) \in \mathbb{R} \times X$, there exists a (local) solution $x: [t_0, \tau) \rightarrow X$ with $x(t_0) = x^0$, and every such solution can be extended into a solution on $[t_0, \infty)$.

(ii) *Uniform boundedness of solutions.* For each $\delta > 0$, there exists $d(\delta) > 0$ such that $x(t) \in d(\delta)\mathcal{B}_X$, for all t on every solution $x: [t_0, \infty) \rightarrow X$ with $x(t_0) \in \delta\mathcal{B}_X$.

(iii) *Uniform stability of S .* For each $\delta > 0$, there exists $D(\delta) > 0$ such that $x(t) \in S + \delta\mathcal{B}_X$, for all t on every solution $x: [t_0, \infty) \rightarrow X$ with $x(t_0) \in S + D(\delta)\mathcal{B}_X$.

(iv) *Uniform attractivity of S .* For each $\delta > 0$ and $\varepsilon > 0$, there exists $T(\delta, \varepsilon) \geq 0$ such that $x(t) \in S + \varepsilon\mathcal{B}_X$, for all $t \geq t_0 + T(\delta, \varepsilon)$ on every solution $x: [t_0, \infty) \rightarrow X$ with $x(t_0) \in S + \delta\mathcal{B}_X$.

Remarks. (i) Ideally, a generalized feedback \mathcal{F} is sought that renders $S = \{0\}$ g.u.a.s.; it will be shown (by construction) that such a feedback exists if $\mathcal{G}_r \equiv \{0\}$ (i.e., in the absence of residual uncertainty). If $\mathcal{G}_r \neq \{0\}$, then, in essence, \mathcal{F} is sought such that (3) exhibits a g.u.a.s. compact set S (containing the origin) with acceptably small diameter.

(ii) If compact S is a g.u.a.s. set for (3), then (3) is globally uniformly ultimately bounded (in the sense of [5]) within every open set containing S .

5. Proposed generalized feedback. In this section, a generalized feedback \mathcal{F} is constructed which, in the absence of residual uncertainty (i.e., system (3) with $\mathcal{G}_r \equiv \{0\}$), renders the zero state of (3) g.u.a.s. The approach is akin to that of [13] and draws on concepts from (a) variable structure systems theory [14], and (b) Lyapunov-based theory [5]. In particular, the concept of an invariant $(n-m)$ -dimensional subspace $W \subset X$ is adopted from variable structure systems theory. More specifically, a set $\Lambda = \{\lambda_1, \dots, \lambda_{n-m}\} \subset \mathbb{C}^-$ (the open left half complex plane) of $(n-m)$ *ideal model* eigenvalues is selected that implicitly determines $W \subset X$. The feedback is then designed (via Lyapunov based analysis) to ensure the following: (i) W is invariant under $\mathcal{H}_{\mathcal{F}}$ and the flow on W (in the absence of residual uncertainty) is equivalent to that of a linear system (*the ideal model*) with prescribed spectrum Λ ; (ii) W is globally finite-time attractive in the sense that W is ultimately attained on every solution of (3). Thus, in the context of Definition 1, the proposed feedback \mathcal{F} guarantees in the absence of residual uncertainty that $\{0\}$ is a g.u.a.s. set. It is important to note that, in addition to asymptotic stability, the proposed feedback \mathcal{F} ensures that the flow (in the absence of residual uncertainty) exhibits other desired features, namely that *all* trajectories $x(\cdot)$ of (3) ultimately coincide with a trajectory of a linear system with prescribed spectrum Λ ; more precisely, with $L_1 \in \mathbb{R}^{(n-m) \times n}$ and $M^* \in \mathbb{R}^{n \times (n-m)}$ defined as in § 5.1, $x(t) = M^*w(t)$ for all t sufficiently large, where $w(\cdot)$ satisfies the linear equation $\dot{w} = (L_1AM^*)w$, with spectrum $\sigma(L_1AM^*) = \Lambda$. If residual uncertainty is present (i.e., if $\mathcal{G}_r \neq \{0\}$) then, under assumption (A2) below, invariance and global finite-time attractivity of W are preserved. However, ideal model motion on W cannot be guaranteed; instead, g.u.a.s. of a calculable compact set $S \subset W$ only is assured, the diameter of which is determined by bounds on the values of \mathcal{G}_r .

5.1. The subspace W . Let $L_1 \in \mathbb{R}^{(n-m) \times n}$ be such that $\ker L_1 = \text{im } B$. Define $L_2 := (B^TB)^{-1}B^T$ and $L := \begin{bmatrix} L_1 \\ L_2 \end{bmatrix}$ with inverse $L^{-1} = R = \begin{bmatrix} R_1 \\ B \end{bmatrix}$, then

$$LAR = \begin{bmatrix} L_1AR_1 & L_1AB \\ L_2AR_1 & L_2AB \end{bmatrix}, \quad LB = \begin{bmatrix} 0 \\ I \end{bmatrix}.$$

Let $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_{n-m}\} \subset \mathbb{C}^-$ be the ideal model spectrum. Noting that (L_1AR_1, L_1AB) is a controllable pair (since, by (A1), (A, B) is controllable), there exists $M \in \mathbb{R}^{m \times (n-m)}$ such that

$$\sigma(L_1AR_1 + L_1ABM) = \sigma(L_1AM^*) = \Lambda,$$

where, for notational convenience, $M^* := R_1 + BM$. The $(n-m)$ -dimensional subspace

$W \subset X$ is now defined as

$$(4) \quad W := \ker \bar{M}, \quad \bar{M} := L_2 - ML_1.$$

Suppose initially that there is no uncertainty in the system (i.e., suppose that $\mathcal{G}_r \equiv \{0\}$, $\mathcal{G}_m \equiv \{0\}$ and $\beta \equiv 0$). Clearly, if $F \in \mathbb{R}^{m \times n}$ is such that (i) $\sigma(A + BF) \subset \mathbb{C}^-$, and (ii) W is $(A + BF)$ -invariant, then asymptotic stability of the origin and ideal model motion on W is assured under the feedback $u(t) = Fx(t)$. One such F is the following:

$$(5) \quad F := C\bar{M} - \bar{M}A \quad \text{with } \sigma(C) \subset \mathbb{C}^-$$

(in which case it is readily verified that (i) $\sigma(A + BF) = \sigma(L_1AM^*) \cup \sigma(C) \subset \mathbb{C}^-$, and (ii) for each $x \in W$, $\bar{M}(A + BF)x = \bar{M}(A - B\bar{M}A)x = 0$).

Remark. Clearly, there is considerable scope for judicious choice of the operators M and C . For example, it may be desirable to attempt to maximize the stability radius [11]–[12] of $L_1AM^* = L_1AR_1 + L_1ABM$ over the set of matrices $M \in \mathbb{R}^{m \times (n-m)}$ that assign the spectrum Λ to L_1AM^* .

In the ensuing section, the linear feedback operator F is augmented by an appropriate multifunction to yield a generalized feedback \mathcal{F} that preserves the property of asymptotic stability of the origin in the presence of matched uncertainty.

5.2. The multifunction \mathcal{F} . Let $P_1 \in \mathbb{R}^{(n-m) \times (n-m)}$ and $P_2 \in \mathbb{R}^{m \times m}$ denote the unique positive-definite symmetric solutions of the Lyapunov equations

$$(6) \quad P_1 L_1 A M^* + (L_1 A M^*)^T P_1 + I = 0$$

and

$$(7) \quad P_2 C + C^T P_2 + I = 0.$$

Define multifunctions $\mathcal{P}_1: X \rightrightarrows \mathbb{R}^{n-m}$ and $\mathcal{P}_2: X \rightrightarrows U$ by

$$(8) \quad x \mapsto \mathcal{P}_1(x) := \{v \in \mathbb{R}^{n-m} : \langle v, P_1 L_1 x \rangle \geq 0\}$$

and

$$(9) \quad x \mapsto \mathcal{P}_2(x) := \{u \in U : \langle u, P_2 \bar{M}x \rangle \geq 0\}.$$

The final assumption on the residual uncertainty is now imposed.

$$(A2) \quad \xi(L_1 \mathcal{G}_r(t, x) \cap \mathcal{P}_1(x)) \leq \kappa_1 \|L_1 x\| + \kappa_2 \|\bar{M}x\| + \kappa_3 \text{ for all } (t, x), \text{ where } \kappa_1, \kappa_2, \text{ and } \kappa_3 \text{ are nonnegative constants with } \kappa_1 < \frac{1}{2} \|P_1\|^{-1}.$$

Define multifunction $\mathcal{D}: U \rightrightarrows U$ as

$$(10) \quad u \mapsto \mathcal{D}(u) := \begin{cases} \|u\|^{-1} u, & u \neq 0, \\ \bar{\mathcal{B}}_U, & u = 0 \end{cases}$$

which is clearly upper semicontinuous ($\mathcal{D}(u)$ is the generalized gradient [4] of $\|\cdot\|: U \rightarrow \mathbb{R}^+$ at u).

The proposed feedback is now given by the multifunction

$$(11a) \quad (t, x) \mapsto \mathcal{F}(t, x) := Fx - \rho(t, x) \mathcal{D}(P_2 \bar{M}x)$$

where F is given by (5) and ρ is any continuous functional on $\mathbb{R} \times X$ satisfying

$$(11b) \quad \rho(t, x) \geq \rho_0(t, x) := (1 - \beta(t))^{-1} \cdot [\beta(t) \|Fx\| + \xi((\mathcal{G}_m(t, x) - ML_1 \mathcal{G}_r(t, x)) \cap \mathcal{P}_2(x)) + \gamma]$$

where $\gamma > 0$ is a design parameter. Note that the continuity of ρ and the upper semicontinuity of \mathcal{D} ensure (by Proposition 3) that \mathcal{F} is upper semicontinuous and clearly takes convex and compact values; thus, \mathcal{F} qualifies as a generalized feedback. Note further that \mathcal{F} is singleton-valued except on the set $\Gamma_{\mathcal{F}} = \mathbb{R} \times W$.

Remark. The intersection $(\mathcal{G}_m(t, x) - ML_1 \mathcal{G}_r(t, x)) \cap \mathcal{P}_2(x)$ is adopted in (11b) in order to economize on control gain by exploiting the possible occurrence of “stability enhancing” uncertainties.

6. Stability properties of the feedback system.

6.1. Existence, continuation, and uniform boundedness of solutions. For each $(t_0, x^0) \in \mathbb{R} \times X$, the existence of at least one local solution $x: [t_0, \tau) \rightarrow X$ with $x(t_0) = x^0$ will first be established. To this end, the following lemma (essentially Theorem 3 of [1, Chap. 2]) will be invoked.

LEMMA 1. *If $\mathcal{H}_{\mathcal{F}}$ is upper semicontinuous with convex and compact values, then, for each (t_0, x^0) , there exists a local solution $x: [t_0, \tau) \rightarrow X$ of (3) with $x(t_0) = x^0$.*

Thus, it suffices to show that $\mathcal{H}_{\mathcal{F}}$, given by (3) and (11), satisfies the hypotheses of the lemma. Since $\beta \mathcal{G}_c$ and \mathcal{F} are upper semicontinuous with compact values, it follows (by Propositions 1 and 2) that $(t, x) \mapsto \beta(t)(\mathcal{G}_c \circ \mathcal{F})(t, x)$ is also upper semicontinuous with compact values; thus, $\mathcal{H}_{\mathcal{F}}$ is the sum of upper semicontinuous multifunctions with compact values, and hence is itself upper semicontinuous with compact values. Since $\beta \mathcal{G}_c$ is convex valued and $\mathcal{F}(t, x)$ is either a singleton or a closed ball (of radius $\rho(t, x)$), it follows that $(t, x) \mapsto \beta(t)(\mathcal{G}_c \circ \mathcal{F})(t, x)$ is convex valued; thus, $\mathcal{H}_{\mathcal{F}}$ is the sum of convex-valued multifunctions, and hence is itself convex valued. For each (t_0, x^0) , the existence of a local solution $x: [t_0, \tau) \rightarrow X$ with $x(t_0) = x^0$ now follows by Lemma 1.

To establish that every such solution can be extended into a solution on $[t_0, \infty)$, consider the behaviour, along local solutions, of the function $V: X \rightarrow \mathbb{R}^+$ given by

$$(12) \quad x \mapsto V(x) := \frac{1}{2} \left\langle x, \begin{bmatrix} L_1^T & \bar{M}^T \end{bmatrix} Q_{\zeta} \begin{bmatrix} L_1 \\ \bar{M} \end{bmatrix} x \right\rangle$$

where $Q_{\zeta} := \begin{bmatrix} P_1 & 0 \\ 0 & \zeta P_2 \end{bmatrix}$ and ζ is a positive real number (to be specified).

Along each (local) solution $x: [t_0, \tau) \rightarrow X$

$$(13a) \quad \dot{V}(x(t)) \in \mathcal{L}(t, x(t)) = \mathcal{L}_1(t, x(t)) + \mathcal{L}_2(t, x(t)) \quad \text{a.e.}$$

where

$$(13b) \quad \mathcal{L}_1(t, x) := \langle L_1 \mathcal{H}_{\mathcal{F}}(t, x), P_1 L_1 x \rangle \subset \mathbb{R},$$

$$(13c) \quad \mathcal{L}_2(t, x) := \zeta \langle \bar{M} \mathcal{H}_{\mathcal{F}}(t, x), P_2 \bar{M} x \rangle \subset \mathbb{R}.$$

Thus, if it can be shown that there exists $s \in \mathbb{R}^+$ such that

$$(14) \quad \mathcal{L}(t, x) \cap [0, \infty) = \emptyset \quad \forall (t, x) \in [t_0, \infty) \times (X \setminus s\bar{\mathcal{B}}_X),$$

then it follows that every solution $x: [t_0, \tau) \rightarrow X$ with $x(t_0) = x^0$ evolves within the compact set

$$\left\{ x: V(x) \leq \max \left\{ V(x^0), \frac{1}{2} \left\| \begin{bmatrix} L_1^T & \bar{M}^T \end{bmatrix} Q_{\zeta} \begin{bmatrix} L_1 \\ \bar{M} \end{bmatrix} \right\| s^2 \right\} \right\},$$

and hence can be continued indefinitely. The property of uniform boundedness of solutions readily follows. It remains to establish the existence of $s \in \mathbb{R}^+$ such that (14) holds. Now, using (6), we have

$$\mathcal{L}_1(t, x) = -\frac{1}{2} \|L_1 x\|^2 + \langle L_1 A B \bar{M} x, P_1 L_1 x \rangle + \langle L_1 \mathcal{G}_r(t, x), P_1 L_1 x \rangle,$$

whence, in view of (A2),

$$\max \mathcal{L}_1(t, x) \leq -\frac{1}{2}(1 - 2\kappa_1 \|P_1\|) \|L_1 x\|^2 + (\|P_1 L_1 A B\| + \kappa_2 \|P_1\|) \|L_1 x\| \|\bar{M}x\| + \kappa_3 \|P_1\| \|L_1 x\|.$$

Using (7), we have

$$\begin{aligned} \mathcal{L}_2(t, x) = & \zeta[-\frac{1}{2}\|\bar{M}x\|^2 - \rho(t, x)\|P_2 \bar{M}x\| + \langle \mathcal{G}_m(t, x) - M L_1 \mathcal{G}_r(t, x) \\ & + \beta(t)(\mathcal{G}_c \circ \mathcal{F})(t, x), P_2 \bar{M}x \rangle], \end{aligned}$$

whence, in view of (11b),

$$\max \mathcal{L}_2(t, x) \leq -\frac{1}{2}\zeta \|\bar{M}x\|^2.$$

Combining the above, we obtain

$$(15a) \quad \max \mathcal{L}(t, x) \leq -\frac{1}{2} \left\langle \begin{bmatrix} \|L_1 x\| \\ \|\bar{M}x\| \end{bmatrix}, E_\zeta \begin{bmatrix} \|L_1 x\| \\ \|\bar{M}x\| \end{bmatrix} \right\rangle + \kappa_3 \|P_1\| \|L_1 x\|$$

where

$$(15b) \quad E_\zeta := \begin{bmatrix} 1 - 2\kappa_1 \|P_1\| & -(\|P_1 L_1 A B\| + \kappa_2 \|P_1\|) \\ -(\|P_1 L_1 A B\| + \kappa_2 \|P_1\|) & \zeta \end{bmatrix}.$$

By virtue of (A2) and when we choose ζ such that

$$\zeta > (1 - 2\kappa_1 \|P_1\|)^{-1} [\|P_1 L_1 A B\| + \kappa_2 \|P_1\|]^2,$$

the first term on the right-hand side of (15a) is a negative definite quadratic form in x , which clearly ensures the existence of $s \in \mathbb{R}^+$ such that (14) holds. Note that, if $\kappa_3 = 0$, then $s = 0$ suffices.

The analysis above is summarized in the following lemma and corollary.

LEMMA 2. *The feedback-controlled differential inclusion system (3) and (11) exhibits the properties of (i) existence and continuation of solutions, and (ii) uniform boundedness of solutions.*

COROLLARY. *If $\kappa_3 = 0$, then the zero state is a globally uniformly asymptotically stable equilibrium of system (3) and (11).*

6.2. Positive invariance and finite-time attractivity of W . To establish that every solution of the feedback controlled differential inclusion attains the subspace W in finite time and thereafter remains in W , the behaviour of the function $V_2: X \rightarrow \mathbb{R}^+$, $x \mapsto V_2(x) := \frac{1}{2}\langle x, \bar{M}^T P_2 \bar{M}x \rangle$, along solutions is considered.

LEMMA 3. *For each $(t_0, x^0) \in \mathbb{R} \times X$, the subspace W is attained in finite time:*

$$t_f < \gamma^{-1} [2\|P_2^{-1}\| V_2(x^0)]^{1/2}.$$

Proof. Along every solution $x: [t_0, \infty) \rightarrow X$ with $x(t_0) = x^0$ the following holds almost everywhere:

$$(16) \quad \begin{aligned} \dot{V}_2(x(t)) & \leq -\gamma \|P_2 \bar{M}x(t)\| \\ & \leq -\gamma [2\|P_2^{-1}\|^{-1} V_2(x(t))]^{1/2}, \end{aligned}$$

which, on integration, ensures that $V_2(x(t_f + t_0)) = 0$ ($\Leftrightarrow x(t_f + t_0) \in W$) for some $t_f < \gamma^{-1} [2\|P_2^{-1}\| V_2(x^0)]^{1/2}$; moreover, $V_2(x(t)) = 0 \Leftrightarrow x(t) \in W$, for all $t \geq t_f + t_0$, i.e., the subspace W is positively $\mathcal{H}_{\mathcal{F}}$ -invariant. \square

Remark. Note that the upper bound on the time required to attain W is inversely proportional to the controller parameter $\gamma > 0$.

6.3. Motion in W . In view of the $\mathcal{H}_{\mathcal{F}}$ -invariance of W , $\bar{M}x(\cdot) \equiv 0$ on every solution $x: [t_0, \infty) \rightarrow X$ of the feedback system with $x(t_0) = x^0 \in W$. Thus, motion in W is governed by

$$(17a) \quad \dot{w}(t) \in L_1 A M^* w(t) + L_1 \mathcal{G}_r(t, M^* w(t)), \quad w(t) \in \mathbb{R}^{n-m},$$

$$(17b) \quad w(t_0) = L_1 x^0$$

in the sense that $x(\cdot): [t_0, \infty) \rightarrow W$, $x(t_0) = x^0 \in W$, is a trajectory of the feedback controlled system (3) and (11) if and only if $x(\cdot) = M^* w(\cdot)$, where $w(\cdot): [t_0, \infty) \rightarrow \mathbb{R}^{n-m}$ solves the initial value problem (17).

In the absence of residual uncertainty, i.e., for $\mathcal{G}_r \equiv \{0\}$, (17a) reduces to the *asymptotically stable linear ideal model* equation

$$(18) \quad \dot{w}(t) = L_1 A M^* w(t), \quad \sigma(L_1 A M^*) = \Lambda \subset \mathbb{C}^-;$$

Lemma 4 immediately follows.

LEMMA 4. *Let $\mathcal{G}_r \equiv \{0\}$. Then, for each $(t_0, x^0) \in \mathbb{R} \times W$, the feedback-controlled differential inclusion, given by (3) and (11), admits a unique solution $x: [t_0, \infty) \rightarrow W$, with $x(t_0) = x^0$, given by*

$$x(\cdot) = M^* \exp [L_1 A M^* (\cdot - t_0)] L_1 x^0$$

where $\sigma(L_1 A M^*) = \Lambda \subset \mathbb{C}^-$.

If residual uncertainty is present, i.e., if $\mathcal{G}_r \neq \{0\}$, then uniqueness of solutions and ideal model motion in W cannot be guaranteed. However, the feedback-controlled system has an attractive compact set S (the diameter of which is proportional to the parameter κ_3), quantified in the following lemma and corollary.

LEMMA 5. *Define $V_1: \mathbb{R}^{n-m} \rightarrow \mathbb{R}^+$, $w \mapsto \frac{1}{2} \langle w, P_1 w \rangle$. Then*

$$(19) \quad \Sigma := \{w: V_1(w) \leq \frac{1}{2} r^2 \|P_1\|\}, \quad r := 2\kappa_3 [1 - 2\kappa_1 \|P_1\|]^{-1} \|P_1\|$$

is a globally uniformly asymptotically stable set for system (17).

Proof. When we invoke (A2) the following holds along every solution $w(\cdot)$ of (17):

$$\dot{V}_1(w(t)) \leq -\frac{1}{2} [1 - 2\kappa_1 \|P_1\|] \|w(t)\|^2 + \kappa_3 \|P_1\| \|w(t)\| \quad \text{a.e.}$$

from which the required result may be deduced by standard arguments. \square

COROLLARY. *$S := \{x \in X: L_1 x \in \Sigma\} \cap W$ is a uniformly attractive set for motion in W of the feedback controlled system (3) and (11).*

Proof. An immediate consequence of the above lemma and the positive $\mathcal{H}_{\mathcal{F}}$ -invariance of W . \square

Finally, for $(t_0, x^0) \in \mathbb{R} \times W$, let $\Delta x(\cdot)$ denote the deviation of a trajectory $x(\cdot)$ of the feedback system (with $x(t_0) = x^0$) from ideal model motion. Specifically,

$$(20a) \quad \Delta x(t) = M^* y(t), \quad t \geq t_0$$

where

$$(20b) \quad y(t) := L_1 x(t) - \exp [L_1 A M^* (t - t_0)] L_1 x^0.$$

Noting that $y(\cdot)$ solves the initial value problem

$$(21) \quad \begin{aligned} \dot{y}(t) &\in L_1 A M^* y(t) + L_1 \mathcal{G}_r(t, x(t)), \\ y(t_0) &= 0, \end{aligned}$$

and again invoking (A2), we have

$$(22) \quad \dot{V}_1(y(t)) \leq -\frac{1}{2} \|y(t)\|^2 + \|P_1\| \|y(t)\| [\kappa_1 \|L_1 x(t)\| + \kappa_3] \quad \text{a.e.}$$

Moreover, since $L_1 x(\cdot)$ solves (17),

$$(23) \quad \dot{V}_1(L_1 x(t)) \leq -\frac{1}{2} [1 - 2\kappa_1 \|P_1\|] \|L_1 x(t)\|^2 + \kappa_3 \|P_1\| \|L_1 x(t)\| \quad \text{a.e.}$$

Combining (22) and (23), we readily deduce the following bound on the deviation $\Delta x(\cdot)$:

LEMMA 6.

$$\|\Delta x(t)\| \leq 2\|M^*\| \|P_1\| [\|P_1^{-1}\| \|P_1\|]^{1/2} [\kappa_1 \|P_1^{-1/2}\| \tilde{r}(L_1 x^0) + \kappa_3]$$

where

$$\tilde{r}(L_1 x^0) := \begin{cases} \sqrt{2} V_1^{1/2}(L_1 x^0), & L_1 x^0 \notin \Sigma, \\ r \|P_1\|^{1/2}, & L_1 x^0 \in \Sigma. \end{cases}$$

6.4. Summary of main results. For *matched* uncertainty, the following result is a direct consequence of Lemmas 2–4.

THEOREM 1. If $\mathcal{G}_r \equiv \{0\}$, then the generalized feedback \mathcal{F} , given by (11), renders the zero state of the differential inclusion system (3) globally uniformly asymptotically stable. Moreover, the dynamic behaviour of the feedback-controlled system ultimately corresponds to that of the ideal model in the sense that $x(t) = M^* w(t)$ for all $t \geq t_0 + \gamma^{-1} [\|P_2^{-1}\| \|P_2\|]^{1/2} \delta$ on every solution $x(\cdot)$ with $\|\bar{M}x(t_0)\| \leq \delta$, where $w(\cdot)$ is a solution of the linear ideal model $\dot{w} = L_1 A M^* w$ with prescribed spectrum $\sigma(L_1 A M^*) = \Lambda \subset \mathbb{C}^-$.

In the general case (i.e., in the presence of *unmatched* uncertainty), the following is a direct consequence of Lemmas 2, 3, and 5 (and the corollary).

THEOREM 2. The generalized feedback \mathcal{F} , given by (11), renders the compact set $S = \{x \in X : L_1 x \in \Sigma\} \cap W$, with Σ given by (19), globally uniformly asymptotically stable.

Remark. Note that, if the unmatched uncertainty parameter κ_3 is zero, then $S = \{0\}$.

7. Example. Consider the controlled nonlinear uncertain system

$$\ddot{z}(t) + z(t) - \mu \dot{z}(t)[1 - z^2(t)][1 - z^4(t)] - [1 - b]u(t) = 0, \quad z(t), u(t) \in \mathbb{R}$$

where $0 \leq \mu \leq \frac{1}{3}$ and $-\frac{1}{4} \leq b \leq \frac{1}{4}$ are unknown parameters (assumed here to be constant for notational convenience). For example, for $\mu = \frac{1}{3}$, the phase portrait of the *uncontrolled* system is depicted in Fig. 1 (unstable equilibrium, two limit cycles—one stable and one unstable).

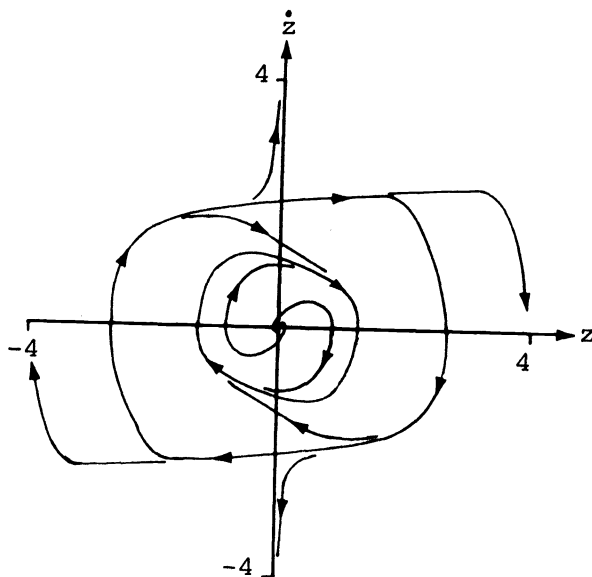


FIG. 1. Phase portrait of uncontrolled system ($\mu = \frac{1}{3}$).

When we write $x = [x_1 \ x_2]^T = [z \ \dot{z}]^T$, it is readily verified that (A1) holds with

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathcal{G}_r \equiv \{0\},$$

$$\mathcal{G}_m(t, x) = \{\mu x_2(1 - x_2^2)(1 - x_2^4) : 0 \leq \mu \leq \tfrac{1}{3}\} \subset \mathbb{R}, \quad \beta \equiv \tfrac{1}{4}.$$

When we adopt the ideal model spectrum $\Lambda = \{-1\}$, the subspace W is given by

$$W = \{x : x_1 + x_2 = 0\}.$$

When we select $C = -1$, the linear feedback operator F in (5) becomes $F = [0 \ -2]$. P_1 and P_2 are now determined as $P_1 = \tfrac{1}{2} = P_2$. The associated multifunction \mathcal{P}_2 is

$$x \mapsto \mathcal{P}_2(x) = \begin{cases} [0, \infty), & x_1 + x_2 > 0, \\ \mathbb{R}, & x_1 + x_2 = 0, \\ (-\infty, 0], & x_1 + x_2 < 0. \end{cases}$$

Hence, the function $(t, x) \mapsto \xi((\mathcal{G}_m(t, x) - L_1 M \mathcal{G}_r(t, x)) \cap \mathcal{P}_2(x))$, in (11b), reduces to the function

$$x \mapsto \begin{cases} f^+(x_2), & x_1 + x_2 > 0, \\ |f(x_2)|, & x_1 + x_2 = 0, \\ |f^-(x_2)|, & x_1 + x_2 < 0 \end{cases}$$

where f^+ and f^- denote the positive and negative parts of the function $f : x_2 \mapsto \tfrac{1}{3}x_2(1 - x_2^2)(1 - x_2^4)$.

For a design parameter value $\gamma = 0.1$, Fig. 2 depicts a typical family of trajectories of the feedback-controlled system, wherein the attractivity of the subspace W is clearly evident.

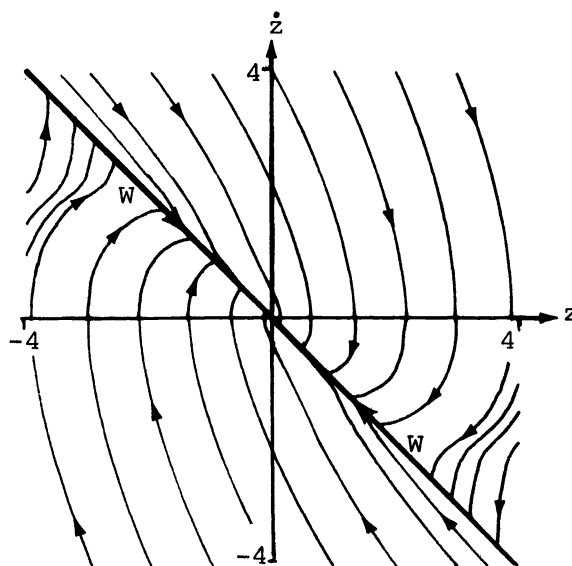


FIG. 2. Invariant subspace W and typical trajectories of the feedback controlled system.

REFERENCES

- [1] J. P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, Berlin, New York, 1984.
- [2] B. R. BARMISH, M. CORLESS, AND G. LEITMANN, *A new class of stabilizing controllers for uncertain dynamical systems*, SIAM J. Control Optim., 21 (1983), pp. 246–255.
- [3] B. R. BARMISH AND G. LEITMANN, *On ultimate boundedness control of uncertain systems in the absence of matching conditions*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 153–157.
- [4] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
- [5] M. CORLESS AND G. LEITMANN, *Continuous state feedback guaranteeing uniform ultimate boundedness for uncertain dynamic systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 1139–1141.
- [6] ———, *Adaptive control of systems containing uncertain functions and unknown functions with uncertain bounds*, J. Optim. Theory Appl., 41 (1983), pp. 155–168.
- [7] D. P. GOODALL AND E. P. RYAN, *Controlled differential inclusions and stabilization of uncertain dynamical systems*, in Proc. 25th IEEE Conference on Decision and Control, Athens, Greece, 1986, pp. 31–33.
- [8] S. GUTMAN, *Uncertain dynamical systems, a Lyapunov min-max approach*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 437–443.
- [9] S. GUTMAN AND Z. PALMOR, *Properties of min-max controllers in uncertain dynamical systems*, SIAM J. Control Optim., 20 (1983), pp. 850–861.
- [10] G. LEITMANN, *Guaranteed asymptotic stability for some linear systems with bounded uncertainties*, Trans. ASME Ser. G. J. Dynamic Systems Measurement Control, 101 (1979), pp. 212–216.
- [11] D. HINRICHSSEN AND A. J. PRITCHARD, *Stability radii of linear systems*, Systems Control Lett., 7 (1986), pp. 1–10.
- [12] ———, *Stability radius for structured perturbations and the algebraic Riccati equation*, Systems Control Lett., 8 (1986), pp. 105–113.
- [13] E. P. RYAN AND M. CORLESS, *Ultimate boundedness and asymptotic stability of a class of uncertain dynamical systems via continuous and discontinuous feedback control*, IMA J. Math. Control Inform., 1 (1984), pp. 223–242.
- [14] V. I. UTKIN, *Sliding Modes and their Application in Variable Structure Systems*, MIR, Moscow, 1978.

ON NONCONVERGENCE OF ADJOINT SEMIGROUPS FOR CONTROL SYSTEMS WITH DELAYS*

JOHN BURNS†, KAZUFUMI ITO‡, AND GEORG PROPST§

Abstract. It is shown that the adjoints of a spline-based approximation scheme for delay equations do not converge strongly.

Key words. optimal control, numerical analysis, semigroups

AMS(MOS) subject classifications. 93, 93C25, 93C55

1. Introduction. The primary goal of this paper is to illustrate, by a simple problem, the necessity of conducting a careful analysis of numerical schemes (developed for open-loop simulation) when these schemes are to be applied to an optimization-based control design problem. We feel that many distributed parameter control systems (viscoelastic structures, fluid flow control, etc.) are such that “standard finite element/finite difference” schemes might lead to numerical difficulties in certain control design methods, if care is not taken to ensure that these approximation schemes have convergence properties essential for the intended application.

