

THE FIRST ORDER NECESSARY CONDITIONS FOR NONSMOOTH VARIATIONAL AND CONTROL PROBLEMS*

HALINA FRANKOWSKA†

Abstract. The first order necessary conditions for nonsmooth problems of Lagrange, Bolza, multiple integral and optimal control are given. The necessary conditions are expressed in terms of differential inclusions. We deal with some new differentials that generalize the differential of Fréchet. In particular the Pontryagin maximum principle is proved without the assumption of continuous differentiability.

Key words. control theory, generalized derivative, maximum principle, nonsmooth problems, variational principle

1. Introduction. The study of nonsmooth problems in calculus of variations, mathematical programming and control theory led to various extensions of the notion of derivative (see for example [2], [3], [10], [13], [16] and [18]).

The generalizations of this notion are related to the method of attacking the problem. The dual theory of Rockafellar [16], the reduction technique of Clarke [2], [3], [4], the approximative method of Warga [18], [19], the interior mapping theorem of Halkin [10] suggested different extensions of the usual differential.

In the present paper we apply Ekeland's variational principle to several nonsmooth problems. For our purposes we need to introduce some new objects, generalizing the Fréchet differential.

The following theorem is a version of Ekeland's variational principle and can be proved exactly in the same way as [7, Thm. 2.2].

THEOREM 1.1. *Let E be a Banach space, $f: E \rightarrow \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous (l.s.c.) function, and $f(v_0) \leq \inf f + \varepsilon$, where $\varepsilon > 0$, $v_0 \in E$. Assume that for some $\lambda > 0$, f is Gâteaux differentiable on the open ball $B(v_0, \lambda)$ of centre v_0 and radius λ . Then there exists $v \in B(v_0, \lambda)$ such that*

$$(1.2) \quad f(v) \leq f(v_0),$$

$$(1.3) \quad \|f'(v)\|^* \leq 2\varepsilon/\lambda$$

where f' denotes the Gâteaux derivative of f and $\|\cdot\|^*$ is the (usual) norm in the dual space E^* .

From the above theorem we obtain certain "approximate" necessary conditions. A limit procedure then leads to the necessary conditions in the form of inclusions. Minimal elements in the family of sets obtained by such a process are proposed as a generalization of the differential. In general, minimal elements are not unique. But in the case of a Fréchet differentiable (or convex) function $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$, our "derivatives" coincide with the usual differential (subdifferential).

We generalize some results from [1], [2], [3], [5], [8] and [16]. The plan of the paper is as follows. In §2 preliminary results are proved and some notions are introduced. Their application to the multiple integral problem and to the problems of Bolza and Lagrange is given in §3. Section 4 is devoted to the maximum principle of Pontryagin.

2. Preliminary results. Consider a Banach space $(E, \|\cdot\|)$ and its dual $(E^*, \|\cdot\|^*)$. The canonical bilinear form on $E^* \times E$ will be denoted by $\langle \cdot, \cdot \rangle$; $B(v, r)(\bar{B}(v, r))$ will

* Received by the editors February 9, 1982, and in revised form September 29, 1982.

† International School for Advanced Studies (ISAS), Trieste, Italy. On leave of absence from Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland.

denote the open (respectively closed) ball in E of centre v and radius r . We denote by V a closed subset of E .

In this paper we deal with functions f on V assuming values either in $\bar{R} = R \cup \{+\infty\}$ or in R^m , $m \geq 1$, and satisfying the following condition for a fixed $v_0 \in V$:

Condition 2.1. There exist $\lambda_i > 0$, $f_i: E \rightarrow \bar{R} (R^m)$, $i = 1, 2, \dots$, such that f_i is Gâteaux differentiable on $B(v_0, \lambda_i)$, $f_i(v_0) = f(v_0)$, $\lambda_i \rightarrow 0$ and for some $\varepsilon_i > 0$ satisfying $\varepsilon_i/\lambda_i \rightarrow 0$ either

- (a) $f_i: E \rightarrow \bar{R}$ is l.s.c., $f(v) \leq f_i(v) + \varepsilon_i$ for $v \in V \cap \bar{B}(v_0, \lambda_i)$, or
- (b) f_i is continuous, $|f_i(v) - f(v)| \leq \varepsilon_i$ on $V \cap \bar{B}(v_0, \lambda_i)$

holds for each i .

The sets

$$\bigcap_{N \geq 1} \text{cl} \bigcup_{i \geq N} \{f'_i(v): v \in B(v_0, \lambda_i) \cap V\},$$

$$\bigcap_{N \geq 1} \text{clconv} \bigcup_{i \geq N} \{f'_i(v): v \in B(v_0, \lambda_i) \cap V\}$$

will be denoted by $\mathcal{P}_af(v_0)$, $\mathcal{CP}_af(v_0)$ if (a), or by $\mathcal{P}_bf(v_0)$, $\mathcal{CP}_bf(v_0)$ if (b) holds, respectively. Clearly if $m = 1$ then every set $\mathcal{P}_bf(v_0) (\mathcal{CP}_bf(v_0))$ is also $\mathcal{P}_af(v_0) (\mathcal{CP}_af(v_0))$. The above sets depend on the particular sequences $\{f_i\}$, $\{\lambda_i\}$ which were chosen.

In general it may happen that the above sets are empty. But if $f: V \rightarrow R^m$ is F-differentiable (Fréchet differentiable) at $v_0 \in \text{Int } V$, then we may put $f_i(v) = f(v_0) + \langle f'(v_0), v - v_0 \rangle$ for each i and fix any $\lambda_i \rightarrow 0$. Then $\{f_i\}$, $\{\lambda_i\}$ define a set $\mathcal{P}_bf(v_0)$ equal to $\{f'(v_0)\}$. This proves:

PROPOSITION 2.2. *If $f: V \rightarrow R^m$ is F-differentiable at $v_0 \in \text{Int } V$ then $\{f'(v_0)\}$ is a set $\mathcal{P}_bf(v_0)$, $\mathcal{CP}_bf(v_0)$.*

Assume $f: V \rightarrow \bar{R}$ has a local minimum at v_0 , i.e. $f(v_0) = \min \{f(v): v \in B(v_0, r) \cap V\}$ for some $r > 0$. Set $V_1 = \bar{B}(v_0, r/2) \cap V$. V_1 is closed and the restriction $f|_{V_1}$ of f to V_1 achieves its minimum at v_0 . Therefore it is enough to study the necessary conditions for the global minimum only.

THEOREM 2.3. *Assume $f: V \rightarrow \bar{R}$ achieves its minimum at $v_0 \in \text{Int } V$. Then every set $\mathcal{P}_af(v_0)$ contains 0.*

Proof. Let $\{f_i\}$, $\{\lambda_i\}$, $\{\varepsilon_i\}$ be as in Condition 2.1(a). Set $\bar{f}_i(v) = f_i(v)$ for $v \in V \cap \bar{B}(v_0, \lambda_i)$, $\bar{f}_i(v) = +\infty$ otherwise. Then $\bar{f}_i(v_0) = f(v_0) = \min_V f \leq \inf \bar{f}_i + \varepsilon_i$. From Theorem 1.1 it follows that for large i and some $v_i \in B(v_0, \lambda_i)$ conditions (1.2), (1.3) hold with f , v , λ , ε replaced by f_i , v_i , λ_i , ε_i respectively. Hence the result.

COROLLARY 2.3.1. *If $f: V \rightarrow \bar{R}$ is F-differentiable at $v_0 \in \text{Int } V$ then $f'(v_0)$ is contained in every set $\mathcal{P}_af(v_0)$.*

Proof. Without any loss of generality we may assume $f(v_0) = 0$, $f'(v_0) = 0$. Set $f_1(v) = |f(v)|$ for $v \in V$. Then f_1 achieves the minimum at v_0 . Fix $\{f_i\}$, $\{\lambda_i\}$, $\{\varepsilon_i\}$ as in Condition 2.1(a). We may assume $B(v_0, \lambda_i) \subset V$ for every i . Set $\bar{f}_i(v) = f_i(v)$ for $v \in \bar{B}(v_0, \lambda_i)$, $\bar{f}_i(v) = +\infty$ otherwise. Then $f_1(v) \leq f(v) + 2|f(v)| \leq \bar{f}_i(v) + \varepsilon_i + 2 \sup \{|f(v)|: v \in \bar{B}(v_0, \lambda_i)\}$. $\{\bar{f}_i\}$, $\{\lambda_i\}$ define a set $\mathcal{P}_{af_1}(v_0) = \mathcal{P}_af(v_0)$. An application of Theorem 2.3 completes the proof.

Assume V is convex, $f: V \rightarrow R$ is convex, $v_0 \in V$. Set subgrad $f(v_0) = \{p \in E^*: f(v) - f(v_0) \geq \langle p, v - v_0 \rangle \text{ for } v \in V\}$.

COROLLARY 2.3.2. *Assume V is convex, $f: V \rightarrow R$ is convex, $v_0 \in \text{Int } V$. Then subgrad $f(v_0)$ is contained in every set $\mathcal{P}_af(v_0)$.*

Proof. If $p \in \text{subgrad } f(v_0)$ then the function $\phi(v) = f(v) - f(v_0) - \langle p, v - v_0 \rangle$ achieves its minimum at v_0 . Fix $\{f_i\}$, $\{\lambda_i\}$ as in Condition 2.1(a) and set $\phi_i(v) =$

$f_i(v) - f(v_0) - \langle p, v - v_0 \rangle$. By Theorem 2.3, $0 \in \mathcal{P}_a \phi(v_0) = \mathcal{P}_a f(v_0) - p$. So the proof is complete.

The proof of Corollary 2.3.2 is similar to the proof of [19, Thm. 2.10].

Fix $v \in V$. Every set $W = W(v, V)$ consisting of such $w \in E$ that $v + tw \in V$ for all $t \in [0, \delta]$, where $\delta = \delta(w) > 0$, is called a cone of admissible directions at v .

To study the necessary conditions for minimum at a boundary point $v_0 \in V$ we assume

Condition 2.4. There exist a nonempty $W \subset E$ and $r > 0$ such that for every $v \in B(v_0, r) \cap V$ the set W is a cone of admissible directions at v .

THEOREM 2.5. Assume $f: V \rightarrow \bar{R}$ achieves the minimum at v_0 , Conditions 2.1 and 2.4 are satisfied and for every choice of $v_i \in B(v_0, \lambda_i) \cap V$, $i = 1, 2, \dots$, the sequence $\{f'_i(v_i)\}$ has a weak cluster point. Then there exist $v_i \in B(v_0, \lambda_i) \cap V$ such that $\{f'_i(v_i)\}$ has a weak cluster point g , $\langle g, w \rangle \geq 0$ for all $w \in W$. Moreover $g \in \mathcal{CP}_a f(v_0)$ if (a), or $g \in \mathcal{CP}_b f(v_0)$ if (b) holds, respectively.

Proof. As in [7, proof of Thm. 2.2], we verify the existence of $v_i \in B(v_0, \lambda_i) \cap V$ such that $\langle f'_i(v_i), w \rangle \geq -\varepsilon_i 2\|w\|/\lambda_i$ for $w \in W$. By assumptions, there exists a subsequence of indexes $\{i_j\}$ such that $f'_{i_j}(v_{i_j}) \rightarrow g$ (weakly). Clearly $\langle g, w \rangle \geq 0$ for $w \in W$. An application of Mazur's lemma [8] completes the proof.

From now until the end of the section we assume $E = R^n$. Let $f: V \rightarrow R^m$, $v_0 \in \text{Int } V$. It follows from the definition of the Halkin screen [10] that every nonempty set $\mathcal{P}_b f(v_0)$ generated by $f_i \in C^1(B(v_0, \lambda_i))$ is a screen and the closure of any screen of f at v_0 contains a set $\mathcal{P}_b f(v_0)$. Hence every set $\Lambda f(v_0)$ corresponding to a derivative container of Warga (see [19]) contains some $\mathcal{P}_b f(v_0)$. In particular, this holds for Clarke's generalized gradient of a locally Lipschitzian function [3].

Consider $f: V \rightarrow \bar{R}(R^m)$, $v_0 \in V$ and the family of all sets $\mathcal{P}_a f(v_0)(\mathcal{P}_b f(v_0), \mathcal{CP}_a f(v_0), \mathcal{CP}_b f(v_0))$ ordered by the relation of inclusion.

LEMMA 2.6. For each linearly ordered set $\{\mathcal{P}_\tau\}_{\tau \in T}$ of elements $\mathcal{P}_a f(v_0)(\mathcal{P}_b f(v_0), \mathcal{CP}_a f(v_0), \mathcal{CP}_b f(v_0))$ there is an element $\mathcal{P}_a^0 f(v_0)(\mathcal{P}_b^0 f(v_0), \mathcal{CP}_a^0 f(v_0), \mathcal{CP}_b^0 f(v_0))$ contained in $\mathcal{P} = \bigcap_{\tau \in T} \mathcal{P}_\tau$.