We shall concentrate on a nonconvergence result for an optimal control problem governed by a delay differential equation. Although we feel that the technical details required to prove nonconvergence are interesting, we hope that this proof is not viewed as the major contribution of the paper. Indeed, we hope that the reader will be motivated to think about similar problems for more complex distributed parameter systems.

During the past 15 years, considerable attention has been devoted to the construction of finite-dimensional approximations of distributed parameter systems. Much of this work is based on algorithms first developed primarily for simulation. However, it is not clear that finite-dimensional models developed for (open-loop) simulations will also be suitable for certain optimization-based control design techniques. Moreover, there may be several reasons that a numerical scheme developed specifically for simulation does not perform well when applied to a control design problem.

In this paper, we concentrate on the use of the so-called spline scheme developed by Banks and Kappel [2] as an approximation method for calculating optimal feedback gains for control systems governed by functional differential equations. This problem is simple enough to be addressed in a short paper and yet it still illustrates how a specific numerical scheme, when combined with a particular control design approach, can lead to numerical difficulties.

* Received by the editors August 24, 1987; accepted for publication (in revised form) March 6, 1988. This research was partly supported by National Aeronautics and Space Administration contract NAS1-18107 while the authors were in residence at the Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, Virginia 23665.

† Interdisciplinary Center for Applied Mathematics, Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. This work was supported in part by Air Force Office of Scientific Research grant AFOSR-85-0287, SDIO under contract F49620-87-C-008 and by the Defense Advanced Research Projects Agency under contract F49620-87-C-0116.

‡ Center for Control Sciences, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This work was supported in part by Air Force Office of Scientific Research contract F-49620-86-C-011, by National Aeronautics and Space Administration grant NAG-1-517, and by National Science Foundation grant UINT-8521208.

§ Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, Virginia 23665.

To motivate the reader, and to describe the main technical contribution of the paper, we first review some known results.

Let H and U be real Hilbert spaces and $S(t): H \rightarrow H$ denote a C_0 -semigroup of bounded linear operators with generator A . We assume that $B: U \rightarrow H$, $W: H \rightarrow H$, and $R: U \rightarrow U$ are bounded linear operators with W and R being self-adjoint and nonnegative and R satisfying $R \geq mI > 0$. The (LQ) optimal control problem is to minimize

$$(1.1) \quad J(u) = \int_0^T [\langle Wz(s), z(s) \rangle + \langle Ru(s), u(s) \rangle] ds,$$

where $z(t)$ is defined by

$$(1.2) \quad z(t) = S(t)z_0 + \int_0^t S(t-s)Bu(s) ds$$

for $0 \leq t < T$ and given $z_0 \in H$. If $T = +\infty$, then we have the linear quadratic regulator problem.

Now assume that there exists a sequence of C_0 -semigroups $S^N(t)$ on H and positive constants M, β such that

$$(1.3) \quad \|S^N(t)\| \leq M e^{\beta t}, \quad t \geq 0, \quad N \geq 1,$$

$$(1.4) \quad S^N(t) \rightarrow S(t) \quad \text{strongly as } N \rightarrow +\infty$$

and the convergence in (1.4) is uniform in t on bounded intervals. Denote by A^N the generator of $S^N(t)$, and also assume the existence of bounded linear operators $B^N: U \rightarrow H$, $W^N: H \rightarrow H$ with W^N self-adjoint and $W^N \geq 0$ satisfying

$$(1.5) \quad \|B^N - B\| \rightarrow 0, \quad W^N \rightarrow W \quad \text{strongly as } N \rightarrow +\infty.$$

Note that the assumption of uniform convergence (1.5) is stronger than required by Gibson (see [4, p. 113]). It is well known that the optimal control (if it exists) is given by state feedback

$$(1.6) \quad u^*(t) = -R^{-1}B^*\Pi(t)z^*(t),$$

where the bounded self-adjoint operator $\Pi(\cdot)$ satisfies the Riccati (integral) equation

$$(1.7) \quad \Pi(t)z = \int_t^T S^*(s-t)[W - \Pi(s)BR^{-1}B^*\Pi(s)]S(s-t) ds.$$

Let $\Pi^N(t)$ be the solution to the "approximating" Riccati equation

$$(1.8) \quad \Pi^N(t)z = \int_t^T S^{N*}(s-t)[W^N - \Pi^N(s)B^NR^{-1}B^N\Pi^N(s)]S^N(s-t) ds$$

and observe that (1.8) would determine the suboptimal gains if we used the approximating system (A^N, B^N) with weights W^N .

The following theorem is a direct consequence of Gibson's more general results [4, Thms. 6.1, 6.2].

THEOREM 1.1. *If conditions (1.1)–(1.5) hold, then for $0 \leq t \leq T$, $\Pi^N(t)$ converges weakly to $\Pi(t)$ and the convergence is uniform on $[0, T]$. If, in addition,*

$$(1.9) \quad S^{N*}(t) \rightarrow S^*(t) \quad \text{strongly},$$

then $\Pi^N(t)$ converges strongly to $\Pi(t)$ and the convergence is uniform for $t \in [0, T]$.

COROLLARY 1.2. *Let $K(t) = R^{-1}B^*\Pi(t)$ and $K^N(t) = R^{-1}B^{N*}\Pi^N(t)$ denote the feedback gain operators. Assume that $U = \mathbb{R}^m$ (i.e., is finite-dimensional). If $\Pi^N(t)$ converges strongly to $\Pi(t)$, then as $N \rightarrow +\infty$,*

$$(1.10) \quad \|K^N(t) - K(t)\| \rightarrow 0.$$

We emphasize the point that if the control space is finite-dimensional, then uniform convergence is assured, provided the numerical scheme is stable and consistent *and* (1.9) holds. If we are concerned only with simulation, then stability and consistency are sufficient for most numerical problems. Moreover, it can be shown that many standard numerical schemes developed for simulation of self-adjoint partial differential equations do satisfy (1.9). Therefore, it is not surprising that until Gibson “needed” (1.9) to establish the uniform convergence of optimal gain operators, the question of whether a numerical scheme satisfied (1.9) received little attention. Indeed, even after Gibson published his result, it was still not obvious that condition (1.9) was anything more than a technical assumption needed in Gibson’s proof.

In [3], Banks, Rosen, and Ito apply a convergent spline-based scheme to an optimal control problem governed by a delay differential equation. The numerical results in [3] seemed to show that $K^N(t)$ did not converge uniformly to $K(t)$, and these numerical results have often been used as evidence that (1.9) did not hold for this particular scheme. Moreover, several new schemes have since been generated specifically to ensure that (1.3), (1.4), and (1.9) are valid. Still, it was not known whether condition (1.9) held for the spline scheme used in [3]. We shall provide a proof that (1.9) fails for this scheme. We also show that this spline scheme is stable and consistent to A^* on a dense subset of $D(A^*)$.

2. Spline approximations of hereditary systems. Consider the delay differential equation

$$(2.1) \quad \dot{x}(t) = A_0 x(t) + A_1 x(t-r) + \int_{-r}^0 A(s)x(t+s) ds$$

with initial data

$$(2.2) \quad x(0) = \eta, \quad x(s) = \phi(s), \quad -r \leq s < 0,$$

where $x(t) \in \mathbb{R}^n$ and the elements of $A(\cdot)$ are square-integrable on $[-r, 0]$. It is well known (e.g., [1]) that for $\eta \in \mathbb{R}^n$ and $\phi \in L^2(-r, 0; \mathbb{R}^n)$ there exists a unique solution of (2.1), (2.2) $x: [-r, +\infty) \rightarrow \mathbb{R}^n$ such that $x \in W^{1,2}(0, T; \mathbb{R}^n)$ for all $T > 0$. If we define the solution map $S(t)$, $t \geq 0$ on the product space $Z = \mathbb{R}^n \times L^2(-r, 0; \mathbb{R}^n)$ by

$$(2.3) \quad S(t)(\eta, \phi(\cdot)) = (x(t), x(t+\cdot)),$$

where x is the solution to (2.1), (2.2), then $\{S(t)\}_{t \geq 0}$ is a strongly continuous semigroup (i.e., a C_0 -semigroup) on Z . The infinitesimal generator A is the operator defined on the domain

$$(2.4) \quad D(A) = \{(\eta, \phi(\cdot)) \in Z \mid \phi(\cdot) \in W^{1,2}(-r, 0; \mathbb{R}^n), \phi(0) = \eta\}$$

by

$$(2.5) \quad A(\eta, \phi(\cdot)) = \left(A_0 \eta + A_1 \phi(-r) + \int_{-r}^0 A(s)\phi(s) ds, \dot{\phi}(\cdot) \right).$$

The adjoint operator A^* generates the adjoint semigroup $S^*(t)$ and it is easy to show that (see [4], [7])

$$(2.6) \quad D(A^*) = \{(\xi, \psi) \in Z \mid \psi \in W^{1,2}(-r, 0; \mathbb{R}^n), \psi(-r) = A_1^T \xi\}$$

and, for $(\xi, \psi) \in D(A^*)$,

$$(2.7) \quad A^*(\xi, \psi) = (\psi(0) + A_0^T \xi, [A^T(\cdot)\xi - \dot{\psi}(\cdot)]).$$

As in [2], we define the linear spline-based approximation for $S(t)$. Let $\{B_i^N\}_{i=0}^N$ denote the usual linear B-splines defined on the interval $[-r, 0]$ by

$$(2.8) \quad B_i^N(s) = \begin{cases} \frac{N}{r}(s - \tau_{i+1}^N), & s \in [\tau_{i+1}^N, \tau_i^N], \\ \frac{N}{r}(\tau_{i-1}^N - s), & s \in [\tau_i^N, \tau_{i-1}^N], \\ 0 & \text{otherwise,} \end{cases}$$

where $\tau_i^N = -ir/N$, $i = 0, 1, \dots, N$, $\tau_{N+1}^N = -r$, and $\tau_{-1}^N = 0$. For each $N = 1, 2, \dots$, let Z^N denote the linear subspace of Z defined by

$$(2.9) \quad Z^N = \left\{ z \in Z \mid z = \sum_{k=0}^N a_k (B_k^N(0), B_k^N(\cdot)), a_k \in \mathbb{R}^n \right\}$$

and let P^N denote the orthogonal projection of Z onto Z^N . This subspace can be identified with $\mathbb{R}^{n(N+1)}$ by the prolongation $i^N: \mathbb{R}^{n(N+1)} \rightarrow Z$ defined by

$$(2.10) \quad i^N a = \left(a_0, \sum_{k=0}^N a_k B_k^N(\cdot) \right),$$

where $a = (a_0^T, a_1^T, \dots, a_N^T)^T \in \mathbb{R}^{n(N+1)}$. The space $\mathbb{R}^{n(N+1)}$ is normed with the induced inner product

$$(2.11) \quad \langle a, b \rangle_N = a^T Q^N b,$$

where $a, b \in \mathbb{R}^{n(N+1)}$ and Q^N is defined by

$$(2.12) \quad Q^N = \begin{bmatrix} I & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & & & 0 \end{bmatrix} + \frac{r}{N} \begin{bmatrix} \frac{1}{3}I & \frac{1}{6}I & \cdots & 0 \\ \frac{1}{6}I & \frac{2}{3}I & \ddots & \\ \vdots & \ddots & \ddots & \frac{2}{3}I & \frac{1}{6}I \\ 0 & & & \frac{1}{6}I & \frac{1}{3}I \end{bmatrix}.$$

The adjoint operator $[i^N]^*: Z \rightarrow \mathbb{R}^{n(N+1)}$ is given by

$$(2.13) \quad [i^N]^*(\eta, \phi(\cdot)) = [Q^N]^{-1} \begin{bmatrix} \eta + \phi_0^N \\ \phi_1^N \\ \vdots \\ \phi_N^N \end{bmatrix},$$

where $\phi_i^N = \int_{-r}^0 \phi(s) B_i^N(s) ds$. Moreover, it is easy to show that

$$(2.14) \quad \begin{aligned} [i^N]^* i^N &= I \quad \text{the identity on } \mathbb{R}^{n(N+1)}, \\ i^N [i^N]^* &= P^N, \end{aligned}$$

and, for $z, w \in Z^N$,

$$(2.15) \quad \langle z, w \rangle_Z = \langle [i^N]^* z, [i^N]^* w \rangle_N.$$

In order to construct the standard Galerkin approximation of A , we note that $Z^N \subseteq D(A)$ and define A^N by $A^N = P^N A P^N$. Observe that A^N (and hence $[A^N]^*$) is continuous and although $P^N Z \subseteq D(A)$, P^N does not map all of Z into $D(A^*)$. It is shown in [2] that

$$(2.16) \quad A^N = i^N [Q^N]^{-1} H^N [i^N]^*,$$

where

$$(2.17) \quad H^N = \begin{bmatrix} A_0^N & A_1^N & \cdots & A_N^N \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} I & -I & & 0 \\ I & 0 & \ddots & \\ & \ddots & 0 & -I \\ 0 & & I & -I \end{bmatrix}$$

and

$$\begin{aligned} A_0^N &= A_0 + \int_{-r}^0 A(s) B_0^N(s) \, ds, \\ A_k^N &= \int_{-r}^0 A(s) B_k^N(s) \, ds, \quad 1 \leq k \leq N-1, \\ A_N^N &= A_1 + \int_{-r}^0 A(s) B_N^N(s) \, ds. \end{aligned}$$

If $S^N(t)$ denotes the C_0 -semigroup generated by A^N , then it is shown in [2] that for each $z \in Z$,

$$(2.18) \quad \|S^N(t)z - S(t)z\| \rightarrow 0,$$

where the convergence is uniform in t on bounded intervals. The convergence (2.18) was established by proving that the Galerkin approximations A^N satisfy the Trotter-Kato Theorem [6], and hence provide a stable and consistent approximation scheme for A . We shall prove that the convergence statement above does not hold for the sequence of adjoint semigroups $[S^N]^*(t)$. Moreover, this convergence fails even though the adjoint operators $[A^N]^*$ are stable and consistent to A^* on a dense subset of Z (i.e., there is a dense set $C \subseteq D(A^*) \subseteq Z$ such that $[A^N]^*z \rightarrow A^*z$ for all $z \in C$).

3. Convergence of the adjoint generators. We shall follow the approach given in [5] and consider the $n(N+1) \times n$ matrix

$$\gamma_+^N = [Q^N]^{-1}(0 \cdots 0, I)^T.$$

Define the operators $\delta_+^N: \mathbb{R}^n \rightarrow Z^N$ by

$$\delta_+^N x = (\phi_x(0), \phi_x(\cdot)),$$

where

$$\phi_x(\cdot) = [B_0^N(\cdot)I, B_1^N(\cdot)I, \dots, B_N^N(\cdot)I]\gamma_+^N x.$$

It follows that for all $(\eta, \phi) = z \in Z$ and $x \in \mathbb{R}^n$

$$\begin{aligned} [[i^N]^* \delta_+^N x]^T Q^N [i^N]^* P^N z &= [\gamma_+^N x]^T Q^N [i^N]^* z \\ &= x^T [B_0^N(-r)I, \dots, B_N^N(-r)I][i^N]^* z, \end{aligned}$$

and (2.14), (2.15) imply

$$(3.1) \quad \langle \delta_+^N x, z \rangle_Z = x^T [\phi^N(-r)],$$

where $P^N z = P^N(\eta, \phi) = (\phi^N(0), \phi^N(\cdot)) \in Z^N$. Furthermore, if λ_{\min}^N denotes the smallest eigenvalue of Q^N , then $(r/6N) \leq \lambda_{\min}^N$ and it follows from (2.11) that

$$(3.2) \quad \|\delta_+^N\| \leq (6N/r)^{1/2} \quad \text{for all } N.$$

We also need the following representation.

LEMMA 3.1. *The operators $[A^N]^*: Z \rightarrow Z^N$ are given by*

$$[A^N]^*(\xi, \psi) = P^N(\psi^N(0) + A_0^T \psi^N(0), [A^T(\cdot)\psi^N(0) - \dot{\psi}^N(\cdot)]) \\ + \delta_+^N(A_1^T \psi^N(0) - \psi^N(-r)),$$

where $P^N(\xi, \psi) = (\psi^N(0), \psi^N(\cdot))$.

Proof. Assume that $z = (\eta, \phi)$ and $w = (\xi, \psi)$ belong to Z and let $P^N z = (\phi^N(0), \phi^N(\cdot))$ and $P^N w = (\psi^N(0), \psi^N(\cdot))$ denote the orthogonal projections of z and w , respectively. The identity (3.1) implies that

$$\langle (\psi^N(0) + A_0^T \psi^N(0), [A^T(\cdot)\psi^N(0) - \dot{\psi}^N(\cdot)]), P^N z \rangle_Z + \langle \delta_+^N(A_1 \psi^N(0) - \psi^N(-r)), z \rangle_Z \\ = [\psi^N(0)]^T [\phi^N(0)] + [\psi^N(0)]^T A_0 [\phi^N(0)] - \int_{-r}^0 \langle \dot{\psi}^N(s), \phi^N(s) \rangle ds \\ + [\psi^N(0)]^T \int_{-r}^0 A(s) \psi^N(s) ds + [\psi^N(0)]^T A_1 [\phi^N(-r)] - [\psi^N(-r)]^T [\phi^N(-r)].$$

When we integrate by parts, the boundary terms cancel with the first and last term in the sum. Therefore, it follows that if $[A^N]^*$ is defined as above, then

$$\langle [A^N]^* w, z \rangle = \langle P^N w, A P^N z \rangle = \langle w, A^N z \rangle$$

and this completes the proof. \square

It should be noted that $[A^N]^* w = P^N A^* P^N w$ if and only if $A_1^T \psi^N(0) - \psi^N(-r) = 0$, i.e., if and only if $P^N w \in D(A^*)$. Also, if

$$D = D(A) \cap D(A^*),$$

$$C = \{(\xi, \psi) \in D \mid \psi \in C^2(-r, 0; \mathbb{R}^n)\},$$

then D and C are dense in Z . Moreover, we have the following convergence result.

LEMMA 3.2. *If C and D are defined as above, then*

(a) $[A^N]^* w \rightarrow A^* w$ for all $w \in C$;

(b) For all $\lambda \in \mathbb{R}$, $(\lambda I - A^*)D$ is not dense in Z .

Proof. Let $w \in C \subseteq D(A) \cap D(A^*)$. Note that $w = (\psi(0), \psi(\cdot))$, where $\psi(-r) = A_1^T \psi(0)$ and $P^N w = (\psi^N(0), \psi^N(\cdot)) \in D(A)$. It follows from (2.7) and Lemma 3.1 that

$$\|[A^N]^* w - A^* w\| \leq \|P^N A^* w - A^* w\| \\ + \|(\psi^N(0) + A_0^T \psi^N(0), [A^T(\cdot)\psi^N(0) - \dot{\psi}^N(\cdot)]) \\ - (\psi(0) + A_0^T \psi(0), [A_0^T(\cdot)\psi(0) - \dot{\psi}(\cdot)])\| \\ + \|\delta_+^N[A_1^T \psi^N(0) - \psi^N(-r)]\| \\ = F_1^N + F_2^N + F_3^N.$$

Since $w = (\psi(0), \psi(\cdot)) \in C \subseteq D(A) \cap D(A^*)$, and $\psi(\cdot) \in C^2(-r, 0; \mathbb{R}^n)$, it follows from standard estimates on interpolating splines that (see [2, Eqs. (4.1)–(4.3)])

$$|\psi^N(0) - \psi(0)| \leq O(1/N^2),$$

$$|\psi^N(s) - \psi(s)| \leq O(1/N), \quad -r \leq s \leq 0,$$

$$\|\dot{\psi}^N(\cdot) - \dot{\psi}(\cdot)\| \leq O(1/N).$$

The first term $F_1^N \rightarrow 0$, since $w \in D(A^*)$ and $P^N z \rightarrow z$ for all $z \in Z$. The second term is estimated by

$$F_2^N \leq |\psi^N(0) + A_0^T \psi^N(0) - \psi(0) - A_0^T \psi(0)| + \|A^T(\cdot)(\psi^N(0) - \psi(0))\| + \|\dot{\psi}^N(\cdot) - \dot{\psi}(\cdot)\| \\ \leq (1 + |A_0| + \|A(\cdot)\|)|\psi^N(0) - \psi(0)| + O(1/N).$$

Therefore, $F_2^N \rightarrow 0$ as $N \rightarrow \infty$. Applying (3.2), we estimate the last term by

$$\begin{aligned} F_3^N &\leq (6N/r)^{1/2} |A_1^T \psi(0) - \psi^N(-r)| \\ &\leq (6N/r)^{1/2} (|A_1| |\psi^N(0) - \psi(0)| + |\psi(-r) - \psi^N(-r)|), \end{aligned}$$

and hence $F_3^N \rightarrow 0$, which establishes part (a) of the lemma.

Turning to part (b), we let $w = (\psi(0), \psi) \in D$, $\xi, x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$. A straightforward calculation yields

$$\begin{aligned} &\langle (\xi, e^{\lambda \cdot} x), [\lambda I - A^*](\psi(0), \psi(\cdot)) \rangle \\ &= \psi^T(0)[(\lambda - 1)I - A_0]\xi + \psi^T(0) \left[I - e^{-\lambda r} A_1 - \int_{-r}^0 e^{\lambda s} A^T(s) ds \right] x. \end{aligned}$$

If $(\lambda - 1) \in \sigma(A_0)$ and $e^{\lambda r} \in \sigma(A_1 + \int_{-r}^0 e^{(\lambda+r)s} A^T(s) ds)$, then there exist $\xi \neq 0$ and $x \neq 0$ such that

$$\langle (\xi, e^{\lambda \cdot} x), [\lambda I - A^*](\psi(0), \psi(\cdot)) \rangle = 0.$$

If $(\lambda - 1) \notin \sigma(A_0)$, let $x \neq 0$ and define ξ by

$$\xi = -[(\lambda - 1)I - A_0]^{-1} \left(I - e^{-\lambda r} A_1 - \int_{-r}^0 e^{\lambda s} A^T(s) ds \right) x,$$

or if $e^{\lambda r} \notin \sigma(A_1 + \int_{-r}^0 e^{(\lambda+r)s} A^T(s) ds)$, let $\xi \neq 0$ and define x by

$$x = - \left[I - e^{-\lambda r} A_1 - \int_{-r}^0 e^{\lambda s} A^T(s) ds \right]^{-1} [(\lambda - 1)I - A_0] \xi.$$

In any case, there always exists an element $z = (\xi, \psi(\cdot)) = (\xi, e^{\lambda \cdot} x) \neq 0$, such that

$$\langle z, (\lambda I - A^*)w \rangle = 0$$

for all $w = (\psi(0), \psi(\cdot)) \in D$. Consequently, $(\lambda I - A^*)D$ is not dense in Z . \square

Since $\|[S^N]^*(t)\| = \|S^N(t)\| \leq e^{\omega t}$ for some ω independent of N (see [2]), the existence of a set $E \subseteq D(A^*)$ and a $\lambda \in \mathbb{R}$ such that $(\lambda I - A^*)E$ is dense in Z and $[A^N]^*w \rightarrow A^*w$ for $w \in E$ would imply strong convergence of the semigroups $[S^N]^*(t)$ to $S^*(t)$ ([6, Thm. 4.5, § III]). Although $[A^N]^*w \rightarrow A^*w$ on the dense set C , $(\lambda I - A^*)C$ is not dense in Z . We shall establish that there does not exist a set $E \subseteq D(A^*)$ with the specified properties. In fact, we shall show that $[S^N]^*(t)$ does not converge strongly to $S^*(t)$.

4. Calculation of $P^N([A^N]^*)^{-1}P^N$. In this section, we present several technical lemmas that will be needed in § 5. The proof of Lemma 4.1 is straightforward and hence is omitted. Lemma 4.3 can be found in [4] and [7].

LEMMA 4.1. *The operator $[A^N]^*$ can be decomposed as*

$$(4.1) \quad [A^N]^* = i^N [Q^N]^{-1} [H^N]^T [i^N]^*.$$

LEMMA 4.2. *Assume H^N is invertible. If $z = (\eta, \phi(\cdot)) \in Z$, then $P^N([A^N]^*)^{-1}P^N z$ exists and*

$$(4.2) \quad P^N([A^N]^*)^{-1}P^N z = i^N a^N,$$

where a^N is the solution of

$$(4.3) \quad [H^N]^T a^N = \begin{bmatrix} \eta + \phi_0^N \\ \phi_1^N \\ \vdots \\ \phi_N^N \end{bmatrix}.$$

Proof. Consider the equation $A_w^{N*} = P^N z$ for $w \in Z^N$. By Lemma 4.1 and from (2.14), we have

$$i^N (Q^N)^{-1} H^{NT} i^{N*} w = i^N i^{N*} z.$$

Multiplying with i^{N*} and from (2.13) and (2.14), we have

$$H^{NT} i^{N*} w = \begin{bmatrix} \eta + \phi_0^N \\ \phi_1^N \\ \vdots \\ \phi_N^N \end{bmatrix}.$$

The lemma now follows from the fact that $w = P^N w = i^N i^{N*} w = i^N a^N$. \square

LEMMA 4.3. If $\Delta = A_0 + A_1 + \int_{-r}^0 A(s) ds$, then $0 \in \rho(A^*)$ if and only if Δ is invertible and $(\xi, \psi(\cdot)) = (A^*)^{-1}(\eta, \phi(\cdot)) \in D(A^*)$ is given by

$$(4.4) \quad \xi = [\Delta^T]^{-1} \left(\eta + \int_{-r}^0 \phi(s) ds \right),$$

$$(4.5) \quad \psi(s) = \left(A_1^T + \int_{-r}^s A^T(\theta) d\theta \right) \xi - \int_{-r}^s \phi(\theta) d\theta.$$

Consider (4.3). It follows from (2.17) that (4.3) is equivalent to the system

$$(4.6) \quad \frac{1}{2}(a_{k+1}^N - a_{k-1}^N) + [A_k^N]^T a_0^N = \phi_k^N, \quad 1 \leq k \leq N-1,$$

$$(4.7) \quad -\frac{1}{2}(a_N^N + a_{N-1}^N) + [A_N^N]^T a_0^N = \phi_N^N,$$

$$(4.8) \quad \frac{1}{2}(a_0^N + a_1^N) + [A_0^N]^T a_0^N = \eta + \phi_0^N.$$

A straightforward induction argument yields

$$(4.9) \quad a_{2k}^N = a_0^N - 2 \sum_{i=1}^k ([A_{2i-1}^N]^T a_0^N - \phi_{2i-1}^N) \quad \text{if } 2k \leq N,$$

$$(4.10) \quad a_{2k+1}^N = a_1^N - 2 \sum_{i=1}^k ([A_{2i}^N]^T a_0^N - \phi_{2i}^N) \quad \text{if } 2k+1 \leq N.$$

Thus it follows from (4.7) that

$$(4.11) \quad -\frac{1}{2}(a_0^N - a_1^N) + \sum_{k=1}^N [A_k^N]^T a_0^N = \sum_{k=1}^N \phi_k^N.$$

Moreover, (4.8) and (4.11) imply that

$$\sum_{k=0}^N [A_k^N]^T a_0^N = \eta + \sum_{j=0}^N \phi_j^N = \eta + \int_{-r}^0 \phi(s) ds,$$

where

$$\sum_{k=0}^N [A_k^N]^T = \left[A_0 + A_1 + \int_{-r}^0 A(s) ds \right]^T = \Delta^T.$$

If we assume that $0 \in \rho(A^*)$, then by Lemma 4.3 it follows that Δ^T is invertible and

$$(4.12) \quad a_0^N = \Delta^{-T} \left(\eta + \int_{-r}^0 \phi(s) ds \right).$$

Observe that (4.12) implies that a_0^N is independent of N . By (4.4) it follows that $a_0^N = \xi$, where $(\xi, \psi(\cdot)) = [A^*]^{-1}(\eta, \phi(\cdot))$. Equation (4.8) yields the identity

$$(4.13) \quad a_1^N = -a_0^N + 2 \sum_{k=1}^N ([A_k^N]^T a_0^N - \phi_k^N),$$

and hence it now follows from (4.10) that

$$(4.14) \quad \begin{aligned} a_{2k+1}^N &= -a_0^N + 2 \sum_{k=1}^N ([A_k^N]^T a_0^N - \phi_k^N) - 2 \sum_{i=1}^k ([A_{2i}^N]^T a_0^N - \phi_{2i}^N) \\ &= -a_0^N + 2 \sum_{j=2k+1}^N ([A_j^N]^T a_0^N - \phi_j^N) + 2 \sum_{i=1}^k ([A_{2i-1}^N]^T a_0^N - \phi_{2i-1}^N). \end{aligned}$$

Note that for $1 \leq k \leq N-1$

$$|\phi_k^N|^2 = \left| \int_{-r}^0 \phi(s) B_k^N(s) ds \right|^2 \leq \left(\frac{2r}{3N} \right) \int_{\tau_{k+1}^N}^{\tau_{k-1}^N} |\phi(s)|^2 ds$$

and similarly,

$$|\phi_0^N|^2 \leq \left(\frac{r}{3N} \right) \int_{\tau_1^N}^0 |\phi(s)|^2 ds, \quad |\phi_N^N|^2 \leq \left(\frac{r}{3N} \right) \int_{-r}^{\tau_{N-1}^N} |\phi(s)|^2 ds.$$

Therefore, (4.9) and (4.10) imply that

$$\begin{aligned} |a_{2k}^N| &\leq |a_0^N| (1 + 2|A_1|) + 2\sqrt{r/3} (\|A\|_{L_2} |a_0^N| + \|\phi\|_{L_2}), \\ |a_{2k+1}^N| &\leq |a_1^N| + 2|A_1| |a_0^N| + 2\sqrt{r/3} (\|A\|_{L_2} |a_0^N| + \|\phi\|_{L_2}), \end{aligned}$$

respectively. The identity (4.12) yields the estimate

$$|a_0^N| \leq |\Delta^{-T}| (1 + \sqrt{r}) \|(\eta, \phi(\cdot))\|_Z,$$

and (4.13) leads to the bound

$$\begin{aligned} |a_1^N| &= |-(1 + 2[A_0^N]^T) a_0^N + 2(\eta + \phi_0^N)| \\ &\leq |a_0^N| + 2|\eta| + 2\sqrt{r/3N} (\|A\|_{L_2} |\phi_0^N| + \|\phi\|_{L_2}). \end{aligned}$$

Combining these estimates, we have for $0 \leq k \leq N$

$$|a_k^N| \leq M \|(\eta, \phi(\cdot))\|_Z,$$

where $M \geq 0$ is independent of N . An application of Lemma 4.2 yields the estimate

$$\|P^N [A^{N*}]^{-1} P^N z\|_Z = \|i^N a^N\| \leq (\sqrt{1+r}) M \|z\|_Z$$

for all $z \in Z$. We summarize these results in the following theorem.

THEOREM 4.4. *If $0 \in \rho(A)$, then*

$$P^N [A^{N*}]^{-1} P^N (\eta, \phi(\cdot)) = i^N a^N,$$

where $a^N = \text{col}([a_0^N]^T, [a_1^N]^T, \dots, [a_N^N]^T)$ is given by

$$a_0^N = \Delta^{-T} \left(\eta + \int_{-r}^0 \phi(s) ds \right)$$

and for $0 \leq 2k \leq N$

$$a_{2k}^N = a_0^N - 2 \sum_{i=1}^k ([A_{2i-1}^N]^T a_0^N - \phi_{2i-1}^N),$$

while for $1 \leq 2k+1 \leq N$

$$a_{2k+1}^N = -a_0^N + 2 \sum_{j=2k+1}^N ([A_j^N]^T a_0^N - \phi_j^N) + 2 \sum_{i=1}^k ([A_{2i-1}^N]^T a_0^N - \phi_{2i-1}^N).$$

Moreover, $\|P^N[A^{N*}]^{-1}P^N\|_{\mathcal{L}(Z)}$ is uniformly bounded in N .

Let P_{AVE}^N denote the “averaging” orthogonal projection on Z defined by

$$P_{\text{AVE}}^N(\eta, \phi(\cdot)) = \left(\eta, \sum_{i=1}^N \frac{N}{r} \left(\int_{\tau_i^N}^{\tau_{i-1}^N} \phi(s) ds \right) \chi_{\tau_i^N, \tau_{i-1}^N} \right),$$

where χ_I denotes the characteristic function for the interval I (see [1] for details).