Proof. Let $\tau(0) \in T$. For all $j \geq 1$ let $\tau(j)$ be such that $\tau(j) \geq \tau(j-1)$, $\mathcal{P}_{\tau(j)} = \mathcal{P}_j$, $\mathcal{P}_j \cap \bar{B}(0, j) \subset \{p \in \bar{B}(0, j) : \text{dist}(p, \mathcal{P}) \leq 1/j\}$. (We assume $\text{dist}(p, \mathcal{P}) = 1$ when $\mathcal{P} = \emptyset$.) Then $\mathcal{P}_1 \supset \mathcal{P}_2 \supset \dots$. It is enough to show $\bigcap_{\tau \in T} \mathcal{P}_\tau = \bigcap_{j \geq 1} \mathcal{P}_j$ contains some $\mathcal{P}_a^0 f(v_0)(\mathcal{P}_b^0 f(v_0), \mathcal{CP}_a^0 f(v_0), \mathcal{CP}_b^0 f(v_0))$. Let $\{f'_i\}$, $\{\lambda'_i\}$, $\{\varepsilon'_i\}$ define \mathcal{P}_j and let $i(j)$ be such that $\varepsilon'_{i(j)}/\lambda'_{i(j)} < 1/2^j$, $\{(f'_{i(j)})'(v) : v \in B(v_0, \lambda'_{i(j)}) \cap V\} \cap B(0, j) \subset \{x \in R^{nm} : \text{dist}(x, \mathcal{P}_j f(v_0)) < 1/2^j\} \cap B(0, j)$. We may assume $\lambda'_{i(j)} \rightarrow 0$ as $j \rightarrow \infty$ and set $\bar{\lambda}_j = \lambda'_{i(j)}$, $\bar{\varepsilon}_j = \varepsilon'_{i(j)}$, $\bar{f}_j = f'_{i(j)}$. Then $\{\bar{f}_j\}$, $\{\bar{\lambda}_j\}$, $\{\bar{\varepsilon}_j\}$ define the required set.

From the Kuratowski-Zorn lemma we deduce

COROLLARY 2.6.1. In the family of all sets $\mathcal{P}_a f(v_0)(\mathcal{P}_b f(v_0), \mathcal{CP}_a f(v_0), \mathcal{CP}_b f(v_0))$ there exist minimal elements.

DEFINITION. We shall denote minimal elements in the family of all sets $\mathcal{P}_a f(v_0), \mathcal{P}_b f(v_0), \mathcal{CP}_a f(v_0), \mathcal{CP}_b f(v_0)$ by $\mathcal{A}f(v_0), \mathcal{B}f(v_0), \mathcal{C}f(v_0), \mathcal{D}f(v_0)$, respectively.

In general there may be more than one such minimal element, but from Proposition 2.2 and Corollary 2.3.1. follows

COROLLARY 2.6.2. If $f: V \rightarrow R$ is F-differentiable at $v_0 \in \text{Int } V$, then $\mathcal{A}f(v_0) = \mathcal{B}f(v_0) = \mathcal{C}f(v_0) = \mathcal{D}f(v_0) = \{f'(v_0)\}$.

Let $V \subset R^n$ be convex, $f: V \rightarrow R$ be convex, continuous, $v_0 \in \text{Int } V$. By [15, Prop. 7.3] there exist locally uniformly Lipschitzian $\{f_i\}$ and $\{\lambda_i\}$ generating $\mathcal{P}_b f(v_0) \subset \text{subgrad } f(v_0)$. From this and Corollary 2.3.2 follows

COROLLARY 2.6.3. Under the above assumptions $\mathcal{A}f(v_0) = \mathcal{B}f(v_0) = \mathcal{C}f(v_0) = \mathcal{D}f(v_0) = \text{subgrad } f(v_0)$.

Note that if a nonempty set $\mathcal{C}f(v_0)(\mathcal{D}f(v_0))$ is bounded then we may assume that it is generated by $\{f_i\}$, $\{\lambda_i\}$ of bounded $\{f'_i(v): v \in B(v_0, \lambda_i) \cap V, i \geq 1\}$. The following two lemmas characterize the F-differentiability.

LEMMA 2.7. *Let $f: V \rightarrow \mathbb{R}^m$, $v_0 \in \text{Int } V$ and A be an $(m \times n)$ -matrix. Assume $\mathcal{D}f(v_0) = \{A\}$ is generated by $\{f_i\}$, $\{\lambda_i\}$ and $\lambda_{i+1} \leq \lambda_i \leq \rho \lambda_{i+1}$ for all i and some $\rho > 0$. Then A is the F-differential of f at v_0 .*

Proof. By assumption, $\text{dist}(A, \{f'_i(v): v \in B(v_0, \lambda_i) \cap V\}) = d_i \rightarrow 0$ as $i \rightarrow \infty$ and $\lim_{i \rightarrow \infty} \varepsilon_i / \lambda_{i+1} = 0$. For all large i and $v \in B(v_0, \lambda_i) \setminus B(v_0, \lambda_{i+1})$ we have $f(v) - f(v_0) \in f_i(v) - f_i(v_0) + B(0, \varepsilon_i) = \int_0^1 f'(v_0 + \theta(v - v_0))(v - v_0) d\theta + B(0, \varepsilon_i) \subset A(v - v_0) + B(0, d_i|v - v_0|) + B(0, |v - v_0|\varepsilon_i / \lambda_{i+1})$. Hence the result.

LEMMA 2.8. *If $f: V \rightarrow \mathbb{R}^m$ is F-differentiable at $v_0 \in \text{Int } V$, then $\mathcal{D}f(v_0) = \{f'(v_0)\}$.*

Proof. Assume $\{f_i\}$, $\{\lambda_i\}$ generate $\mathcal{D}f(v_0)$. From Proposition 2.2 it follows that it is enough to show that $f'(v_0) \in \mathcal{D}f(v_0)$. We adopt the idea of [20, proof of Thm. 4]. Let A be an arbitrary $(m \times n)$ -matrix and let a_j be its j th row.

Let f^j be the j th coordinate of f and set $\theta = 1/(1 + \max_j |a_j|)$, $\phi(t) = \sum_{j=1}^m f^j(v_0 + ta_j^T)$ for $t \in \mathbb{R}^1$. The sequence of functions $\mathbb{R}^1 \ni t \rightarrow \sum_{j=1}^m f^j_i(v_0 + ta_j^T)$ and $\{\lambda_i \theta\}$ define a set $\mathcal{P}_b \phi(0)$. By Corollary 2.3.1, $\phi'(0) = \sum_{j=1}^m \langle a_j, (f^j)'(v_0) \rangle \in \mathcal{P}_b \phi(0)$. Hence for every $(m \times n)$ -matrix A $A \odot f'(v_0) \subset A \odot \mathcal{P}_b \phi(0) \subset A \odot \mathcal{D}f(v_0)$, where \odot denotes the scalar product of two $(m \times n)$ -matrices viewed as elements of \mathbb{R}^{nm} . From the theorem about the separation of disjoint convex sets it follows that $f'(v_0) \in \mathcal{D}f(v_0)$.

COROLLARY 2.8.1. *Let $f: V \rightarrow \mathbb{R}^m$, $v_0 \in \text{Int } V$. Then the intersection of any two bounded nonempty sets $\mathcal{D}f(v_0)$ defined by $\{f_i\}$, $\{\lambda_i\}$ and $\{\tilde{f}_i\}$, $\{\tilde{\lambda}_i\}$ is nonempty.*

Proof. Set $g_i = f_i - \tilde{f}_i$. Then $\{g_i\}$, $\{\lambda_i\}$ satisfy Condition 2.1(b) for $f = 0$. Our conclusion follows from Lemma 2.8.

Remark 2.9. A question arises if to a given $\{\lambda_i\}$ as in Condition 2.1 corresponds at most one bounded set $\mathcal{D}f(v_0)$. The answer is positive in the case where $V \subset \mathbb{R}^1$. The proof follows from Corollary 2.8.1 by simple contradiction arguments and will appear elsewhere. We do not know the answer in the general situation.

3. Applications to the multiple integral and other problems. Let an open set $\Omega \subset \mathbb{R}^n$ and a function $L: \Omega \times \mathbb{R}^d \times \mathbb{R}^{dn} \rightarrow \bar{\mathbb{R}}$ be given. Consider the multiple integral problem

$$(\mathcal{M}) \quad \text{minimize} \left\{ F(v) = \int_{\Omega} L(x, v(x), \nabla v(x)) dx : v \in V \right\}$$

where V is a closed subset of the Sobolev space $H_1^1(\Omega)$ (see [14]). Here we assume that F is well defined. Our tool will be certain admissible families of sets $\mathcal{CP}_a L(x, \cdot, \cdot)$.

DEFINITION 3.1. Let $v \in H_1^1(\Omega)$ be such that $x \rightarrow L(x, v(x), \nabla v(x))$ is in $L^1(\Omega)$. The family of sets $\{\mathcal{C}L(x): x \in \Omega\}$ is called *admissible along v* if there exist $\lambda_i > 0$, integrable $\varepsilon_i: \Omega \rightarrow \mathbb{R}_+$, $L_i: \Omega \times \mathbb{R}^d \times \mathbb{R}^{dn} \rightarrow \bar{\mathbb{R}}$, $i = 1, 2, \dots$, such that for some $k \in L^1(\Omega)$

$$(i) \quad \int_{\Omega} \varepsilon_i(x) dx / \lambda_i \rightarrow 0 \quad \text{as } i \rightarrow \infty,$$

(ii) L_i is measurable in x , $L_i(x, \cdot, \cdot)$ is Lipschitzian of constant $k(x)$ on $B((v(x), \nabla v(x)), \lambda_i)$,

(iii) $\{L_i(x, \cdot, \cdot), \{\lambda_i\}, \{\varepsilon_i(x)\}$ define a set $\mathcal{CP}_a L(x, v(x), \nabla v(x)) = \mathcal{C}L(x)$ for almost every $x \in \Omega$. Here V of Condition 2.4 is the space \mathbb{R}^{d+dn} ,

(iv) $\{\mathcal{C}L(x): x \in \Omega\}$ is minimal, i.e. for every choice of $\{L_i\}$, $\{\lambda_i\}$, $\{\varepsilon_i\}$ satisfying (i)–(iii) that define sets $\mathcal{CP}_a L(x, v(x), \nabla v(x)) \subset \mathcal{C}L(x)$ we have $\mathcal{CP}_a L(x, v(x), \nabla v(x)) = \mathcal{C}L(x)$ almost everywhere in Ω .

Remark 3.2. (a) Condition (iv) makes sense. Indeed, as in § 2 it is enough to show that, for each linearly ordered set T and the family $\{\mathcal{P}_\tau\}_{\tau \in T}$ of elements $\{C\mathcal{P}_a^\tau L(x, v(x), \nabla v(x)) : x \in \Omega\}$ defined by sequences $\{L_i^\tau\}, \{\lambda_i^\tau\}, \{\varepsilon_i^\tau\}$ satisfying (i)–(iii) for all $\tau \in T$ and such that $C\mathcal{P}_a^\tau L(x, v(x), \nabla v(x)) \subset C\mathcal{P}_a^\chi L(x, v(x), \nabla v(x))$ in Ω for all $\tau < \chi$, there are $\{L_j^0\}, \{\lambda_j^0\}, \{\varepsilon_j^0\}$ as in (i)–(iii) and

$$C\mathcal{P}_a^0 L(x, v(x), \nabla v(x)) \subset \bigcap_{\tau \in T} C\mathcal{P}_a^\tau L(x, v(x), \nabla v(x)) \quad \text{a.e. in } \Omega.$$

For every integer $j \geq 1$ let $A_j \subset \Omega$, $\tau(j) \in T$, be such that the Lebesgue measure of A_j $\mu(A_j) < 1/j$ and $\text{dist}(\bigcap_{\tau \in T} C\mathcal{P}_a^\tau L(x, v(x), \nabla v(x)), C\mathcal{P}_a L(x, v(x), \nabla v(x))) < 1/j$ for all $x \in \Omega \setminus A_j$. Let $i(j)$, $B_i \subset \Omega \setminus A_j$ be such that $\mu(B_i) < 1/j$,

$$\int_{\Omega} \varepsilon_{i(j)}^{\tau(j)}(x) dx / \lambda_{i(j)}^{\tau(j)} < 1/j, \quad \lambda_{i(j)}^{\tau(j)} < 1/j \quad \text{and}$$

$$\{\partial L_{i(j)}^{\tau(j)}(x, s, w) : (s, w) \in B((v(x), \nabla v(x)), \lambda_{i(j)}^{\tau(j)})\} \subset C\mathcal{P}_a^{\tau(j)} L(x, v(x), \nabla v(x)) + B(0, 1/j)$$

for almost every $x \in \Omega \setminus (B_i \cup A_j)$. To achieve the proof it is enough to set

$$L_j^0 = L_{i(j)}^{\tau(j)}, \quad \lambda_j^0 = \lambda_{i(j)}^{\tau(j)}, \quad \varepsilon_j^0 = \varepsilon_{i(j)}^{\tau(j)}.$$

(b) If $L(x, \cdot, \cdot)$ has an F-differential $\partial L(x, v(x), \nabla v(x))$ at $(v(x), \nabla v(x))$ almost everywhere and for some $k \in L^1(\Omega)$ and all $(s, w) \in B(0, r)$ the function $x \rightarrow L(x, v(x) + s, \nabla v(x) + w)$ is measurable and $|L(x, v(x) + s, \nabla v(x) + w) - L(x, v(x), \nabla v(x))| \leq k(x)|s| + |w|$, then $\{\partial L(x, v(x), \nabla v(x)) : x \in \Omega\}$ is the only admissible family along v .

(c) If L is measurable in x and for some $r > 0$ and $k \in L^1(\Omega)$, $L(x, \cdot, \cdot)$ is Lipschitzian of constant $k(x)$ on $B((v(x), \nabla v(x)), r)$, then from the proof of [19, Thm. 25] follows the existence of an admissible $\{\mathcal{C}L(x) : x \in \Omega\}$ along v such that $\mathcal{C}L(x)$ is contained in Clarke's generalized gradient of $L(x, \cdot, \cdot)$ at $(v(x), \nabla v(x))$.

Assume Ω is bounded and denote by $C^1(\bar{\Omega})$ the Banach space of continuously differentiable functions $v : \bar{\Omega} \rightarrow \mathbb{R}^d$ with the norm $\|v\| = \sup_{x \in \bar{\Omega}} |v(x)| + \sup_{x \in \bar{\Omega}} |\nabla v(x)|$, where $\bar{\Omega}$ is the closure of Ω .

If v_0 solves (\mathcal{M}) then we may always assume that $V \subset v_0 + C^1(\bar{\Omega})$.

THEOREM 3.3. Assume Ω is bounded, $V \subset v_0 + C^1(\bar{\Omega})$ and $F(v_0) = \min_V F < \infty$. Suppose that Condition 2.4 holds and $\{\mathcal{C}L(x) : x \in \Omega\}$ is admissible along v_0 . Then there exist $\xi^0, \xi \in L^1(\Omega)$ such that

$$(3.4) \quad (\xi^0, \xi)(x) \in \mathcal{C}L(x) \text{ a.e.,}$$

$$(3.5) \quad \int_{\Omega} (\xi^0(x)w(x) + \xi(x)\nabla w(x)) dx \geq 0 \quad \text{on } W$$

where W is the cone of admissible directions of Condition 2.4.

Proof. Without any loss of generality we may assume $v_0 = 0$. Thus $V \subset C^1(\bar{\Omega})$. Let $\{L_i\}, \{\lambda_i\}$ define $\{\mathcal{C}L(x) : x \in \Omega\}$. Set $F_i(v) = \int_{\Omega} L_i(x, v(x), \nabla v(x)) dx$ for $\|v\| \leq \lambda_i$, $F_i(v) = +\infty$ otherwise. Then $F_i : C^1(\bar{\Omega}) \rightarrow \mathbb{R}$ is Gâteaux differentiable on the open ball $B(0, \lambda_i) \subset C^1(\bar{\Omega})$, and

$$\langle F_i'(v), w \rangle = \int_{\Omega} \partial_2 L_i(x, v(x), \nabla v(x))w(x) dx + \int_{\Omega} \partial_3 L_i(x, v(x), \nabla v(x))\nabla w(x) dx$$

for $v \in B(0, \lambda_i)$, $w \in C^1(\bar{\Omega})$. By the Dunford–Pettis criterion on weak convergence in L^1 , Mazur's lemma (see [8]) and Theorem 2.5 there exist $\xi^0, \xi \in L^1(\Omega)$ such that (3.4)

holds and $(\xi^0, \xi)(x) \in \bigcap_{N \geq 1} \text{clconv} \bigcup_{i \geq N} \{\partial L(x, s, v) : (s, v) \in B((0, 0), \lambda_i)\} = \mathcal{E}L(x)$ almost everywhere in Ω .

COROLLARY 3.3.1. *Let Ω be an open set, $V \subset H_1^1(\Omega)$, $F(v_0) = \min_V F$. Suppose that Condition 2.4 holds with $E = H_1^1(\Omega)$, $W = C_0^\infty(\Omega)$. If $\{\mathcal{E}L(x) : x \in \Omega\}$ is admissible along v_0 then there exists $\xi \in L^1(\Omega)$ such that $\text{div } \xi \in L^1(\Omega)$ and*

$$(\text{div } \xi, \xi)(x) \in \mathcal{E}L(x) \quad \text{a.e. in } \Omega.$$

Proof. If Ω is bounded, then by Theorem 3.3 $\langle \xi^0, w \rangle - \langle \text{div } \xi, w \rangle \geq 0$ for all $w \in W$ (in the sense of distributions). Thus $\xi^0 = \text{div } \xi$ almost everywhere. Otherwise consider a sequence of bounded open sets $\Omega_1 \subset \Omega_2 \subset \dots$ such that $\bigcup_{i \geq 1} \Omega_i = \Omega$. Then $\int_{\Omega_i} L(x, v_0(x), \nabla v_0(x)) dx = \min \{ \int_{\Omega_i} L(x, v(x), \nabla v(x)) dx : v \in (v_0 + H_{1,0}^1(\Omega_i)) \cap V \}$. By the first part of the proof there exists $\xi_i \in L^1(\Omega_i)$ such that $(\text{div } \xi_i, \xi_i)(x) \in \mathcal{E}L(x)$ almost everywhere in Ω_i . Set $\xi_i(x) = 0$ on $\Omega \setminus \Omega_i$. By the Dunford–Pettis criterion, some subsequences $\{\xi_{i_j}\}$, $\{\text{div } \xi_{i_j}\}$ converge weakly to ξ , $\text{div } \xi \in L^1(\Omega)$ respectively. An application of Mazur's lemma completes the proof.

The theorem from [5] follows from Corollary 3.3.1 and Remark 3.2(c).

COROLLARY 3.3.2. *Under all assumptions of Theorem 3.3 assume $V = \{v \in v_0 + C^1(\bar{\Omega}) : (v(x), \nabla v(x)) \in U(x) \text{ for all } x \in \Omega\}$, where U is a multifunction from Ω into closed subsets of \mathbb{R}^{d+dn} . Further assume that for any open set $\Omega' \subset \{x \in \Omega : (v_0(x), \nabla v_0(x)) \in \text{Int } U(x)\}$ the set $C_0^\infty(\Omega') \subset W$. Then the assertion of Theorem 3.3 is valid with $\xi^0(x) = \text{div } \xi(x)$ for $x \in \text{Int } \{y \in \Omega : (v_0(y), \nabla v_0(y)) \in \text{Int } U(y)\}$.*

[1, Thm. 1] follows from the above corollary.

Next we consider the so-called Lagrange problem. Let \mathcal{L} be a closed subset of \mathbb{R}^{2n} . Assume in (\mathcal{M}) that $\bar{\Omega} = [0, 1]$ and V is the closed set $\{v \in H_1^1([0, 1]) : (v(0), v(1)) \in \mathcal{L}\}$.

THEOREM 3.6. *Assume v_0 solves (\mathcal{M}) , where V is as described above and $\{\mathcal{E}L(x) : x \in [0, 1]\}$ is admissible along v_0 . Let $W = W((v_0(0), v_0(1)), \mathcal{L})$ be the cone of all admissible directions for \mathcal{L} at $(v_0(0), v_0(1))$. Then there exists an absolutely continuous (a.c.) function $\xi : [0, 1] \rightarrow \mathbb{R}^n$ such that*

$$(3.7) \quad (\dot{\xi}(x), \xi(x)) \in \mathcal{E}L(x) \quad \text{a.e.},$$

$$(3.8) \quad \langle (-\xi(0), \xi(1)), y \rangle \geq 0$$

for all $y \in W$.

Proof. By Theorem 3.1 relations (3.4) and (3.5) hold for some $\xi^0, \xi \in L^1([0, 1])$. Hence, in particular, $\int_0^1 (\xi^0(x)w(x) + \xi(x)\dot{w}(x)) dx = 0$ for all $w \in C_0^\infty([0, 1])$. Integrating by parts we obtain $\int_0^1 (\xi(x) - \int_0^x \xi^0(t) dt) \times \dot{w}(x) dx = 0$ for all $w \in C_0^\infty([0, 1])$. Therefore $\xi(x) = \xi(0) + \int_0^x \xi^0(t) dt$ almost everywhere for some $\xi(0) \in \mathbb{R}^n$. We may assume that the last equality holds everywhere and thus (3.7) is valid. Moreover $\int_0^1 (\xi(x)w(x) + \xi(x)\dot{w}(x)) dx = \xi(1)w(1) - \xi(0)w(0) = \langle (-\xi(0), \xi(1)), (w(0), w(1)) \rangle \geq 0$ for all $w \in C^1([0, 1])$ with $(w(0), w(1)) \in W$.

Next we apply the above results to the Bolza problem:

$$(B) \quad \text{minimize} \left\{ F(v) = l(v(0), v(1)) + \int_0^1 L(x, v(x), \dot{v}(x)) dx : v \in A \right\}$$

where $l : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$ and A is the set of all a.c. functions.

THEOREM 3.9. *Assume v_0 solves (B) , $F(v_0) < +\infty$, L has an admissible $\{\mathcal{E}L(x) : x \in [0, 1]\}$ along v_0 defined by $\{L_i\}$, $\{\lambda_i\}$ and l has a set $\mathcal{A}l(v_0(0), v_0(1))$ generated by $\{l_i\}$, $\{\lambda_i\}$ with bounded $\{l'_i(y) : y \in \mathbb{R}^{2n}, |y - (v_0(0), v_0(1))| \leq \lambda_i, i \geq 1\}$. Then for some*

a.c. $\xi: [0, 1] \rightarrow R^n$, (3.7) holds and

$$(3.10) \quad (\xi(0), -\xi(1)) \in \mathcal{A}(v_0(0), v_0(1)).$$

Proof. We may assume that $v_0 = 0$. Then $F(0) = \min \{F(v): v \in C^1([0, 1])\}$. Set $F_i(v) = l_i(v(0), v(1)) + \int_0^1 L_i(x, v(x), \dot{v}(x)) dx$ for $\|v\| \leq \lambda_i$ and $F_i(v) = +\infty$ otherwise. As in Theorem 3.3 we verify the existence of $\xi^0, \xi \in L^1([0, 1])$ satisfying (3.4) and $\eta^0, \eta^1 \in R^n$ satisfying $(\eta^0, \eta^1) \in \mathcal{A}(0, 0)$ such that $\langle (\eta^0, \eta^1), (w(0), w(1)) \rangle + \int_0^1 (\xi^0(x)w(x) + \xi(x)\dot{w}(x)) dx \geq 0$ for all $w \in C^1([0, 1])$. As in the proof of Theorem 3.6 we may assume that ξ is a.c. and $\xi^0 = \dot{\xi}$ almost everywhere. Thus $\langle \eta^0, w(0) \rangle + \langle \eta^1, w(1) \rangle + \langle \xi(1), w(1) \rangle - \langle \xi(0), w(0) \rangle \geq 0$ for all $(w(0), w(1)) \in R^{2n}$, which implies $\eta^0 = \xi(0), \eta^1 = -\xi(1)$.

Using the reduction technique from the proof of [2, Thm. 2.4, case 3], and Theorem 3.6 it is possible to prove

THEOREM 3.11. Assume v_0 solves (\mathcal{B}) , $F(v_0) < \infty$, l is l.s.c. and L has an admissible $\{\mathcal{C}L(x): x \in [0, 1]\}$ along v_0 . Then there exists an a.c. ξ satisfying (3.7) such that

$$(3.12) \quad \langle (-\xi(0), \xi(1), 1), y \rangle \geq 0$$

for all $y \in R^{2n+1}$ for which $((v_0(0), v_0(1)), l(v_0(0), v_0(1))) + \delta y \in \text{epi}(l)$ (epigraph) for all small $\delta > 0$.

Remark 3.13.(a) If, for some $r > 0$, $k \in L^1([0, 1])$ on the ball $B((v_0(x), v_0(x)), r)$, the function $L(x, \cdot, \cdot)$ is continuous and $|L(x, \cdot, \cdot)|$ is bounded by $k(x)$, then in Theorem 3.6 the set \mathcal{L} need not be assumed closed. Indeed, if v_0 solves the problem of Lagrange with some \mathcal{L} then it also solves the problem with \mathcal{L} replaced by $\bar{\mathcal{L}}$. Under such additional assumptions, the hypothesis of l.s.c. of l in Theorem 3.11 may also be omitted.

(b) [3, Thm. 2.4, Case 1] is covered by our results. In [3] another technique is applied to the problem of Lagrange. In general, the transversality conditions (3.8), (3.10), (3.12) are different from those proposed in [3].

4. The maximum principle of Pontryagin. Let U be a set valued function from $[0, 1]$ into subsets of R^m . By a control we will mean any measurable $u: [0, 1] \rightarrow R^m$ such that $u(t) \in U(t)$ almost everywhere (measurable selector of U). On the set of all controls \mathcal{U} consider the metric $d(u_1, u_2) = \mu(\{t: u_1(t) \neq u_2(t)\})$. Arguing as in [7, Lemma 7.2], it is possible to verify that (\mathcal{U}, d) is a complete metric space.

Let $\mathcal{W} \in \mathcal{U}$ be closed and

(H₁) For any $u_1, u_2 \in \mathcal{W}$ and $0 \leq t_1 \leq t_2 \leq 1$ the function

$$U_3(t) = \begin{cases} u_1(t) & \text{if } t \in [t_1, t_2], \\ u_2(t) & \text{otherwise} \end{cases}$$

belongs to \mathcal{W} .

Let $C \subset R^n$, $f: [0, 1] \times R^n \times R^m \rightarrow R^n$, $g: R^n \rightarrow \bar{R}$ be given. Denote by X the set of all solutions to the system

$$(4.1) \quad \begin{aligned} \dot{x} &= f(t, x, u(t)), & u &\in \mathcal{W}, \\ x(0) &\in C \end{aligned}$$

defined on $[0, 1]$. We shall study the following problem:

$$(4.2) \quad \text{minimize } \{g(x(1)): x \in X\}.$$

It is well known that the control problem of minimizing $\int_0^1 f^0(t, x(t), u(t)) dt$ over all $x \in X$ and corresponding control u can be reduced to (4.2) (see, for example, [3, proof of Cor. 2]).