COROLLARY 4.5. *There exists a constant $K \geq 0$ (independent of N) such that for all $z \in Z$*

$$\|P_{\text{AVE}}^N(P^N[A^{N*}]^{-1}P^N z) - [A^*]^{-1}z\| \leq \left(\frac{K}{N}\right) \|z\|.$$

Proof. A direct application of Theorem 4.4 yields the identities

$$\begin{aligned} \frac{1}{2}(a_{2k+1}^N + a_{2k}^N) &= \sum_{i=2k+1}^N ([A_i^N]^T \xi - \phi_i^N), \quad 0 \leq 2k \leq N-1, \\ \frac{1}{2}(a_{2k-1}^N + a_{2k}^N) &= \sum_{i=2k}^N ([A_i^N]^T \xi - \phi_i^N), \quad 2 \leq 2k \leq N. \end{aligned}$$

On the other hand, since $(\xi, \psi(\cdot)) = [A^*]^{-1}(\eta, \phi(\cdot))$, Lemma 4.3 implies that

$$\begin{aligned} \xi &= a_0^N = \Delta^{-T} \left(\eta + \int_{-r}^0 \phi(s) ds \right), \\ \psi(s) &= \left(A^T + \int_{-r}^s A^T(\theta) d\theta \right) \xi - \int_{-r}^s \phi(\theta) d\theta. \end{aligned}$$

Therefore, if $s_j^N = (\tau_j^N + \tau_{j-1}^N)/2$, then

$$\begin{aligned} &|(a_{2k+1}^N + a_{2k}^N)/2 - \psi(s_{2k+1}^N)| \\ &= \left| \int_{\tau_{2k+1}^N}^{\tau_{2k+1}^N} (B_{2k+1}^N(s) - 1)(A^T(s)\xi - \phi(s)) ds \right. \\ &\quad \left. + \int_{s_{2k+1}^N}^{\tau_{2k}^N} (B_{2k+1}^N(s)(A^T(s)\xi - \phi(s)) ds \right| \\ &\leq (r/\sqrt{12}N) \left\{ \left(\int_{\tau_{2k+1}^N}^{\tau_{2k}^N} |A(s)|^2 ds \right)^{1/2} |\xi| + \left(\int_{\tau_{2k+1}^N}^{\tau_{2k}^N} |\phi(s)|^2 ds \right)^{1/2} \right\}, \end{aligned}$$

and similarly,

$$\left| \frac{a_{2k}^N + a_{2k-1}^N}{2} - \psi(s_{2k}^N) \right| \leq (r/\sqrt{12}N) \left\{ \left(\int_{\tau_{2k}^N}^{\tau_{2k-1}^N} |A(s)|^2 ds \right)^{1/2} |\xi| + \left(\int_{\tau_{2k}^N}^{\tau_{2k-1}^N} |\phi(s)|^2 ds \right)^{1/2} \right\}.$$

It follows that

$$\left\| P_{\text{AVE}}^N i^N a - \left(\xi, \sum_{j=1}^N \psi(s_j^N) \chi_{[\tau_j^N, \tau_{j-1}^N)} \right) \right\| \leq (r/\sqrt{12}N) (\|A\|_{L_2} |\xi| + \|\phi\|_{L_2}).$$

Since $\psi \in W^{1,2}(-r, 0; \mathbb{R}^n)$ and

$$\dot{\psi}(s) = A^T(s)\xi - \phi(s) = A^T(s)\Delta^{-T} \left(\eta + \int_{-r}^0 \phi(\theta) d\theta \right) - \phi(s),$$

there exists a constant \hat{K} such that

$$\left\| \psi - \sum_{j=1}^N \psi(s_j^N) \chi_{[\tau_j^N, \tau_{j-1}^N)} \right\|_{L_2} \leq \left(\frac{\hat{K}}{N} \right) \|(\eta, \phi(\cdot))\|_Z.$$

This estimate, combined with the previous inequality, establishes the proof. \square

Now we provide proof that the approximating adjoint semigroups constructed above do not converge strongly to the adjoint semigroup generated by A^* .

5. Nonstrong convergence. Let A, A^*, A^N , and A^{N*} be defined by (2.4), (2.5), and (2.6), (2.7), (2.16), and (4.1), respectively. The corresponding semigroups will be denoted by $S(t), S^*(t), S^N(t)$, and $S^{N*}(t)$. Recall that for $z \in Z$ (see [2])

$$(5.1) \quad \langle A^N z, z \rangle = \langle AP^N z, P^N z \rangle \leq \omega \|P^N z\|^2 \leq \omega \|z\|^2,$$

where $\omega = (1 + 2|A_0| + |A_1|^2 + 2\|A\|_{L_2})/2$ and for $t \geq 0$

$$(5.2) \quad \|S^N(t)\| \leq e^{\omega t}, \quad \|S^{N*}(t)\| \leq e^{\omega t}.$$

The following result is a special case of Theorem 4.2 of [6, Chap. 3].

THEOREM 5.1. *The following are equivalent:*

(a) *For every $z \in Z$ and $\lambda \in \mathcal{C}$ with $\operatorname{Re} \lambda > \omega$*

$$(\lambda I - A^{N*})^{-1} z \rightarrow (\lambda I - A^*)^{-1} z \quad \text{as } N \rightarrow \infty.$$

(b) *For every $z \in Z$ and $t \geq 0$*

$$S^{N*}(t)z \rightarrow S^*(t)z \quad \text{as } N \rightarrow +\infty,$$

the convergence being uniform in t on bounded intervals.

We shall also need the following technical lemmas.

LEMMA 5.2. *Suppose condition (b) of Theorem 5.1 holds, $\lambda \in \rho(A^*)$, and $\|P^N(\lambda I - A^{N*})^{-1}P^N\|$ is uniformly bounded in N . Then for every $z \in Z$*

$$P^N(\lambda I - A^{N*})^{-1}P^N z \rightarrow (\lambda I - A^*)^{-1}z.$$

Proof. From Theorem 5.1, for $\lambda_0 > \omega$ and $z \in Z$

$$P^N(\lambda_0 I - A^{N*})^{-1}P^N z \rightarrow (\lambda_0 I - A^*)^{-1}z \quad \text{as } N \rightarrow \infty.$$

Note that for $z \in Z$

$$\begin{aligned} & P^N(\lambda I - A^{N*})^{-1}P^N z - P^N(\lambda_0 I - A^{N*})^{-1}P^N z \\ &= (\lambda_0 - \lambda)P^N(\lambda I - A^{N*})^{-1}P^N(\lambda_0 I - A^{N*})^{-1}P^N z, \end{aligned}$$

and similarly,

$$(\lambda I - A^*)^{-1}(I + (\lambda - \lambda_0)(\lambda_0 I - A^*)^{-1}) = (\lambda_0 I - A^*)^{-1}.$$

Hence, if $w = z + (\lambda - \lambda_0)(\lambda_0 I - A^*)^{-1}z$, then

$$\begin{aligned} & P^N(\lambda I - A^{N*})^{-1}P^N w - (\lambda I - A^*)^{-1}w \\ &= (\lambda_0 - \lambda)P^N(\lambda I - A^{N*})^{-1}P^N((\lambda_0 I - A^{N*})^{-1}P^N z - (\lambda_0 I - A^*)^{-1}z) \\ &+ P^N(\lambda_0 I - A^{N*})^{-1}P^N z - (\lambda_0 I - A^*)^{-1}z. \end{aligned}$$

This implies that for all $z \in Z$

$$P^N(\lambda I - A^{N*})^{-1}P^N w \rightarrow (\lambda I - A^*)^{-1}w.$$

But since $I + (\lambda - \lambda_0)(\lambda_0 I - A^*)^{-1}$ is onto, the above statement holds for all $w \in Z$.

We turn now to a special case where $\tilde{A}_0 = A_0 + \alpha I$, $\alpha \in \mathbb{R}^1$, $\tilde{A}(s) \equiv 0$, $-r \leq s \leq 0$, $\tilde{\Delta} = \tilde{A}_0 + A_1$ and denote by $\tilde{A}^*, \tilde{A}^{N*}$, and $\tilde{S}^*(t), \tilde{S}^{N*}(t)$ the corresponding infinitesimal generators and semigroups.

LEMMA 5.3. *If statement (b) of Theorem 5.1 holds for $S^{N*}(t)$, then it holds for $\tilde{S}^{N*}(t)$.*

Proof. Note that

$$\tilde{A}^* = A^* + E \quad \text{and} \quad \tilde{A}^{N*} = A^{N*} + P^N E P^N,$$

where $E: Z \rightarrow Z$ is the bounded linear operator defined by

$$E(\eta, \phi(\cdot)) = (\alpha\eta, -A^T(\cdot)\eta).$$

It follows from (5.2) that $\|\tilde{S}^{N*}(t)\| \leq e^{\tilde{\omega}t}$, where $\tilde{\omega} = \omega + |\alpha| + \|A\|_{L_2}$. Consequently, if $\lambda > \tilde{\omega}$, then

$$(5.3) \quad (\lambda I - \tilde{A}^{N*})^{-1} = (\lambda I - A^{N*})^{-1} + (\lambda I - A^{N*})^{-1} P^N E P^N (\lambda I - \tilde{A}^{N*})^{-1},$$

$$(5.4) \quad (\lambda I - \tilde{A}^*)^{-1} = (\lambda I - A^*)^{-1} + (\lambda I - A^*)^{-1} E (\lambda I - \tilde{A}^*)^{-1}.$$

Theorem 5.1 implies that for all $z \in Z$

$$(\lambda I - A^{N*})^{-1} z \rightarrow (\lambda I - A^*)^{-1} z,$$

and, since the ranks of E and E^* are finite, it follows that

$$\|P^N E P^N - E\| \rightarrow 0.$$

It now follows from (5.3)–(5.4) that

$$(\lambda I - \tilde{A}^{N*})^{-1} z \rightarrow (\lambda I - \tilde{A}^*)^{-1} z$$

for all $z \in Z$, and this completes the proof. \square

By Lemma 5.3, without loss of generality, we can assume that $A(\cdot) = 0$ and $\Delta = A_0 + A_1$ is invertible in what follows. We will show that there exists an element $z \in Z$ such that $P^N(A^{N*})^{-1}P^N z$ does not converge to $(A^*)^{-1}z$. First we consider the case when A_1 is not the identity. From Lemma 4.3, if $(\xi, \psi(\cdot)) = [A^*]^{-1}(\eta, 0)$, where $0 \neq \eta \in \mathbb{R}^n$, then

$$\xi = \Delta^{-T} \eta \quad \text{and} \quad \psi(s) \equiv A_1^T \xi, \quad -r \leq s \leq 0.$$

Applying Theorem 4.4, we obtain

$$P^N[A^{N*}]^{-1}P^N(\eta, 0) = i^N a^N,$$

where

$$a_{2k}^N = \xi, \quad 0 \leq 2k \leq N,$$

$$a_{2k+1}^N = -\xi + 2A_1^T \xi, \quad 0 \leq 2k+1 \leq N.$$

For illustration, we have Fig. 1 for the case $N = 4$ and $n = 1$, where the solid line stands for $(A^*)^{-1}(\eta, 0)$ and the dashed line for $i^N a^N$. Since $a_{2k}^N - A_1^T \xi = A_1^T \xi - a_{2k+1}^N = \xi - A_1^T \xi = \varepsilon \in \mathbb{R}^n$ (independent of N), it is easy to show that

$$(5.5) \quad \|P^N(A^{N*})^{-1}P^N(\eta, 0) - (A^*)^{-1}(\eta, 0)\|_Z^2 = 2N \int_0^{r/2N} \left| \frac{2N}{r} s\varepsilon \right|^2 ds = \frac{r}{3} |\varepsilon|^2 \neq 0.$$

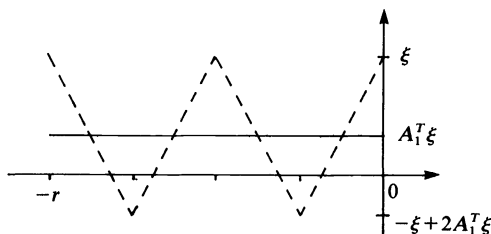


FIG. 1

Next we consider the case $A_1 = I$. Let $\phi(s) = x \neq 0$ (constant vector in R^n). Then, from Lemma 4.3 $(\xi, \psi(\cdot)) = (A^*)^{-1}(0, \phi(\cdot))$ is given by

$$(5.6) \quad \xi = r\Delta^{-T}x \quad \text{and} \quad \psi(s) = \xi - (s+r)x, \quad -r \leq s \leq 0.$$

And, from Theorem 4.4, $P^N(A^{N*})^{-1}P^N(0, \phi(\cdot)) = i^N a^N$, where

$$(5.7) \quad \begin{aligned} a_0^N &= \xi, \\ a_{2k}^N &= \xi + \sum_{\substack{j \text{ odd} \\ 1 \leq j \leq 2k-1}} \left(\frac{r}{N}\right)x = \xi + \frac{2k}{N}rx, \quad 0 \leq 2k \leq N, \\ a_{2k+1}^N &= \xi - \frac{2N-1}{N}rx + 2 \sum_{\substack{j \text{ even} \\ 2 \leq j \leq 2k}} \left(\frac{r}{N}\right)x = \xi - \left(\frac{2N-2k-1}{N}\right)rx, \quad 0 \leq 2k < N. \end{aligned}$$

Since $(a_{2k}^N + a_{2k+1}^N)/2 = \psi((\tau_{2k}^N + \tau_{2k+1}^N)/2)$ and $(a_{2k-1}^N + a_{2k}^N)/2 = \psi((\tau_{2k-1}^N + \tau_{2k}^N)/2)$, it follows from (5.6) and (5.7) that

$$(5.8) \quad \begin{aligned} &\|P^N(A^{N*})^{-1}P^N(0, \phi(\cdot)) - (A^*)^{-1}(0, \phi(\cdot))\|_Z^2 \\ &= 2N \int_0^{r/2N} |2Nsx|^2 ds = \frac{r^2}{3}|x|^2 \neq 0. \end{aligned}$$

Now we may state the main theorem.

THEOREM 5.4. *There exists an element $z \in Z$ and $t > 0$ such that*

$$S^{N*}(t)z \quad \text{does not converge to } S^*(t)z.$$

Proof. If for every $z \in Z$ and $t > 0$, $S^{N*}(t)z$ converges to $S^*(t)z$, then it follows from Theorem 4.4 and Lemma 5.2 that for every $z \in Z$, $P^N(A^{N*})^{-1}P^N z \rightarrow (A^*)^{-1}z$, where by Lemma 5.3 we can assume that $A(\cdot) = 0$ and $\Delta = A_0 + A_1$ is invertible. This contradicts the facts (5.5) and (5.8).

Remarks. The proof of Theorem 5.4 in essence indicates why the numerical computations in [3] exhibit a “zigzag” behavior. The theorem does not provide a description of the set of elements z for which $S^{N*}(t)z$ converges to $S^*(t)z$. In general, we conjecture that the only element z for which $S^{N*}(t)z$ converges to $S^*(t)z$ uniformly on bounded t -intervals is the zero element.

REFERENCES

- [1] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: numerical methods based on averaging approximation*, SIAM J. Control Optim., 16 (1978), pp. 169–208.
- [2] H. T. BANKS AND F. KAPPEL, *Spline approximations for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496–522.
- [3] H. T. BANKS, G. I. ROSEN, AND K. ITO, *A spline based technique for computing Riccati operators and feedback controls in regulator problems for delay equations*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 830–855.
- [4] J. S. GIBSON, *Linear quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 21 (1983), pp. 95–139.
- [5] F. KAPPEL AND D. SALAMON, *Spline approximations for retarded systems and the Riccati equation*, SIAM J. Control Optim., 25 (1987), pp. 1082–1117.
- [6] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.
- [7] R. B. VINTER, *On the evolution of the state of linear differential delay equations in M^2 : Properties of the generator*, J. Inst. Math. Appl., 21 (1978), pp. 13–23.

ON THE CONVERGENCE RATE OF ANNEALING PROCESSES*

CHIANG TZUU-SHUH[†] AND CHOW YUNSHYONG[†]

Abstract. For the class of inhomogeneous Markov processes arising from simulated annealing, it is shown that $\lim_{t \rightarrow \infty} P(X_t = i) / \exp(-u(i)/T(t))$ exists and is positive for each state i , where $T(t)$ is the temperature and $u(i)$ is the energy level at i (assuming that $\min_i u(i) = 0$). The method used is to consider the Forward equations associated with such processes.

Key words. simulated annealing, Forward equations, Perron-Frobenius Theorem, convergence rate

AMS(MOS) subject classifications. primary 60J27, 60J99; secondary 15A51, 15A18, 90B40

0. Introduction. Let $S = \{1, 2, \dots, N\}$ be a finite set and let u be a real-valued function defined on S . For a positive function P (called transition) defined on $S \times S$, we have a continuous-time inhomogeneous Markov process X_t with state space S and transition rate as follows:

$$(0.1) \quad Q_{ij}(t) = \begin{cases} P(i, j) \exp[-(u(j) - u(i))^+ / T(t)] & \text{for } j \neq i, \\ -\sum_{k \neq i} Q_{ik}(t) & \text{for } j = i, \end{cases}$$

where $T(t)$ is a positive function converging to zero as $t \rightarrow \infty$. Markov processes of this type naturally arise from simulated annealing problems, where u is called the energy function and T is the temperature. The reader is referred to [3] and [5] for the motivation and applications. Under some mild conditions on P , $u(\cdot)$, and T , it is proved in [5] that $\lim_{t \rightarrow \infty} P(X_t \in B(S)) = 1$, where $B(S) = \{i \in S : u(i) = \min_{j \in S} u(j)\}$ is the bottom of S . Without loss of generality, we assume that u is integer-valued and $u(i) = 0$ for $i \in B(S)$. In this paper, we shall study the rate of convergence of $P(X_t = i)$ and show that $\lim_{t \rightarrow \infty} P(X_t = i) / \exp(-u(i)/T(t))$ exists and is positive for each $i \in S$ under slightly stronger conditions on $T(t)$ than [5]. Our method is to consider the Forward equations associated with X_t , i.e.,

$$F'(t) = Q^T(t) \cdot F(t),$$

where $Q(t) = (Q_{ij}(t))$ is defined in (0.1) and $F(t) = (F_i(t); i \in S)^T = (P(X_t = i); i \in S)^T$. Note that for each $t \in [0, \infty)$,

$$(0.2) \quad \sum_{i \in S} F_i(t) = 1.$$

In particular, we have the following trivial estimate:

$$(0.3) \quad F_i(t) = O(1) \quad \text{for each } i \in S.$$

In order to describe our results more precisely, we begin with some necessary definitions. A state j is reachable from i if there are states $i_0 = i, i_1, \dots, i_m = j$ such that $P(i_k, i_{k+1}) > 0$ for each $k = 0, \dots, m-1$. Moreover, we say that j is reachable from i at height h if $u(i_k) \leq h$ for each $k = 0, \dots, m$. A state i is called a local minimum if it is not a global minimum (i.e., $i \notin B(S)$), and no state j with $u(j) < u(i)$ can be reached from i at height $u(i)$. Let $h(i, j)$ be the smallest number h such that j can be

* Received by the editors May 25, 1987; accepted for publication (in revised form) February 16, 1988. This work was initiated when the authors were visiting the Institute for Mathematics and Its Applications at the University of Minnesota, and partly supported by the National Science Council, Republic of China.

[†] Institute of Mathematics, Academia Sinica, Taipei, Taiwan.

reached from i at height $u(i) + h$. By convention, we let $h(i, i) = 0$. The depth $d(i)$ of a state i is defined as follows:

$$d(i) = \begin{cases} \min_{j: u(j) < u(i)} h(i, j), & i \notin B(S), \\ \max_{j \in B(S)} h(i, j), & i \in B(S). \end{cases}$$

Note that $d(i) = 0$ if i is not a local or global minimum.

There are two constants associated with the depths of states in S :

$$d_H = \max_{i \in S \setminus B(S)} d(i), \quad d_V = \max_{i \in S} d(i).$$

Obviously, $0 \leq d_H \leq d_V$.

Roughly speaking, d_H is the minimal amount of energy required for any local minimum state to reach $B(S)$ and d_V is the minimal energy required for any minimum (global or local) to reach a state (other than itself) in $B(S)$. The constant d_H was first introduced in Hajek [5] in treating simulated annealing problems and played an essential role there, while d_V was introduced in Ventcel [9] in studying the asymptotic behavior of the second eigenvalue of the transition rate matrix $Q(t)$ and assumes the present form due to a theorem in [1]. Gidas [4] related d_V to the simulated annealing problems.

Let $\lambda(t) = \exp(-1/T(t))$. Obviously, $\lim_{t \rightarrow \infty} \lambda(t) = 0$. We make the following conditions on the transition P and λ .

- (A1) *Irreducibility.* i is reachable from j for any two states i and j in S ;
- (A2) *Weak reversibility.* If i is reachable from j at height h , j is also reachable from i at height h ;
- (A3) $\int_0^\infty \lambda^{d_V}(t) dt = \infty$;
- (A3') $\int_0^\infty \lambda^{d_H}(t) dt = \infty$;
- (A4) $\lambda'(t)/\lambda(t) = o(\lambda^{d_H}(t))$ as $t \rightarrow \infty$;
- (A4') $\lambda'(t)/\lambda(t) = \begin{cases} O(\lambda^{d_H}(t)) & \text{if } d_H > 0 \\ \text{unrestricted} & \text{if } d_H = 0. \end{cases}$ as $t \rightarrow \infty$.

Note that (A3'), (A4') are weaker than (A3) and (A4), respectively.

The main result of this paper, which will be proved in § 3, can now be stated as follows.

THEOREM 0.1. *Let X_t be an annealing process satisfying (A1)–(A4). Then there exist positive constants β_i such that $\lim_{t \rightarrow \infty} P(X_t = i)/\lambda^{u(i)}(t) = \beta_i$ holds for each $i \in S$. As a consequence, we have*

- (i) $P(X_t \in B(S)) = 1 + O(\lambda^a(t))$, where $a = \min_{i \notin B(S)} u(i)$;
- (ii) $\sum_{i \in B(S)} \beta_i = 1$ and $\{\beta_i; i \in B(S)\}$ is the limit distribution of X_t on $B(S)$.

It will be apparent from the proof that the constants β_i are independent of the initial distribution of X_0 and can be obtained through solving systems of linear equations.

Condition (A3) implies that X_t can “communicate” between any two states in $B(S)$. If we are interested only in the rate of convergence of X_t to $B(S)$, then, as will be shown in § 3, (A3) can be replaced by (A3'), which guarantees that X_t can “escape” from any local minimum state to reach $B(S)$.

COROLLARY 0.2. *Let X_t be as in Theorem 0.1. Then (0.4) holds if (A3) is replaced by the weaker condition (A3').*

If we are interested in showing that $P(X_t \in B(S))$ converges to 1, then (A4) can be replaced by (A4'). This will be proved in § 4.

COROLLARY 0.3. *Assume (A1), (A2), (A3'), and (A4') hold. Then*

$$(0.5) \quad P(X_t \in B(S)) = 1 + o(1).$$

A remarkable theorem in [5] asserts that if X_t satisfies (A1), (A2), and (A3') then (0.5) holds. An extra regularity condition on λ (such as (A4')) is needed for our proof because we must solve linear differential equations. In the most interesting case, where $T(t) = c/(\ln t)$, i.e., $\lambda(t) = t^{-1/c}$, it is easy to check that (A3') holds if and only if $c \geq d_H$, and thus (A3') implies (A4'). Thus we have the following result.

COROLLARY 0.4. *Assume (A1), (A2) hold and $\lambda(t) = t^{-1/c}$ with $c > 0$. Then we have*

- (i) *Theorem 0.1 holds if $c \geq d_V$ and $c > d_H$;*
- (ii) *Equation (0.5) holds if and only if $c \geq d_H$.*

Note that the "only if" part of Corollary 0.4(ii) is proved in [5]. Theorem 0.1 has been obtained in [2] for the special case $d_H = 0$. This is not an interesting example because S has a local minimum state if and only if

$$(0.6) \quad d_H > 0.$$

The method used to prove Theorem 0.1 is technically complicated but conceptually simple. To make it clear, we give an example in § 5.

A generalized Ventcel-Freidlin's cycle method, which is different from our approach, was used in [6] to treat similar problems.

We shall adopt the following definitions and notation in the rest of the paper:

$$(0.7) \quad \begin{aligned} F_i &= F_i(t) = P(X_t = i), \\ F_A &= (F_i; i \in A)^T \text{ is a column vector,} \\ F'_A &= (F'_i(t); i \in A)^T, \\ P_i(A) &= \sum_{j \in A} P(i, j), \\ P_{i,L} &= P_i(\{j \in S: j \neq i \text{ and } u(j) \leq u(i)\}), \\ P(A, B) &= (P(i, j); i \in A, j \in B) \text{ is an } |A| \times |B| \text{ matrix,} \\ Q_A &= (q_A(i, j); i, j \in A)^T \text{ is a matrix of order } |A|, \end{aligned}$$

where $q_A(i, i) = -P_{i,L}$ and $q_A(i, j) = P(i, j)$ if $i \neq j$.

DEFINITION 0.1. A cup C at level E is a collection of states in S such that any two states in C can be reached from each other at height E , and any state that can be reached from some state in C at height E is in C .

The maximum M_C , minimum m_C , rim $R(C)$, and bottom $B(C)$ of a cup C are defined as follows:

$$\begin{aligned} M_C &= \max_{i \in C} u(i), & m_C &= \min_{i \in C} u(i), \\ R(C) &= \{i \in C: u(i) = M_C\}, & B(C) &= \{i \in C: u(i) = m_C\}. \end{aligned}$$

DEFINITION 0.2. A cup C is said to be in order $\{r_i\}_{m_C}^{M_C}$ if

- (i) $r_i \leq r_{i+1} \leq r_i + 1$ for each $m_C \leq i \leq M_C - 1$;
- (ii) $F_s = O(\lambda^{r_i})$ if $s \in C$ and $u(s) = i$;
- (iii) $F_s = O(\lambda^{r_{m_C}})$ for any state s with $u(s) > M_C$.

The cup C is said to be in strictly increasing order $\{r_i\}_{m_C}^{M_C}$ if $r_{i+1} = r_i + 1$ for each $m_C \leq i \leq M_C - 1$.

Since $u(\cdot)$ is assumed integer-valued, *throughout this paper we require that all r_i be integers*. Note that r_i can be negative.

DEFINITION 0.3. A cup C in order $\{r_i\}_{m_C}^{M_C}$ is said to have a strictly increasing top from level k if $r_{i+1} = r_i + 1$ for $i \geq k$. Moreover, k is called a bottleneck of C if $r_k = r_{k-1}$.

DEFINITION 0.4. An $n \times n$ matrix $A = (a_{ij})$ is called a transition rate matrix if

- (i) The column sums of A are nonpositive, i.e., $\sum_{i=1}^n a_{ij} \leq 0$ for each j .
- (ii) $a_{ij} \geq 0$ for all $i \neq j$ and $a_{ii} \leq 0$ for each i .

1. Some preliminary estimates. The basic idea behind our approach is first, to show that the cup S is in strictly increasing order $\{r_i\}_0^{M_S}$ with $r_0 = 0$ and then, to establish the exact order estimate for each state.

Since the cup S is obviously in order $\{r_i\}_0^{M_S}$, where $r_i = 0$ for each i , we thus want to improve our estimate level by level starting from $R(S)$. In this section we shall show that this improvement can continue until we hit the highest level where there is a local minimum. We first prove a few technical lemmas that will be used repeatedly throughout the paper.

LEMMA 1.1. *Let $f(t)$ be a complex function and α a complex number with $\operatorname{Re} \alpha > 0$. Suppose*

$$f'(t) = -\alpha f(t) \cdot \lambda^E(t) + O(\lambda^F(t)) \quad (o(\lambda^F(t)), \text{ respectively}),$$

where $E \geq 0$ and F is a real number. Then

$$(1.0) \quad f(t) = O(\lambda^{F-E}(t))(o(\lambda^{F-E}(t)), \text{ respectively})$$

if

- (i) $0 \leq E < d_H$ and $(A3')$, $(A4')$ hold, or if
- (ii) $E = d_H$ and $(A3')$, $(A4)$ hold.

Furthermore, $\int_0^\infty \lambda^E(t) dt = \infty$ is a necessary condition for (1.0) to be true.

Proof. We consider only the $O(\lambda^F(t))$ case. The other case can be treated similarly.

Let $h(t) = \exp[\int_a^t \alpha \cdot \lambda^E(s) ds]$. A simple computation shows $(f(t)h(t))' = h(t) \cdot O(\lambda^F(t))$. Thus $f(t) = [f(a) + \int_a^t h(s) \cdot O(\lambda^F(s)) ds]/h(t)$.

Suppose it is already known that

$$(1.1) \quad \lim_{t \rightarrow \infty} \int_a^t |h(s)| \lambda^F(s) ds = \lim_{t \rightarrow \infty} \lambda^{F-E}(t) |h(t)| = \infty.$$

Since $|h(t)| = \exp[\int_a^t (\operatorname{Re} \alpha) \cdot \lambda^E(s) ds]$, by L'Hôpital's rule and $(A4)$ (or $(A4')$,

$$\begin{aligned} \overline{\lim}_{t \rightarrow \infty} |f(t)| / \lambda^{F-E}(t) &\leq O\left(\overline{\lim}_{t \rightarrow \infty} \left\{ \int_a^t |h(s)| \lambda^F(s) ds / (\lambda^{F-E}(t) |h(t)|) \right\}\right) \\ &= O(\overline{\lim}_{t \rightarrow \infty} \{ |h(t)| \lambda^F(t) / [\lambda^F(t) |h(t)| ((\operatorname{Re} \alpha) \\ &\quad + (F-E) \lambda'(t) / \lambda^{E+1}(t))] \}) \\ &= O(1/(\operatorname{Re} \alpha)). \end{aligned}$$

Note that this can also be shown using integration by parts. It remains to verify (1.1).

By $(A4)$ (or $(A4')$, $[\log(\lambda^F(t)|h(t)|)]' = (F \cdot \lambda'(t)/\lambda(t)) + (\operatorname{Re} \alpha) \cdot \lambda^E(t) \approx (\operatorname{Re} \alpha) \cdot \lambda^E(t)$. Since $E \leq d_H$, it follows from $(A3')$ that $\lim_{t \rightarrow \infty} \lambda^F(t)|h(t)| = \infty$. Formula (1.1) can then be proved easily.

If $\int_0^\infty \lambda^E(t) dt < \infty$ then $h(\infty)$ is finite, and thus

$$\lim_{t \rightarrow \infty} f(t) = \left[f(a) + \int_a^\infty h(s) \cdot O(\lambda^F(s)) ds \right] / h(\infty) \neq 0.$$

This completes the proof. \square

With a similar technique we can prove Lemma 1.2.

LEMMA 1.2. *Let f, α be given as in Lemma 1.1. Suppose*

$$f'(t) = -\alpha f(t) \cdot \lambda^E(t) + o(\lambda^E(t)).$$

Then $f(t) = o(1)$ if

- (i) $0 \leq E \leq d_h$ and (A3') holds, or if
- (ii) $0 \leq E \leq d_v$ and (A3) holds.

Proof. We need only to note that $\int_a^t |h(s)| \lambda^E(s) ds = (\operatorname{Re} \alpha)^{-1} (|h(t)| - 1)$ and $\lim |h(t)| = \infty$. \square

The next lemma is essentially a simple consequence of the Perron-Frobenius Theorem [8] and we omit its proof.

LEMMA 1.3. *Let $A = (a_{ij})$ be a transition rate matrix of order m . If $A^{-1} = (b_{ij})$ exists, then we have the following:*

- (i) *All the eigenvalues of A have negative real parts.*
- (ii) $b_{ii} \leq (\min_i a_{ii})^{-1}$ and $b_{ii} \leq b_{ij} \leq 0$ for all i, j .
- (iii) $b_{ij} < 0$ if and only if i is reachable from j , i.e., there exist $i_0 = i, i_1, \dots, i_k = j$ such that $a_{i_n i_{n+1}} > 0$ for each $0 \leq n < k$.

On the other hand, if A is noninvertible but irreducible, then we have the following:

- (iv) *Zero is an eigenvalue with multiplicity 1 and all other eigenvalues of A have negative real parts.*
- (v) $\sum_{i=1}^m a_{ij} = 0$ for each $1 \leq j \leq m$.
- (vi) *For any proper subset B of $\{1, 2, \dots, m\}$, the principal minor $A_B = (a_{ij}; i, j \in B)$ is an invertible transition rate matrix.*

The following result shows that the trivial estimate (0.3) can be improved from the highest level down until we hit a local minimum.

THEOREM 1.4. *Besides (A1), (A2), and (A3'), assume: (i) Formulae (0.6) and (A4') hold; or (ii) $d_H = 0$ and (A4) holds. Let C be a cup in order $\{r_i\}_{m_C}^{M_C}$. Let E_l be the highest energy level with a local minimum in C . If there exists an integer $h > 0$ such that $F_s(t) = O(\lambda^h(t))$ for any state s with $u(s) > M_C$, then the cup C is in order $\{\tilde{r}_i\}_{m_C}^{M_C}$, where $\tilde{r}_i = r_i$ for $i \leq E_l$ and $\tilde{r}_i = \min(i - E_l + r_{E_l}, h)$ for $i > E_l$.*

Proof. It is enough to show that if $r_{M_C-1} = r_{M_C} = h - 1$ then we have a better estimate $F_s(t) = O(\lambda^h(t))$ for each $s \in R(C)$. Otherwise we can find a subcup $C' \subseteq C$ with $r_{M_C'-1} = r_{M_C'} < r_{M_C'+1}$ and improve the estimate for each state in $R(C')$. This procedure can be repeated over and over until we get the desired estimate.