Assume $u_* \in \mathcal{W}$ generates $x_* \in X$. Suppose

(H₂) The function $t \rightarrow f(t, s, u(t))$ is measurable for all $s \in \mathbb{R}^n$ and $u \in \mathcal{W}$.

(H₃) There are $r > 0$, $k \in L^1$ such that, for all $u \in \mathcal{W}$ and almost every $t \in [0, 1]$, $f(t, \cdot, u(t))$ is Lipschitzian of constant $k(t)$ on $B(x_*(t), r)$.

As in § 3 our tool will be certain admissible families of sets along the solution.

DEFINITION 4.3. The family of sets $\{\mathcal{D}f(t): t \in [0, 1]\}$ is called *admissible along* (x_*, u_*) if there exist $\lambda_i > 0$, integrable $\varepsilon_i: [0, 1] \rightarrow \mathbb{R}_+$, $f_i: [0, 1] \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ such that:

$$(i) \quad \int_0^1 \varepsilon_i(t) dt / \lambda_i \rightarrow 0 \quad \text{as } i \rightarrow \infty;$$

(ii) For some $r > 0$, $k \in L^1$ and all i , f_i satisfies (H₂), (H₃) and for all $u \in \mathcal{W}$ and almost every $t \in [0, 1]$, $f_i(t, \cdot, u(t))$ is F-differentiable on $B(x_*(t), \lambda_i)$ and $|f_i(t, \cdot, u(t)) - f(t, \cdot, u(t))| \leq \varepsilon_i(t)$;

(iii) $\{f_i(t, \cdot, u_*(t))\}$, $\{\lambda_i\}$, $\{\varepsilon_i(t)\}$ define a set

$$\mathcal{CP}_b f(t, x_*(t), u_*(t)) = \mathcal{D}f(t) \quad \text{a.e. with } V = \mathbb{R}^{n \times n};$$

(iv) $\{\mathcal{D}f(t): t \in [0, 1]\}$ is minimal, i.e. for every $\{\tilde{f}_i\}$, $\{\tilde{\lambda}_i\}$, $\{\tilde{\varepsilon}_i\}$ of the above properties defining a family of sets $\mathcal{CP}_b f(t, x_*(t), u_*(t)) \subset \mathcal{D}f(t)$ almost everywhere we have the equality in the last inclusion.

Remark 4.4. As in Remark 3.2, we have (a) (iv) makes sense;

(b) If f satisfies (H₂), (H₃) and $f(t, \cdot, u_*(t))$ is F-differentiable at $x_*(t)$, then the family $\{(\partial/\partial x)f(t, x_*(t), u_*(t)): t \in [0, 1]\}$ is the only one admissible.

(c) If f satisfies (H₂), (H₃), then from [19, proof of Thm. 2.5] follows the existence of an admissible $\{\mathcal{D}f(t): t \in [0, 1]\}$ such that $\mathcal{D}f(t)$ is contained in Clarke's generalized gradient of $f(t, \cdot, u_*(t))$ at $x_*(t)$.

THEOREM 4.5. Assume (x_*, u_*) solves (4.2), (H₁)–(H₃) hold, and C is closed and satisfies Condition 2.4 at $x_*(0)$. Let $\{\mathcal{D}f(t): t \in [0, 1]\}$ be admissible along (x_*, u_*) and defined by $\{f_i\}$, $\{\lambda_i\}$ and suppose that g has a set $\mathcal{A}g(x_*(1))$ defined by $\{g_i\}$, $\{\lambda_i\}$, $V = \mathbb{R}^n$ with bounded $\{g'_i(y); y \in B(x_*(1), \lambda_i), i \geq 1\}$. Then there exist a measurable matrix-function $A(t)$ and a.c. $p: [0, 1] \rightarrow \mathbb{R}^n$ such that

$$(4.6) \quad \dot{p}(t) = -p(t)A(t), \quad A(t) \in \mathcal{D}f(t) \text{ a.e.},$$

$$(4.7) \quad \langle p(t), f(t, x_*(t), u(t)) \rangle \leq \langle p(t), f(t, x_*(t), u_*(t)) \rangle \quad \text{a.e. for all } u \in \mathcal{W},$$

$$(4.8) \quad p(1) \in -\mathcal{A}g(x_*(1)),$$

$$(4.9) \quad \langle p(0), w \rangle \leq 0 \quad \text{for all } w \in W$$

where W is the cone referred to in Condition 2.4.

From Remark 4.4 and Corollary 2.6.2 we obtain

COROLLARY 4.5.1. Assume in Theorem 4.5 that $f(t, \cdot, u_*(t))$ is F-differentiable at $x_*(t)$. Then there exists an a.c. $p: [0, 1] \rightarrow \mathbb{R}^n$ such that (4.7)–(4.9) hold and

$$(4.6') \quad \dot{p}(t) = -p(t) \frac{\partial f}{\partial x}(t, x_*(t), u_*(t)) \quad \text{a.e.}$$

Furthermore if g is F-differentiable at $x_*(1)$ then

$$(4.8') \quad p(1) = -g'(x_*(1)).$$

From [3, Prop. 1 and Lemma 1] follows

COROLLARY 4.5.2. Assume in Theorem 4.5 that $f(\cdot, x, \cdot)$ and the graph of U is $L \times B^m$ measurable (see [3]) and $\mathcal{W} = \mathcal{U}$. Then the assertions of the theorem are valid with (4.7) replaced by

$$(4.7') \quad \sup_{u \in U(t)} \langle p(t), f(t, x_*(t), u) \rangle = \langle p(t), f(t, x_*(t), u_*(t)) \rangle \quad a.e.$$

We precede the proof of the theorem with several lemmas. Note that it is not restrictive to assume that for some $K > 0$

$$(4.10) \quad |f(t, x_*(t), u(t)) - f(t, x_*(t), u_*(t))| \leq K \quad a.e.$$

for all $u \in \mathcal{W}$. Indeed let $\mathcal{W}_j = \{u \in \mathcal{W} : |f(t, x_*(t), u(t)) - f(t, x_*(t), u_*(t))| \leq j \text{ a.e.}\}$. Then (H_1) holds for \mathcal{W}_j and u_* solves (4.2) with \mathcal{W} replaced by \mathcal{W}_j . If we prove that for each \mathcal{W}_j some p_j satisfies all the assertions of the theorem, then from the Dunford–Pettis criterion, (H_1) and the Mazur lemma (see [8]), for some subsequence $\{p_{j_k}\}$ and a.c. p , we have $p_{j_k} \rightarrow p$ uniformly, $\dot{p}_{j_k} \rightarrow \dot{p}$ (weakly in L^1) and p is a required function.

From now on we assume (4.10). The proof of the theorem is a slight modification of [7, proof of Thm. 7.1].

From the generalized lemma of Gronwall [6] we deduce

PROPOSITION 4.11. For some $M > 0$ and all large i , $\bar{x} \in B(x_*(0), \lambda_i/M) \cap C$, any $u \in \mathcal{W}$ with $d(u, u_*) \leq \lambda_i/M$ generates $x \in X$ with $x(0) = \bar{x}$ and a solution x_i to

$$(4.12) \quad \dot{x} = f_i(t, x, u(t)), \quad x(0) = \bar{x}$$

defined on $[0, 1]$ and satisfying $\|x_i - x_*\| \leq \lambda_i/2$, $\|x_i - x\| = o(\lambda_i)$.

Let $u \in \mathcal{W}$, $d(u, u_*) \leq \lambda_i/2M$ and let x be the solution of (4.12) where $|\bar{x} - x_*(0)| \leq \lambda_i/M$. Fix $u_0 \in \mathcal{W}$ and let $t_0 \in (0, 1)$ be such that the function $t \rightarrow \int_0^t f_i(s, x(s), u_0(s)) ds$ has the derivative $f_i(t_0, x(t_0), u_0(t_0))$ at t_0 and $\dot{x}(t_0) = f_i(t_0, x(t_0), u(t_0))$. For $\tau \in [0, t_0]$, let

$$u_\tau(t) = \begin{cases} u_0(t) & \text{if } t \in [t_0 - \tau, t_0], \\ u(t) & \text{otherwise.} \end{cases}$$

By Proposition 4.11, u_τ generates a solution x_τ of (4.12) and $\|x_\tau - x_*\| < \lambda_i/2$, when $\tau \in [0, \lambda_i/2M]$. Denote by $Z(t, t_0)$ the (fundamental) matrix of solutions to

$$(4.13) \quad \dot{z} = \frac{\partial f_i}{\partial x}(t, x(t), u(t))z, \quad Z(t_0, t_0) = \text{Id}.$$

LEMMA 4.14. The function $[0, \lambda_i/2M] \ni \tau \rightarrow x_\tau(1)$ is differentiable at $\tau = 0$ and

$$\frac{d}{d\tau} x_\tau(1) = Z(1, t_0)(f_i(t_0, x(t_0), u_0(t_0)) - f_i(t_0, x(t_0), u(t_0))).$$

Proof. For $0 \leq t \leq t_0 - \tau$ we have $x(t) = x_\tau(t)$. When $t > t_0 - \tau$ we have

$$\begin{aligned} |x(t) - x_\tau(t)| &= \left| \int_{t_0 - \tau}^t (f_i(s, x(s), u(s)) - f_i(s, x_\tau(s), u_\tau(s))) ds \right| \\ &\leq \left| \int_{t_0 - \tau}^{t_0} (f_i(s, x(s), u(s)) - f_i(s, x(s), u_0(s))) ds \right| \\ &\quad + \int_{t_0 - \tau}^t k(s) |x(s) - x_\tau(s)| ds. \end{aligned}$$

By Gronwall's lemma and the choice of t_0 there exists $M_1 = M_1(t_0) > 0$ such that

$$(4.15) \quad \|x - x_\tau\| \leq M_1 \tau.$$

By Taylor's formula $x(t_0 - \tau) = x(t_0) - \tau f_i(t_0, x(t_0), u(t_0)) + o(\tau)$.

$$\begin{aligned} \int_{t_0-\tau}^{t_0} f_i(s, x_\tau(s), u_0(s)) ds &= \int_{t_0-\tau}^{t_0} \left(f_i(s, x(s), u_0(s)) \right. \\ &\quad \left. + \frac{\partial f_i}{\partial x}(s, x(s), u_0(s))(x_\tau(s) - x(s)) + o(|x_\tau(s) - x(s)|) \right) ds, \\ \int_{t_0}^t f_i(s, x_\tau(s), u(s)) ds &= \int_{t_0}^t \left(f_i(s, x(s), u(s)) \right. \\ &\quad \left. + \frac{\partial f_i}{\partial x}(s, x(s), u(s))(x_\tau(s) - x(s)) + o(|x_\tau(s) - x(s)|) \right) ds. \end{aligned}$$

Adding the above equalities, taking into account (4.15) and the relation $x(t) = x(t_0) + \int_{t_0}^t f_i(s, x(s), u(s)) ds$, we obtain

$$\begin{aligned} x_\tau(t) &= x(t_0 - \tau) + \int_{t_0-\tau}^t f_i(s, x_\tau(s), u_\tau(s)) ds \\ &= x(t) - \tau(f_i(t_0, x(t_0), u(t_0)) - f_i(t_0, x(t_0), u_0(t_0))) \\ &\quad + \int_{t_0}^t \frac{\partial f_i}{\partial x}(s, x(s), u(s))(x_\tau(s) - x(s)) ds + o(\tau). \end{aligned}$$

The solution of (4.13) satisfying $z(t_0) = -\tau(f_i(t_0, x(t_0), u(t_0)) - f_i(t_0, x(t_0), u_0(t_0)))$ is equal to $Z(t, t_0)z(t_0)$. Since $x_\tau(t_0) - x(t_0) - z(t_0) = o(\tau)$ we have $|x_\tau(t) - x(t) - z(t)| < o(\tau) + \int_{t_0}^t \|\partial f_i / \partial x(s, x(s), u(s))\| |x_\tau(s) - x(s) - z(s)| ds$. By Gronwall's lemma $x_\tau(t) - x(t) = z(t) + o(\tau)$. Hence the result.

LEMMA 4.16. *For all large i there exist $\delta_i > 0$, $\bar{x} \in B(x_*(0), \lambda_i/2M)$ and $u_i \in \mathcal{W}$ with $d(u_i, u_*) \leq \lambda_i/2M$ generating the solution x_i of (4.12) such that $\lim_{i \rightarrow \infty} \delta_i = 0$, and for all $u \in \mathcal{W}$, $w \in W$*

$$\begin{aligned} \langle f_i(t, x_i(t), u(t)), p_i(t) \rangle &\leq \langle f_i(t, x_i(t), u_i(t)), p_i(t) \rangle + \delta_i \quad a.e., \\ \langle p_i(0), w \rangle &\leq \delta_i |w| \end{aligned}$$

where $p_i: [0, 1] \rightarrow \mathbb{R}^n$ is a.c. and

$$(4.17) \quad \dot{p}_i(t) = -p_i(t) \frac{\partial f_i}{\partial x}(t, x_i(t), u_i(t)) \quad a.e.,$$

$$(4.18) \quad p_i(1) = -g'_i(x_i(1)).$$

Proof. Let $V = C \times \mathcal{W}$ and $\bar{d}((x_1 u_1), (x_2 u_2)) = |x_1 - x_2| + d(u_1, u_2)$ for $(x_j, u_j) \in V$, $j = 1, 2$. Then (V, \bar{d}) is a complete metric space. Let $\bar{\lambda}_i = \lambda_i/M$ and $\{\bar{\varepsilon}_i\}$ be as in the definition of $\mathcal{A}g(x_*(1))$. Set $F_i(\bar{x}, u)$ to be $g_i(x(1))$ if $\bar{d}((\bar{x}, u), (x_*(0), u)) \leq \bar{\lambda}_i$ and if x is the solution of (4.12), and let $F_i(\bar{x}, u) = +\infty$ otherwise. By Proposition 4.11, for large i , F_i is l.s.c. and $F_i(x_*(0), u_*) \leq \inf F_i + \bar{\varepsilon}_i + o(\lambda_i)$. By [7, Thm. 1.1] there is $(x_i(0), u_i) \in V$ such that $\bar{d}((x_i(0), u_i), (x_*(0), u_*)) \leq \lambda_i/2M$ and

$$(4.19) \quad F_i(x(0), u) - F_i(x_i(0), u_i) \geq -\delta_i \bar{d}((x(0), u), (x_i(0), u_i))$$

for some $\{\delta_i\}$ converging to zero.