The Forward equation of a state $i \in R(C)$ is

$$\begin{aligned} F'_i &= -P_{i,L} F_i - \sum_{j: u(j) > u(i)} P(i, j) \lambda^{u(j)-u(i)} F_j + \sum_{\substack{j \in R(C) \\ j \neq i}} P(j, i) F_j \\ &\quad + \sum_{j: u(j) > M_C} P(j, i) F_j + \sum_{j: u(j) < u(i)} P(j, i) \lambda^{u(i)-u(j)} F_j \\ &= I_1 + I_2 + I_3 + I_4 + I_5. \end{aligned} \tag{1.2}$$

Obviously, $I_2 = O(\lambda^h)$. I_4 is of order $O(\lambda^h)$ because each F_j with $u(j) > M_C$ is. I_5 is also of order $O(\lambda^h)$ because C is a cup in order $\{r_i\}_{m_C}^{M_C}$ (see Definition 0.2) and $r_{M_C-1} = h - 1$. In view of (0.7) the Forward equations of states in $R(C)$ take the following matrix form:

$$F'_{R(C)} = Q_{R(C)} \cdot F_{R(C)} + O(\lambda^h). \tag{1.3}$$

Since $Q_{R(C)}$ is an invertible transition rate matrix, all the eigenvalues of $Q_{R(C)}$ have negative real parts by Lemma 1.3. Let α be an eigenvalue of $Q_{R(C)}$ and v be a

left eigenvector corresponding to α . We have from (1.3)

$$(v \cdot F_{R(C)})' = v \cdot F'_{R(C)} = \alpha(v \cdot F_{R(C)}) + O(\lambda^h).$$

Thus $v \cdot F_{R(C)} = O(\lambda^h)$ by Lemma 1.1. If w is a vector such that $w \cdot Q_{R(C)} = \alpha w + v$, then similarly

$$\begin{aligned} (w \cdot F_{R(C)})' &= \alpha(w \cdot F_{R(C)}) + v \cdot F_{R(C)} + O(\lambda^h) \\ &= \alpha(w \cdot F_{R(C)}) + O(\lambda^h). \end{aligned}$$

Therefore, $w \cdot F_{R(C)} = O(\lambda^h)$ by Lemma 1.1 again. Applying Jordan's Decomposition Theorem to $Q_{R(C)}$, we thus have a basis $\{v_1, v_2, \dots, v_{|R(C)|}\}$ of $\mathbb{C}^{|R(C)|}$ such that $v_j \cdot F_{R(C)} = O(\lambda^h)$ for each $j \leq |R(C)|$. Hence $F_{R(C)} = (F_i; i \in R(C))^T = O(\lambda^h)$. This completes the proof. \square

Remark. It is clear from the proof of Theorem 1.4 that, by first changing the error term to $o(1)$ and then applying Lemma 1.2, $F_i = o(1)$ holds for $i \notin B(S)$ under the condition $d_H = 0$. This verifies (0.5) for the case $d_H = 0$.

2. Merging. In this section we shall show (see Theorem 2.1) that a cup in strictly increasing order can be "viewed" as a "single" state in some sense.

Unless otherwise specified, we assume throughout this section that (0.6), (A3'), and (A4') hold.

THEOREM 2.1. *Let C be a cup in strictly increasing order $\{r_i\}_{m_C}^{M_C}$ such that*

$$(2.1) \quad F_s = O(\lambda^{r_{m_C}+1}) \text{ for any state } s \text{ with } u(s) > M_C.$$

Then for any two states i, j in C , there exists a positive constant $\theta(i, j)$ such that

$$(2.2) \quad F_i \lambda^{M_C - u(i)} = \theta(i, j) F_j \lambda^{M_C - u(j)} + \Delta \cdot \lambda^{r_{m_C}},$$

where

$$\Delta = \begin{cases} O(\lambda) & \text{if } m_C > 0 \text{ or } m_C = 0, r_0 < 0, \text{ and (A4) holds,} \\ o(1) & \text{if } m_C = r_0 = 0 \text{ and (A3), (A4) hold, whether } d_H > 0 \text{ or not.} \end{cases}$$

Remark. As explained in the beginning of § 1, our aim is to show that S is in strictly increasing order $\{\bar{r}_i\}_0^{M_S}$ with $\bar{r}_0 = 0$. We do not expect r_0 to be positive in the case $m_C = 0$, i.e., where C contains a global minimum state. This is why we only consider the case $r_0 \leq 0$ if $m_C = 0$ in Theorem 2.1.

We need some preliminary results before we can prove Theorem 2.1.

LEMMA 2.2. *Let C be a cup as in Theorem 2.1 (whether $m_C = 0$ or not). Then*

$$(2.3) \quad F'_i = O(\lambda^{r_{m_C}+1}) \text{ for each } i \in B(C).$$

Moreover, if $B(C)$ is connected in the sense that for any two states i, j in $B(C)$, i can reach j through states in $B(C)$ only, then there exist positive constants $\theta(i, j)$ such that $F_i = \theta(i, j) F_j + O(\lambda^{r_{m_C}+1})$.

Remark. Lemma 2.2 is valid if $d_H = 0$ and (A4) holds.

Proof. Without loss of generality, we assume that $B(C)$ is connected. Otherwise, we can consider each connected component of $B(C)$ separately. Then by (A2), $B(C)$ is irreducible, i.e., any two states in $B(C)$ can reach each other through states in $B(C)$ only. The Forward equation of a state $i \in B(C)$ is

$$\begin{aligned} F'_i &= -P_{i,L} F_i - \sum_{j: u(j) > u(i)} P(i, j) \lambda^{u(j) - u(i)} F_i + \sum_{j \in B(C)} P(j, i) F_j \\ &\quad + \sum_{j: u(j) > u(i)} P(j, i) F_j \\ &= -P_{i,L} F_i + \sum_{j \in B(C)} P(j, i) F_j + O(\lambda^{r_{m_C}+1}). \end{aligned}$$

Or, in matrix form,

$$(2.4) \quad F'_{B(C)} = Q_{B(C)} \cdot F_{B(C)} + O(\lambda^{r_{m_C}+1}).$$

Since $Q_{B(C)}$ is an irreducible transition rate matrix, it follows from Lemma 1.3 that all the eigenvalues of $Q_{B(C)}$ have negative real parts except zero, which is an eigenvalue with multiplicity 1 and $e = (1, 1, \dots, 1)$ as its left eigenvector. By (2.4),

$$(2.5) \quad e \cdot F'_{B(C)} = O(\lambda^{r_{m_C}+1}).$$

Using the same technique as in the proof of Theorem 1.4, we can obtain a basis $\{v_1 = e, v_2, \dots, v_{|B(C)|}\}$ such that for each $2 \leq j \leq |B(C)|$, $v_j \cdot F_{B(C)} = O(\lambda^{r_{m_C}+1})$. Then, in view of (2.4), we have $v_j \cdot F'_{B(C)} = O(\lambda^{r_{m_C}+1})$ for $2 \leq j \leq |B(C)|$. These, together with (2.5), imply (2.3). For the remaining part, we fix a state $i_0 \in B(C)$ and let $A = B(C) \setminus \{i_0\}$. The following linear equation can be obtained from (2.4) and (2.3):

$$Q_A \cdot F_A = -(P(i_0, i); i \in A)^T \cdot F_{i_0} + O(\lambda^{r_{m_C}+1}).$$

By Lemma 1.3(vi), Q_A^{-1} , the inverse of Q_A , exists. Moreover, by Lemma 1.3(ii), all the entries of $(-Q_A^{-1})$ are nonnegative and the diagonal elements are positive. Since

$$(2.6) \quad F_A = (-Q_A^{-1}) \cdot (P(i_0, i); i \in A)^T \cdot F_{i_0} + O(\lambda^{r_{m_C}+1}),$$

it is now clear that there exist constants $\theta(i, i_0) \geq 0$ such that $F_i = \theta(i, i_0)F_{i_0} + O(\lambda^{r_{m_C}+1})$, and by Lemma 1.3(iii), $\theta(i, i_0) > 0$ if $P(i_0, i) > 0$. This completes the proof because i_0 is arbitrary and $B(C)$ is connected. \square

LEMMA 2.3. Let C be a cup as in Theorem 2.1 (whether $m_C = 0$ or not). Then

$$(2.7) \quad F'_s = O(\lambda^{r_i+1}) \quad \text{if } s \in C \quad \text{and} \quad u(s) = i.$$

Proof. Obviously, any subcup C' of C is also in strictly increasing order. Lemma 2.2 shows that (2.7) holds for $s \in B(C')$. By mathematical induction, it is enough to check that (2.7) holds for $s \in R(C)$ under the assumption that

$$(2.8) \quad \text{Formula (2.7) holds for any } s \in A = C \setminus R(C).$$

Since I_2 and I_4 in (1.2) are $O(\lambda^{r_{m_C}+1})$, the Forward equations of states in $R(C)$ take the following matrix form (see (0.7) for notation):

$$(2.9) \quad F'_{R(C)} = Q_{R(C)} \cdot F_{R(C)} + H(t) + O(\lambda^{r_{m_C}+1}),$$

where $H(t) = P(A, R(C))^T \cdot (\lambda^{M_C - u(j)} F_j; j \in A)^T$.

By Lemma 1.3(vi) and (i), $Q_{R(C)}$ is an invertible transition rate matrix and all its eigenvalues have negative real parts. Let v_1 be a left eigenvector corresponding to the eigenvalue α of $Q_{R(C)}$. Then we have, from (2.9),

$$v_1 \cdot F'_{R(C)} = \alpha[v_1 \cdot F_{R(C)} + \alpha^{-1}v_1 \cdot H] + O(\lambda^{r_{m_C}+1}).$$

Since C is in strictly increasing order, $F_j = O(\lambda^{r_{u(j)}}) = O(\lambda^{r_{m_C} - M_C + u(j)})$. By using (0.6), (A4'), and (2.8) it is easy to check that

$$(2.10) \quad H' = O(\lambda^{r_{m_C}+1}).$$

Then, with $f = v_1 \cdot F_{R(C)} + \alpha^{-1}v_1 \cdot H$, we have

$$(2.11) \quad f' = \alpha f + O(\lambda^{r_{m_C}+1}).$$

Therefore, $f = O(\lambda^{r_{m_C}+1})$ by Lemma 1.1, and then $f' = O(\lambda^{r_{m_C}+1})$ in view of (2.11). Now, using the definition of f and (2.10), we obtain

$$(2.12) \quad v_1 \cdot F'_{R(C)} = O(\lambda^{r_{m_C}+1}).$$

Let v_2 be a vector such that $v_2 \cdot Q_{R(C)} = \alpha v_2 + v_1$. From (2.9),

$$v_2 \cdot F'_{R(C)} = \alpha[v_2 \cdot F_{R(C)} + \alpha^{-1} v_2 \cdot H + \alpha^{-1} v_1 \cdot F_{R(C)}] + O(\lambda^{r_{M_C}+1}).$$

By using the same argument and (2.12) we have $v_2 \cdot F'_{R(C)} = O(\lambda^{r_{M_C}+1})$. By Jordan's Decomposition Theorem we can continue this process and finally obtain a basis $\{v_1, \dots, v_{|R(C)|}\}$ of $\mathbb{C}^{|R(C)|}$ such that (2.12) holds with v_1 replaced by v_k , $k \leq |R(C)|$. Hence (2.7) holds for $s \in R(C)$ and the proof is completed. \square

Since $Q_{R(C)}$ is invertible, we immediately obtain the following result from (2.9) and Lemma 2.3.

COROLLARY 2.4. *Let C be a cup as in Theorem 2.1 and let $A = C \setminus R(C)$. Then $F_{R(C)} = -Q_{R(C)}^{-1} \cdot P(A, R(C))^T \cdot (\lambda^{M_C - u(j)} F_j; j \in A)^T + O(\lambda^{r_{M_C}+1})$.*

Remark. If $d_H = 0$ and (A4) holds, then the conclusions in Lemma 2.3 and Corollary 2.4 hold, with the error terms there replaced by $o(\lambda^{r_i})$ and $o(\lambda^{r_{M_C}})$, respectively, because in this case the right-hand side of (2.10) is $o(\lambda^{r_{M_C}})$ instead of $O(\lambda^{r_{M_C}+1})$.

We now prove Theorem 2.1 by an induction argument. Let C be as in Theorem 2.1 and let $A = C \setminus R(C)$. We decompose A into a disjoint union of subcups with height $M_C - 1$ indexed by their depths, i.e., $A = \bigcup_{m=1}^{M_C - m_C} \bigcup_{n=1}^{N_m} C_{m,n}$, where $C_{m,n} \subseteq C$ is a subcup with height $M_C - 1$ and its bottom has energy level $M_C - m$. If $A = \emptyset$, then $C = B(C)$ is connected by Definition 0.1 and Theorem 2.1 follows from Lemma 2.2 and the remark after it. This sets up the induction procedure. Thus it is enough to show that (2.2) holds, given that

(2.13) Equation (2.2) holds for every subcup $C_{m,n}$.

Note that any $C_{m,n}$ is also in strictly increasing order and (2.13) means that for any two states i, j in $C_{m,n}$, there exists $\theta(i, j) > 0$ such that

$$F_i \lambda^{M_{C_{m,n}} - u(i)} = \theta(i, j) F_j \lambda^{M_{C_{m,n}} - u(j)} + \Delta \cdot \lambda^{r_{M_{C_{m,n}}}},$$

where Δ is $o(1)$ or $O(\lambda)$, depending on whether or not $C_{m,n}$ is in Case (iii). By multiplying the previous equation by λ and using $M_{C_{m,n}} = M_C - 1$, we obtain

$$(2.14) \quad F_i \lambda^{M_C - u(i)} = \theta(i, j) F_j \lambda^{M_C - u(j)} + \Delta \cdot \lambda^{r_{M_C}}.$$

We shall use this form henceforth.

Under the assumptions in Theorem 2.1 and (2.13) we can improve Lemma 2.3.

LEMMA 2.5. *For any subcup $C_{m,n}$ of C ,*

$$\left(\sum_{i \in C_{m,n}} F_i \right)' = \begin{cases} O(\lambda^{r_{M_C}+1}) & \text{if } m_C > 0, \text{ or } m_C = 0, r_0 < 0 \text{ and (A4) hold,} \\ o(\lambda^{r_{M_C}}) & \text{if } m_C = r_0 = 0, \text{ (A3), and (A4) hold, whether or not } d_H > 0. \end{cases}$$

Proof. Let $b_{m,n}$ be a fixed state in $B(C_{m,n})$. We shall adopt the following notation. For $m = 1, 2, \dots, M_C - m_C$,

$$I(m) = \{(m, 1), (m, 2), \dots, (m, N_m)\},$$

$$J(m) = \bigcup_{k < m} I(k)$$

$$= \{(1, 1), \dots, (1, N_1), (2, 1), \dots, (2, N_2), \dots, (m-1, 1), \dots, (m-1, N_{m-1})\},$$

$$\begin{aligned}
L(m) &= \bigcup_{k > m} I(k) \\
&= \{(m+1, 1), \dots, (m+1, N_{m+1}), \dots, (M_C - m_C, 1), \dots, (M_C - m_C, N_{M_C - m_C})\}, \\
(2.15) \quad C_m &= \{b_{k,n} : (k, n) \in J(m)\} = \{b_{k,n} : k < m\}, \\
F_{m,n} &= \sum_{i \in C_{m,n}} F_i, \\
w_{m,n} &= \left[\sum_{i \in B(C_{m,n})} \theta(i, b_{m,n}) \right]^{-1}, \\
q_{m,n}(j) &= \left[\sum_{i \in C_{m,n}} \theta(i, b_{m,n}) P(i, j) \right] w_{m,n}, \\
q_{m,n}(B) &= \sum_{j \in B} q_{m,n}(j) \quad \text{for any set } B.
\end{aligned}$$

Case (i). $m_C > 0$. Then $m_{C_{k,n}} \geq m_C \geq 1$, because $u(\cdot)$ is assumed to be integer-valued. Since C is in strictly increasing order, $\{r_i\}_{m_C}^{M_C}$, $F_i = O(\lambda^{r_{M_C} - M_C + u(i)})$ for $i \in C$. Under (2.13) it is easy to check that for any subcup $C_{k,n}$ and $j \in R(C)$,

$$\begin{aligned}
\lambda^k F_{b_{k,n}} &= w_{k,n} \lambda^k F_{k,n} + O(\lambda^{r_{M_C} + 1}) \sum_{i \in C_{k,n}} P(i, j) \lambda^{M_C - u(i)} F_i \\
(2.16) \quad &= \left[\sum_{i \in C_{k,n}} P(i, j) \theta(i, b_{k,n}) \right] \lambda^k F_{b_{k,n}} + O(\lambda^{r_{M_C} + 1}) \\
&= q_{k,n}(j) \lambda^k F_{k,n} + O(\lambda^{r_{M_C} + 1}),
\end{aligned}$$

and then

$$(2.17) \quad \lambda^k F_{k,n} = O(\lambda^{r_{M_C}}) \quad \text{for any subcup } C_{k,n}.$$

By using the Forward equations for states in $C_{k,n}$ (see (0.7) for notation),

$$\begin{aligned}
F'_{k,n} &= \sum_{i \in C_{k,n}} F'_i \\
(2.18) \quad &= - \sum_{i \in C_{k,n}} P_i(R(C)) \lambda^{M_C - u(i)} F_i + \sum_{j \in R(C)} P_j(C_{k,n}) F_j + O(\lambda^{r_{M_C} + 1}) \\
&= -q_{k,n}(R(C)) \lambda^k F_{k,n} + \sum_{j \in R(C)} P_j(C_{k,n}) F_j + O(\lambda^{r_{M_C} + 1}).
\end{aligned}$$

Thus, from (2.17) we have

$$(2.19) \quad F'_{k,n} = O(\lambda^{r_{M_C}}) \quad \text{for any subcup } C_{k,n}.$$

Assume temporarily that

$$(2.20) \quad \text{The lemma is true for all } k \leq m-1.$$

Then we have the following equation:

$$\begin{aligned}
(2.21) \quad &\begin{bmatrix} Q_{R(C)} & Q_{J(m)}(R(C)) \\ P_{R(C)}(J(m)) & \tilde{Q}_{J(m)}(R(C)) \end{bmatrix} \begin{bmatrix} F_{R(C)} \\ G_{J(m)} \end{bmatrix} \\
&+ \begin{bmatrix} Q_{I(m)}(R(C)) & Q_{L(m)}(R(C)) \\ O_1 & O_2 \end{bmatrix} \begin{bmatrix} G_{I(m)} \\ G_{L(m)} \end{bmatrix} = O(\lambda^{r_{M_C} + 1}),
\end{aligned}$$

where $Q_{R(C)}$ is a $|R(C)| \times |R(C)|$ matrix defined in (0.7), O_1 and O_2 are zero matrices of proper size, $P_{R(C)}(J(m))$ is a matrix of order $|J(m)| \times |R(C)|$ with $(P_i(C_{k,n}); (k, n) \in J(m))^T$ as its i th column, $\tilde{Q}_{J(m)}(R(C))$ is a $|J(m)| \times |J(m)|$ diagonal matrix with

$q_{k,n}(R(C))$, $(k, n) \in J(m)$ as its diagonal elements, and for $D = I(m)$, $J(m)$ or $L(m)$, $Q_D(R(C))$ is a matrix of order $|R(C)| \times |D|$ with $(q_{k,n}(i); (k, n) \in D)$ as its i th row, and $G_D = (\lambda^k(t)F_{k,n}(t); (k, n) \in D)^T$ is a column vector of length $|D|$. Indeed, the first $|R(C)|$ equations are the Forward equations for states in $R(C)$ and follow from Lemma 2.3 and (2.9). Note that the summation over A in $H(t)$ of (2.9) is decomposed into many subsummations, each of which sums over a subcup $C_{k,n}$, and is then transformed to a single term by using (2.16). The remaining $|J(m)|$ equations follow from (2.18) and (2.20).

Call the first constant matrix in (2.21) Q_m . It is clear from the definition of a cup that $Q_{M_C - m_C + 1}$ is an irreducible transition rate matrix with all its column sums equal to zero. Since $m \leq M_C - m_C$, Q_m is an invertible transition rate matrix by Lemma 1.3(vi) and thus

$$(2.22) \quad \begin{bmatrix} F_{R(C)} \\ G_{J(m)} \end{bmatrix} = -Q_m^{-1} \cdot \begin{bmatrix} Q_{I(m)}(R(C)) & Q_{L(m)}(R(C)) \\ O_1 & O_2 \end{bmatrix} \begin{bmatrix} G_{I(m)} \\ G_{L(m)} \end{bmatrix} + O(\lambda^{r_{M_C}+1}).$$

This equation means that functions $F_{R(C)}$ and $\lambda^k F_{k,n}$ for subcups $C_{k,n}$ with $k < m$ can be expressed in terms of those corresponding functions for deeper subcups, i.e., $\lambda^k F_{k,n}$ with $k \geq m$.

Observe that $|C_m| = |J_m|$ and for $j \in C_m$, $1 \leq n \leq N_m$,

$$(2.23) \quad P_j(C_{m,n}) = 0.$$

By first adding $(P_j(C_{m,n}); j \in C_m) \cdot G_{J(m)} \equiv 0$ to the second term on the right-hand side of (2.18) and then substituting (2.2) into (2.18), we have

$$\begin{aligned} F'_{m,n} &= -q_{m,n}(R(C))\lambda^m F_{m,n} + (P_j(C_{m,n}); j \in R(C) \cup C_m) \cdot \begin{pmatrix} F_{R(C)} \\ G_{J(m)} \end{pmatrix} + O(\lambda^{r_{M_C}+1}) \\ &= -q_{m,n}(R(C))\lambda^m F_{m,n} + (P_j(C_{m,n}); j \in R(C) \cup C_m) \cdot (-Q_m^{-1}) \\ &\quad \cdot \left\{ \begin{bmatrix} Q_{I(m)}(R(C)) \\ O_1 \end{bmatrix} \cdot G_{I(m)} + \begin{bmatrix} Q_{L(m)}(R(C)) \\ O_2 \end{bmatrix} \cdot G_{L(m)} \right\} + O(\lambda^{r_{M_C}+1}). \end{aligned}$$

Hence the Forward equations of $F_{m,n}$, where $n = 1, \dots, N_m$, have the following matrix form:

$$\begin{aligned} (2.24) \quad (F'_{m,n}; n \leq N_m)^T &= P_m \cdot G_{I(m)} + R_m \cdot G_{L(m)} + O(\lambda^{r_{M_C}+1}) \\ &= P_m \cdot (\lambda^m F_{m,n}; n \leq N_m)^T + R_m \cdot (\lambda^k F_{k,n}; n \leq N_k, k > m)^T \\ &\quad + O(\lambda^{r_{M_C}+1}), \end{aligned}$$

where P_m is an $N_m \times N_m$ matrix that on the diagonal, $P_m(n, n) = -q_{m,n}(R(C)) + (P_j(C_{m,n}); j \in R(C) \cup C_m) \cdot (-Q_m)^{-1} \cdot (q_{m,n}(j); j \in R(C) \cup C_m)^T$; and off the diagonal, $P_m(l, n) = (P_j(C_{m,l}); j \in R(C) \cup C_m) \cdot (-Q_m)^{-1} \cdot (q_{m,n}(j); j \in R(C) \cup C_m)^T$. Note that $q_{m,n}(j) = 0$ for $j \in C_m$. We claim that

$$(2.25) \quad P_m \text{ is a transition rate matrix;}$$

$$(2.26) \quad P_m \text{ is invertible unless } m = M_C - m_C, \text{ in which case } P_m \text{ is irreducible and all its column sums are zero.}$$

Indeed, the n th column sum of P_m is

$$\sum_{l=1}^{N_m} P_m(l, n) = -q_{m,n}(R(C)) + Y_m \cdot (q_{m,n}(j); j \in R(C) \cup C_m)^T,$$

where $Y_m = (y_i; i \in R(C) \cup J_m) = (P_j(\bigcup_{k=1}^{N_m} C_{m,k}); j \in R(C) \cup C_m) \cdot (-Q_m)^{-1}$. Since

$$(2.27) \quad \left(\begin{array}{c|c} Q_m & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ \hline (P_j(\bigcup_{k=1}^{N_m} C_{m,k}); j \in R(C) \cup C_m) & \begin{matrix} -Q_m^{-1} \\ 0 \\ 1 \end{matrix} \end{array} \right) = -I,$$

all $y_i \leq 1$ by Lemma 1.3(ii). Therefore,

$$(2.28) \quad \begin{aligned} \sum_{l=1}^{N_m} P_m(l, n) &\leq -q_{m,n}(R(C)) + (1, \dots, 1) \cdot (q_{m,n}(j); j \in R(C) \cup C_m)^T \\ &= -q_{m,n}(R(C)) + q_{m,n}(R(C)) = 0. \end{aligned}$$

This verifies (2.25).

It is clear from (2.15) that, for $j \in R(C)$, the condition $q_{m,n}(j) > 0$ means that (some state in) subcup $C_{m,n}$ can reach j in one step. Write $-Q_m^{-1} = (\eta_{ij})$. Then by using Lemma 1.3(iii),

$$(2.29) \quad P_m(l, n) = \sum_{j,k=1}^{|R(C)|} P_j(C_{m,l}) \eta_{jk} q_{m,n}(k) > 0 \text{ if and only if } C_{m,l} \text{ can be reached from } C_{m,n} \text{ through } R(C) \text{ and subcups of less depth, i.e., } C_{k,l} \text{ with } k < m.$$

Suppose P_m is not invertible. By the Perron-Frobenius Theorem there exists a nonzero vector $Z_m = (z_i; 1 \leq i \leq N_m)$ such that all $z_i \geq 0$ and $Z_m \cdot P_m = 0$. Let $\bar{D} = \{i \in R(C) \cup J(m): y_i = 1\}$ and $D = \{i: 1 \leq i \leq N_m \text{ and } z_i \geq z_j \text{ for all } j \leq N_m\}$. Assume $n \in D$. By computing the n th component of $Z_m \cdot P_m$, we have the following:

- (i) $l \in D$ if $P_m(l, n) > 0$;
- (ii) Equality holds in (2.28);
- (iii) If $q_{m,n}(i) > 0$ then $i \in R(C) \cap \bar{D}$, and moreover, by comparing the i th component on both sides of the equation

$$(2.30) \quad \left(P_j \left(\bigcup_{l=1}^{N_m} C_{m,l} \right); j \in R(C) \cup C_m \right) = Y_m \cdot (-Q_m)$$

and noting that

$$P_i \left(\bigcup_{l=1}^{N_m} C_{m,l} \right) = P_{i,L} - P(i, R(C) \setminus \{i\}) - P_i \left(\bigcup_{k < m} \bigcup_{l=1}^{N_k} C_{k,l} \right) - P_i \left(\bigcup_{k > m} \bigcup_{l=1}^{N_k} C_{k,l} \right)$$

we can easily show the following:

- (iv) $P_i(\bigcup_{k > m} \bigcup_{l=1}^{N_k} C_{k,l}) = 0$;
- (v) $j \in \bar{D}$ if $P(i, j) > 0$ and $j \in R(C)$;
- (vi) $(k, l) \in \bar{D}$ if $P_i(C_{k,l}) > 0$ and $(k, l) \in J(m)$. Then by comparing the (k, l) th component on both sides of (2.30), $j \in \bar{D}$ if $q_{k,l}(j) > 0$.

We can conclude from the conditions above that if $i \in R(C)$ can be reached from $C_{m,n}$ through $R(C)$ and subcups of depth $\leq m$, then (iv) holds, i.e., any subcup of depth greater than m cannot be reached from i in one step. Since by Definition 0.1 subcups $C_{k,l}$ of C can be reached from each other within C , a contradiction follows unless there is no subcup of depth greater than m . In that case it is clear from (2.29) that P_m is irreducible. It remains to show that when $m = M_C - m_C$ all the column sums

of P_m are zero. Let w_i , $1 \leq i < |R(C) \cup C_m|$, be the i th column vector of the first matrix in (2.27). By changing the multiplication order of those two matrices in (2.27) we have

$$(2.31) \quad (Y_m, 1) \cdot w_i = 0 \quad \text{for all } i.$$

Observe that when $m = M_C - m_C$ all w_i have components' sum equal to zero. Since Y_m is unique by (2.27), it follows from (2.31) and the observation above that when $m = M_C - m_C$, $Y_m = (1, 1, 1, \dots, 1)$, and thus all the column sums of P_m are zero. This completes the proof of (2.26). In particular,

$$(2.32) \quad P_m \text{ is the scalar zero if } m = M_C - m_C \text{ and } N_m = 1.$$

Let v_1 be a left eigenvector corresponding to an eigenvalue α of P_m with $\operatorname{Re} \alpha < 0$. From (2.24),

$$(2.33) \quad v_1 \cdot (F'_{m,n}; n \leq N_m)^T = \alpha \lambda^m [v_1 \cdot (F_{m,n}; n \leq N_m)^T + \tilde{H}(t)] + O(\lambda^{r_{M_C}+1}),$$

where $\tilde{H}(t) = \alpha^{-1} v_1 \cdot R_m \cdot (\lambda^{k-m} F_{k,n}; n \leq N_k, k > m)^T$.

Since $m_C \geq 1$ implies that any state in $B(C)$ must climb at least a height of $M_C - m_C + 1$ to get to some lower state, we have $m + 1 \leq M_C - m_C + 1 \leq d_H$. In view of (A4') and (2.17),

$$(2.34) \quad \lambda^{k-m} (\lambda'/\lambda) F_{k,n} = O(\lambda^{r_{M_C}+1}),$$

and then from (2.19),

$$(2.35) \quad \begin{aligned} \tilde{H}'(t) &= \alpha^{-1} v_1 \cdot R_m \cdot (\lambda^{k-m} [(k-m)(\lambda'/\lambda) F_{k,n} + F'_{k,n}]; n \leq N_k, k > m) \\ &= O(\lambda^{r_{M_C}+1}). \end{aligned}$$

Let f be the sum inside the square bracket of (2.33). From (2.33) and (2.35),

$$(2.36) \quad f' = \alpha \lambda^m f + O(\lambda^{r_{M_C}+1}).$$

Then by Lemma 1.1, (2.36), and (2.35), both f' and $v_1 \cdot (F'_{m,n}; n \leq N_m)^T$ are $O(\lambda^{r_{M_C}+1})$. If v_2 is a vector such that $v_2 \cdot P_m = \alpha v_2 + v_1$, then, using the same technique as in the proof of Lemma 2.3, we obtain $v_2 \cdot (F'_{m,n}; n \leq N_m)^T = O(\lambda^{r_{M_C}+1})$. If zero is an eigenvalue of P_m , then by (2.26), $m = M_C - m_C$, and the corresponding left eigenvector is $e = (1, \dots, 1)$. Since in this case the second term on the right-hand side of (2.24) disappears, $e \cdot (F'_{m,n}; n \leq N_m)^T = O(\lambda^{r_{M_C}+1})$. Thus in either case we have a basis $\{v_1, \dots, v_{N_m}\}$ of \mathbb{C}^{N_m} such that $v_i \cdot (F'_{m,n}; n \leq N_m)^T = O(\lambda^{r_{M_C}+1})$ for each i . Hence,

$$(2.37) \quad F'_{m,n} = O(\lambda^{r_{M_C}+1}) \quad \text{for each } n \leq N_m$$

under condition (2.20). This condition is unnecessary, in fact. When $m = 1$, $J(1) = \emptyset$ and there are only $|R(C)|$ equations in (2.21). The previous argument shows that (2.37) holds for $m = 1$. This sets up, in some sense, the induction procedure. Therefore, (2.37) holds for all $m \leq M_C - m_C$. This completes the proof for Case (i).

Case (ii). $m_C = 0$, $r_0 < 0$, and (A4) holds. That is, C contains a global minimum state but is not yet in the order we want. Since every step in Case (i) works well for $m < d_H$, we need only to consider the case that $m \geq d_H$. Note that the definition of d_H implies that there is no subcup $C_{k,l}$ with $d_H < k < M_C$. When $m = M_C$, the second term on the right-hand side of (2.24) disappears. The first term is $O(\lambda^{r_{M_C}+1})$, because by (0.3),

$$(2.38) \quad \lambda^{M_C} F_{M_C,n} = O(\lambda^{M_C}) = O(\lambda^{r_{M_C}-r_0}) = O(\lambda^{r_{M_C}+1}).$$

Hence, (2.37) holds for $m = M_C$. In the case $m = d_H$, applying (2.38) to (2.24), we have

$$(2.39) \quad (F'_{d_H,n}; n \leq N_{d_H})^T = P_{d_H} \cdot (\lambda^{d_H} F_{d_H,n}; n \leq N_{d_H})^T + O(\lambda^{r_{M_C}+1}).$$

Since P_{d_H} is an invertible transition rate matrix by (2.25) and (2.26), (2.37) still holds for $m = d_H$ by Lemma 1.1(ii) and the same argument as before. Note that this is the only place we need (A4). If $r_{M_C} + 1 \leq d_H$ then $\lambda^{d_H} F_{d_H,n} = O(\lambda^{d_H}) = O(\lambda^{r_{M_C}+1})$ by (0.3), and thus (2.37) holds for $m = d_H$. Therefore, for any subcup $C_{m,n}$,

$$(2.40) \quad F'_{m,n} = O(\lambda^{r_{M_C}+1}) \quad \text{if } r_{M_C} + 1 \leq d_H, r_0 < 0 \quad \text{and} \quad (A3'), (A4') \text{ hold.}$$

If $r_{M_C} = d_H$, (2.40) may no longer hold. However, we have

$$(2.41) \quad F_{d_H,n} = o(1) \quad \text{if } r_{M_C} = d_H, r_0 < 0 \quad \text{and} \quad (A3'), (A4') \text{ hold.}$$

In particular, $F_i = o(1)$ for $i \in B(C_{d_H,n})$. Then, as in the proof of Theorem 1.4,

$$(2.42) \quad F_i = o(\lambda^{d_H}) \quad \text{for any } i \in R(C).$$

Condition (2.41) can be shown by the same technique if we first replace the error term in (2.39) by $o(\lambda^{d_H})$ and then apply Lemma 1.2(i). Conditions (2.41) and (2.42) will be used in the proof of Corollary 0.3. Note that (2.41) and (2.42) hold with (2.1) replaced by

$$(2.1') \quad F_s = o(\lambda^{d_H}) \quad \text{for any state } s \text{ with } u(s) > M_C.$$

Case (iii). $m_C = r_0 = 0$ and (A3), (A4) hold. First we consider the case $d_H > 0$. With $O(\lambda^{r_{M_C}+1})$ replaced by $o(\lambda^{r_{M_C}})$, the same argument in Case (i) works for $m \leq d_H$. In particular, we have

$$(2.24') \quad (F'_{m,n}; n \leq N_m)^T = P_m \cdot (\lambda^m F_{m,n}; n \leq N_m)^T + R_m \cdot G_{L(m)} + o(\lambda^m), m \leq M_C.$$

Note that when $m = M_C$ the term $R_m \cdot G_{L(m)}$ in (2.24') disappears. If $N_{M_C} \geq 2$, i.e., if there are at least two subcups of depth M_C , then $d_V \geq M_C$. Now apply Lemma 1.2(ii) and repeat the same argument in Case (i). If $N_{M_C} = 1$, then $F'_{M_C,1}$ is already of the desired order by (2.32) and (2.24').