Let $Z(t, t_0)$ be the matrix solution to $\dot{Z} = (\partial f_i / \partial x)(t, x_i(t), u_i(t))Z$, $Z(t_0, t_0) = \text{Id}$. Let $w \in W$. If i is large and $\tau > 0$ is small then by Proposition 4.11, there is a solution x_τ of (4.12) where $\bar{x} = x_i(0) + \tau w$, $u = u_i$. As we did in Lemma 4.16, we verify that $\frac{d}{d\tau} x_\tau(1)|_{\tau=0} = Z(1, 0)w$. Thus by (4.19), $\langle g'_i(x_i(1)), \frac{d}{d\tau} x_\tau(1)|_{\tau=0} \rangle = \langle Z^T(1, 0)g'_i(x_i(1)), w \rangle = \langle -p_i(0), w \rangle \geq -\delta_i|w|$. As in [7, proof of Thm. 7.1] we verify, using Lemma 4.14 and (4.13), that $\{u_i\}$ satisfy the required property. We apply the Ascoli–Arzela theorem, the Dunford–Pettis criterion and Mazur’s lemma to complete the proof of the theorem.

Remark 4.20. In [3], a more general problem

$$(4.2'') \quad x(1) \in \text{Fr} \{x(1): x \in X\}$$

(where Fr denotes the boundary) is considered. The proof is based on the reduction of the problem through the Ekeland principle to (4.2) (see [3, p. 1087]). Exploiting this idea and Corollary 4.2 it is possible to derive [3, Thm. 1]. To treat the problem

$$(4.2''') \quad x(1) \in \text{Fr} \{\phi(x(1)): x \in X\}$$

where $\phi: R^n \rightarrow R^n$ is locally Lipschitzian, one has to apply the same idea to a family of problems,

$$x(1) \in \text{Fr} \{\phi_j(x(1)): x \in X\}$$

where $\phi_j \in C^1$ are locally uniformly Lipschitzian and approximate ϕ uniformly on a ball of centre $x_*(1)$.

Acknowledgments. The author would like to thank Professor A. Cellina for reading the manuscript and for critical comments, Professor C. Olech for many useful suggestions and Professor R. T. Rockafellar for a discussion which stimulated this research.

REFERENCES

- [1] V. BARBU, *Necessary conditions for multiple integral problem in the calculus of variations*, Preprint Series in Mathematics, 27/1981, Bucuresti, 1981.
- [2] F. H. CLARKE, *The Euler–Lagrange differential inclusion*, J. Differential Equations, 19 (1975), pp. 80–90.
- [3] ———, *The maximum principle under minimal hypothesis*, this Journal, 14 (1976), pp. 1078–1091.
- [4] ———, *Admissible relaxation in variational and control problems*, J. Math. Anal. Appl., 51 (1975), pp. 557–576.
- [5] ———, *Multiple integral of Lipschitz functions in the calculus of variations*, Proc. Amer. Math. Soc., 64 (1977), pp. 260–264.
- [6] R. CONTI AND G. SANSONE, *Nonlinear Differential Equations*, Pergamon Press, Oxford, 1964, pp. 11–14.
- [7] I. EKELAND, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.
- [8] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.
- [9] I. EKELAND, *Nonconvex minimization problems*, Bull. Amer. Math. Soc. (New Ser.), 1 (1979), pp. 443–474.
- [10] H. HALKIN, *Interior mapping theorem with set-valued derivatives*, J. Analyse Math., 30 (1976), pp. 200–207.
- [11] ———, *Mathematical programming without differentiability*, in Calculus of Variations and Control Theory, D. L. Russel, ed., Academic Press, New York, 1976, pp. 279–287.
- [12] ———, *Necessary conditions for optimal control problems with differentiable or nondifferentiable data*, in Mathematical Control Theory, W. A. Coppel, ed., Lecture Notes in Mathematics 680, Springer-Verlag, Berlin, 1978, pp. 77–118.
- [13] A. D. IOFFE, *Approximate subdifferentials of nonconvex functions*, Cahiers de mathématiques de la décision, CEREMADE, 8120 (1981), Paris.

- [14] C. MORREY, *Multiple Integrals in the Calculus of Variations*, MR34#2380, Springer-Verlag, Berlin, 1966.
- [15] B. POURCIAU, *Analysis and optimization of Lipschitz continuous mappings*, J. Optim. Theory Appl., 22 (1977), pp. 311–351.
- [16] R. T. ROCKAFELLAR, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 32 (1970), pp. 174–222.
- [17] ———, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 174–222.
- [18] J. WARGA, *Necessary conditions without differentiability assumptions in optimal control*, J. Differential Equations, 18 (1975), pp. 41–62.
- [19] ———, *Derivative containers, inverse functions and controllability*, in *Calculus of Variations and Control Theory*, D. L. Russel, ed., Academic Press, New York, 1976, pp. 13–45.
- [20] ———, *Fat homeomorphisms and unbounded derivative containers*, J. Math. Anal. Appl., 81 (1981), pp. 545–560.

AN INVARIANT MEASURE APPROACH TO THE CONVERGENCE OF STOCHASTIC APPROXIMATIONS WITH STATE DEPENDENT NOISE*

HAROLD J. KUSHNER† AND ADAM SHWARTZ‡

Abstract. A new method is presented for quickly getting the ODE (ordinary differential equation) associated with the asymptotic properties of the stochastic approximation $X_{n+1} = X_n + a_n f(X_n, \xi_n)$ (or the projected algorithm for the constrained problem). Either $a_n \rightarrow 0$, or a_n can be constant, in which case the analysis is on the sequence obtained when $a \rightarrow 0$. The method requires that $\{X_n, \xi_{n-1}\}$ be Markov with a “Feller” transition function, but little else. The simplest result requires that if $X_n \equiv x$, the corresponding noise process $\{\xi_n(x), n \geq 0\}$ have a unique invariant measure; but the “nonunique” case can also be treated. No mixing condition is required, nor the construction of averaged test functions, and $f(\cdot, \cdot)$ need not be continuous. A detailed analysis of the way that $\{\xi_n\}$ varies with $\{X_n\}$ is not required. For the class of sequences treated, the conditions seem easier to verify than for other methods. There are extensions to the non-Markov case. Two examples illustrate the power and ease of use of the approach. Aside from the advantages of the method in treating standard problems, it seems to be particularly useful for handling the type of iterative algorithms which arise in adaptive communication theory, where the dynamics are often discontinuous and the “noise” is often state-dependent due to the effects of feedback. If the noise $\{\xi_n\}$ is not “state-dependent,” then the Markov assumption can be dropped, and the method is even easier to use.

Key words. stochastic approximation, recursive algorithms, adaptive control

1. Introduction. We consider stochastic approximations of the form

$$(1.1) \quad X_{n+1} = X_n + a_n f(X_n, \xi_n),$$

where $f(\cdot, \cdot)$ might be discontinuous, and the evolution of $\{\xi_n\}$ depends on $\{X_n\}$ in the sense that, in general,

$$P\{\xi_{n+1} \in A | \xi_i, i \leq n\} \neq P\{\xi_{n+1} \in A | X_i, \xi_i, i \leq n\}.$$

We also treat the following “projected” version of (1.1). Let G be a bounded set of the form $G = \{x: q_i(x) \leq 0, i = 1, \dots, s\}$, where $q_i(\cdot)$ are continuously differentiable, and G is the closure of its interior. Let $\pi_G(y)$ denote any closest point in G to y . Then the projected algorithm is

$$(1.2) \quad X_{n+1} = \pi_G(X_n + a_n f(X_n, \xi_n)).$$

Several so-called ordinary differential equations (ODE) methods for proving convergence of $\{X_n\}$ have been developed in recent years. (See references [1]–[4], and [5], a more polished form of [4] with weaker conditions.) The aim of these methods is to get an ODE, which we write symbolically (for (1.1)) as

$$(1.3) \quad \dot{x} = E^x f(x, \xi) \equiv \int f(x, \xi) P^x(d\xi),$$

where (loosely speaking) $P^x(\cdot)$ is the stationary distribution of the sequence $\{\xi_n\}$, when $X_n \equiv x$. The idea is that $\{X_n\}$ in (1.1) varies much more slowly (for large n) than $\{\xi_n\}$

* Received by the editors June 28, 1982, and in revised form December 26, 1982. This research was supported in part by the Air Force Office of Scientific Research under contract AFOSR 81-0116, in part by the National Science Foundation under contract NSF Eng. 77-12946-A04 and in part by the Office of Naval Research under contract N00014-76-C-0279-P0004.

† Division of Applied Mathematics and Engineering, Brown University, Providence, Rhode Island 02912.

‡ Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

does and that some sort of averaging method or law of large numbers can be used to show that the asymptotic properties of $\{X_n\}$ are the same as those of (1.3), with a proper definition of $P^x(\cdot)$.

We now take some notation from [2]. Define $t_n = \sum_{i=0}^{n-1} a_i$ and $m(t) = \max \{n: t_n \leq t\}$. Thus $m(t_n) = n$. Let $x^0(\cdot)$ denote the piecewise linear interpolation of the function with value X_n at t_n . Define the shifted function $x^n(\cdot)$ by $x^n(t) = x^0(t + t_n)$, $t \geq 0$. Thus $x^n(0) = X_n$, and the asymptotic properties (as $t \rightarrow \infty$) of any limit (as $n \rightarrow \infty$) of $\{x^n(\cdot)\}$ yield the asymptotic behavior of $\{X_n\}$. The convergence of $x^n(\cdot)$ to a limit $x(\cdot)$ is in the sense of weak convergence of a sequence of probability measures (see below). We give the differential equations which $x(\cdot)$ satisfies. Using this differential equation and the properties of weak convergence, one can analyze the asymptotic behavior of X_n as in [2]. See that reference for details. Here, we concentrate on proving representations for the differential equations.

The methods in [1]–[3] are very useful, but they are often difficult to apply when the noise is state-dependent, in the sense that the conditions are often either hard to verify or do not hold in many important cases of interest. References [4] and [5] present an “averaging method” which works quite well for such problems, although one would like to avoid the work associated with constructing the “averaged test functions” and verifying the conditions on them. The results in [4] and [5] are for w.p.1. convergence and also prove stability and similar properties for $\{X_n\}$ sequences which are not artificially bounded. But generally, past methods required what is often a difficult analysis of the way $\{\xi_n\}$ depends on $\{X_n\}$.

In this paper, the essential assumption for the validity of (1.3) is that $\{\xi_n\}$ depends on $\{X_n\}$ in such a way that if $X_n \equiv x$, a constant, then the corresponding $\{\xi_n\}$ process possesses a unique stationary measure. Such an assumption, either implicitly or explicitly, was used in much past work on the state-dependent noise case. If the stationary measures are not unique, then a very similar result (2.9) holds. The conditions required here are generally weaker and much easier to check than those in previous papers and are useful even when the noise does not depend on the state. As amply shown by the examples, the method is easy to use. The techniques used are new for the class of problems treated.

We concentrate on the case $a_n \rightarrow 0$. The same proofs work (even more easily) when $a_n \equiv a$, a constant. Then, we get that the limit (as $a \rightarrow 0$) of $x^a(\cdot)$ satisfies (1.3) or (2.12) (in the constrained case), where $x^a(\cdot)$ is the piecewise linear function with values X_n at na . This approach is advantageous for treating many standard problems because the conditions are relatively easy to verify. They are particularly useful for treating the type of algorithms which appear in adaptive communication theory, where the dynamics are often discontinuous and the noise is often state-dependent owing to the role of feedback. In such cases, one normally has $a_n \equiv a$.

In § 2, we discuss the case where $\{X_n, \xi_{n-1}, n \geq 1\}$ is a Markov process, or where the “state-noise” pair can be “Markovianized.” This is the case which is most fully understood and easiest to use. A class of non-Markov processes is dealt with in § 3, and in § 4 we illustrate the power and ease of use of the method via two examples.

Weak convergence: definitions [6]. Let $C'[0, \infty)$ denote the space of E' -valued continuous functions on $[0, \infty)$, with the topology of uniform convergence on bounded intervals. A sequence of measures $\{P_n\}$ on a topological space A is *tight* if for each $\varepsilon > 0$, there is a compact K_ε such that $P_n(K_\varepsilon) \geq 1 - \varepsilon$ for all n . If A is complete and separable, then tightness is equivalent to weak compactness; i.e., for each subsequence $\{P_n\}$ there are a further subsequence $\{P_n\}$ and a P such that for each continuous,

bounded and real-valued $f(\cdot)$, $\int f(x) dP_n(x) \rightarrow \int f(x) dP(x)$. Let $A = C'[0, \infty)$ and suppose that P_n converges weakly to P . Let $x^n(\cdot)$ and $x(\cdot)$ denote the corresponding processes. Then, the Skorokhod imbedding theorem [7, Thm. 3.1.1] states that we can choose an underlying probability space $(\tilde{\Omega}, \tilde{P}, \tilde{B})$ such that $x^n(\cdot) \rightarrow x(\cdot)$ uniformly on bounded intervals w.p.1. (in the sense that there are processes on $(\tilde{\Omega}, \tilde{P}, \tilde{B})$ with the same distributions as $x^n(\cdot)$, $x(\cdot)$, and where the indicated convergence occurs).

2. Limit theorems with $\{X_n, \xi_{n-1}\}$ Markov. In this section, we are concerned with the Markov case. In very many applications the system is either Markovian or the actual physical noise can be Markovianized, perhaps leading to an abstract-valued process. Below, it is assumed that $\{X_n\}$ is either tight or lies in a compact set. This is not a very serious restriction, since practical algorithms tend to use various truncation devices. In any case, the use of the projection method (1.2) guarantees the compactness when X_n lies in a Euclidean space.

Some assumptions. Assumptions A2.4, A2.6 and the first part of A2.1 will be weakened later, as will the form given for $f(\cdot, \cdot)$.

A2.1. $\{X_n, \xi_{n-1}, n \geq 0\}$ is a Markov process with a (possibly nonhomogeneous) transition function $P(x, \xi, n, l, A) = P\{(X_{n+l}, \xi_{n+l-1}) \in A | X_n = x, \xi_{n-1} = \xi\}$. The X_n take values in a compact subset H of a Euclidean space E^r .

A2.2. $\{\xi_n\}$ is tight in S , a metric space.

A2.3. For each Borel $B \subset S$, define the one step "transition function" $P_x(\xi, 1, B) = P\{\xi_n \in B | \xi_{n-1} = \xi, X_n = x\}$, and suppose that it does not depend on n . Let $P_x(\xi, 1, \cdot)$ be weakly¹ continuous in (x, ξ) .

For each x , we now define a Markov process $\{\xi_n(x), n \geq 0\}$ on S via the transition function $P_x(\xi, l, \cdot)$, where $P_x(\xi, l, B) = \int P_x(\xi, l-k, d\xi') P_x(\xi', k, B)$ is defined recursively, starting with the given $P_x(\xi, 1, d\xi')$.

A2.4. For each $x \in H$, let $\{\xi_n(x), n \geq 0\}$ have a unique invariant measure $P^x(\cdot)$, and let $\{P^x(\cdot), x \in H\}$ be tight.

A2.5. $\sum_n |a_{n+1} - a_n| < \infty$, $0 < a_n \rightarrow 0$, $\sum_n a_n = \infty$.

A2.6. $f(\cdot)$ is bounded.

A2.7. There is an integer $c \geq 0$ such that $\int P_x(\xi, c+1, d\xi') f(x, \xi')$ is continuous in (x, ξ) . It equals $\lim_j \int P(x, \xi, j-c, c, dx', d\xi') P_{x'}(\xi', 1, d\xi'') f(x', \xi'') = \lim_j E[f(X_j, \xi_j) | X_{j-c} = x, \xi_{j-c-1} = \xi]$, where the limit is uniform on compact (x, ξ) sets.

Condition A2.8 will be discussed below: either A2.8a or A2.8b will be used. Condition A2.8b always holds for $N(K) = 1$ if S is compact. Let $I_K(\xi)$ denote the indicator of the set where $\xi \in K$.

A2.8a. Either S is compact or $\{\xi_n\}$ is mutually independent, or

A2.8b. For each compact K , there is an integer $N(K) < \infty$ such that for each T the set

$$\{P(x, \xi, m, j-m, \cdot) I_K(\xi), \text{ all compact } K, \text{ all } x \in H, \text{ all } m, j$$

$$\text{such that } j-m \geq N(K) \text{ and } t_j - t_m \leq T\}$$

is tight. (If $\xi \notin K$, then $P(\cdot) I_K(\xi)$ is the zero measure.)

Remark on A2.7. If $f(\cdot, \cdot)$ is continuous, then A2.3 implies that we can take $c = 0$. If $c = 0$, then the second sentence of A2.7 is redundant. Even if $f(\cdot, \cdot)$ is not continuous, $c = 0$ is often enough to get the required smoothing. See, for example, the applications in § 4. Even if $c > 0$ is needed, the second sentence of A2.7 does not seem to be particularly restrictive, since $|X_j - X_{j-c}| \rightarrow 0$ as $j \rightarrow \infty$ implies that the measure

¹ That is $\int P_x(\xi, 1, d\xi') g(\xi)$ is continuous in (x, ξ) if $g(\cdot)$ is bounded and continuous.

in the \lim_j expression is essentially $P_x(\xi, c+1, \cdot)$ for large j . The assumption is stated as it is for technical reasons. In all applications that we are aware of now, if the first sentence of A2.7 holds, so does the second sentence.

Despite its seemingly complicated structure, A2.8b is quite natural and is often easy to verify. See, for example, the application in § 4b, and the example below. It is motivated by the following consideration. If S is not compact, it is possible that

$$(*) \quad \{P\{\xi_{j-1} \in \cdot | X_{m_l}(\omega), \xi_{m_l-1}(\omega)\}, j \geq m_l, \omega\}$$

is not tight. Suppose for example that $\{\xi_n\}$ is a stationary scalar-valued Gaussian Markov process, not depending on $\{X_n\}$ (in the sense that the inequality below (1.1) is an equality), and whose correlation function $\rho(\cdot)$ tends to zero. Then the set $(*)$ is not tight, since arbitrarily large initial conditions $\xi_{m_l-1}(\omega)$ are allowed. But, if the $\xi_{m_l-1}(\omega)$ were all confined to a bounded set, then $(*)$ would be tight. As K increases, with $\xi_{m_l-1}(\omega)$ taking an arbitrary value in K , it might take longer for the “future” ξ_j ($j > m_l$) to “settle down.” This is why we allow $N(K)$ steps for this “settling down,” where $N(K)$ increases with K . In this example, if $K = \{\xi: |\xi| \leq k\}$, then any $N(K)$ satisfying $\rho(N(K)) \cdot k \leq \text{constant}$ is satisfactory.

Notation. Some additional notation is required for the next theorem.

Let $0 < \delta_n \rightarrow 0$ as $n \rightarrow \infty$ such that $\lim_n \sup \{a_j: j \geq n\}/\delta_n = 0$. For each n choose an increasing sequence $n = m(n, 1) < m(n, 2) < \dots$ such that $\sum_{m(n,l)}^{m(n,l+1)-1} a_j = \delta_n$, modulo an “end” value of a_j . Thus $(t_{m(n,l+1)} - t_{m(n,l)})/\delta_n \rightarrow 1$ as $n \rightarrow \infty$, uniformly in l . For notational convenience we henceforth suppress the n in $m(l, n)$ and write simply $m(n, l) = m_l$. For each ω, l, n define the measure on the Borel sets of S :

$$Q(\omega, l, n, \cdot) = \frac{1}{\delta_n} \sum_{m_l}^{m_{l+1}-1} a_j P\{\xi_{j-1} \in \cdot | X_{m_l}(\omega), \xi_{m_l-1}(\omega)\}.$$

Define $Q_K(\omega, l, n, \cdot) = Q(\omega, l, n, \cdot) I_K(\xi_{m_l-1})$. Thus, if $\xi_{m_l-1}(\omega) \notin K$, the measure is the zero measure.

A remark on the proof of Theorem 1. By A2.6, there is a constant K such that $|x^n(t+s) - x^n(t)| \leq Ks$ for all $t, s > 0$ and n . Thus, since $\{x^n(0)\} = \{X_n\}$, $\{x^n(\cdot)\}$ is tight in $C'[0, \infty)$. Fix and work with a weakly convergent subsequence, also indexed by n . Let $h(\cdot)$ be a smooth function with compact support and let $t_1 < t_2 < \dots < t_k < t < t + \tau$. Since $a_n \rightarrow 0$,

$$Eh(x^n(t_i), i \leq k) \left[x^n(t + \tau) - x^n(t) - \sum_{m(t_n+t)}^{m(t_n+t+\tau)-1} a_j f(X_j, \xi_j) \right] \rightarrow 0.$$

Divide $[0, t + \tau]$ into subintervals of width δ_n , and let $f(\cdot, \cdot)$ be continuous. By the properties of conditional expectation, we can replace the sum above by

$$\sum_l \delta_n \left[\frac{1}{\delta_n} E_{m_l} \sum_{i=m_l}^{m_{l+1}-1} a_i f(X_i, \xi_i) \right],$$

where E_{m_l} denotes conditioning on X_{m_l}, ξ_{m_l-1} and l runs over integers such that $\sum_n^{m_l} a_i \in [t, t + \tau]$.

Consider the l th bracketed term, and let $t_{m_l} \rightarrow s$. The proof shows that the X_i can be replaced by $x(s)$ and that the set (indexed by ω, n, l) of (slightly altered) measures defining the conditional expectation is tight. Then a “continuity” argument is used to show that the limit of any weakly convergent subsequence is actually $P^{x(s)}(\cdot)$, the invariant measure associated with $x = x(s)$. This characterization of the limit is the key point in the proof.

THEOREM 1. Under A2.1 to A2.8, $\{x^n(\cdot)\}$ is tight and any weak limit $x(\cdot)$ satisfies

$$(2.1) \quad \dot{x} = E^x f(x, \xi) = \int f(x, \xi) P^x(d\xi) \quad \text{w.p.1,}$$

where $x(0) \in H$. The right-hand side of (2.1) is continuous in x .

Proof. The continuity is a consequence of the tightness and uniqueness A2.4.

Now, by the tightness of $\{\xi_n\}$, we can choose $\delta_n \rightarrow 0$ and nondecreasing compact K_α such that

$$P\{\xi_{\nu_\alpha} \in K_\alpha\} \rightarrow 1 \quad \text{for any sequence } \{\nu_\alpha\},$$

$$\frac{1}{\delta_n} \sum_{m_l}^{m_l + N(K_n)} a_l \rightarrow 0 \quad \text{uniformly in } l \text{ (where } m_l \geq n), \text{ as } n \rightarrow \infty,$$

$$\frac{1}{\delta_n} \sum_{j \geq n} |a_{j+1} - a_j| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Define the piecewise constant function $\tilde{f}_n(\omega, t)$ by

$$\tilde{f}_n(\omega, t) = \frac{1}{\delta_n} \sum_{m_l}^{m_{l+1}-1} a_j E_{m_l} f(X_j, \xi_j) I_{K_n}(\xi_{m_l-1}) \quad \text{on } [t_{m_l} - t_n, t_{m_{l+1}} - t_n),$$

and set $\tilde{F}_n(\omega, t) = \int_0^t \tilde{f}_n(\omega, s) ds$.

By A2.6, $\{x^n(\cdot), \tilde{F}_n(\cdot), n \geq 0\}$ is tight in $C^{2r}[0, \infty)$. Henceforth, we work with a weakly convergent subsequence (called subsequence 1), also indexed by n , and with limit $(x(\cdot), \tilde{F}(\cdot))$, or with a subsequence of it. We use Skorokhod imbedding wherever convenient with no notational change. Thus, we can assume where convenient, that w.p.1. $(x^n(\cdot), \tilde{F}_n(\cdot))$ converges to $(x(\cdot), \tilde{F}(\cdot))$ uniformly on bounded time intervals. Suppose that

$$(2.2) \quad \tilde{F}(t) = \int_0^t E^{x(u)} f(x(u), \xi) du$$

and that for arbitrary k, t, s and $s_1 < s_2 < \dots < s_k < t < t+s$ and bounded and continuous $h(\cdot)$,

$$(2.3) \quad Eh(x(s_j), \tilde{F}(s_j), j \leq k)[x(t+s) - x(t) - (\tilde{F}(t+s) - \tilde{F}(t))] = 0.$$

Then $M(t) = x(t) - x(0) - \tilde{F}(t)$ is a continuous martingale with $M(0) = 0$. By A2.6, $M(\cdot)$ satisfies a Lipschitz condition. Thus its quadratic variation is zero. Hence $M(t) \equiv 0$ w.p.1, and (2.1) holds. So, we only need to prove (2.2) and (2.3).

For smooth $h(\cdot)$,

$$(2.4) \quad Eh(x^n(s_j), \tilde{F}_n(s_j), j \leq k) \left[x^n(t+s) - x^n(t) - \sum_{m(t_n+t)}^{m(t_n+t+s)-1} a_j f(X_j, \xi_j) \right] \equiv \varepsilon_n,$$

$$(2.5) \quad Eh(x^n(s_j), \tilde{F}_n(s_j), j \leq k) \left[x^n(t+s) - x^n(t) - \int_t^{t+s} \tilde{f}_n(s) ds \right] \equiv \varepsilon'_n,$$

where ε_n and ε'_n go to zero as $n \rightarrow \infty$.

We now prove

$$(2.6) \quad \tilde{f}_n(s) \rightarrow E^{x(s)} f(x(s), \xi) \text{ in probability for each } s.$$

The limit (2.6) implies that $\tilde{F}_n(\cdot)$ converges in measure ($\omega \times t$) to the right side of (2.2). This and the weak convergence of $(x^n(\cdot), \tilde{F}_n(\cdot))$ to $(x(\cdot), \tilde{F}(\cdot))$ and (2.5) yield (2.3), with $\tilde{F}(\cdot)$ defined by (2.2). So, only (2.6) needs to be proved.