In case $d_H = 0$ it is clear from the remarks after Lemma 2.2 and Corollary 2.4 that all formulas hold with $O(\lambda^{r_{M_C}+1})$ replaced by $o(\lambda^{r_{M_C}})$. Thus the same argument for $d_H > 0$ works as well now.

This completes the proof of the lemma. \square

We can now prove (2.2) in C under assumption (2.13). Remember that $A = C \setminus R(C) = \bigcup_{n=1}^{M_C - m_C} C_{m,n}$. Let $h = M_C - m_C$. From (2.24) (or (2.24')) and Lemma 2.5, we have

$$P_h \cdot (\lambda^h F_{h,n}; n \leq N_h)^T = \Delta \cdot \lambda^{r_{M_C}},$$

where $\Delta = o(1)$ or $O(\lambda)$, depending on whether C is in Case (iii) or not. By (2.25), (2.26), and Lemma 1.3(iv), P_h is an irreducible matrix of rank $N_h - 1$. Thus

$$(2.43) \quad \lambda^h F_{h,n} = \tilde{\theta}(n, l) \lambda^h F_{h,l} + \Delta \cdot \lambda^{r_{M_C}},$$

where $\tilde{\theta}(n, l) > 0$ by the same argument as in the proof of Lemma 2.2. Note that if we let $\tilde{\theta}(1, 1) = 1$, then (2.43) automatically holds for the case $N_h = 1$. For $k < h$, we have from (2.24) (or (2.24'))

$$(2.44) \quad (\lambda^k F_{k,n}; n \leq N_k)^T = -P_k^{-1} \cdot R_k \cdot (\lambda^m F_{m,n}; n \leq N_m, m > k)^T + \Delta \cdot \lambda^{r_{M_C}}.$$

It obviously follows from (2.43), (2.44), and (2.14) that (2.2) holds for any two states in A . For the states in $R(C)$, (2.2) is now an easy consequence of Corollary 2.4. The theorem is thus proved by induction. \square

3. Proof of Theorem 0.1 and Corollary 0.2. In this section we shall show that under (A3') and (A4) the cup S is of strictly increasing order $\{r_i\}_0^{M_S}$ with $r_0 = 0$ (note

that $m_s = 0$ is assumed for brevity), and then, as a consequence, we establish our main result. First we need a lemma.

LEMMA 3.1. *Assume that (A1), (A2), (A3'), and (A4) hold. Let C be a cup in order $\{r_i\}_{m_C}^{M_C}$ with a bottleneck at k . If $F_s = O(\lambda^{r_{M_C}+1})$ for any s with $u(s) > M_C$, then C is a cup in order $\{\bar{r}_i\}_{m_C}^{M_C}$ with a bottleneck at $k+1$, where $\bar{r}_i = r_i$ for $i \neq k$ and $\bar{r}_k = r_k + 1$.*

Proof. *Case 1.* $d_H > 0$. Let $s \in S$ with $u(s) = k$. Obviously, all we need to show is that $F_s = O(\lambda^{r_k+1})$. As in § 2, decompose $C = R(C) \cup A = R(C) \cup (\cup_{m=1}^h \cup_{n=\frac{m}{2}}^{N_m} C_{m,n})$, where $h = M_C - m_C$ and $C_{m,n}$ is a subcup with height $M_C - 1$ and bottom energy $M_C - m$. If s is not a local minimum, then the lemma follows from applying Theorem 1.4 to some subcup of C having s in its rim. Let $s \in C_{d,n}$ be a local minimum. Certainly $d < h$. Without loss of generality we assume that $s \in B(C_{d,n})$. Otherwise, replace C by $C_{d,n}$ and start the same procedure over. Note that now $k = M_C - d$. Obviously C is in strictly increasing order $\{r'_i\}_{m_C}^{M_C}$, where $r'_i = r_i$ for $i \geq k$ and $r'_{k-i} = r_k - i$ for $1 \leq i \leq k - m_C$. In case $m_C = 0$, we have $r'_0 \leq r_0 - 1 < 0$ (see the remark after Theorem 2.1). Thus Theorem 2.1 is applicable with $\Delta = O(\lambda)$ in (2.2). From (2.44), we have

$$(3.1) \quad (\lambda^d F_{d,n}; n \leq N_d)^T = -P_d^{-1} \cdot R_d \cdot (\lambda^m F_{m,n}; n \leq N_m, m > d) + O(\lambda^{r_{M_C}+1}).$$

Since k is a bottleneck of C , $r_{M_C-m} \geq r_{M_C} - m + 1$ holds for $m > M_C - k = d$. By assumption $F_{b_{m,n}} = O(\lambda^{r_{M_C-m}})$. Now, by (2.16),

$$(3.2) \quad \lambda^m F_{m,n} = (1/w_{m,n}) \lambda^m F_{b_{m,n}} + O(\lambda^{r_{M_C}+1}) = O(\lambda^{r_{M_C}+1}) \quad \text{for } m > d.$$

Thus, by (3.1) and (3.2),

$$F_s \leq \sum_{i \in C_{d,n}} F_i = F_{d,n} = O(\lambda^{r_{M_C}+1-d}) = O(\lambda^{r_k+1}).$$

Case 2. $d_H = 0$. Since C has no local minimum in this case, it is already treated in Case 1.

This completes the proof. \square

We are now ready to prove Theorem 0.1 and Corollary 0.2. Assume (A3') holds. Obviously, S is in order $\{r'_i\}_0^{M_S}$, where all $r'_i = 0$. By successive applications of Lemma 3.1, we then have S in strictly increasing order $\{r_i\}_0^{M_S}$ with $r_0 = 0$. (This proves Corollary 0.2. If (A3) also holds, then by Theorem 2.1 (Case (iii)), (2.2) is true with $C = S$ and $\Delta = o(1)$. Note that $r_{M_S} = M_S$. In view of (0.2), Theorem 0.1 follows from a simple computation. \square

4. Proof of Corollary 0.3.

Case (i). $d_H = 0$. See the remark after the proof of Theorem 1.4.

Case (ii). $d_H > 0$. It is enough to show that $F_i = o(1)$ for any $i \notin B(S)$. Let $D = \{i \in S: d(i) = d_H, \text{ i.e., state } i \text{ has the maximum depth}\}$ and let $D_1 = \{i \in D: u(i) \geq u(j) \text{ for all } j \in D\}$. In view of (2.40), Lemma 3.1 can be applied repeatedly to improve the trivial estimate (0.3) until we hit a state in D_1 ; that is, if $u(D_1) = k$ then S is in order $\{r_i\}_0^{M_S}$ with $r_i = (i - k)^+$, or in strictly increasing order $\{r'_i\}_0^{M_S}$ with $r'_i = i - k$. Note that $M_S - k \geq d_H$.

As in § 2, decompose $S = R(S) \cup A = R(S) \cup (\cup_{m=1}^{M_S} \cup_{n=\frac{m}{2}}^{N_m} C_{m,n})$, where $C_{m,n}$ is a subcup with height $M_S - 1$ and bottom energy $M_S - m$. Without loss of generality we may assume $M_S - k = d_H$, i.e., each state in D_1 is in the bottom of some $C_{d_H,n}$. Otherwise, replace S by cups at level $k + d_H$ (see Definition 0.1). Thus

$$(4.1) \quad \sum_{i \in R(S)} F_i + \sum_{m < d_H} \sum_n F_{m,n} = O(\lambda) = o(1).$$

Since $r_{M_C} = d_H$ and $r'_0 < 0$, we have from (2.41) and (2.42),

$$(4.2) \quad \sum_{i \in D_1} F_i \leq \sum_n F_{d_H, n} = o(1),$$

$$(4.3) \quad F_i = o(\lambda^{d_H}) \quad \text{for any } i \text{ with } u(i) \geq k + d_H.$$

Note that (4.1)–(4.3) hold under (2.1').

From (4.1) and (4.2), it remains to consider $C_{M_S, n}$. Take $C = C_{M_S, 1}$, for example. Since (2.1') holds for C by (4.3), the previous argument can be repeated, by appealing to Lemma 3.1 and Theorem 1.4, to treat states in $D_2 = \{i \in D \setminus D_1: u(i) \geq u(j) \text{ for all } j \in D \setminus D_1\}$, and so on. Thus we can obtain

$$\sum_{i \in D} F_i = o(1), \quad \sum_{i \notin D \cup B(S)} F_i = O(\lambda) = o(1).$$

This completes the proof. \square

5. An example. In Fig. 1, $S = \{1, 2, 3, \dots, 9\}$ and $B(S) = \{1, 9\}$. It is easy to check that $d(5) = d(7) = 1$, $d(3) = 2$, and $d(1) = d(9) = 4$. Hence $d_H = 2$ and $d_V = 4$.

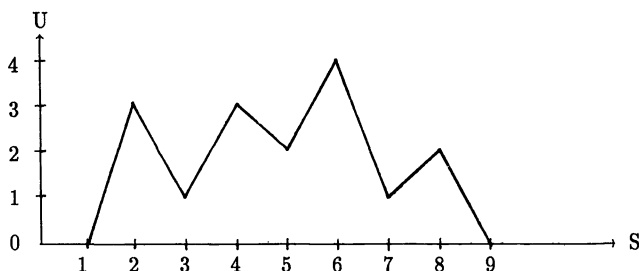


FIG. 1

Step 1. By applying Theorem 1.4 to S , (0.3) can be improved from the top level to the highest local minimum state five, i.e.,

$$F_6 = O(\lambda^2), \quad F_2, F_4 = O(\lambda).$$

Similarly, by applying Theorem 1.4 to the cup $C_1 = \{7, 8, 9\}$, we find that

$$F_8 = O(\lambda).$$

Step 2. Write $F_5 = O(\lambda^0)$, $F_3 = O(\lambda^{-1})$, and $F_1 = O(\lambda^{-2})$ so that cup $C_2 = \{1, 2, 3, 4, 5\}$ is in strictly increasing order. By applying Theorem 2.1 to C_2 , we have $F_5 \cdot \lambda = \theta(5, 1) F_1 \cdot \lambda^3 + O(\lambda^2)$. Since $F_1 = O(1)$, we immediately get $F_5 = O(\lambda)$. By using Lemma 3.1 twice we have

$$(5.1) \quad F_5, F_8 = O(\lambda), \quad F_2, F_4 = O(\lambda^2), \quad F_6 = O(\lambda^3).$$

Step 3. Write $F_3 = O(\lambda^0)$ and $F_1 = O(\lambda^{-1})$. Repeating the same procedure in Step 2 we can improve (5.1) to

$$(5.2) \quad F_3 = O(\lambda), \quad F_5 = O(\lambda^2), \quad F_2, F_4 = O(\lambda^3).$$

Step 4. Cup $C_1 = \{7, 8, 9\}$ can be treated by the same way to yield

$$(5.3) \quad F_7 = O(\lambda), \quad F_8 = O(\lambda^2).$$

Then Lemma 3.1 can be used to show that

$$(5.4) \quad F_6 = O(\lambda^4).$$

Step 5. Equations (5.2)–(5.4) demonstrate that cup S is in the desired strictly increasing order. Theorem 0.1 then follows from Theorem 2.1.

Acknowledgment. The authors are grateful to the Institute for Mathematics and Its Applications at the University of Minnesota for its hospitality during their visit.

REFERENCES

- [1] T. S. CHIANG AND Y. CHOW, *On eigenvalues and annealing rates*, Math. Oper. Res., to appear.
- [2] ———, *On the convergence rate of a special class of annealing processes*, Soochow J. Math., 13 (1987), pp. 23–30.
- [3] S. GEMAN AND D. GEMAN, *Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Mach. Intelligence, 6 (1984), pp. 721–741.
- [4] B. GIDAS, *Global optimization via the Langevin equation*, in Proc. 24th IEEE Conference on Decision and Control, Fort Lauderdale, FL, 1985.
- [5] B. HAJEK, *Cooling schedules for optimal annealing*, preprint.
- [6] C. R. HWANG AND S. J. SHEU, *Large time behaviors of perturbed diffusion Markov processes with applications III: Simulated annealing*, preprint.
- [7] S. KIRKPATRICK, C. GEBATT, AND M. VECCHI, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671–680.
- [8] E. SENETA, *Nonnegative Matrices*, John Wiley, New York, 1973.
- [9] A. D. VENTCEL, *On the asymptotics of eigenvalues of matrices with elements of order $\exp(-V_{ij}/(2\varepsilon^2))$* , Dokl. Akad. Nauk SSSR, 202 (1972), pp. 65–68.

HIGHER-ORDER CONDITIONS FOR CONICAL CONTROLLABILITY*

J. WARGA†

Abstract. Let Q be a convex subset of a linear space, $\mathcal{U} \subset Q$, \mathcal{X} a topological vector space, $C \subset \mathcal{X}$ convex with a nonempty interior, $\bar{q} \in Q$, $\varphi = (\varphi_1, \varphi_2): Q \rightarrow \mathbb{R}^m \times \mathcal{X}$, and $\varphi_2(\bar{q}) \in C$. Also, let $\mathcal{A} := \{u \in \mathcal{U} \mid \varphi_2(u) \in C\}$. This framework can be used to model many optimization problems including control problems with unilateral (state inequality) constraints. Sufficient conditions are derived for $\varphi_1(\mathcal{A})$ to cover the interior of a convex set containing $\varphi_1(\bar{q})$ and an open neighborhood of a point $\varphi_1(\bar{q}) + w$ for $w \in \mathbb{R}^m$. These sufficient conditions can be used to provide information about the attainable sets of control problems and, in some cases, to show that $\varphi_1(\mathcal{A})$ contains a neighborhood of $\varphi_1(\bar{q})$. Several illustrative examples are also discussed.

Key words. controllability, conical controllability, sufficient conditions, inclusion restrictions, equality restrictions, attainable sets

AMS(MOS) subject classifications. 49B27, 49E15, 49E30

1. Introduction. Let \mathcal{X} be a real vector space, Q a convex subset of \mathcal{X} , $\bar{q} \in Q$, and $\varphi_1: Q \rightarrow \mathbb{R}^m$. We wish to study the controllability of the function φ_1 at \bar{q} which, in its simplest form, concerns the image $\varphi_1(Q)$ in the vicinity of $\varphi_1(\bar{q})$. Typical of such problems is the investigation of the attainable set of a control system

$$\dot{x}(t) = f(t, x(t), u(t)) \quad \text{a.e. in } [0, 1], \quad x(0) = 0,$$

where the control functions u are selected from a convex set Q , $\varphi_1(u) = x(u)(1)$, and $t \rightarrow x(t) = x(u)(t)$ is the absolutely continuous solution of the system. In more sophisticated versions of the problem, we study the set $\varphi_1(\mathcal{A})$ in the vicinity of $\varphi_1(\bar{q})$, where

$$\mathcal{A} := \{u \in \mathcal{U} \mid \varphi_2(u) \in C\},$$

\mathcal{U} is an “abundant” subset of Q , $\varphi_2: Q \rightarrow \mathcal{X}$, \mathcal{X} is a topological vector space, C is a convex subset of \mathcal{X} with a nonempty interior, and $\varphi_2(\bar{q}) \in C$. Such problems are exemplified by unilateral control systems in which the restriction $\varphi_2(u) \in C$ models a state inclusion restriction such as $x(t) \in A$ for all $t \in [0, 1]$, Q may represent the set of relaxed controls, and \mathcal{U} a set of original controls closed under finite measurable concatenations.

If the set $\varphi_1(\mathcal{A})$ covers a nhd (a set containing an open neighborhood) of $\varphi_1(\bar{q})$ then we say that φ_1 is *controllable at \bar{q}* (with respect to given \mathcal{U} , φ_2 , and C). The problem of controllability is closely related to that of optimization, say, of minimizing φ_1^1 on the set

$$\{u \in \mathcal{A} \mid \varphi_1^i(u) = 0 \quad \text{for } i = 2, \dots, m\},$$

where $\varphi_1 = (\varphi_1^1, \dots, \varphi_1^m)$. Clearly, if φ_1 is controllable at \bar{q} , then \bar{q} is ruled out as a candidate for a restricted minimum. We refer the reader to [4, pp. 715, 716] for some further comments on the relation of higher-order sufficient conditions for controllability to necessary conditions for optimality, and on the role that the presence of Lagrange coefficients plays in the former. We should also mention Bernstein’s paper [1], which has influenced our recent work in this area.

* Received by the editors September 30, 1987; accepted for publication (in revised form) February 29, 1988. This research was partially supported by National Science Foundation grant DMS-8619002.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.

In [4] we formulated some sufficient higher-order conditions for controllability. In the present paper we shall use similar arguments to establish sufficient conditions for *conical controllability*, that is, for $\varphi_1(\mathcal{A})$ to cover the interior of a convex set containing $\varphi_1(\bar{q})$ and some nhd of a point $\varphi_1(\bar{q}) + w$. If these sufficient conditions apply to $w = (-\varepsilon, 0, \dots, 0)$ for some $\varepsilon > 0$, then this alone rules out \bar{q} as a candidate for minimum. If these sufficient conditions apply to $w = \gamma_z z$ for all $z \in \mathbb{R}^m$ with $|z| = 1$ and some corresponding $\gamma_z > 0$ then this implies that φ_1 is “nearly” controllable at \bar{q} (and we shall provide a simple example in which controllability at \bar{q} can be established in this manner while the results of [4] are inapplicable). However, even when conical controllability does not imply controllability, it may still shed some light on the structure of $\varphi_1(\mathcal{A})$ and, in particular, of the attainable sets in unilateral control problems in the vicinity of a point determined by an extremal.

The present p th-order controllability theorem (as well as those of [4]) can also be viewed as a form of a stability theorem under perturbations. In particular, it provides criteria for functions on restricted domains with polynomial components (p th order Taylor approximations) to cover certain conical nhds even when perturbed by arbitrary higher-order terms (the size of the nhds obviously depends on the perturbations).

Our main results are presented in § 2. Section 3 contains some examples, and § 4 contains the proofs.

2. Sufficient conditions for conical controllability. Let

$$\begin{aligned}\varphi &:= (\varphi_1, \varphi_2), \\ \mathcal{T}_k &:= \left\{ (\theta_1, \dots, \theta_k) \in \mathbb{R}^k \mid \theta_j \geq 0, \sum_{j=1}^k \theta_j \leq 1 \right\}, \\ Y &:= \{y_1, \dots, y_k\} \subset Q - \bar{q}.\end{aligned}$$

We shall say that φ has a p th order Taylor approximation at \bar{q} with respect to Y if the function

$$(2.0.1) \quad \theta \rightarrow \psi(\theta; Y) := \psi(\theta) := \varphi \left(\bar{q} + \sum_{j=1}^k \theta_j y_j \right) : \mathcal{T}_k \rightarrow \mathbb{R}^m \times \mathcal{X}$$

is continuous and admits a p th-order Taylor approximation at zero, i.e., there exists, for each $n = 1, \dots, p$, an n -linear symmetric operator $\psi^{(n)}(0) : (\mathbb{R}^k)^n \rightarrow \mathbb{R}^m \times \mathcal{X}$ such that

$$\lim_{\theta \rightarrow 0} |\theta|^{-p} \left\{ \psi(\theta) - \left[\psi(0) + \psi^{(1)}(0)\theta + \dots + \frac{1}{p!} \psi^{(p)}(0)\theta^p \right] \right\} = 0 \quad \text{as } \theta \rightarrow 0, \quad \theta \in \mathcal{T}_k \setminus \{0\}.$$

Condition 2.1.1. Let $P \subset \mathbb{R}^s$ and $\hat{q} : P \rightarrow Q$. We shall say that \hat{q} satisfies Condition 2.1.1 if, for every choice of $p \in P$, there exists a sequence $(u_n(p))$ in \mathcal{U} such that

$$\lim_n \varphi(u_n(p)) = \varphi(\hat{q}(p)) \quad \text{uniformly for } p \in P$$

and

$$p \rightarrow \varphi(\hat{q}(p)) \text{ and } p \rightarrow \varphi(u_n(p)) \text{ are continuous for each } n = 1, 2, \dots.$$

We shall say that $Y := \{y_1, \dots, y_k\} \subset Q - \bar{q}$ satisfies Condition 2.1.1 if the function

$$\theta \rightarrow \bar{q} + \sum_{j=1}^k \theta_j y_j : \mathcal{T}_k \rightarrow Q$$

satisfies Condition 2.1.1.

Remark. We observe that $Y = \{y_1, \dots, y_k\}$ satisfies Condition 2.1.1 in the following three cases of particular interest:

(a) Q is the set of relaxed controls (with its weak norm topology), \mathcal{U} is any set of ordinary controls that is closed under finite measurable concatenations [3, Thm. IV. 3.9, p. 285], and φ is continuous.

(b) $\mathcal{U} = Q$ and φ has a p th-order Taylor approximation at \bar{q} with respect to Y . (Then we can choose $u_n(\theta) = \hat{q}(\theta) = \bar{q} + \sum \theta_j y_j$.)

(c) Q is a convex subset of a Banach space with a nonempty interior Q° , $\mathcal{U} \supset Q^\circ$, and φ is continuous. (Then for any $n \in \{1, 2, \dots\}$ we can determine

$$\bar{q}^n \in Q^\circ \subset \mathcal{U} \quad \text{and} \quad y_j^n \in Q^\circ - \bar{q} \subset \mathcal{U} - \bar{q}$$

such that

$$|\bar{q}^n - \bar{q}| < \frac{1}{n}, \quad |y_j^n - y_j| < \frac{1}{n} \quad \text{for } j = 1, \dots, k.$$

Condition 2.1.1 is then satisfied by

$$u_n(\theta) = \bar{q}^n + \sum_{j=1}^k \theta_j y_j^n \in Q^\circ \subset \mathcal{U}.$$

DEFINITION 2.1.2. We shall say that φ_1 is *w-conically controllable at \bar{q}* (with respect to \mathcal{U} , φ_2 , and C , which are assumed given) if $w \in \mathbb{R}^m$ and there exist $\gamma_0 > 0$ and nhds G_1 , G_2 of zero in \mathbb{R}^m , \mathcal{X} such that

$$(2.1.2.1) \quad \varphi_1(\bar{q}) + \gamma(w + G_1) \subset \{\varphi_1(u) \mid u \in \mathcal{U}, \varphi_2(u) + \gamma G_2 \subset C\} \quad \forall \gamma \in (0, \gamma_0].$$

We shall write \hat{C} for $C - \varphi_2(\bar{q})$, $B(x, r)$, respectively, $\bar{B}(x, r)$ for the open, respectively, closed ball of center x and radius r , and A° , respectively, ∂A for the interior, respectively, boundary, of A . Our general higher-order sufficient conditions for conical controllability are stated in Theorem 2.2 below (itself a generalization of [4, Thm. 2.2, p. 717]) from which we also derive more specific first- and second-order conditions.

THEOREM 2.2. Let $H \subset \mathbb{R}^k$ be open, $\alpha_1 > 0$, $p \in \{1, 2, \dots\}$, $\bar{x} \in H$, $w \in \mathbb{R}^m$, and let the functions $f^n := (f_1^n, f_2^n): H \rightarrow \mathbb{R}^m \times \mathcal{X}$ be continuous. Assume that:

(a) $f^n(x) \in \{0\} \times \hat{C}$ for $n = 1, 2, \dots, p-1$, $x \in H$;

(b) $f^p(\bar{x}) \in \{w\} \times \hat{C}^\circ$;

(c) There exists a function $(x, \alpha) \rightarrow \hat{q}(x, \alpha): H \times [0, \alpha_1] \rightarrow Q$ satisfying Condition 2.1.1 and such that

$$\varphi(\hat{q}(x, \alpha)) = \varphi(\bar{q}) + \sum_{n=1}^p \frac{\alpha^n}{n!} f^n(x) + \alpha^p d(x, \alpha),$$

where $\lim_{\alpha \rightarrow 0+} d(x, \alpha) = 0$ uniformly for $x \in H$, and either

(d) f_1^p is continuously differentiable and the set $\{\partial f_1^p(\bar{x})/\partial x_j \mid j = 1, \dots, k\}$ spans \mathbb{R}^m , or

(d') there exist a nhd V of 0 in \mathbb{R}^m and a continuous $a := (a_1, \dots, a_k): V \rightarrow H$ such that the function $\theta \rightarrow f_1^p(a(\theta)): V \rightarrow f_1^p(a(V))$ is a homeomorphism and $a(0) = \bar{x}$.

Then φ_1 is *w-conically controllable at \bar{q}* .

COROLLARY 2.3 [4, Thm. 2.2, p. 717]. If there exists $\bar{x} \in H$ satisfying the assumptions of Theorem 2.2 for $w = 0$, then there exist nhds G_1 , G_2 of 0 in \mathbb{R}^m , \mathcal{X} such that

$$\varphi_1(\bar{q}) + G_1 \subset \{\varphi_1(u) \mid u \in \mathcal{U}, \varphi_2(u) + G_2 \subset C\}.$$

Proof. This follows directly from Theorem 2.2 if we set $\gamma = \gamma_0$ in relation (2.1.2.1) and replace $\gamma_0 G_1$, $\gamma_0 G_2$ with G_1 , G_2 .

We next observe that conical controllability sometimes implies controllability.

THEOREM 2.4. Assume that for each $z \in \mathbb{R}^m$ with $|z| = 1$ there exist $\gamma_z > 0$ and $q_z \in Q$ such that $\varphi(q_z) = \varphi(\bar{q})$ and φ_1 is w -conically controllable at q_z for $w = \gamma_z z$. Then there exists a nhd G_1 of O in \mathbb{R}^m such that

$$(2.4.1) \quad \varphi_1(\bar{q}) + G_1 \subset \{\varphi_1(u) \mid u \in \mathcal{U}, \varphi_2(u) \in C\} \cup \{\varphi_1(\bar{q})\}.$$

In particular, if $\bar{q} \in \mathcal{U}$ then φ_1 is controllable.

Proof. This is a direct consequence of the compactness of $\partial B(0, 1)$ in \mathbb{R}^m .

Remark. Relation (2.4.1) cannot be improved upon by eliminating the singlet $\{\varphi_1(\bar{q})\}$ from its right side. This is shown by Example 5 in § 3. Example 3 shows that Theorem 2.4 is sometimes applicable in situations when Corollary 2.3 (which is equivalent to [4, Thm. 2.2, p. 717]) does not hold.

As an application of Theorem 2.2 we obtain the first- and second-order conditions below. In a similar manner we can derive a generalization and a slight simplification of the third-order conditions of Theorem 2.5 of [4, p. 718]. (This can be obtained by replacing the right side of relation (d) in Theorem 2.5 of [4] by $\{w\} \times \hat{C}^0$ and the right side of the last display of that theorem by $f_1^3(a(\theta), b(\theta), c(\theta))$.)

We define the function ψ as in (2.0.1) and denote by $\mathbf{e}_1, \dots, \mathbf{e}_k$ the columns of the unit $k \times k$ matrix.

THEOREM 2.5 (First-order conditions). Let $Y := \{y_i \mid i = 1, \dots, k\} \subset Q - \bar{q}$ and $w \in \mathbb{R}^m$. Assume that Y satisfies Condition 2.1.1, φ has a first-order Taylor approximation at \bar{q} with respect to Y , and

- (a) $\psi'(0) \sum_i \mathbf{e}_i \in \{w\} \times \hat{C}^0$;
- (b) the set $\{\psi'_i(0)\mathbf{e}_i \mid i = 1, 2, \dots, k\}$ spans \mathbb{R}^m .

Then φ is w -conically controllable at \bar{q} .

THEOREM 2.6 (Second-order conditions). Let

$$I := \{1, 2, \dots, i_1\}, \quad J := \{i_1 + 1, \dots, k\},$$

$$Y := \{y_i \mid i = 1, \dots, k\} \subset Q - \bar{q}, \quad w \in \mathbb{R}^m$$

and assume that Y satisfies Condition 2.1.1, φ has a second-order Taylor approximation at \bar{q} with respect to Y , and, for all $i \in I$ and $j \in J$,

- (a) $\psi'(0)\mathbf{e}_i \in \{0\} \times \hat{C}$,
- (b) $\psi''(0)(\sum_{i \in I} \mathbf{e}_i)^2 + 2\psi'(0) \sum_{j \in J} \mathbf{e}_j \in \{w\} \times \hat{C}^0$,

and either

- (c) the vectors

$$\psi''_1(0) \left(\sum_{i \in I} \mathbf{e}_i \right) \mathbf{e}_s, \quad \psi'_1(0)\mathbf{e}_j \quad \text{for } s \in I, \quad j \in J$$

span \mathbb{R}^m , or

(c') there exists a nhd V of O in \mathbb{R}^m and continuous $a_i, b_j: V \rightarrow (0, \infty)$ such that the function

$$\theta := (\theta_1, \dots, \theta_m) \rightarrow \psi''_1(0) \left(\sum_{i \in I} a_i(\theta) \mathbf{e}_i \right)^2 + 2\psi'_1(0) \sum_{j \in J} b_j(\theta) \mathbf{e}_j$$

is a homeomorphism and $a_i(0) = b_j(0) = 1$ for i and j .

Then φ_1 is w -conically controllable at \bar{q} .

3. Examples. In all the examples below we shall assume that $\mathcal{X} = C = \mathbb{R}$ and $\varphi_2(q) = 0$ for $q \in Q$. This enables us to neglect the restriction $\varphi_2(q) \in C$, which is then automatically satisfied for all $q \in Q$, as are the other restrictions in Theorems 2.2–2.6 in the space \mathcal{X} . For simplicity of notation we shall write φ for φ_1 .

The first two examples deal with control problems, Example 1 involving only original (i.e., ordinary) controls while Example 2 involves both original and relaxed controls.

Example 1. Let Q be the set of all Lebesgue measurable functions $u: [0, 1] \rightarrow [-1, 1]$, $\mathcal{U} = Q$, and $h: \mathbb{R}^2 \rightarrow \mathbb{R}$ be bounded and C^2 , with $h(0) > 0$. Let $\bar{q} = 0 \in Q$, and let $\varphi: Q \rightarrow \mathbb{R}^2$ be defined by $\varphi(u) = x(u)(1)$, where $t \rightarrow x(u)(t) = x(t) = (x_1, x_2)(t)$ is the absolutely continuous solution of the differential equation

$$\begin{aligned}\dot{x}_1 &= x_2^2 h(x), \quad \dot{x}_2 = u \quad \text{a.e. in } [0, 1], \\ x(0) &= (0, 0).\end{aligned}$$

(It is easy to see that \bar{q} is an extremal of this control problem.)

We verify that, for $y, z \in Q = Q - \bar{q}$, the function $\alpha \rightarrow \sigma(\alpha) := \varphi(\alpha y + \alpha^2 z)$ for sufficiently small $|\alpha|$ is C^2 and we have

$$\varphi(\alpha y + \alpha^2 z) = \alpha f^1 + \frac{1}{2} \alpha^2 f^2 + \alpha^2 o(1) = \alpha \psi'(0) \mathbf{e}_1 + \frac{1}{2} \alpha^2 [\psi''(0) \mathbf{e}_1^2 + 2\psi'(0) \mathbf{e}_2] + \alpha^2 o(1)$$

where

$$f^i = (f_1^i, f_2^i) = d^i \sigma / d\alpha|_{\alpha=0} \quad \text{and} \quad \psi(\theta_1, \theta_2) = \varphi(\theta_1 y + \theta_2 z).$$

Thus, setting

$$Y(t) := \int_0^t y(s) \, ds,$$

we have

$$f_1^1 = 0, \quad f_2^1 = Y(1).$$

$$f_1^2 = 2h(0) \int_0^1 [Y(s)]^2 \, ds, \quad f_2^2 = 2 \int_0^1 z(s) \, ds,$$

$$\psi'(0) \mathbf{e}_1 = f^1, \quad \psi''(0) \mathbf{e}_1^2 = (f_1^2, 0), \quad \psi'(0) \mathbf{e}_2 = \left(0, \int_0^1 z(s) \, ds\right).$$

The assumptions of Theorem 2.6 will be satisfied for a point $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2) \in \mathbb{R}^2$ if we can choose corresponding $y, z \in Q$ such that

$$Y(1) = 0, \quad \int_0^1 [Y(s)]^2 \, ds = \frac{w_1}{2h(0)}, \quad \int_0^1 z(s) \, ds = \frac{1}{2} w_2$$

and the vectors

$$\left(\int_0^1 [Y(s)]^2 \, ds, 0 \right), \quad \left(0, \int_0^1 z(s) \, ds \right)$$

span \mathbb{R}^2 . We can do this when

$$0 < w_1 < h(0)/6, \quad 0 < |w_2| < 2$$

by letting

$$\bar{y}(t) = \begin{cases} 1 & \text{for } 0 \leq t < \frac{1}{2}, \\ -1 & \text{for } \frac{1}{2} \leq t \leq 1, \end{cases} \quad \bar{z}(t) = 1 \quad \text{for } t \in [0, 1]$$

and choosing

$$y = [6w_1/h(0)]^{1/2} \bar{y}, \quad z = \frac{1}{2} w_2 \bar{z}.$$

Example 2. We define \mathcal{U} , as in Example 1, as the set of all Lebesgue measurable functions $u: [0, 1] \rightarrow [-1, 1]$, but Q as the corresponding set of relaxed controls (i.e., measurable functions $q: [0, 1] \rightarrow \text{rpm}([-1, 1])$, where $\text{rpm}([-1, 1])$ is the set of Radon probability measures on $[-1, 1]$ with the weak-star topology of $C([-1, 1])^*$; see [3, Chap. IV]). We assume given a bounded C^2 function $h: \mathbb{R}^2 \rightarrow \mathbb{R}$ with $h(0) > 0$, let $\bar{q}(t) := \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1}$, where δ_r is the Dirac measure at r , and define $\varphi: Q \rightarrow \mathbb{R}^2$ by $\varphi(q) = x(q)(1)$, where $t \rightarrow x(q)(t) = x(t) = (x_1, x_2)(t)$ is the absolutely continuous solution of the differential equation

$$\begin{aligned} \dot{x}_1 &= x_2^2[h(x) + 1 - \int r^2 q(t)(dr)], \quad \dot{x}_2 = \int r q(t)(dr) \quad \text{a.e.} \\ x(0) &= (0, 0). \end{aligned}$$

(Again, we observe that \bar{q} is an extremal of this relaxed control problem.)

The arguments of Example 1 will apply to this problem, with f^i, φ', φ'' satisfying the same formulas in which $y(t), z(t), Y(t)$ are replaced by

$$\int ry(t)(dr), \quad \int rz(t)(dr), \quad \int_0^t ds \int ry(s)(dr)$$

for $y, z \in Q - \bar{q}$, and with $\bar{y}(t), \bar{z}(t)$ now defined by

$$\bar{y}(t) = \begin{cases} \delta_1 - \bar{q} = \frac{1}{2}\delta_1 - \frac{1}{2}\delta_{-1} & \text{for } 0 \leq t < \frac{1}{2}, \\ \delta_{-1} - \bar{q} = \frac{1}{2}\delta_{-1} - \frac{1}{2}\delta_1 & \text{for } \frac{1}{2} \leq t \leq 1, \end{cases} \quad \bar{z}(t) = \delta_1 - \bar{q} = \frac{1}{2}\delta_1 - \frac{1}{2}\delta_{-1} \quad \text{for all } t.$$

These arguments thus show, using Theorem 2.6 and Remark (a) following Condition 2.1.1, that φ is w -conically controllable at \bar{q} for

$$0 < w_1 < h(0)/6, \quad 0 < |w_2| < 2.$$

(This implies that, for each such w , there exist $\gamma_0 > 0$ and a nhd G_1 of O in \mathbb{R}^2 such that $\gamma(w + G_1) \subset \varphi(\mathcal{U})$ for $0 < \gamma \leq \gamma_0$, where \mathcal{U} is the set of ordinary control functions.)

In the remaining Examples 3–6 we consider $Q \subset \mathcal{X} = \mathbb{R}^s$ for some positive integer s , $\varphi = g + o(|q|^p)$, where the components g_1, \dots, g_m of g are homogeneous polynomials of degree p , $\bar{q} = 0$, and $\hat{q}(x, \alpha) = \alpha x$ for $x \in H$.

Example 3 (An application of Theorem 2.4). Let $\mathcal{U} = Q = \mathbb{C}$ be the complex plane identified with \mathbb{R}^2 , $p \in \{2, 3, \dots\}$, and $g(z) = (1/p!)z^p$. Then $\varphi(\alpha x) = \alpha^p x^p + \alpha^p o(1)$, where $o(1) \rightarrow 0$ as $\alpha \rightarrow 0+$ uniformly for all x in a bounded set and, in the notation of Theorem 2.2, $f^n = 0$ for $n < p$, $f^p(x) = x^p$. We set $H = B(0, 2) \subset \mathbb{C}$ and, for an arbitrary $w \in \mathbb{C}$ with $|w| = 1$, we choose $\bar{x} = x_w \in H$ such that $x_w^p = w$. Then, for $x = x_1 + ix_2$, we have

$$\begin{aligned} \frac{\partial f^p(\bar{x})}{\partial x_1} &= p\bar{x}^{p-1}, & \frac{\partial f^p(\bar{x})}{\partial x_2} &= ip\bar{x}^{p-1}, \\ \det(f^p)'(\bar{x}) &= p^2|\bar{x}|^{2p-2} \neq 0. \end{aligned}$$

It follows, by Theorems 2.2 and 2.4, that φ is an open mapping at O .

On the other hand, to be applicable the results of [4, Thm. 2.2] (the simplified version of which is Corollary 2.3 above) would require the existence of a continuous mapping $\theta \rightarrow a(\theta)$ of a nhd of O in \mathbb{C} that maps O into O and such that $\theta \rightarrow a(\theta)^p$ is a homeomorphism. I am indebted to Terence J. Gaffney [2] for providing a short demonstration, based on homology arguments, that such a mapping cannot exist.

Example 4 (w -conical controllability “nearly everywhere”). Let $\mathcal{U} = Q = \mathbb{R}^2 = \mathbb{C}$, $p = 8$, and $g: \mathbb{C} \rightarrow \mathbb{C}$ be defined by $g(x_1, x_2) = g(x_1 + ix_2) = (1/8!)(x_1^2 + ix_2^2)^4$. For any $w \in \mathbb{C}$ there exists $\bar{x} = \bar{x}_1 + i\bar{x}_2$ such that $g(\bar{x}) = w$. Furthermore, $\varphi(\alpha x) = \alpha^8 f^8(x) + \alpha^8 o(1)$ (where $f^8 = g$) and

$$\frac{\partial f^8(x)}{\partial x_1} = 8(x_1^2 + ix_2^2)^3 x_1, \quad \frac{\partial f^8(x)}{\partial x_2} = 8(x_1^2 + ix_2^2)^3 ix_2;$$

hence $\det (f^8)'(x) = 64x_1x_2(x_1^4 + x_2^4)^3$. Thus the vectors $\partial f^8(x)/\partial x_j$ span \mathbb{R}^2 if and only if $x_1x_2 \neq 0$, and Theorem 2.2 shows that φ is w -conically controllable at O if $w = (w_1, w_2) = (x_1^2 + ix_2^2)^4$ for $x_1x_2 \neq 0$, i.e., if $w \notin [0, \infty) + 0i$.

In fact, this result is the best possible because $x \rightarrow \varphi(x_1, x_2) = (x_1^2 + ix_2^2)^4 - (x_1^2 + ix_2^2)^5$ is not an open mapping. (Its image does not contain a set of the form $\{(w_1, w_2) \mid h(w_1) < w_2 < 0, w_1 > 0\}$, where $\lim_{w_1 \rightarrow 0+} h(w_1) = 0$.)

For $\mathcal{U} = Q = [0, \infty)^2 \subset \mathbb{R}^2 = \mathbb{C}$, $p = 4$, and $g(x_1, x_2) = (x_1 + ix_2)^4$, the conclusions are similar.

Example 5 (Counterexample related to Theorem 2.4). This example is intended to show that relation (2.4.1) cannot be improved by deleting the set $\{\varphi_1(\bar{q})\}$ on the right side of the inclusion.

Let $Q = [0, 1]^2 \subset \mathbb{R}^2 = \mathbb{C}$, $\mathcal{U} = Q \setminus \{(0, 0)\}$, $p = 5$, and $\varphi(x) = g(x) = (1/5!)x^5$. We observe that $g(Q)$ contains an nhd of O and, since $\mathcal{U} \supset Q^\circ$, Condition 2.1.1 is satisfied. We have $\varphi(\alpha x) = \alpha^5 x^5$; hence $f^5(x) = x^5$ and (as in Example 3) $\det (f^5)'(x) = 25|x|^8$. We set $H = (0, 1)^2 \subset Q^\circ$. For an arbitrary $w \in \mathbb{C}$ with $|w| = \frac{1}{2}$ we can choose a fifth root $\bar{x} = x_w$ of w such that $\bar{x} \in H$. Since $\det (f^5)'(\bar{x}) \neq 0$, the assumptions of Theorem 2.2 and of Theorem 2.4 are satisfied. Therefore there exists a nhd G of O in \mathbb{C} such that $G \subset \varphi(\mathcal{U}) \cup \varphi(0)$. Since $x^5 = 0$ only for $x = 0$, $\varphi(\mathcal{U})$ does not contain G .

Example 6. This example demonstrates that the type of situation described in Example 4 may occur even for quadratic mappings.

Let $H = B(0, 1)$,

$$\begin{aligned} \mathcal{U} = Q = \mathbb{R}^3, \quad p = 2, \quad m = 3, \quad g = (g_1, g_2, g_3), \quad x = (x_1, x_2, x_3), \\ g_1(x) = x_1^2 - x_2^2 + x_3^2, \quad g_2(x) = x_2(x_1 + x_2), \quad g_3(x) = x_2x_3. \end{aligned}$$

We can easily verify that

$$g(\mathbb{R}^3) = \{w = (w_1, w_2, w_3) \mid 2w_2 + w_1 > 0\} \cup \{(0, 0, 0)\}$$

and

$$\det g'(x) = 2x_2[(x_1 + x_2)^2 + x_3^2].$$

Thus, by Theorem 2.2, φ is w -conically controllable at O for all w such that $2w_2 + w_1 > 0$ and $(w_2, w_3) \neq (0, 0)$.

We observe that $\gamma(1, 0, 0) \in \text{Interior } g(\mathbb{R}^3)$ for $\gamma > 0$ but $g(x) = \gamma(1, 0, 0)$ implies $\det g'(x) = 0$. Thus it remains an open question whether the open interval $(0, \gamma) \times \{(0, 0)\}$ is contained in $\varphi(\mathbb{R}^3)$ for all $\varphi(x) = g(x) + o(|x|^2)$ and some corresponding $\gamma = \gamma_\varphi > 0$.

4. Proofs.

LEMMA 4.1. Let V be a nhd of O in \mathbb{R}^m , $w \in \mathbb{R}^m$, $\alpha_1 > 0$ and, for $n = 1, 2, \dots$,

$$\begin{aligned} g &= (g_1, g_2): V \rightarrow \mathbb{R}^m \times \mathcal{X}, \\ e &= (e_1, e_2): V \times [0, \alpha_1] \rightarrow \mathbb{R}^m \times \mathcal{X}, \\ h^n &= (h_1^n, h_2^n): V \times [0, \alpha_1] \rightarrow \mathbb{R}^m \times \mathcal{X}. \end{aligned}$$

Assume that $g_1: V \rightarrow g_1(V)$ is a homeomorphism, that g_2 , $e(\cdot, \alpha)$ and $h^n(\cdot, \alpha)$ are continuous for each α and n , and that

$$g(0) \in \{w\} \times \hat{C}^\circ, \quad \lim_{\alpha \rightarrow 0+} e(x, \alpha) = 0 \quad \text{uniformly for } x \in V,$$

$$\forall \alpha > 0 \quad \lim_{n \rightarrow \infty} h^n(x, \alpha) = 0 \quad \text{uniformly for } x \in V.$$

Then there exist $\alpha_0 \in (0, \log 2]$, nhds V_1, G_1, G_2 of O in $\mathbb{R}^m, \mathbb{R}^m, \mathcal{X}$, and $\alpha \rightarrow N(\alpha): (0, \alpha_0] \rightarrow \{1, 2, \dots\}$ such that, for all $\alpha \in (0, \alpha_0]$,

$$w + G_1 \subset \{g_1(x) + e_1(x, \alpha) + h_1^{N(\alpha)}(x, \alpha) \mid x \in V_1\},$$

$$g_2(x) + e_2(x, \alpha) + h_2^{N(\alpha)}(x, \alpha) + G_2 \subset \hat{C}^\circ \quad \forall x \in V_1.$$

Proof. Let $r, s > 0$, and a nhd G_2 of O in \mathcal{X} be such that

$$\bar{B}(0, r) \subset V, \quad g_2(\bar{B}(0, r)) + G_2 + G_2 + G_2 \subset \hat{C}^\circ, \quad \bar{B}(w, s) \subset g_1(\bar{B}(0, r)).$$

Let $\alpha_0 \in (0, \min(\alpha_1, \log 2)]$ be sufficiently small so that

$$|e_1(x, \alpha)| \leq \frac{1}{3}s, \quad e_2(x, \alpha) \in G_2 \quad \forall x \in V, \quad \alpha \in (0, \alpha_0].$$

Finally, for each $\alpha \in [0, \alpha_0]$, let $N(\alpha)$ be sufficiently large so that

$$|h_1^{N(\alpha)}(x, \alpha)| \leq \frac{1}{3}s, \quad h_2^{N(\alpha)}(x, \alpha) \in G_2 \quad \forall x \in V.$$

Next we choose an arbitrary $u \in \mathbb{R}^m$ with $|u| \leq \frac{1}{3}s$. The equation

$$(1) \quad g_1(x) + e_1(x, \alpha) + h_1^{N(\alpha)}(x, \alpha) = u + w$$

is equivalent to

$$(2) \quad x = g_1^{-1}(u + w - e_1(x, \alpha) - h_1^{N(\alpha)}(x, \alpha)).$$

Since the right-hand side of (2) is a continuous function of x mapping $\bar{B}(0, r)$ into itself, there exists a fixed point $\bar{x} \in \bar{B}(0, r)$ that must clearly satisfy (1). Thus

$$w + G_1 := \bar{B}(w, \frac{1}{3}s) \subset \{g_1(x) + e_1(x, \alpha) + h_1^{N(\alpha)}(x, \alpha) \mid x \in \bar{B}(0, r)\} \quad \forall \alpha \in (0, \alpha_0].$$

Furthermore, for each $x \in \bar{B}(0, r)$ and $\alpha \in (0, \alpha_0]$, we have

$$g_2(x) + e_2(x, \alpha) + h_2^{N(\alpha)}(x, \alpha) + G_2 \subset g_2(x) + G_2 + G_2 + G_2 \subset \hat{C}^\circ.$$

Thus the lemma holds with $V_1 = \bar{B}(0, r)$. \square

Proof of Theorem 2.2. We shall first assume that assumption (d) holds. Let e_1, \dots, e_m be a base in \mathbb{R}^m . By assumption (d), there exist numbers $a_{\mu j}$ for $\mu = 1, \dots, m$, $j = 1, \dots, k$ such that

$$e_\mu = \sum_{j=1}^k a_{\mu j} \frac{\partial f^p(\bar{x})}{\partial x_j}.$$

Let

$$\bar{x} := (\bar{x}_1, \dots, \bar{x}_k), \quad \theta := (\theta_1, \dots, \theta_m),$$

$$a_j(\theta) := \bar{x}_j + \sum_{\mu} a_{\mu j} \theta_\mu, \quad a(\theta) := (a_1, \dots, a_k)(\theta),$$

and let \hat{V} be a sufficiently small nhd of O in \mathbb{R}^m so that $a(\theta) \in H$ for $\theta \in \hat{V}$. We set

$$g(\theta) := (g_1, g_2)(\theta) := f^p(a(\theta)) \quad \forall \theta \in \hat{V}$$

and observe that

$$\frac{\partial g_1(0)}{\partial \theta_\mu} = \sum_{j=1}^k a_{\mu j} \frac{\partial f_1^p(\bar{x})}{\partial x_j} = e_\mu.$$

Thus the matrix $g'_1(0)$ is invertible and, by the inverse function theorem, $\theta \rightarrow g_1(\theta)$ is a homeomorphism of some nhd V of O in \mathbb{R}^m onto $g_1(V)$. This shows that assumption (d) implies (d') so that we may assume in all cases that (d') is valid.

Condition 2.1.1 implies that, for all $(\theta, \alpha) \in V \times (0, \alpha_1]$, there exists a sequence $(u_n(\theta, \alpha))$ in \mathcal{U} such that

$$\lim_n \varphi(u_n(\theta, \alpha)) = \varphi(\hat{q}(a(\theta), \alpha)) \quad \text{uniformly for all } (\theta, \alpha) \in V \times (0, \alpha_1],$$

and $(\theta, \alpha) \rightarrow \phi(u_n(\theta, \alpha))$ and $(\theta, \alpha) \rightarrow \varphi(\hat{q}(a(\theta), \alpha))$ are continuous for each n . For $(\theta, \alpha) \in V \times (0, \alpha_1]$, we set

$$\begin{aligned} g(\theta) &:= f^p(a(\theta)), \\ h^n(\theta, \alpha) &:= p! \alpha^{-p} [\varphi(u_n(\theta, \alpha)) - \varphi(\hat{q}(a(\theta), \alpha))], \\ e(\theta, \alpha) &:= (e_1, e_2)(\theta, \alpha) := p! d(a(\theta), \alpha), \end{aligned}$$

and observe that g , e , and h^n satisfy the assumptions of Lemma 4.1. It follows that there exist

$$\alpha_0 \in (0, \log 2], \quad \alpha \rightarrow N(\alpha) : (0, \alpha_0] \rightarrow \{1, 2, 3, \dots\},$$

and nhds V_1, G_1, G_2 of O in $\mathbb{R}^m, \mathbb{R}^m, \mathcal{X}$ such that, for all $\alpha \in (0, \alpha_0]$,

$$\begin{aligned} (1) \quad w + G_1 &\subset \{g_1(\theta) + e_1(\theta, \alpha) + h_1^{N(\alpha)}(\theta, \alpha) \mid \theta \in V_1\}, \\ (2) \quad g_2(\theta) + e_2(\theta, \alpha) + h_2^{N(\alpha)}(\theta, \alpha) + G_2 &\subset \hat{C}^\circ \quad \forall \theta \in V_1. \end{aligned}$$

Relation (2) yields

$$p! \alpha^{-p} \left[\varphi_2(u_{N(\alpha)}(\theta, \alpha)) - \varphi_2(\bar{q}) - \sum_{n=1}^{p-1} \frac{1}{n!} \alpha^n f_2^n(a(\theta)) + \frac{\alpha^p}{p!} G_2 \right] \subset \hat{C}^\circ \quad \forall \theta \in V_1.$$

Since

$$\sum_{n=1}^p \frac{1}{n!} \alpha^n < e^\alpha - 1 \leq 1 \quad \text{and} \quad f_2^n(a(\theta)) \in \hat{C} \quad \forall n = 1, \dots, p-1,$$

it follows that

$$(3) \quad \varphi_2(u_{N(\alpha)}(\theta, \alpha)) + \frac{\alpha^p}{p!} G_2 \subset C \quad \forall \theta \in V_1, \quad \alpha \in (0, \alpha_0].$$

Relation (1) and assumption (a) imply that

$$\varphi_1(\bar{q}) + \frac{1}{p!} \alpha^p (w + G_1) \subset \{\varphi_1(u_{N(\alpha)}(\theta, \alpha)) \mid \theta \in V_1\} \quad \forall \alpha \in (0, \alpha_0]$$

which, together with (3), yields

$$\varphi_1(\bar{q}) + \gamma(w + G_1) \subset \{\varphi_1(u) \mid u \in \mathcal{U}, \varphi_2(u) + \gamma G_2 \subset C\}$$

for all $\gamma \in (0, \gamma_0]$, where $\gamma_0 = (1/p!) \alpha_0^p$. \square

Proof of Theorem 2.5. This follows directly from Theorem 2.2 for $p = 1$ by choosing for H an open bounded nhd of the point $(1, 1, \dots, 1)$ in $(0, \infty)^k$ and setting

$$\hat{q}(x, \alpha) = \bar{q} + \alpha \sum_{i=1}^k x_i y_i \quad \text{for } x = (x_1, \dots, x_k) \in H.$$

Proof of Theorem 2.6. For H we choose an open bounded nhd of $\bar{x} = (1, 1, \dots, 1) \in \mathbb{R}^k$. Then, for a sufficiently small $\alpha_1 > 0$ and $\alpha \in [0, \alpha_1]$,

$$\hat{q}(x, \alpha) := \bar{q} + \alpha \sum_{i \in I} x_i y_i + \alpha^2 \sum_{j \in J} x_j y_j \in Q$$

and

$$\begin{aligned}\varphi(\bar{q}(x, \alpha)) &= \varphi(\bar{q}) + \alpha \sum_{i \in I} x_i \psi'(0) \mathbf{e}_i \\ &\quad + \frac{1}{2} \alpha^2 \left[\sum_{i, s \in I} x_i x_s \psi''(0) \mathbf{e}_i \mathbf{e}_s + \sum_{j \in J} x_j \psi'(0) \mathbf{e}_j \right] + \alpha^2 d(x, \alpha),\end{aligned}$$

where $\lim_{\alpha \rightarrow 0+} d(x, \alpha) = 0$ uniformly for $x \in H$. Then the assumptions of Theorem 2.6 imply those of Theorem 2.2, whence our conclusion follows. \square

REFERENCES

- [1] D. S. BERNSTEIN, *A systematic approach to higher order necessary conditions in optimization theory*, SIAM J. Control Optim., 22 (1984), pp. 49–60.
- [2] T. J. GAFFNEY, private communication.
- [3] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [4] ———, *Higher order conditions with and without Lagrange multipliers*, SIAM J. Control Optim., 24 (1986), pp. 715–730.

ON REACHABLE SETS AND EXTREMAL REARRANGEMENTS OF CONTROL FUNCTIONS*

G. R. BURTON† AND E. P. RYAN†

Abstract. A characterization is provided for the set $\mathcal{A}(x^0; \mathcal{V})$ of states reachable from a given state x^0 along trajectories of a linear control system with controls in the set \mathcal{V} of rearrangements of a prescribed function v . This reachable set is shown to be convex and compact. A necessary and sufficient condition (maximum principle) for extremal controls (that is, controls generating trajectory endpoints in the boundary $\partial\mathcal{A}(x^0; \mathcal{V})$) is provided. Finally, some results on existence of solutions to Mayer-type optimal control problems are deduced.

Key words. extremal controls, maximum principle, normal systems, optimal control, reachable sets, rearrangements of functions

AMS(MOS) subject classifications. 49E15, 49A36, 93B05, 46A55

1. Introduction. This paper provides a characterization of the set $\mathcal{A}(x^0; \mathcal{U})$ of points reachable from x^0 along trajectories of the linear control system

$$(1) \quad \dot{x} = Ax + Bu, \quad x(t) \in \mathbb{R}^n, \quad u(t) \in \mathbb{R}^m$$

with controls u in the set \mathcal{U} of admissible controls taken here to be either the set \mathcal{V} of rearrangements (in the componentwise sense of Definition 1 below) of a prescribed function $v: [0, t_f] \rightarrow \mathbb{R}^m$, or the weak closure $\bar{\mathcal{V}}$ of \mathcal{V} . Without loss of generality, $t_f = 1$ will be assumed and I will denote the underlying time interval $[0, 1]$. Throughout, $1 \leq p < \infty$ and $L_p^m(I)$ will denote the space of p -integrable \mathbb{R}^m -valued functions on I ($L_p^1(I) \equiv L_p(I)$); q will be taken to be the conjugate exponent of p , and $\langle \cdot, \cdot \rangle$ will denote the usual bilinear form on $L_q^m(I) \times L_p^m(I)$.

DEFINITION 1. Functions $y, z \in L_p^m(I)$ are (componentwise) rearrangements of one another if

$$\mu(y_i^{-1}[\alpha, \infty)) = \mu(z_i^{-1}[\alpha, \infty)) \quad \forall \alpha \in \mathbb{R}, \quad i = 1, 2, \dots, m$$

where $y_i, z_i \in L_p(I)$ denote the i th components of the \mathbb{R}^m -valued functions y and z , respectively, and μ denotes Lebesgue measure on \mathbb{R} .

For example, suppose that system (1) is controlled via a scalar “on-off” device with prescribed total “on-time” $\tau < 1$. In this case the set of admissible controls may be interpreted as the set of rearrangements of the piecewise constant function

$$t \mapsto \begin{cases} 1, & 0 \leq t \leq \tau, \\ 0, & \tau < t \leq 1. \end{cases}$$

In the general context, let $v \in L_p^m(I)$ be a prescribed function and let \mathcal{V} (with weak closure $\bar{\mathcal{V}}$) denote the set of rearrangements of v . For $\mathcal{U} = \mathcal{V}$ or $\bar{\mathcal{V}}$, the *reachable set* is defined as

$$\mathcal{A}(x^0; \mathcal{U}) := (\exp A)[x^0 + L(\mathcal{U})] = \{(\exp A)[x^0 + L(u)] \mid u \in \mathcal{U}\} \subset \mathbb{R}^n$$

* Received by the editors October 5, 1987; accepted for publication (in revised form) March 14, 1988.

† School of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, U.K. The first author's research was performed under a grant from the University of Bath while the author was on leave from University College, London.

where the compact linear operator $L: L_p^m(I) \rightarrow \mathbb{R}^n$ is given by

$$L: u \mapsto \int_0^1 (\exp - At)Bu(t) dt.$$

We summarize the main results (which are reminiscent of results in linear time-optimal control [1]) as follows. First, it will be shown that $\mathcal{A}(x^0; \mathcal{V}) = \mathcal{A}(x^0; \bar{\mathcal{V}})$, and more generally that $T(\mathcal{V}) = T(\bar{\mathcal{V}})$ for any bounded linear map $T: L_p^m(I) \rightarrow \mathbb{R}^n$. It is then deduced from known properties of $\bar{\mathcal{V}}$ that $\mathcal{A}(x^0; \mathcal{V})$ is convex and compact. This latter result is somewhat surprising (although a parallel may be drawn with the conventional bang-bang principle [1]), for it shows that the *convex* set $\mathcal{A}(x^0; \mathcal{V})$ is a translation of a linear image of the (generically) *nonconvex* set \mathcal{V} . Next, a necessary and sufficient condition (maximum principle) for a control $u^* \in \bar{\mathcal{V}}$ to be *extremal* (that is, to generate a trajectory endpoint $x^* = (\exp A)[x^0 + L(u^*)]$ in the boundary $\partial \mathcal{A}(x^0; \mathcal{V})$ of the reachable set) will be established. Under an assumption of *normality* of the pair (A, B) , it will be shown that $\mathcal{A}(x^0; \mathcal{V})$ is relatively strictly convex, and that every control in $\bar{\mathcal{V}}$ generating a relative boundary point of $\mathcal{A}(x^0; \mathcal{V})$ must belong to \mathcal{V} . Finally, existence of optimal rearrangements is deduced for a class of Mayer-type control problems. Some other aspects of the control of a linear system by a set of rearrangements have been considered by Tahraoui [11].

Our results are readily derived using the general theory of rearrangements and maximization of linear functionals developed in [2]–[7]. For convenience, specific aspects of this theory of particular relevance in the present context are reiterated below. The proofs, for the case $m = 1$, may be found in [2]–[7]; the cases $m \geq 2$ follow easily. In particular, Lemma 1 follows from the results on doubly stochastic operators in [2], [7] and a direct proof is given in [3]. Lemma 2 appears in [3], [4]. Lemma 3 summarises results from [5]–[7]; some parts were only proved for $p = 1$, but the general case is easily deduced.

LEMMA 1. *Let $v = (v_1, \dots, v_m) \in L_p^m(I)$, and let \mathcal{V} denote the set of (component-wise) rearrangements of v . Then the weak closure $\bar{\mathcal{V}} \subset L_p^m(I)$ of \mathcal{V} is convex, so $\bar{\mathcal{V}} = \overline{\text{conv}} \mathcal{V}$.*

Remark. It should be noted that, with the exception of the case of a *constant* function v , the set \mathcal{V} is not convex.

LEMMA 2. *Let v , \mathcal{V} , and $\bar{\mathcal{V}}$ be as in Lemma 1, and let $g = (g_1, \dots, g_m) \in L_q^m(I)$.*

(i) *If there exist increasing functions $\varphi_1, \dots, \varphi_m$ such that $v^* := (\varphi_1 \circ g_1, \dots, \varphi_m \circ g_m) \in \mathcal{V}$, then v^* is the unique maximizer of the linear functional $\langle g, \cdot \rangle$ relative to $\bar{\mathcal{V}}$.*

(ii) *If every level set of each function g_1, \dots, g_m has zero measure, then such increasing functions $\varphi_1, \dots, \varphi_m$ exist.*

If u is a real measurable function on I , we denote by u^* the (essentially unique) increasing rearrangement of u .

LEMMA 3. *Let $m = 1$, let v , \mathcal{V} and $\bar{\mathcal{V}}$ be as in Lemma 1, and let*

$$\mathcal{W} := \left\{ w \in L_1(I) \left| \int_0^s w^* \geq \int_0^s v^*, 0 < s < 1; \int_0^1 w = \int_0^1 v \right. \right\}.$$

Then \mathcal{W} is a weakly compact convex subset of $L_p(I)$, and the set of extreme points of \mathcal{W} is \mathcal{V} . Consequently $\mathcal{W} = \bar{\mathcal{V}}$.

We now prove two further lemmas. Lemma 5 is a refinement of Lemma 3. The most significant change in the proof is the use of Lyapunov's Theorem; in addition to this, we have adopted an approach that clarifies the proof and avoids doubly stochastic operators.

LEMMA 4. Let $0 < s \leq 1$; then $u \mapsto -\int_0^s u^*$ defines a continuous convex functional on $L_p(I)$.

Proof. Let f be the increasing function

$$f(t) = \begin{cases} -1, & 0 \leq t < s, \\ 0, & s \leq t \leq 1. \end{cases}$$

Let \mathcal{F} be the set of all rearrangements of f on I , and let $\chi: L_q(I) \rightarrow \bar{\mathbb{R}}$ be the characteristic function of \mathcal{F} , that is,

$$\chi(u) = \begin{cases} 0, & u \in \mathcal{F}, \\ \infty, & u \notin \mathcal{F}. \end{cases}$$

The conjugate convex function Ψ of χ is given by

$$(2) \quad \Psi(w) := \sup \{ \langle w, u \rangle - \chi(u) \mid u \in L_q(I) \} = \sup \{ \langle w, u \rangle \mid u \in \mathcal{F} \}$$

for $w \in L_p(I)$. Then Ψ is convex [9, pp. 17–19], and from (2) it follows that

$$|\Psi(w) - \Psi(y)| \leq \|w - y\|_p \|f\|_q \quad \forall w, y \in L_p(I),$$

so Ψ is continuous. It is shown in [3, § 5] that

$$\Psi(u) = \int_0^1 u^* f^* \quad \forall u \in L_p(I).$$

Since $f^* = f$, for $u \in L_p(I)$ we now have

$$\Psi(u) = \int_0^1 u^* f = - \int_0^s u^*.$$

□

LEMMA 5. Let v , \mathcal{V} , and $\tilde{\mathcal{V}}$ be as in Lemma 1, let $g_1, \dots, g_n \in L_q^m(I)$, let $\alpha_1, \dots, \alpha_n \in \mathbb{R}$, and let

$$\begin{aligned} \mathcal{W} &= \left\{ w = (w_1, \dots, w_m) \in L_1^m(I) \mid \int_0^s w_i^* \right. \\ &\quad \left. \geq \int_0^s v_i^*, 0 < s < 1, i = 1, \dots, n; \int_0^1 w = \int_0^1 v \right\}, \\ \mathcal{G} &= \left\{ w \in L_p^m(I) \mid \int_0^1 w g_i = \alpha_i, i = 1, \dots, n \right\} \end{aligned}$$

(so $\mathcal{G} = L_p^m(I)$ if $n = 0$). Then \mathcal{W} is a weakly compact convex subset of $L_p^m(I)$, the set of extreme points of $\mathcal{W} \cap \mathcal{G}$ is $\mathcal{V} \cap \mathcal{G}$, and $\tilde{\mathcal{V}} \cap \mathcal{G} = \overline{\text{conv}}(\mathcal{V} \cap \mathcal{G})$.