Fix s and $\{m_l\}$ such that² $s \in [t_{m_l} - t_n, t_{m_l+1} - t_n)$. Let N_0 denote the null set on which $(x^n(\cdot), \tilde{F}_n(\cdot))$ does not converge uniformly to $(x(\cdot), \tilde{F}(\cdot))$ on bounded intervals (under the Skorokhod imbedding). It is enough to show that (for each s) each subsequence of subsequence 1 contains a further subsequence for which the limit in (2.6) holds in probability. Select a subsequence of subsequence 1, indexed also by n but called subsequence 2, such that³ $P\{\xi_{m_l-1} \in K_n, \text{ all large } n\} = 1$. Let $N(s)$ denote the exceptional null set. Fix $\omega \notin N_0 \cup N(s)$, and extract a weakly convergent subsequence (a subsequence of subsequence 2) of the set of measures (tight by A2.8 and the properties of δ_n) $Q = \{Q_{K_n}(\omega, l, n, \cdot) : n \in \text{subsequence 2}, s \text{ fixed as above}\}$, with limit $\bar{P}_\omega(\cdot)$. The limits in (2.7) below are on this subsequence. Let $g(\cdot)$ be bounded and continuous and set $G(x, \xi) = \int P_x(\xi, 1, d\xi') g(\xi')$. Then by A2.3, A2.5, and A2.8

$$\begin{aligned} & \lim_n \sum_{m_l}^{m_{l+1}-1} \frac{a_j}{\delta_n} \int P(X_{m_l}, \xi_{m_l-1}, m_l, j - m_l, d\xi) g(\xi) I_{K_n}(\xi_{m_l-1}) \\ &= \lim_n \sum_{m_l}^{m_{l+1}-1} \frac{a_j}{\delta_n} \int P(X_{m_l}, \xi_{m_l-1}, m_l, j - m_l - 1, d\xi', dx') \\ & \quad \cdot P_x(\xi', 1, d\xi) g(\xi) I_{K_n}(\xi_{m_l-1}) \\ (2.7) \quad &= \lim_n \int Q_{K_n}(\omega, l, n, d\xi') G(X_{m_l}, \xi') \\ &= \int \bar{P}_\omega(d\xi') G(x(s), \xi') \\ &= \int \bar{P}_\omega(d\xi') P_x(\xi', 1, d\xi) g(\xi). \end{aligned}$$

Similarly, the limit in the first line is $\int \bar{P}_\omega(d\xi) g(\xi)$. In going from the second to the third line of (2.7), we used the facts that $G(\cdot, \xi')$ is uniformly continuous on compact ξ' , that $\sup_{m_{l+1} \geq j \geq m_l} |X_j - X_{m_l}| \rightarrow 0$ as $n \rightarrow \infty$, that $X_{m_l} \rightarrow x(s)$ and the tightness of the set Q .

Due to the arbitrariness of $g(\cdot)$, and the uniqueness A2.4 and to the equality of the last line of (2.7) with $\int \bar{P}_\omega(d\xi) g(\xi)$, we have $\bar{P}_\omega(\cdot) = P^{x(s)}(\cdot)$. Again, by the uniqueness, the limit does not depend on the chosen subsequence of Q . Thus, if $\omega \notin N_0 \cup N(s)$, $Q_{K_n}(\omega, l, n, \cdot) \rightarrow P^{x(s)}(\cdot)$ weakly as $n \rightarrow \infty$, where n now indexes the second chosen “subsequence 2” of the theorem.

² That is, for each n , choose $m_l = m(l, n)$ such that s is in the indicated interval. Keep in mind that l depends on n , and that we suppress the n -dependence of $m(l, n)$ in the notation.

³ By the tightness of $\{\xi_j\}$, we can always choose such a subsequence.

Using A2.7 we now have (limits are on the “subsequence 2”)

$$\begin{aligned}
 & \lim_n \sum_{m_l}^{m_{l+1}-1} \frac{a_j}{\delta_n} \int P(X_{m_l}, \xi_{m_l-1}, m_l, j - m_l, d\xi', dx) P_x(\xi', 1, d\xi) f(x, \xi) I_{K_n}(\xi_{m_l-1}) \\
 &= \lim_n \sum_{m_l}^{m_{l+1}-1} \frac{a_j}{\delta_n} \int P(X_{m_l}, \xi_{m_l-1}, m_l, j - m_l - c, d\xi', dx') \\
 & \quad \cdot P(x', \xi', j - c, c, d\xi, dx) P_x(\xi, 1, d\xi'') f(x, \xi'') I_{K_n}(\xi_{m_l-1}) \\
 &= \lim_n \sum_{m_l}^{m_{l+1}-1} \frac{a_j}{\delta_n} \int P(X_{m_l}, \xi_{m_l-1}, m_l, j - m_l - c, d\xi', dx) \\
 & \quad \cdot P_x(\xi', c + 1, d\xi) f(x, \xi) I_{K_n}(\xi_{m_l-1}) \\
 (2.8) \quad &= \lim_n \sum_{m_l}^{m_{l+1}-1} \frac{a_j}{\delta_n} \int P(X_{m_l}, \xi_{m_l-1}, m_l, j - m_l - c, d\xi') I_{K_n}(\xi_{m_l-1}) \\
 & \quad \cdot [P_{X_{m_l}}(\xi', c + 1, d\xi) f(X_{m_l}, \xi)] \\
 &= \lim_n \int Q_{K_n}(\omega, l, n, d\xi') P_{x(s)}(\xi', c + 1, d\xi) f(x(s), \xi) \\
 &= \int \bar{P}_\omega(d\xi') P_{x(s)}(\xi', c + 1, d\xi) f(x(s), \xi) \\
 &= \int P^{x(s)}(d\xi) f(x(s), \xi).
 \end{aligned}$$

In going from the 3rd to the 4th and then to the 5th line we used the continuity in (x, ξ) of $\int P_x(\xi, c + 1, d\xi') f(x', \xi')$ and the facts concerning convergence cited below (2.7). In going from the next to last step to the last step, we used the fact that $\bar{P}_\omega(\cdot) = P^{x(s)}(\cdot)$ is an invariant measure for the transition function $P_{x(s)}(\xi, j, \cdot)$ for each $x(s)$. The equality of the first and last lines of (2.8) for $\omega \notin N_0 \cup N(s)$ yields the desired result (2.6) for “subsequence 2.” But, since each subsequence of subsequence 1 contains a further subsequence satisfying the requirements of subsequence 2 (but with perhaps a different $N(s)$) (2.6) holds for subsequence 1 also. Furthermore, the limits for all possible subsequence 1's differ only in the initial condition $x(0)$. \square

The proof of the theorem also implies the following corollary:

COROLLARY. Assume A2.1–A2.8 but let X_n and $f(\cdot, \cdot)$ take values in H , a compact subset of a linear metric space \bar{H} , with an invariant metric $d(\cdot)$. Then $\{x^n(\cdot)\}$ is tight on $C^H([0, \infty))$ and if $x(\cdot)$ is the limit of any weakly convergent subsequence, for each continuous linear functional λ on \bar{H} ,

$$\lambda(x(t) - x(0)) = \int_0^t ds \int \lambda f(x(s), \xi) P^{x(s)}(d\xi) \quad \text{w.p.1.}$$

Some more detail on the abstract case is in [12].

Extension. In many problems of interest (see § 4), the algorithm (1.1) takes the form $X_{n+1} = X_n + a_n f_n(\omega)$, where

$$E[f_n(\omega) | X_i, \xi_{i-1}, i \leq n] \equiv F_n(X_n, \xi_{n-1}),$$

and $F_n(x, \xi) \rightarrow F(x, \xi)$, a continuous function, uniformly in (x, ξ) on compact sets. Then Theorem 1 still holds. This extension is useful when $f_n(\cdot)$ depends on variables other

than (X_n, ξ_n) ; for example, it might depend on a "choice" or "logical" variable Z_n , where $P(Z_n = 1 | X_i, \xi_{i-1}, i \leq n) = q(X_n, \xi_{n-1})$, for some continuous function $q(\cdot)$.

Nonunique $P^x(\cdot)$. A very similar result to Theorem 1 can be obtained when the uniqueness in (A2.4) is dropped. Let $\mathcal{P}^x = \{P_\alpha^x(\cdot), \alpha \in \text{some set } A(x)\}$ denote the set of invariant measures for the transition function $P_x(\xi, j, \cdot)$. Assume that $\mathcal{P} = \{\mathcal{P}^x, x \in \text{a compact set}\}$ is tight. For each x , \mathcal{P}^x is convex and weakly compact. Define the set

$$C(x) = \left\{ y : y = \int P_\alpha^x(d\xi) f(x, \xi), \alpha \in A(x) \right\}.$$

Then $C(x)$ is closed and convex. The sets \mathcal{P}^x and $C(x)$ are upper semi-continuous in x in the Hausdorff topology (with the metrized weak topology on the space of distributions and the metric topology on H). This is a consequence of the fact that under the tightness of \mathcal{P} , if $x_n \rightarrow x$ and $P^{x_n}(\cdot) \in \mathcal{P}^{x_n}$, then $\{P^{x_n}(\cdot)\}$ is tight, and all weak limits are in \mathcal{P}^x by A2.3.

THEOREM 2. *Assume A2.1–A2.8, with A2.4 altered as above. Then $\{x^n(\cdot)\}$ is tight and any weak limit $x(\cdot)$ satisfies*

$$(2.9) \quad \dot{x} \in C(x) \quad \text{for almost all } \omega, t.$$

Remarks on the proof. The proof is essentially the same as that of Theorem 1, and we only remark on a couple of points. By the argument of Theorem 1, if $\omega \notin N_0 \cup N(s)$ is fixed and n indexes a weakly convergent subsequence of the set of measures Q defined above (2.7), then we must have

$$(2.10) \quad \lim_n \sum_{m_l}^{m_{l+1}-1} \frac{a_j}{\delta_n} \int P(X_{m_l}, \xi_{m_l-1}, m_l, j - m_l, dx', d\xi') f(x', \xi') I_{K_n}(\xi_{m_l-1}) \\ = \lim_n \tilde{f}_n(s) = \int f(x(s), \xi) P_\alpha^{x(s)}(d\xi) \in C(x(s)),$$

for some $\alpha \in A(x(s))$, perhaps depending on ω and s and on the selected subsequence. Under the weak convergence of the selected subsequence and the Skorokhod imbedding, $\tilde{F}_n(\cdot) \rightarrow \tilde{F}(\cdot)$ (which is absolutely continuous) uniformly on bounded time intervals w.p.1, but $\tilde{f}_n(s)$ does not necessarily converge in probability to $\tilde{f}(s) = \tilde{F}(s)$ as it did in Theorem 1. But, for each fixed $\omega \notin N_0$ and each $T < \infty$, $\tilde{f}_n(\cdot)$ (considered as a function on $[0, T]$) converges (along any "subsequence 1") to $\tilde{f}(\cdot)$ weakly when these functions are considered as elements of $L_1[0, T]$. Thus for each $\omega \notin N_0$ there are $\{\beta_{ni}, i \leq n\}$ such that $0 \leq \beta_{ni}$, $\sum_{i=1}^n \beta_{ni} = 1$, $\beta_{ni} \rightarrow 0$ as $n \rightarrow \infty$ for each i , and $\sum_{i=1}^n \beta_{ni} \tilde{f}_i(\cdot) \rightarrow \tilde{f}(\cdot)$ in the norm of $L_1[0, T]$. This convergence, together with the limit (2.10), the convexity and closure of $C(x)$ and the upper semi-continuity cited above Theorem 2 imply that $\tilde{f}(s) \in C(x(s))$ for almost all (ω, s) . See [12] for more detail.

The projection algorithm (1.2). Recall the definition of π_G from § 1. Let $\bar{\pi}(h(\cdot))$ denote the (not necessarily unique) projection of the vector field $h(\cdot)$ onto G ; i.e.,

$$(2.11) \quad \bar{\pi}(h(x)) = \text{set of limits } \lim_{\Delta \downarrow 0} \frac{[\pi_G(x + \Delta h(x)) - x]}{\Delta}.$$

THEOREM 3. *Assume (1.2) and the conditions above it instead of (1.1), and assume A2.1–A2.8. Then $x^n(\cdot)$ is tight and if $x(\cdot)$ is the limit of a weakly convergent subsequence, $x(\cdot)$ satisfies the "projected" equation*

$$(2.12) \quad \dot{x} = \bar{\pi}(E^x f(x, \xi)) \quad \text{for almost all } \omega, t.$$

Recall the extension of Theorem 1 to the algorithm $X_{n+1} = X_n + a_n f_n(\omega)$, cited after Theorem 1. If $\pi_G(X_n + a_n f_n(\omega))$ is used, then Theorem 3 holds with $\dot{x} = \tilde{\pi}(E^x F(x, \xi))$.