Proof. First we show that every extreme point of $\mathcal{W} \cap \mathcal{G}$ lies in $\mathcal{V} \cap \mathcal{G}$. Suppose initially that $m = 1$. Let $w \in \mathcal{W} \cap \mathcal{G}$, and define continuous convex functions F, G on I by

$$F(s) := \int_0^s w^*, \quad G(s) := \int_0^s v^*.$$

Then $F(s) \geq G(s)$ for all $s \in I$, and $w \in \mathcal{V} \cap \mathcal{G}$ if and only if $F \equiv G$. Therefore, we shall suppose $F(s) > G(s)$ for some s , and show that w is not an extreme point of $\mathcal{W} \cap \mathcal{G}$. Let S denote the open set of points $s \in I$ for which $F(s) > G(s)$, and consider a maximal open interval $(s_0, s_1) \subset S$, so

$$(3) \quad F(s_0) = G(s_0),$$

$$(4) \quad F(s_1) = G(s_1),$$

$$(5) \quad F(s) > G(s) \quad \text{if } s_0 < s < s_1.$$

Since F and G are convex and $F \geq G$, it follows from (3), (4) that the subdifferentials obey

$$(6) \quad \partial F(s_0) \subset \partial G(s_0) \quad \text{if } s_0 > 0,$$

$$(7) \quad \partial F(s_1) \subset \partial G(s_1) \quad \text{if } s_1 < 1.$$

Let l_0 and l_1 denote the one-sided limits $l_0 = v^*(s_0+)$ and $l_1 = v^*(s_1-)$. By (5) and the convexity of F, G we have, for $s_0 < s < s_1$,

$$l_0 \leq (G(s) - G(s_0))/(s - s_0) < (F(s) - F(s_0))/(s - s_0) \leq w^*(s),$$

$$l_1 \geq (G(s_1) - G(s))/(s_1 - s) > (F(s_1) - F(s))/(s_1 - s) \geq w^*(s).$$

Thus

$$(9) \quad l_0 < w^*(s) < l_1 \quad \text{if } s_0 < s < s_1.$$

From (6), (7) we also have

$$(9) \quad w^*(s_0-) \leq l_0 \quad \text{if } s_0 > 0,$$

$$(10) \quad w^*(s_1+) \geq l_1 \quad \text{if } s_1 < 1.$$

By (8) and the definitions of l_0 and l_1 we can choose, in the following order, numbers $\tau_0, \tau_1, a, b, \varepsilon, \sigma_0, \sigma_1$, satisfying

$$(11) \quad s_0 < \sigma_0 < \tau_0 < \tau_1 < \sigma_1 < s_1,$$

$$l_0 \leq v^*(\sigma_0) < a < a + \varepsilon < w^*(\tau_0) \leq w^*(\tau_1) < b - \varepsilon < b < v^*(\sigma_1) \leq l_1.$$

Let $\Sigma = w^{-1}(a + \varepsilon, b - \varepsilon)$, so $\mu(\Sigma) \geq \tau_1 - \tau_0 > 0$. By Lyapunov's Theorem [8, 16.1.i] we can partition Σ into two sets Σ_1, Σ_2 of equal measure, such that

$$(12) \quad \int_{\Sigma_1} g_j = \int_{\Sigma_2} g_j, \quad j = 1, \dots, n.$$

Let

$$u(s) = \begin{cases} 1, & s \in \Sigma_1, \\ -1, & s \in \Sigma_2, \\ 0, & s \in I \setminus \Sigma. \end{cases}$$

From (12) we have $w \pm \xi u \in \mathcal{G}$ for all positive ξ . Next we show that $w \pm \xi u \in \mathcal{W}$, provided $\xi > 0$ is sufficiently small; hence $w = \frac{1}{2}(w + \xi u) + \frac{1}{2}(w - \xi u)$ is a nonextreme point of $\mathcal{W} \cap \mathcal{G}$.

Since $\mu(\Sigma_1) = \mu(\Sigma_2)$ we have

$$(13) \quad \int_0^1 (w \pm \xi u) = \int_0^1 w = \int_0^1 v.$$

Write $t_0 = \mu(w^{-1}(-\infty, a])$ and $t_1 = \mu(w^{-1}(-\infty, b])$, so by (9)-(11),

$$s_0 \leq t_0 \leq t_1 \leq s_1.$$

Now suppose $0 < \xi < \varepsilon$. Then $(w \pm \xi u)^{-1}(a, b) = w^{-1}(a, b)$, and $(w \pm \xi u)(s) = w(s)$ for all $s \in I \setminus w^{-1}(a, b)$. Therefore, $(w \pm \xi u)^*(s) = w^*(s)$ if $0 < s < t_0$ or $t_1 < s < 1$; from this and from (13) we obtain

$$(14) \quad \int_0^s (w \pm \xi u)^* = \int_0^s w^* \geq \int_0^s v^* \quad \text{if } 0 \leq s \leq t_0 \quad \text{or} \quad t_1 \leq s \leq 1.$$

If $t_0 < s < \sigma_0$, then $(w \pm \xi u)^*(s) > a > v^*(s)$. Together with (14) this yields

$$(15) \quad \int_0^s (w \pm \xi u)^* = \int_0^{t_0} w^* + \int_{t_0}^s (w \pm \xi u)^* \geq \int_0^s v^* \quad \text{if } t_0 \leq s \leq \sigma_0.$$

If $\sigma_1 < s < t_1$, then $(w \pm \xi u)^*(s) < b < v^*(s)$. Together with (14) this yields

$$(16) \quad \int_0^s (w \pm \xi u)^* = \int_0^{t_1} w^* - \int_s^{t_1} (w \pm \xi u)^* \geq \int_0^s v^* \quad \text{if } \sigma_1 \leq s \leq t_1.$$

If $\sigma_0 \leq s \leq \sigma_1$, then by Lemma 4

$$\int_0^s (w \pm \xi u)^* \geq \int_0^s w^* + \xi \int_0^s (\pm u)^* \geq \int_0^s w^* - \xi \mu(\Sigma_1);$$

hence

$$(17) \quad \int_0^s (w \pm \xi u)^* \geq \int_0^s v^* \quad \text{if } \sigma_0 \leq s \leq \sigma_1$$

provided that $\xi \mu(\Sigma_1) \leq \delta := \min \{F(t) - G(t) \mid \sigma_0 \leq t \leq \sigma_1\}$. It follows from (13)–(17) that $w \pm \xi u \in \mathcal{W}$, provided $0 < \xi < \min \{\varepsilon, \delta / \mu(\Sigma_1)\}$.

We have now shown that all extreme points of $\mathcal{W} \cap \mathcal{G}$ lie in $\mathcal{V} \cap \mathcal{G}$ if $m = 1$. The generalisation to $m \geq 2$ is deduced as follows. For $i = 1, \dots, m$, let \mathcal{V}_i be the set of rearrangements of v_i on I , let

$$\mathcal{W}_i := \left\{ u \in L_1(I) \mid \int_0^s u^* \geq \int_0^s v_i^*, \quad 0 < s < 1; \quad \int_0^1 u = \int_0^1 v_i \right\},$$

and, for $j = 1, \dots, n$, let $g_j = (g_{1j}, \dots, g_{mj})$. Consider an extreme point $w = (w_1, \dots, w_m)$ of $\mathcal{W} \cap \mathcal{G}$ and, for $i = 1, \dots, m$, write

$$\mathcal{G}_i = \left\{ u \in L_p(I) \mid \int_0^1 u g_{ij} = \int_0^1 w_i g_{ij}, \quad j = 1, \dots, n \right\};$$

then w_i is an extreme point of $\mathcal{W}_i \cap \mathcal{G}_i$, so $w_i \in \mathcal{V}_i \cap \mathcal{G}_i$ by the case $m = 1$. Hence $w \in (\mathcal{V}_1 \cap \mathcal{G}_1) \times \dots \times (\mathcal{V}_m \cap \mathcal{G}_m) \subset \mathcal{V} \cap \mathcal{G}$. Thus all extreme points of $\mathcal{W} \cap \mathcal{G}$ lie in $\mathcal{V} \cap \mathcal{G}$.

We now suppose that $p = 1$, and show that \mathcal{W} is a weakly compact convex set, and that $\bar{\mathcal{V}} \cap \mathcal{G} = \mathcal{W} \cap \mathcal{G} = \overline{\text{conv}}(\mathcal{V} \cap \mathcal{G})$. It is immediate from Lemma 4 that \mathcal{W} is closed and convex; hence \mathcal{W} is weakly closed. If $w = (w_1, \dots, w_m) \in \mathcal{W}$ then, for $i = 1, \dots, m$ and $0 < s < \frac{1}{2}$,

$$\begin{aligned} \int_0^s v_i^* &\leq \int_0^s w_i^* \leq s w_i^*\left(\frac{1}{2}\right) \leq 2s \int_{1/2}^1 w_i^* \leq 2s \int_{1/2}^1 v_i^* \\ \int_{1-s}^1 v_i^* &\geq \int_{1-s}^1 w_i^* \geq s w_i^*\left(\frac{1}{2}\right) \geq 2s \int_0^{1/2} w_i^* \geq 2s \int_0^{1/2} v_i^*, \end{aligned}$$

from which it follows that

$$\lim_{s \rightarrow 0} \int_0^s w_i^* = 0 = \lim_{s \rightarrow 0} \int_{1-s}^1 w_i^*$$

uniformly over $w \in \mathcal{W}$. Therefore

$$\lim_{M \rightarrow \infty} \int_{|w(s)| \geq M} |w(s)| \, ds = 0$$

uniformly over $w \in \mathcal{W}$; hence \mathcal{W} is weakly compact in $L_1^m(I)$. By the Krein-Mil'man Theorem we now have $\mathcal{W} \cap \mathcal{G} \subset \overline{\text{conv}}(\mathcal{V} \cap \mathcal{G})$, and since $\mathcal{V} \subset \mathcal{W}$ we obtain $\mathcal{W} \cap \mathcal{G} = \overline{\text{conv}}(\mathcal{V} \cap \mathcal{G})$. Taking $\mathcal{G} = L_1^m(I)$ gives $\mathcal{W} = \overline{\text{conv}} \mathcal{V}$, and then Lemma 1 yields $\mathcal{W} = \bar{\mathcal{V}}$.

Now suppose that $1 < p < \infty$. Then, by the case already considered, $\mathcal{W} = \overline{\text{conv}} \mathcal{V}$ in $L_1^m(I)$; since $\|u\|_p = \|v\|_p$ for all $u \in \mathcal{V}$, and $\|\cdot\|_p: L_1^m(I) \rightarrow \mathbb{R}$ is a lower-semicontinuous convex functional, it follows that $\|w\|_p \leq \|v\|_p$ for all $w \in \mathcal{W}$. From this and Lemma 4 it follows that \mathcal{W} is a closed bounded convex set in $L_p^m(I)$. We can now apply the Krein-Mil'man Theorem as above to deduce that $\mathcal{W} \cap \mathcal{G} = \overline{\text{conv}}(\mathcal{V} \cap \mathcal{G})$ and $\mathcal{W} = \bar{\mathcal{V}}$.

To complete the proof it will suffice to show that every element of \mathcal{V} is an extreme point of $\bar{\mathcal{V}}$. Let $w = (w_1, \dots, w_m) \in \mathcal{V}$ and let $h = (\tan^{-1} w_1, \dots, \tan^{-1} w_m) \in L_q^m(I)$. It follows from Lemma 2 that w is the unique maximizer relative to $\bar{\mathcal{V}}$ of the linear functional $\langle h, \cdot \rangle$; thus w is an extreme point of $\bar{\mathcal{V}}$. \square

2. The reachable set and extremal controls. Throughout, v is some function in $L_p^m(I)$, $1 \leq p < \infty$, \mathcal{V} is the set of (componentwise) rearrangements of v on I , and $\bar{\mathcal{V}}$ is the weak closure of \mathcal{V} .

THEOREM 1. *Let $T: L_p^m(I) \rightarrow \mathbb{R}^n$ be a bounded linear operator. Then $T(\mathcal{V}) = T(\bar{\mathcal{V}})$.*

Proof. Functions $g_1, \dots, g_n \in L_q^m(I)$ can be chosen such that

$$Tu = \left(\int_0^1 u g_1, \dots, \int_0^1 u g_n \right)$$

for all $u \in L_p^m(I)$. Consider $y = (y_1, \dots, y_n) \in T(\bar{\mathcal{V}})$. Then

$$T^{-1}y = \left\{ w \in L_p^m(I) \mid \int_0^1 w g_i = y_i, i = 1, \dots, n \right\}.$$

It follows from Lemma 5 that $\bar{\mathcal{V}} \cap T^{-1}y = \overline{\text{conv}}(\mathcal{V} \cap T^{-1}y)$; hence $\mathcal{V} \cap T^{-1}y \neq \emptyset$. Therefore $y \in T(\mathcal{V})$. Hence $T(\mathcal{V}) = T(\bar{\mathcal{V}})$. \square

COROLLARY 1. $\mathcal{A}(x^0; \mathcal{V}) = \mathcal{A}(x^0; \bar{\mathcal{V}})$ for the reachable sets of system (1).

Since $\bar{\mathcal{V}}$ is weakly compact and convex we deduce the following corollary.

COROLLARY 2. $\mathcal{A}(x^0; \mathcal{V})$ is compact and convex.

Henceforth, $\partial \mathcal{A}(x^0; \bar{\mathcal{V}})$, $\partial^r \mathcal{A}(x^0; \bar{\mathcal{V}})$, and $\partial^e \mathcal{A}(x^0; \bar{\mathcal{V}})$ will denote the *boundary*, *relative boundary*, and *set of extreme points*, respectively, of $\mathcal{A}(x^0; \bar{\mathcal{V}})$. A control $u^* \in \bar{\mathcal{V}}$ is said to be *extremal* if $x^* := (\exp A)[x^0 + L(u^*)]$ belongs to the set $\partial \mathcal{A}(x^0; \bar{\mathcal{V}})$; $u^* \in \bar{\mathcal{V}}$ is said to be *strongly extremal* if x^* belongs to the set $\partial^r \mathcal{A}(x^0; \bar{\mathcal{V}})$. The following theorem characterizes the set of extremal controls for (1).

THEOREM 2. $u^* \in \bar{\mathcal{V}}$ is extremal if and only if there exists a nontrivial solution $\eta: I \rightarrow \mathbb{R}^n$ of the adjoint equation $\dot{\eta} = -A^T \eta$ such that u^* maximizes the linear functional $\langle B^T \eta, \cdot \rangle$ relative to $\bar{\mathcal{V}}$.

Proof. Let $u^* \in \bar{\mathcal{V}}$ generate the endpoint $x^* := (\exp A)[x^0 + L(u^*)] \in \mathcal{A}(x^0; \bar{\mathcal{V}})$. Observe that the solutions η of the adjoint equation are precisely those functions expressible by

$$\eta: t \mapsto (\exp -A^T(t-1))\eta(1)$$

and that for such a function η , and for any $u \in \bar{\mathcal{V}}$,

$$(18) \quad \langle B^T \eta, u - u^* \rangle = \eta(1)^T (\exp A)[L(u) - L(u^*)] = \eta(1)^T (x_u - x^*)$$

where $x_u = (\exp A)[x^0 + L(u)] \in \mathcal{A}(x^0; \bar{\mathcal{V}})$ is the endpoint generated by u .

Now, by convexity, $x^* \in \partial \mathcal{A}(x^0; \bar{\mathcal{V}})$ if and only if

$$\zeta^T (x - x^*) \leq 0 \quad \forall x \in \mathcal{A}(x^0; \bar{\mathcal{V}})$$

for some nonzero $\zeta \in \mathbb{R}^n$ (more particularly, ζ belongs to the outward normal cone to $\mathcal{A}(x^0; \bar{\mathcal{V}})$ at x^*) and, from (18), this occurs if and only if

$$\langle B^T \eta, u - u^* \rangle \leq 0 \quad \forall u \in \bar{\mathcal{V}}$$

for some nontrivial function η that is a solution of the adjoint equation (with $\zeta = \eta(1)$). \square

COROLLARY 3. *For $x^* \in \partial \mathcal{A}(x^0; \bar{\mathcal{V}})$, let $\mathcal{N}_{\mathcal{A}(x^0; \bar{\mathcal{V}})}(x^*)$ denote the normal cone to $\mathcal{A}(x^0; \bar{\mathcal{V}})$ at x^* . If $u^* \in \bar{\mathcal{V}}$ is extremal, generating endpoint $x^* \in \partial \mathcal{A}(x^0; \bar{\mathcal{V}})$, then, for each $\zeta \in \mathcal{N}_{\mathcal{A}(x^0; \bar{\mathcal{V}})}(x^*) \setminus \{0\}$, u^* maximizes (relative to $\bar{\mathcal{V}}$) the functional $\langle B^T \eta, \cdot \rangle$ with $\eta: t \mapsto (\exp - A^T(t-1))\zeta$.*

We now address the question of whether a strongly extremal control $u^* \in \bar{\mathcal{V}}$ is in fact a rearrangement. First recall the concept of a *normal* system [1]: the pair (A, B) is normal if

$$\text{rank} [b_j : Ab_j : \cdots : A^{n-1}b_j] = n, \quad j = 1, 2, \dots, m$$

where b_j denotes the j th column of $B \in \mathbb{R}^{n \times m}$.

Corollary 1 shows that every point of $\mathcal{A}(x^0; \bar{\mathcal{V}})$ is generated by a set of controls that includes a rearrangement of v . For relative boundary points of the reachable set of a normal system, more is shown by the following theorem.

THEOREM 3. *Let (A, B) be a normal pair. Then $\mathcal{A}(x^0; \bar{\mathcal{V}})$ is relatively strictly convex (that is, $\mathcal{A}(x^0; \bar{\mathcal{V}})$ has no line segments in its relative boundary $\partial^r \mathcal{A}(x^0; \bar{\mathcal{V}})$), every strongly extremal control $u^* \in \bar{\mathcal{V}}$ is a rearrangement, and distinct strongly extremal controls generate distinct trajectory endpoints.*

Proof. For notational convenience, we write $\mathcal{A} = \mathcal{A}(x^0; \bar{\mathcal{V}})$. Relative strict convexity is obvious if $\dim \mathcal{A} = 0$ and in this case $\partial^r \mathcal{A} = \emptyset$, so there are no strongly extremal controls. Suppose now that $\dim \mathcal{A} \geq 1$ (in which case $\partial^r \mathcal{A}(x^0; \bar{\mathcal{V}}) \neq \emptyset$), and consider a control $u^* \in \bar{\mathcal{V}}$ generating trajectory $x(\cdot)$ with endpoint $x^* \in \partial^r \mathcal{A}$. We first show that $\zeta^* \in \mathcal{N}_{\mathcal{A}}(x^*)$ (the normal cone to \mathcal{A} at x^*) can be chosen such that $A^T \zeta^* \neq 0$. This is trivial if $\text{rank } A = n$, so suppose that $\text{rank } A < n$. The assumption of normality then ensures that $Ab_1, \dots, A^{n-1}b_1$ are $n-1$ linearly independent vectors in $\text{im } A$, so $\text{rank } A = n-1$. Further,

$$x^* = x^0 + A \int_0^1 x + B \int_0^1 u^*,$$

and, since $\int_0^1 u^*$ has the same value for all $u^* \in \bar{\mathcal{V}}$, it follows that \mathcal{A} is contained in a translation of the subspace $\text{im } A$. Thus, $\dim \mathcal{A} \leq n-1$, and so $\dim \mathcal{N}_{\mathcal{A}}(x^*) \geq 2$. Since $\text{null } A^T = 1$, we can choose $\zeta^* \in \mathcal{N}_{\mathcal{A}}(x^*) \setminus \ker A^T$, and ζ^* has the required properties.

By Corollary 3, u^* maximizes the functional $\langle B^T \eta, \cdot \rangle$ relative to $\bar{\mathcal{V}}$, with $\eta: t \mapsto \exp(-A^T(t-1))\zeta^*$. To apply Lemma 2, we will establish that every level set of each function $\beta_j(\cdot) := b_j^T \eta(\cdot)$ has zero measure. Noting that β_j is an analytic function and therefore either assumes each value for only finitely many $t \in [0, 1]$ or is constant on $[0, 1]$, we need only show that each β_j is nonconstant. We seek a contradiction and suppose that β_j is constant. Repeated differentiation at $t = 1$ yields $b_j^T (A^T)^k \zeta^* = 0$ for all integers $k \geq 1$. If $\text{rank } A = n-1$ then, by normality, $Ab_j, \dots, A^{n-1}b_j$ span the $(n-1)$ -dimensional subspace $\text{im } A$. Hence, $\zeta^* \in \ker A^T$, contrary to the choice of ζ^* . If $\text{rank } A = n$ then, by the Cayley-Hamilton Theorem, I_n (the identity on \mathbb{R}^n) can be expressed as a linear combination of A, \dots, A^n ; hence $b_j^T \zeta^* = 0 = b_j^T A^T \zeta^* = \dots = b_j^T (A^T)^{n-1} \zeta^*$, and, since $\zeta^* \neq 0$, this contradicts the assumption of normality. Now Lemma 2 shows that $u^* \in \bar{\mathcal{V}}$ and that u^* is the unique maximizer of $\langle B^T \eta, \cdot \rangle$ relative to $\bar{\mathcal{V}}$. It follows that no other control in $\bar{\mathcal{V}}$ generates the same endpoint x^* , and that

no other boundary point of \mathcal{A} has outward normal ζ^* , so $x^* \in \partial^e \mathcal{A}$. Hence, \mathcal{A} is relatively strictly convex. \square

Example. Consider a controlled harmonic oscillator defined via the normal pair

$$(A, B) = \left(\begin{bmatrix} 0 & 2\pi \\ -2\pi & 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right).$$

Let $x^0 = 0$, let the prescribed function $v: I \rightarrow \mathbb{R}$ be given by

$$v(t) = \begin{cases} \pi, & 0 \leq t \leq \frac{1}{2}, \\ 0, & \frac{1}{2} < t \leq 1 \end{cases}$$

and let \mathcal{V} denote the set of rearrangements of v . By Corollary 1, $\mathcal{A}(0; \bar{\mathcal{V}}) = \mathcal{A}(0; \mathcal{V})$ and, by Theorems 2 and 3, $u^* \in \mathcal{V}$ is extremal if and only if, for some scalars κ_1, κ_2 (not both zero),

$$\int_0^1 (\kappa_1 \cos 2\pi t + \kappa_2 \sin 2\pi t)(u^*(t) - u(t)) dt \geq 0 \quad \forall u \in \mathcal{V},$$

whence

$$u^*(t) = \begin{cases} \pi, & \kappa_1 \cos 2\pi t + \kappa_2 \sin 2\pi t \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, by simple calculation, the boundary of the reachable set is given by

$$\partial \mathcal{A}(0; \mathcal{V}) = \left\{ x^* \mid x^* = \pm \begin{bmatrix} \cos 2\pi s \\ \sin 2\pi s \end{bmatrix}; 0 \leq s \leq \frac{1}{2} \right\} = \{x^* \mid |x^*| = 1\} \subset \mathbb{R}^2.$$

Hence, the reachable set $\mathcal{A}(0; \mathcal{V})$ is the closed ball of unit radius in \mathbb{R}^2 .

Lemma 3 can be used to determine $\bar{\mathcal{V}}$ as follows. The set

$$\mathcal{S} = \{w \in L_1(I) \mid 0 \leq w \leq \pi \text{ almost everywhere and } \int_0^1 w = \tfrac{1}{2}\pi\}$$

is closed and convex, and $\mathcal{V} \subset \mathcal{S}$; hence $\bar{\mathcal{V}} \subset \mathcal{S}$. To show that $\mathcal{S} \subset \bar{\mathcal{V}}$, consider $w \in \mathcal{S}$. Since $w \geq 0$ almost everywhere, for $0 < s \leq \frac{1}{2}$

$$\int_0^s w^* \geq 0 = \int_0^s v^*,$$

and since $w \leq \pi$ almost everywhere, for $\frac{1}{2} < s < 1$

$$\int_0^s w^* = \int_0^1 w^* - \int_s^1 w^* \geq \frac{1}{2}\pi - \pi(1-s) = \pi\left(s - \frac{1}{2}\right) = \int_0^s v^*.$$

Moreover,

$$\int_0^1 w = \int_0^1 v$$

and it follows immediately from Lemma 3 that $w \in \bar{\mathcal{V}}$. Thus $\bar{\mathcal{V}} = \mathcal{S}$.

3. Some observations on optimal control problems of Mayer type. The theory of rearrangements has implications for the following Mayer-type optimal control problem \mathcal{P} : maximize $f(x(1))$ subject to (1) with $x(0) = x^0$ (prescribed) and $u \in \bar{\mathcal{V}}$ (the weak closure of the set \mathcal{V} of (componentwise) rearrangements of prescribed function $v \in L_p^m(I)$).

COROLLARY 4. *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be convex. Then there exists a maximizer u^* for the functional $u \mapsto f((\exp A)[x^0 + L(u)])$ relative to $\tilde{\mathcal{V}}$ (that is, a solution of the Mayer problem \mathcal{P}), such that $(\exp A)[x^0 + L(u^*)]$ is an extreme point of $\mathcal{A}(x^0; \tilde{\mathcal{V}})$ and u^* is a rearrangement.*

Proof. By continuity, f attains a maximum relative to compact $\mathcal{A} := \mathcal{A}(x^0; \tilde{\mathcal{V}})$. Since \mathcal{A} is the convex hull of its extreme points, it follows from Theorem 32.2 of [10] that convex f is maximized relative to \mathcal{A} at some point $x^* \in \partial^e \mathcal{A}$. By Corollary 1 there exists $u^* \in \mathcal{V}$ with $x^* = (\exp A)[x^0 + L(u^*)]$. \square

The following is an immediate consequence of Corollary 1.

COROLLARY 5. *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be continuous. Then there is a rearrangement $u^* \in \mathcal{V}$ that maximizes the functional $u \mapsto f((\exp A)[x^0 + L(u)])$ relative to $\tilde{\mathcal{V}}$.*

Acknowledgments. The authors thank one of the referees for a suggestion leading to the removal of an assumption of normality from several of the results. The first author is grateful for the hospitality of the University of Bath.

REFERENCES

- [1] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.
- [2] J. R. BROWN, *Approximation theorems for Markov operators*, Pacific J. Math., 16 (1966), pp. 13–23.
- [3] G. R. BURTON, *Rearrangements of functions, maximization of convex functionals, and vortex rings*, Math. Ann., 276 (1987), pp. 225–253.
- [4] ———, *Variational problems on classes of rearrangements and multiple configurations for steady vortices*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.
- [5] J. V. RYFF, *Orbits of L^1 -functions under doubly stochastic transformations*, Trans. Amer. Math. Soc., 117 (1965), pp. 92–100.
- [6] ———, *Extreme points of some convex subsets of $L^1(0, 1)$* , Proc. Amer. Math. Soc., 18 (1967), pp. 1026–1034.
- [7] ———, *Majorized functions and measures*, Indag. Math., 30 (1968), pp. 431–437.
- [8] L. CESARI, *Optimization—Theory and Applications*, Springer-Verlag, New York, Berlin, 1983.
- [9] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, New York, 1976.
- [10] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [11] R. TAHRAOUI, *Contrôle optimal à réarrangement fixé*, C.R. Acad. Sci. Paris Sér. I Math., 303 (1986), pp. 955–958.

A NEW EXAMPLE OF THE LAVRENTIEV PHENOMENON*

ARTHUR C. HEINRICHER[†] AND VICTOR J. MIZEL[‡]

Abstract. This paper describes a new example of a calculus of variations problem with the Lavrentiev phenomenon, a problem with different value functions for different classes of admissible trajectories. Such problems provide examples of nonuniqueness for Hamilton–Jacobi equations since each value function is a solution. The authors describe these solutions and the corresponding optimal trajectories and focus on how a standard proof of optimality is adapted for these examples.

Key words. Lavrentiev phenomenon, Bolza problems, value functions, Hamilton–Jacobi equations

AMS(MOS) subject classification. 49A05

1. Introduction. This note discusses certain calculus of variations problems that require the exact specification of the class of admissible trajectories. Often, this specification is not critical. Either there is a natural admissible class such as Lipschitz or C^1 functions containing any and all minimizers or, barring this, each minimizer can be approximated by a sequence of elements from such a class, with costs approximating the cost of the minimizer. This approximation property fails for the examples we describe; there is a positive gap between the optimal costs for different admissible classes, even though one class is dense in the other.

Lavrentiev [17] gave the first example of such a problem, where the cost of the absolutely continuous minimizer could not be approximated by “smooth” trajectories. Cesari [7] describes a simpler example due to Mania [21], where the gap is between the cost of the absolutely continuous minimizer and the cost achievable by trajectories with bounded derivative. Ball and Mizel [2],[3] gave the first description of a *fully regular* problem exhibiting the Lavrentiev phenomenon. Angell [1] gives conditions sufficient to preclude this *Lavrentiev phenomenon*. (See also Cesari [7] and Loewen [20].)

The problems we describe are different from the known examples of the Lavrentiev phenomenon because the gap between the minimum values can be demonstrated by explicitly exhibiting the minimizers for the different classes. All other examples demonstrate the Lavrentiev gap without actually exhibiting the different minimizers. (The proofs show that all trajectories in the smaller admissible class must cross a certain region and pay a penalty for this crossing; a minimizer for the smaller admissible class need not exist.) We explicitly solve the Hamilton–Jacobi equation and determine the value functions for the different admissible classes. These value functions provide formulae for the optimal trajectories.

The Hamilton–Jacobi (HJ) equation is a standard tool for studying optimal control or calculus of variations problems. The usual “verification theorem” (see, for example, the text by Fleming and Rishel [11]) states that *the* solution for the (HJ) equation is *the* value function for the optimization problem. The value function de-

* Received by the editors May 18, 1987; accepted for publication (in revised form) February 12, 1988.

[†] Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506-0027. This author’s research was partly supported by the Air Force Office of Scientific Research grant AFOSR-85NM226.

[‡] Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. This author’s research was partly supported by the National Science Foundation grant DMS-8602954.

finds the optimal control, which gives an equation for the optimal trajectory. But “Lavrentiev type” problems have different value functions for different admissible classes and so the associated (HJ) equation¹ should have at least two solutions. We will focus on how the standard verification theorem arguments are adjusted to handle this nonuniqueness.

The problem of nonuniqueness for (HJ) equations is particularly interesting in light of recent studies of generalized solutions by Crandall and Lions [9] (see also Crandall, Evans, and Lions [10] and Lions [19]). Though the value function may be just one of many generalized solutions for the (HJ) equation, it is the unique *viscosity solution*. But in examples exhibiting the Lavrentiev phenomenon, there are two different value functions and both should provide solutions in the viscosity sense. The examples described here violate two crucial hypotheses required for the uniqueness theorems: First, the terminal cost for the optimization problem is discontinuous. Second, the Hamiltonian is discontinuous and unbounded along a certain curve in the state-time plane. Barles and Perthame [4] have considered stopping problems with discontinuous cost and provide a definition of viscosity solutions for such problems as well as uniqueness and stability results. Ishii [15] has considered problems with discontinuous, but bounded Hamiltonians and obtained existence of (possibly discontinuous) viscosity solutions.

We begin by giving a complete formulation of the problem, the cost function and its properties. The minimizers and *pseudo-minimizers* are described in the next section. We then outline the proof of optimality. The method of proof is “standard” and we try to focus on what adjustments must be made because of the Lavrentiev phenomenon. We conclude by discussing some variants of the problem and some extensions to more general problems exhibiting the Lavrentiev phenomenon.

2. Problem formulation. Consider a calculus of variations problem of the Bolza type. There is a running cost up to a fixed final time T and a terminal cost, which is a (discontinuous) function of the final state. For each initial time $s < T$ and each initial position $y \in \mathfrak{R}$, the cost incurred by an admissible trajectory $x(\cdot) = x(\cdot; s, y)$ is given by²

$$J(s, y; x) := \int_s^T e(t, x(t)) |\dot{x}(t)|^\alpha dt + \varphi_N(x(T)),$$

where the terminal cost is

$$\varphi_N(x) := \begin{cases} N, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

(N is a positive constant.)

The admissible trajectories are chosen from

$$\mathcal{A}_p = \mathcal{A}_p(s, y) := \left\{ x \in W^{1,p}(s, T) : x(s) = y \right\}.$$

¹ If the admissible classes are very different, then the form of the (HJ) equation may be different for each problem. This is the case for the impulse control problems of Bensoussan and Lions [6], where the (HJ) equation is replaced by a quasi-variational inequality, or when the admissible controls are allowed to push in only one direction, as in the work of Barron and Jensen [5]. In the examples we describe, the form of the (HJ) equation is the same for the two admissible classes.

² We use the symbol $:=$ to indicate an equality in which the left-hand side is defined by the right-hand side.

The cases $p = 1$ and $p = 2$ are the most important for our discussion.

The following assumptions will be enforced throughout the paper,

Basic assumptions.

$$(A1) \quad \alpha > 2.$$

$$(A2) \quad e(\cdot, \cdot) : (-\infty, T] \times \mathbb{R} \rightarrow \mathbb{R}^+ \text{ is continuous and satisfies the following homogeneity property:}^3$$

$$(1) \quad e(\lambda r + T, \sqrt{\lambda}x) = \lambda^{(\alpha/2)-1} e(r + T, x), \quad r < 0, \quad x \in \mathbb{R}^+.$$

$$(A3) \quad \Gamma := \{(t, x) : e(t, x) = 0\} \neq \emptyset.$$

We refer to Γ as the “free zone” since no running cost is incurred while the trajectory remains in this set. The homogeneity property (A2) implies that Γ has specific structure. If $e(t_0, x_0) = 0$, then $e(t, x) = 0$ for all (t, x) such that $x/\sqrt{T-t} = x_0/\sqrt{T-t_0}$. To see this, take $\lambda = 1/(T-t)$ and $\lambda_0 = 1/(T-t_0)$ in (1) to obtain

$$e(t, x) = \left(\frac{T-t}{T-t_0} \right)^{(\alpha/2)-1} e(t_0, x_0).$$

Hence, Γ consists of “half parabolas” $x = c\sqrt{T-t}$.

To simplify the discussion, we will focus on the case when the free zone consists of a single curve:

$$(2) \quad \Gamma = \{(t, x) : x = \rho_0 \sqrt{T-t}, \quad t < T\},$$

where $\rho_0 > 0$.

We also require that $e(\cdot, \cdot)$ satisfy the following condition.

$$(A4) \quad \text{For each } \rho_1 > \rho_0, \text{ there is a constant } \gamma = \gamma(\rho_1) > 0 \text{ such that}$$

$$\inf \{e(T-1, \rho) : \rho \geq \rho_1\} \geq \gamma.$$

Before proceeding, we note two reductions, which simplify the discussion. First of all, if there is a time $t_0 < T$ such that $x(t_0) \leq 0$, then there is an obvious trajectory attaining zero cost on the remaining interval:

$$x(t) = x(t_0), \quad t_0 \leq t \leq T.$$

In particular, if the initial position y is not positive, then the obvious minimizer is $x(t) \equiv y$. In all the discussion that follows, we will consider only $y > 0$.

Notice also that this problem reduces to a fixed-endpoint Lagrange problem when $N = +\infty$. In this case, any finite cost trajectory must have $x(T) \leq 0$, and so we could consider, without loss, the admissible class

$$\mathcal{A}_1^\infty := \{x \in W^{1,1}(s, T) : x(s) = y, \quad x(T) = 0\}.$$

³ Think of $-r$ as the “time to go.”

We now state two consequences of the homogeneity property (A2). The first is a scale invariance property for the cost function. It is this property that makes the problem (and the Hamilton–Jacobi equation) explicitly solvable.

LEMMA 2.1. *If $s < T$, $y > 0$ and $x(\cdot) \in \mathcal{A}_1(s, y)$, then*

$$J(s, y; x) = J(T - 1, \rho; \tilde{x}),$$

where

$$\rho := y/\sqrt{T - s},$$

and $\tilde{x}(\cdot) \in \mathcal{A}_1(T - 1, \rho)$ is defined by

$$\tilde{x}(t) := \frac{1}{\sqrt{T - s}} x\left((T - s)(t - T) + T\right), \quad t \in [T - 1, T].$$

In particular,

$$\inf \left\{ J(s, y; x); x \in \mathcal{A}_1(s, y) \right\} = \inf \left\{ J(T - 1, \rho; x); x \in \mathcal{A}_1(T - 1, \rho) \right\}$$

where $\rho := y/\sqrt{T - s}$.

The proof of this lemma is simply a change of variables in the cost integral. The lemma says that the cost can always be computed by first shifting the initial data to $(T - 1, \rho)$, and that the value function

$$U^*(s, y) := \inf \left\{ J(s, y; x); x \in \mathcal{A}_1(s, y) \right\}$$

is a function only of the ratio $\rho = y/\sqrt{T - s}$.

The homogeneity property also implies that there is a positive lower bound for $J(s, y; x)$ when (s, y) is above the free zone.

LEMMA 2.2. *Let $y = \rho_1\sqrt{T - s}$ with $\rho_1 > \rho_0$. There is a constant $\delta = \delta(\rho_1, \rho_0) > 0$ such that*

$$J(s, y; x) \geq \delta \quad \text{for all } x \in \mathcal{A}_1(s, y).$$

Proof. The only interesting case is $x(T) = 0$, so let $x(\cdot) \in \mathcal{A}_1(s, y)$ satisfy this condition. By Lemma 2.1, we may assume that $s = T - 1$ and $y = \rho_1$.

Define $\rho_2 := (\rho_1 + \rho_0)/2$ and let

$$\tau_2(x) := \inf \left\{ t \geq T - 1 : x(t) \leq \rho_2 \right\}.$$

Since $x(T) = 0$, $\tau_2 < T$ and we have

$$\begin{aligned} J(T - 1, \rho; x) &\geq \int_{T-1}^{\tau_2(x)} e(t, x(t)) |\dot{x}(t)|^\alpha dt \\ &= \int_{T-1}^{\tau_2(x)} (T - t)^{(\alpha/2)-1} e\left(T - 1, x(t)/\sqrt{T - t}\right) |\dot{x}(t)|^\alpha dt \\ &\geq \gamma(\rho_2) \int_{T-1}^{\tau_2(x)} (T - t)^{(\alpha/2)-1} |\dot{x}(t)|^\alpha dt, \end{aligned}$$

where γ is given by assumption (A4). It is not difficult to compute the minimum for the last integral and show that

$$\int_{T-1}^{\tau_2(x)} (T-t)^{(\alpha/2)-1} |\dot{x}(t)|^\alpha dt \geq 2(\rho_1 - \rho_0)^\alpha \left(\frac{\alpha}{4(\alpha-1)} \right)^{\alpha-1} > 0.$$

So, defining

$$\delta(\rho_1, \rho_0) := 2\gamma(\rho_2) (\rho_1 - \rho_0)^\alpha \left(\frac{\alpha}{4(\alpha-1)} \right)^{\alpha-1},$$

we have the desired lower bound. \square

3. Minimizers and pseudo-minimizers. The absolute minimizer in $\mathcal{A}_1(s, y)$ is not an element of $W^{1,p}(s, T)$ for $p \geq 2$. So, if we restrict the problem to trajectories in \mathcal{A}_2 , this absolute minimizer is no longer admissible. In fact, there is a new minimizer providing a *strictly larger* cost. Following Ball and Mizel [3], we refer to these higher cost minimizers as *pseudo-minimizers*.

The minimizer and the pseudo-minimizer differ in that the first takes advantage of the free zone while the second ignores it completely.

The absolute minimizer is given by the expression

$$(3) \quad x^{(1)}(t) = \begin{cases} y, & y \geq \rho_N^{(1)} \sqrt{T-s}, \\ y \left[\frac{T-t}{T-s} \right]^{\alpha/(2(\alpha-1))} \vee \rho_0 \sqrt{T-t}, & \rho_0 \sqrt{T-s} < y < \rho_N^{(1)} \sqrt{T-s}, \\ y \wedge \rho_0 \sqrt{T-t}, & 0 \leq y \leq \rho_0 \sqrt{T-s}. \end{cases}$$

(The constant $\rho_N^{(1)} > \rho_0$ is determined below.) This defines the unique minimizer in $\mathcal{A}_1(s, y)$, up to the arbitrary choice available when $y = \rho_N^{(1)} \sqrt{T-s}$.

The pseudo-minimizer is identical to the absolute minimizer between the free zone and the critical curve ($y = \rho_N^{(2)} \sqrt{T-s}$, with $\rho_N^{(2)}$ determined below) but it does not (cannot) take advantage of the free zone. In fact, since the curve Γ is not an element of $W^{1,2}(T-\epsilon, T)$ for any $\epsilon > 0$, “following Γ to $x = 0$ ” is not an admissible trajectory. One pseudo-minimizer is given by

$$(4) \quad x^{(2)}(t) = \begin{cases} y, & y > \rho_N^{(2)} \sqrt{T-s}, \\ y \left[\frac{T-t}{T-s} \right]^{\alpha/(2(\alpha-1))}, & 0 \leq y \leq \rho_N^{(2)} \sqrt{T-s}. \end{cases}$$

The minimizer in $\mathcal{A}_2(s, y)$ is not unique. In fact, since motion along Γ is free, it is possible to construct an infinite family of pseudo-minimizers by allowing the trajectory to follow Γ for some interval before exiting to follow another pseudo-minimizer to $x = 0$ at the final time.

Notice that

$$x^{(1)}(\cdot; s, y) \in W^{1,p}(s, T) \text{ for } 1 \leq p < 2,$$

while

$$x^{(2)}(\cdot; s, y) \in W^{1,p}(s, T) \text{ for } 1 \leq p < 2 + \frac{2}{\alpha - 2}.$$

The following theorem summarizes our result.

THEOREM 3.1. *Let $s < T$, $y > 0$, and let $J(s, y; \cdot)$ be defined as above, satisfying conditions (A1)–(A4). Then $x^{(1)}(\cdot; s, y)$ is a minimizer in the admissible class $\mathcal{A}_1(s, y)$ and $x^{(2)}(\cdot; s, y)$ is a minimizer in the admissible class $\mathcal{A}_2(s, y)$. Moreover, for (s, y) such that $0 < y < \rho_N^{(1)} \sqrt{T - s}$, this problem exhibits the Lavrentiev phenomenon:*

$$\inf \left\{ J(s, y; x); x \in \mathcal{A}_1(s, y) \right\} < \inf \left\{ J(s, y; x); x \in \mathcal{A}_2(s, y) \right\}.$$

For (s, y) such that $s < T$ and $y \in \mathfrak{R}$, there is no Lavrentiev gap beyond $\mathcal{A}_2(s, y)$:

$$\inf \left\{ J(s, y; x); x \in \mathcal{A}_2(s, y) \right\} = \inf \left\{ J(s, y; x); x \in \mathcal{A}_\infty(s, y) \right\}.$$

The final statement is proved by constructing a minimizing sequence of trajectories in $W^{1,\infty}(s, T)$. We obtain such a sequence by taking $x^{(2)}(\cdot)$ and replacing the trajectory over a final interval $[T - \epsilon, T]$ with a line segment. It is easy to show that the costs for these approximating trajectories converge to $J(s, y; x^{(2)})$ as $\epsilon \rightarrow 0$.

Example. Take $T = 0$, $\alpha = 6$ and $e(t, x) = (x^2 + t)^2$. Also, for definiteness, fix the initial condition at $(-1, 1)$ and take $N = +\infty$. Then we consider the problem of minimizing

$$J(x) = \int_{-1}^0 (x^2(t) + t)^2 |\dot{x}(t)|^6 dt$$

over the admissible trajectories

$$\mathcal{A} := \left\{ x \in W^{1,1}(-1, 0) : x(-1) = 1, x(0) = 0 \right\}.$$

The formulae for $x^{(1)}(\cdot)$ and $x^{(2)}(\cdot)$ given above reduce to

$$\begin{aligned} x^{(1)}(t) &= \sqrt{-t}, & -1 \leq t \leq 0, \\ x^{(2)}(t) &= (-t)^{3/5}, & -1 \leq t \leq 0. \end{aligned}$$

Notice that

$$J(x^{(1)}) = 0 \text{ while } J(x^{(2)}) = \frac{1}{6} \left(\frac{3}{5} \right)^5.$$

4. On proving optimality. Two different solutions to the Hamilton–Jacobi equation provide the minimizers described above. These two solutions for the (HJ) equation are the value functions when different classes of admissible trajectories are considered for the problem. For the classical verification theory, see the text of Fleming and Rishel [11]. Clarke and Vinter [8] present local optimality conditions under weaker conditions on solutions to the (HJ) equation. We describe below how the classical verification theorem arguments must be adjusted for our examples.

Assume that $U(\cdot, \cdot)$ is a continuously differentiable, nonnegative function satisfying the following conditions:

- (A) $U_s(s, y) + \theta U_y(s, y) + f(s, y, \theta) \geq 0$ for all $\theta \in \mathbb{R}$, $s < T$, $y \in \mathbb{R}$.
- (B) $t \mapsto U(t, x(t)) \in W_{loc}^{1,1}(s, T)$ for any admissible trajectory $x(\cdot)$.
- (C) $\limsup_{t \rightarrow T^-} U(t, x(t)) \leq \phi(x(T))$ for any admissible trajectory $x(\cdot)$.

Conditions (A), (B), and (C) are sufficient to prove that $U(s, y)$ is a lower bound for the cost starting at (s, y) given by

$$I(s, y; x) = \int_s^T f(t, x(t), \dot{x}(t)) dt + \phi(x(T)).$$

To see this, let $x(\cdot)$ be an admissible trajectory and let $s < \delta < T$. It follows from (A) and (B) that

$$\begin{aligned} U(\delta, x(\delta)) - U(s, y) &= \int_s^\delta \left[U_s(t, x(t)) + \dot{x}(t) U_y(t, x(t)) \right] dt \\ &\geq - \int_s^\delta f(t, x(t), \dot{x}(t)) dt, \end{aligned}$$

and hence

$$\begin{aligned} (5) \quad U(s, y) &\leq \int_s^\delta f(t, x(t), \dot{x}(t)) dt + U(\delta, x(\delta)) \\ &= I(s, y; x) - \int_\delta^0 f(t, x(t), \dot{x}(t)) dt \\ &\quad + U(\delta, x(\delta)) - \phi(x(T)). \end{aligned}$$

Taking the upper limit as $\delta \rightarrow T^-$ and using (C), we obtain

$$U(s, y) \leq I(s, y; x).$$

Since $x(\cdot)$ was arbitrary, we have the desired result.

Once we know that $U(\cdot, \cdot)$ provides a lower bound for the optimal cost, any admissible trajectory $x^*(\cdot)$ attaining that lower bound must be optimal. If $U(\cdot, \cdot)$ solves the (HJ) equation, not just the inequality (A), then $U(\cdot, \cdot)$ provides a method of determining $x^*(\cdot)$.

For each (s, y) , let $\theta^*(s, y)$ provide equality in (A). Define a trajectory $x^*(\cdot)$ by

$$\dot{x}^*(t; s, y) = \theta^*(t, x^*(t; s, y)), \quad x^*(s; s, y) = y.$$

Equality holds throughout (5) with $(t, x^*(t), \dot{x}^*(t))$ in place of (s, y, θ) . If condition (C) holds with equality and “lim” in place of “limsup” (and $x^*(\cdot)$ provides finite cost), then we may pass to the limit in (5) and conclude that

$$I(s, y; x^*) = U(s, y).$$

This proof extends to our problem, except for the fact that (C) *does not hold for arbitrary admissible trajectories*. (In fact, (C) would require that the terminal cost be upper semicontinuous, while $\varphi_N(\cdot)$ is *lower* semicontinuous.) We will present two solutions for the Hamilton–Jacobi equation in the next section that satisfy conditions (A) and (B). Condition (C) will hold only when the admissible trajectories are carefully matched to the solution in question (cf. Lemma 4.1).

4.1. Hamilton–Jacobi equation; value functions. We now exhibit two solutions for the Hamilton–Jacobi equation and show that each satisfies a version of condition (C). These solutions are used to construct the minimizers described in §3.

The Hamilton–Jacobi (HJ) equation associated with our minimization problem takes the form

$$(6) \quad U_s(s, y) + \inf_{\theta \in \mathfrak{R}} \left[\theta U_y(s, y) + e(s, y) |\theta|^\alpha \right] = 0,$$

which implies the inequality (A). We seek a solution (in some appropriate sense of the word) satisfying

$$(7) \quad \lim_{t \rightarrow T^-} U(t, x) = \varphi_N(x), \quad x \in \mathfrak{R}.$$

The scale invariance property (Lemma 2.1) implies that the optimal cost starting at (s, y) is a function only of the ratio $\rho = y/\sqrt{T-s}$. Hence it is reasonable to look for solutions of the form

$$U(s, y) = h(y/\sqrt{T-s}) = h(\rho).$$

With this restriction, (6) reduces to the following ordinary differential equation:

$$(8) \quad (1/2)\rho h'(\rho) + \inf_{\theta \in \mathfrak{R}} \left[\theta h'(\rho) + e(T-1, \rho) |\theta|^\alpha \right] = 0.$$

The boundary condition (7) requires

$$(9) \quad h(0) = 0, \quad \lim_{\rho \rightarrow +\infty} h(\rho) = N.$$

One function satisfying (8), with $h(0) = 0$ and $h(+\infty) = +\infty$, is given by

$$(10) \quad h(\rho) := (2c_\alpha)^{1-\alpha} \int_0^\rho \eta^{\alpha-1} e(T-1, \eta) d\eta,$$

where

$$c_\alpha := (\alpha - 1)\alpha^{-\alpha/(\alpha-1)}.$$

This function is continuously differentiable and it satisfies (8) even at $\rho = \rho_0$ since $h'(\rho_0) = 0$.

Two functions satisfying (8) as well as the boundary conditions (9) can be constructed by combining constant solutions with shifts of the function $h(\cdot)$. Define $\rho_N^{(1)}$ by

$$h(\rho_N^{(1)}) = h(\rho_0) + N,$$

and set

$$(11) \quad h_1(\rho) := \begin{cases} 0, & 0 \leq \rho \leq \rho_0, \\ h(\rho) - h(\rho_0), & \rho_0 < \rho \leq \rho_N^{(1)}, \\ N, & \rho_N^{(1)} < \rho. \end{cases}$$

(More simply, $h_1(\rho) = 0 \vee [h(\rho) - h(\rho_0)] \wedge N$.) This function is smooth across $\rho = \rho_0$ (since $h'(\rho_0) = 0$) but there is a jump in the first derivative across $\rho = \rho_N^{(1)}$ (both the right- and left-hand derivatives satisfy (8)).

For a second solution, define $\rho_N^{(2)}$ by $h(\rho_N^{(2)}) = N$ and set

$$(12) \quad h_2(\rho) := \begin{cases} h(\rho), & 0 < \rho < \rho_N^{(2)}, \\ N, & \rho_N^{(2)} < \rho \end{cases}$$

(i.e., $h_2(\rho) = h(\rho) \wedge N$).

We have obtained two solutions for (6):

$$U^{(1)}(s, y) := h_1(y/\sqrt{T-s}) \leq h_2(y/\sqrt{T-s}) =: U^{(2)}(s, y).$$

Condition (C) fails to hold for the larger solution $U^{(2)}$. For example, any trajectory that follows Γ on a final interval $[T - \epsilon, T]$ provides

$$U^{(2)}(t, x(t)) = h_2(\rho_0) > 0 = \varphi_N(x(T)), \quad t \in [T - \epsilon, T],$$

and so (C) fails to hold. But (C) does hold for $U^{(2)}$ when we restrict the admissible class to \mathcal{A}_2 . For solution $U^{(1)}$, (C) once again fails to hold for some $x(\cdot) \in \mathcal{A}_1$. In this case, however, all of the trajectories for which (C) fails must incur infinite cost and therefore do not impede the proof of the verification theorem. These results are summarized in the following lemma.

LEMMA 4.1. *If $x(\cdot) \in \mathcal{A}_2$, then*

$$(C2) \quad \lim_{t \rightarrow T^-} U^{(2)}(t, x(t)) = \varphi_N(x(T)).$$

If $x(\cdot) \in \mathcal{A}_1$ and $J(s, y; x) < +\infty$, then

$$(C1) \quad \lim_{t \rightarrow T^-} U^{(1)}(t, x(t)) = \varphi_N(x(T)).$$

Proof. The only difficulty lies in the case $x(T) = 0$, so consider only such trajectories.

If $x(\cdot) \in \mathcal{A}_2$ and $x(T) = 0$, then

$$\frac{x(t)}{\sqrt{T-t}} \rightarrow 0 \quad \text{as } t \rightarrow T^-.$$

Hence

$$\begin{aligned} \lim_{t \rightarrow T^-} U^{(2)}(t, x(t)) &= \lim_{t \rightarrow T^-} h_2(x(t)/\sqrt{T-t}) \\ &= h_2(0) = 0. \end{aligned}$$

This proves (C2).

Next, let $x(\cdot) \in \mathcal{A}_1$ be given such that $x(T) = 0$ and assume that (C1) fails to hold. Then there is a constant ℓ such that

$$\limsup_{t \rightarrow T^-} U^{(1)}(t, x(t)) \geq \ell > 0 = \varphi_N(x(T)).$$

We may choose an increasing sequence $\{t_n\}$ converging to T such that

$$x_n := x(t_n) \geq \rho_1 \sqrt{T - t_n},$$

where $\rho_1 > \rho_0$ is defined by

$$h_1(\rho_1) = \ell/2.$$

Consider now

$$\begin{aligned} J(s, y; x) &= \int_s^{t_n} e(t, x(t)) |\dot{x}(t)|^\alpha dt + \varphi_N(x(T)) \\ &\quad + \int_{t_n}^T e(t, x(t)) |\dot{x}(t)|^\alpha dt. \end{aligned}$$

The final term is just $J(t_n, x_n; x)$, and by Lemma 2.2, we have a positive lower bound for this term:

$$J(t_n, x_n; x) \geq \delta(\rho_1, \rho_0) > 0 \text{ for all } n.$$

But this implies that the running cost incurred by $x(\cdot)$ must be infinite and this completes the proof. \square

With Lemma 4.1 in hand, we have half of the proof for Theorem 3.1; we can show that $U^{(1)}(\cdot, \cdot)$ and $U^{(2)}(\cdot, \cdot)$ are lower bounds for the optimal cost in the appropriate admissible classes. Since these functions satisfy the (HJ) equation (and not just the inequality (A)), they provide a differential equation for the optimal trajectories.

The minimizing θ in (8) (what we called $\theta^*(s, y)$ in the last section) defines a control policy, one for each of the solutions $h_1(\cdot)$ and $h_2(\cdot)$. The control associated with the larger solution $h_2(\cdot)$ is

$$(13) \quad v^{(2)}(t, x) = \begin{cases} -\frac{\alpha}{2(\alpha-1)} \frac{x}{T-t}, & 0 < x < \rho_N^{(2)} \sqrt{T-t}, \\ 0, & \rho_N^{(2)} \sqrt{T-t} < x. \end{cases}$$

The pseudo-minimizer, $x^{(2)}(\cdot; s, y)$, solves the differential equation

$$(14) \quad \dot{x}(t) = v^{(2)}(t, x(t)), \quad x(s) = y.$$

When we consider the smaller of the similarity solutions, we obtain a different control policy:

$$(15) \quad v^{(1)}(t, x) = \begin{cases} 0, & 0 < x < \rho_0 \sqrt{T-t}, \\ -\frac{\alpha}{2(\alpha-1)} \frac{x}{T-t}, & \rho_0 \sqrt{T-t} < x < \rho_N^{(1)} \sqrt{T-t}, \\ 0, & \rho_N^{(1)} \sqrt{T-t} < x. \end{cases}$$

The absolute minimizer $x^{(1)}(\cdot)$ solves the differential equation

$$\dot{x}(t) = v^{(1)}(t, x(t)), \quad x(s) = y.$$

Notice that the control policies are not defined for all values of (t, x) . The first control $v^{(2)}(\cdot, \cdot)$ is not defined for $x = \rho_N^{(2)}\sqrt{T-t}$. Here, we could choose the limiting value from either side without affecting the cost of the trajectory. (Recall that along this curve the cost of getting to $x = 0$ is exactly equal to the terminal cost.)

The second control is not defined for $x = \rho_N^{(1)}\sqrt{T-t}$ or $x = \rho_0\sqrt{T-t}$. The first curve is handled exactly as for the the previous control; the choice does not affect the cost. The second curve is the free zone Γ and we define

$$v^{(1)}(t, x) := -\frac{1}{2} \frac{x}{T-t}, \quad x = \rho_0\sqrt{T-t}.$$

This choice provides $x^{(1)}(t) \in \Gamma$ after the first intersection. We only remark that the specification of $v^{(1)}(\cdot, \cdot)$ on Γ is not really necessary; the control as defined will not allow the trajectory to leave Γ after they intersect.

We conclude this section by discussing the problem with a more general free zone. As a first extension, let Γ consist of finitely many curves. For $0 < \rho_0 < \rho_1 < \rho_2 < \dots < \rho_n$, define

$$\Gamma_j := \left\{ (t, x) : x = \rho_j\sqrt{T-t}, t < T \right\},$$

so that

$$\Gamma = \bigcup_{j=0}^n \Gamma_j.$$

Following the arguments for $\Gamma = \Gamma_0$, one absolute minimizer in \mathcal{A}_1 is defined exactly as $x^{(1)}(\cdot; s, y)$, with ρ_n taking the place of ρ_0 . This trajectory achieves zero cost for the largest set of initial data (s, y) , (all (s, y) below the top parabola Γ_n).

One minimizer in \mathcal{A}_2 is exactly $x^{(2)}(\cdot; s, y)$, as defined in (4), which ignores all of the free zone curves. Once again, since movement along the free zone incurs no cost, and the cost is constant along the free zone curves, an infinite family of pseudo-minimizers is obtained from $x^{(2)}(\cdot)$ by allowing the trajectory to “stick” to Γ for some interval before exiting to follow another pseudo-minimizer to $x = 0$ at the final time.

The construction that provided $h_1(\cdot)$ and $h_2(\cdot)$ now provides a family of 2^n solutions for the (HJ) equation. At each of $\rho = 0, \rho_0, \rho_1, \dots, \rho_{n-1}$, we have the option of following the constant solution or the increasing similarity solution (given in (10)). (If N , the terminal cost, is small, some of these may be truncated.) All are C^1 since $h'(\rho)$ vanishes at $\rho = \rho_j$, for each j .

The smallest of these solutions is the value function when the admissible class is \mathcal{A}_1 . The largest of these solutions (the one with no constant segments before the last) is the value function for \mathcal{A}_2 . Only the solutions with a single switch from the constant (zero) solution to the increasing solution can reflect costs associated with continuous trajectories. For example, for N such that $h(\rho_0) < N$, let $\rho_N^{(0,1)}$ be defined

by $h(\rho_N^{(0,1)}) = N + h(\rho_1) - h(\rho_0)$ and set

$$(16) \quad h_{0,1}(\rho) := \begin{cases} h(\rho), & 0 \leq \rho \leq \rho_0, \\ h(\rho_0), & \rho_0 < \rho \leq \rho_1, \\ h(\rho) - h(\rho_1) + h(\rho_0), & \rho_1 < \rho \leq \rho_N^{(0,1)}, \\ N, & \rho_N^{(0,1)} < \rho. \end{cases}$$

This would define a cost constant on the strip

$$\rho_0 \sqrt{T-t} \leq x \leq \rho_1 \sqrt{T-t},$$

which could only be attained by *jumping* from Γ_1 to Γ_0 .

5. Remarks and extensions. We conclude by discussing some extensions and variations of the problem described in this paper.

If we alter the terminal cost at $x = 0$ to obtain an upper semicontinuous function, the Lavrentiev phenomenon vanishes. This is the case since condition (C) is now satisfied by *both* solutions $U^{(1)}(\cdot, \cdot)$ and $U^{(2)}(\cdot, \cdot)$ (for arbitrary $x(\cdot) \in \mathcal{A}_1(s, y)$).

Define

$$\varphi_N^*(x) := \begin{cases} N, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

and let $J^*(s, y; \cdot)$ denote the new cost function. The upper semicontinuity of $\varphi_N^*(\cdot)$ implies condition (C), and we have

$$U^{(1)}(s, y) < U^{(2)}(s, y) \leq \inf \left\{ J^*(s, y; x); \ x(\cdot) \in \mathcal{A}_1(s, y) \right\}.$$

This is only possible if the smaller solution is not attainable (or approximable) by trajectories in $\mathcal{A}_1(s, y)$. The larger solution $U^{(2)}(s, y)$ is approximable, though not attained, by trajectories in $\mathcal{A}_1(s, y)$. In fact, a minimizing sequence of Lipschitz approximations can be constructed using the pseudo-minimizer $x^{(2)}(\cdot)$ perturbed so as to have $x(T) < 0$. It follows that $U^{(2)}(s, y)$ is the value function for this problem.

It is possible to adapt the techniques discussed in [3] to show that the Lavrentiev phenomenon persists for a large class of perturbed problems. For example, the cost function

$$J(s, y; x) := \int_s^T [e(t, x(t))|\dot{x}(t)|^\alpha + \epsilon g(\dot{x}(t))] dt + \varphi_N(x(T))$$

will continue to exhibit the phenomenon for smooth, nonnegative functions $g(\cdot)$ (satisfying certain growth conditions at infinity) as long as $\epsilon > 0$ is sufficiently small. (This work will appear in [14].) This is important for two reasons. First, these perturbations provide a *fully regular* problem exhibiting the Lavrentiev phenomenon. Second, since the Hamiltonian is now C^2 , this demonstrates that the nonuniqueness of solutions to the (HJ) equation is not simply a consequence of the discontinuity in the Hamiltonian. The key factor is the discontinuity in the final data. (Recall the remarks concerning viscosity solutions in the Introduction.)

The Lavrentiev phenomenon can be demonstrated for a larger class of Bolza problems, where the variational integrand satisfies a more general homogeneity property [14]. In particular, it is possible to consider cost functions of the form

$$(17) \quad J(s, y; x) := \int_s^T f(t, x(t), \dot{x}(t)) dt + \varphi(x(T)),$$

where the integrand satisfies the homogeneity property

$$(18) \quad f(\lambda t + T, \lambda^\gamma x, \lambda^{\gamma-1} u) = \lambda^{-1} f(t + T, x, u), \quad \lambda > 0$$

for some $\gamma \in (0, 1)$, for $t \leq 0, x \in \mathfrak{R}, u \in \mathfrak{R}$. For this class of problems, it is no longer possible to explicitly exhibit the value functions, but the Lavrentiev phenomenon can be verified using “field theory” arguments and a theorem of E. Noether concerning invariant variational problems. (For this see, for example, [22].)

As a final extension, we note that the problem discussed in this paper extends to an example of a *stochastic* control problem exhibiting the Lavrentiev phenomenon. (These results were announced in [13].) The stochastic problem is obtained by adding a “white noise” perturbation to the equation of state. The admissible trajectories are now stochastic processes

$$x(t) = y + \int_s^{t \wedge \tau_0} u(t) dt + \sigma \int_s^{t \wedge \tau_0} dw(t), \quad s \leq t \leq T,$$

where $w = (w(t) : t \geq s)$ is a standard Brownian motion and τ_0 is the exit time

$$\tau_0 = \tau_0(x) := \inf \{ t > s : x(t) \leq 0 \} \wedge T.$$

(Notice that the process is absorbed when it reaches $x = 0$.) The cost function takes the form

$$J(s, y; x) := E_{sy} \left[\int_s^{\tau_0} e(t, x(t)) |u(t)|^\alpha dt + \varphi_N(x(\tau_0)) \right].$$

The Lavrentiev gap is no longer between the costs for integrable and square integrable control processes, but between the cost for what have been called *singular* and absolutely continuous control policies. (See, for example, [16] and [18].)

REFERENCES

- [1] T. S. ANGELL, *A note on approximation of optimal solutions of free problems of the calculus of variations*, Rend. Circ. Mat. Palermo, 2 (1979), pp. 258–272.
- [2] J. BALL AND V. MIZEL, *Singular minimizers for regular one-dimensional problems in the calculus of variations*, Bull. Amer. Math. Soc., 11 (1984), pp. 143–146.
- [3] ———, *One-dimensional variational problems whose minimizers do not satisfy the Euler–Lagrange equation*, Arch. Rational Mech. Anal., 90 (1985), pp. 325–388.
- [4] G. BARLES AND B. PERTHAME, *Discontinuous solutions for deterministic stopping time problems*, preprint.
- [5] E. N. BARRON AND R. JENSEN, *Optimal control problems with no turning back*, J. Differential Equations, 36 (1980), pp. 223–248.
- [6] A. BENSOUSSAN AND J. L. LIONS, *Nouvelles methodes en contrôle impulsionnel*, Appl. Math. Optim. 1 (1975), pp. 289–312.
- [7] L. CESARI, *Optimization—Theory and Applications*, Springer-Verlag, Berlin, New York, 1983.
- [8] F. H. CLARKE AND R. B. VINTER, *Local optimality conditions and Lipschitzian solutions to the Hamilton–Jacobi equation*, SIAM J. Control Optim., 21 (1983), pp. 856–870.

- [9] M.G. CRANDALL AND P.L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc. 277, pp. 1-42.
- [10] M.G. CRANDALL, L.C. EVANS, AND P.L. LIONS, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc. 282, pp. 487-502.
- [11] W.H. FLEMING AND R.W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, New York, 1975.
- [12] A.C. HEINRICHER, *A singular stochastic control problem arising from a deterministic problem with non-Lipschitzian minimizers*, Ph.D. dissertation, Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA (1986).
- [13] A.C. HEINRICHER AND V.J. MIZEL, *A stochastic control problem with different value functions for singular and absolutely continuous control*, in Proc. 1986 IEEE Conference on Decision and Control, Athens, Greece.
- [14] ———, *The Lavrentiev phenomenon for invariant variational problems*, Arch. Rational Mech. Anal., 102 (1988), pp. 57-93.
- [15] H. ISHII, *Perron's method for Hamilton-Jacobi equations*, preprint.
- [16] I. KARATZAS, *A class of singular stochastic control problems*, Adv. in Appl. Probab., 15 (1983), pp. 225-254.
- [17] M. LAVRENTIEV, *Sur quelques problèmes du calcul des variations*, Ann. Mat. Pura Appl., 4 (1926), pp. 107-124.
- [18] J.P. LEHOCZKY AND S.E. SHREVE, *Absolutely continuous and singular stochastic control*, Stochastics, 17 (1986), pp. 91-109.
- [19] P.L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, MA, 1982.
- [20] P.D. LOEWEN, *On the Lavrentiev Phenomenon*, Canad. Math. Bull., 30 (1987), pp. 102-108.
- [21] B. MANIA, *Sopra un esempio di Lavrentieff*, Bull. Un. Mat. Ital., 13 (1934), pp. 146-153.
- [22] P.J. OLVER, *Applications of Lie Groups to Differential Equations*, Springer-Verlag, Berlin, New York, 1986.