Remarks on the proof. The proof is quite similar to that of Theorem 1, and only a few remarks will be made. Use the partition of [2, (5.3.4)] to write (1.2) in the form

$$(2.13) \quad X_{n+1} = X_n + a_n f(X_n, \xi_n) + a_n d_n,$$

where $a_n d_n = \pi_G(X_n + a_n f(X_n, \xi_n)) - (X_n + a_n f(X_n, \xi_n))$. In [2], $a_n d_n$ is termed τ_n . Using the notation of Theorem 1, define the piecewise constant function $\tilde{d}_n(\omega, t)$ by

$$\tilde{d}_n(\omega, t) = \frac{1}{\delta_n} \sum_{m_l}^{m_{l+1}-1} a_j d_j \quad \text{on } [t_{m_l} - t_n, t_{m_{l+1}} - t_n),$$

and set $\tilde{D}_n(\omega, t) = \int_0^t \tilde{d}_n(\omega, s) ds$. Then, as in Theorem 1, $\{x^n(\cdot), \tilde{F}_n(\cdot), \tilde{D}_n(\cdot)\}$ is tight. Extract a convergent subsequence with limit $(x(\cdot), \tilde{F}(\cdot), \tilde{D}(\cdot))$. Both $\tilde{F}(\cdot)$ and $\tilde{D}(\cdot)$ are absolutely continuous and $\tilde{F}(\cdot)$ satisfies (2.2). Write $\tilde{f}(\cdot) = \dot{\tilde{F}}(\cdot)$ and define $\tilde{d}(\cdot)$ by

$$\tilde{D}(t) = \int_0^t \tilde{d}(s) ds.$$

Write

$$\tilde{f}(t) = \tilde{\pi}(\tilde{f}(t)) + \hat{f}(t), \quad \hat{f}(t) = \tilde{f}(t) - \pi(\tilde{f}(t)),$$

the $\hat{f}(\cdot)$ term being a "projection error." By the method of Theorem 1,

$$(2.14) \quad \begin{aligned} x(t) - x(0) - \tilde{F}(t) - \tilde{D}(t) &= 0 \quad \text{w.p.1.}, \\ \dot{x} &= E^x f(x, \xi) + \tilde{d} = \tilde{f} + \tilde{d}, \\ \dot{x}(t) &= \tilde{\pi}(\tilde{f}(t)) + \hat{f}(t) + \tilde{d}(t) \quad \text{w.p.1.} \end{aligned}$$

Using the ideas of [2, § 5.3], it can be shown that $-\int_0^t \hat{f}(s) ds = \tilde{D}(t)$. This and (2.14) imply (2.12). We omit the rest of the details. We note only that the proof of the last equality uses the facts that if $X_{n+1} \in \partial G$, then d_n is in the cone $-K(X_{n+1})$, and that if $x(t) \in \partial G$, then $\hat{f}(t)$ is in the cone $K(x(t))$, where

$$K(x) = \left\{ y : y = \sum_{i: q_i(x)=0} \lambda_i q_{i,x}(x), \text{ for some set of } \lambda_i \geq 0 \right\}.$$

Unbounded $f(\cdot)$. We will use:

A2.9. There are a $K < \infty$ and a positive-valued function $d(\cdot)$ such that $|f(x, \xi)| \leq K_1(1 + d(\xi))$, and x takes values in the Euclidean space R^r . For some $\alpha > 0$, $\sup_j E|d(\xi_j)|^{1+\alpha} < \infty$.

THEOREM 4. Under A2.9, and the tightness of $\{X_n\}$, both $\{x^n(\cdot)\}$ and $\{\tilde{F}_n(\cdot)\}$ are tight in $C^r[0, \infty)$.

Proof. Both $x^n(\cdot)$ and $\tilde{F}_n(\cdot)$ are sums of terms of the types $a_j f(X_j, \xi_j)$ and $a_j E_{m_l} f(X_j, \xi_j)$, for $j \geq m_l$ respectively. These are bounded by $a_j K_1(1 + d(\xi_j))$ and $a_j K_1(1 + E_{m_l} d(\xi_j))$, respectively. But by A2.9, both $\{d(\xi_j)\}$ and $\{E_{m_l} d(\xi_j)\}$, $l, j: j \geq m_l$ are uniformly integrable, which implies the theorem. \square

Given the tightness, the only further impediment to the result of Theorem 1 for the unbounded $f(\cdot, \cdot)$ case, concerns the meaning of the integrals in (2.8). A truncation and limit argument seems the most natural. We simply take the following natural approach.

Suppose that there is a sequence $\{f_L(x, \xi), L = 1, 2, \dots\}$ each member of which satisfies A2.6, A2.7, where c does not depend on L , and such that $E^x f_L(x, \xi) \rightarrow E^x f(x, \xi)$

uniformly on any compact set, as $L \rightarrow \infty$. Let (see A2.9) $|f(x, \xi)| \leq K(1 + d(\xi))$, and let $f_L(x, \xi) = f(x, \xi)$ when $d(\xi) \leq L$. For each $T < \infty$, let $E \sum_{j=n}^{m(t_n+T)} a_j d(\xi_j) I\{d(\xi_j) \geq L\} \rightarrow 0$ as $L \rightarrow \infty$, uniformly in $n \leq m(t_n + T)$. Then under A2.1–A2.5 and A2.8, the conclusion of Theorems 1 and 3 hold. The condition of the next to last sentence is guaranteed by A2.9 and also implies the tightness.

3. The non-Markov case. The ideas of the last section can be extended to some interesting non-Markov systems where, loosely speaking, if $X_n \equiv x$ (a constant) for all $-\infty < n < \infty$, then $\{\xi_n\}$ is stationary and has certain mixing properties. We next state some assumptions, which are modifications of some in § 2. Then a general convergence theorem is proved. Lastly, it will be shown that the assumptions hold in many cases of interest.

In particular, in Theorem 8 we verify A3.2, A3.3 when $\{\xi_n\}$ is not state-dependent and satisfies a type of ϕ -mixing condition. This case is of interest, since the non-Markov noise and discontinuous dynamics case is usually hard and occurs frequently. In this nonstate dependent case, the measure $P_x(\xi, 1, \cdot) = P(\xi, 1, \cdot)$ below would not depend on x , and would equal the *stationary conditional distribution* $P\{\xi_1 \in \cdot | \xi_0 = \xi\}$ provided that this stationary measure is weakly continuous in ξ .

A3.1. $X_n \in H$, a compact subset of E' and $f(\cdot)$ is bounded.

The limits in (A3.2) and (A3.3) are in the sense of probability.

A3.2. For each x , there is a transition function $P_x(\xi, 1, \cdot)$ which is weakly continuous in (x, ξ) and such that for each bounded and continuous $g(\cdot)$, with $G(x, \xi) = \int P_x(\xi, 1, d\xi') g(\xi')$,

$$\lim_{\substack{j \rightarrow \infty \\ n \rightarrow \infty}} \left[\int P(d\xi_j | X_j, \xi_{j-1}; X_u, \xi_{u-1}, u \leq n) g(\xi_j) - G(X_j, \xi_{j-1}) \right] = 0.$$

A3.3. Define $F(x, \xi) = \int f(x, \xi') P_x(\xi, 1, d\xi')$. Then $F(\cdot, \cdot)$ is continuous and

$$\lim_{\substack{j \rightarrow \infty \\ n \rightarrow \infty}} \left[\int P(d\xi_j | X_j, \xi_{j-1}; X_u, \xi_{u-1}, u \leq n) f(X_j, \xi_j) - F(X_j, \xi_{j-1}) \right] = 0.$$

A3.4. For the Markov process with transition function $P_x(\xi, j, \cdot)$ (which is obtained recursively from $P_x(\xi, 1, \cdot)$ as above A2.4), there is a unique invariant measure $P^x(\cdot)$. The set S is compact. (Hence, $\{P^x(\cdot), x \in H\}$ is tight.)

We define the measure $Q(\omega, l, n, \cdot)$ similarly to that in § 2: i.e., by

$$\int Q(\omega, l, n, d\xi) g(\xi) = \frac{1}{\delta_n} \sum_{j=m_l}^{m_{l+1}-1} a_j \int P(d\xi_j | X_u, \xi_{u-1}, u \leq n) g(\xi_j),$$

where $g(\cdot)$ is an arbitrary bounded measurable function. (Here, we use ξ_j in the sum $Q(\cdot)$; in Theorem 1, ξ_{j-1} was used. The choice is unimportant and is due to notational convenience.)

THEOREM 5. Assume A3.1–A3.4 and A2.5, A2.6. Then $\{x^n(\cdot)\}$ is tight and the limit $x(\cdot)$ of any weakly convergent subsequence satisfies (2.1). If (1.2) is used in lieu of (1.1) and the conditions above (1.2) hold, with X_n in some Euclidean space, then $\{x^n(\cdot)\}$ is tight and the limit of any weakly convergent subsequence satisfies (2.12).

Proof. The proof is close to that of Theorem 1 and we use the same terminology, but with the measure $Q(\omega, l, n, \cdot)$ defined as above and δ_n satisfying the conditions above A2.8 and in the proof of Theorem 1. Since S is compact the truncation factor I_K used in Theorem 1 is not required. By A3.1, $\{x^n(\cdot), \tilde{F}_n(\cdot)\}$ is tight in $C^{2r}[0, \infty)$.

Extract a weakly convergent subsequence (also indexed by n and called subsequence 1) with limit $x(\cdot)$, $\tilde{F}(\cdot)$. We work with this sequence or subsequences of it, henceforth. By the Skorokhod imbedding, there is a null set N_0 such that the limit can be taken to be uniform on bounded intervals, for $\omega \notin N_0$.

Let $g(\cdot)$ be bounded and continuous.

By A3.2, in getting the limit in probability (as $n \rightarrow \infty$) of an expression of the form

$$(3.1) \quad \sum_{m_l}^{m_{l+1}-1} \frac{a_j}{\delta_n} \int P(d\xi_j, dX_j | X_u, \xi_{u-1}, u \leq m_l) g(\xi_j) \\ = \sum_{m_l}^{m_{l+1}-1} \frac{a_j}{\delta_n} \int P(d\xi_{j-1}, dX_j | X_u, \xi_{u-1}, u \leq m_l) \\ \cdot P(d\xi_j | X_j, \xi_{j-1}, X_u, \xi_{u-1}, u \leq m_l) g(\xi_j),$$

we can substitute $G(X_j, \xi_{j-1})$ for $\int P(d\xi_j | X_j, \xi_{j-1}, X_u, \xi_{u-1}, u \leq m_l) g(\xi_j)$, when $m_{l+1} > j \geq m_l$. Fix s . For each n , fix $m_l = m(l, n)$ such that $s \in [t_{m_l} - t_n, t_{m_{l+1}} - t_n)$. Under the Skorokhod imbedding $E|G(X_j, \xi_{j-1}) - G(x(s), \xi_{j-1})| \rightarrow 0$. Thus, by the last three sentences

$$E \left| \int Q(\omega, l, n, d\xi) [g(\xi) - G(x(s), \xi)] \right| \xrightarrow{n} 0.$$

Choose a subsequence (called subsequence 2) for which

$$\int Q(\omega, l, n, d\xi) [g(\xi) - G(x(s), \xi)] \rightarrow 0 \text{ w.p.1}$$

for a countable dense set of bounded and continuous $g(\cdot)$, hence for all bounded and continuous $g(\cdot)$. Denote the exceptional ω -set by $\tilde{N}(s)$. For fixed $\omega \notin N_0 \cup \tilde{N}(s)$, choose a further subsequence (termed subsequence 3) for which $Q(\omega, l, n, \cdot)$ converges weakly to some measure $\bar{P}_\omega(\cdot)$.

Then

$$\int \bar{P}_\omega(d\xi) g(\xi) = \int \bar{P}_\omega(d\xi) G(x(s), \xi) = \int \bar{P}_\omega(d\xi) P_{x(s)}(\xi, 1, d\xi') g(\xi').$$

Thus, by uniqueness $\bar{P}_\omega(\cdot) = P^{x(s)}(\cdot)$, a result which does not depend on the particular subsequence 3 chosen. Hence $Q(\omega, l, n, \cdot) \rightarrow P^{x(s)}(\cdot)$ weakly along subsequence 2, for almost all ω .

Using this last result, A3.3 and a factorization similar to the one used in (2.8), we get that

$$(3.2) \quad \sum_{m_l}^{m_{l+1}-1} \frac{a_j}{\delta_n} \int P(d\xi_j, dX_j | X_u, \xi_{u-1}, u \leq m_l) f(X_j, \xi_j) \rightarrow \int P^{x(s)}(d\xi) f(x(s), \xi)$$

in probability as $n \rightarrow \infty$ along subsequence 2. Hence (3.2) holds in probability as $n \rightarrow \infty$ along subsequence 1.

We omit the details for the projection algorithm. These use an adaptation of the above method which is analogous to the modification of the proof of Theorem 1 which is used in Theorem 3. \square

Remark. Let $X_{n+1} = X_n + a_n f_n(\omega)$ replace (1.1), and suppose that there is a continuous $F(\cdot)$ such that

$$E[f_j(\omega) | X_j, \xi_{j-1}; X_u, \xi_{u-1}, u \leq n] - F(X_j, \xi_{j-1}) \xrightarrow{P} 0$$

as $j - n \rightarrow \infty$ and $n \rightarrow \infty$. Then the theorem continues to hold.

We now examine A3.2, A3.3, under the following ϕ -mixing condition, where the noise does not depend on the state.

A3.5. Let S be compact. Let \mathcal{F}_j , \mathcal{F}_0^m and \mathcal{F}_m^∞ denote the σ -algebras which measure ξ_{j-1} , $\{\xi_{j-1}, j \leq m\}$ and $\{\xi_{j-1}, j \geq m\}$, resp. There are $\{\phi_n\}$ such that $\phi_n \rightarrow 0$ and for each n and any $A \in \mathcal{F}_0^m$, $B \in \mathcal{F}_{n+m}^\infty$, $\phi_n P(B) \geq |P(AB) - P(A)P(B)| \leq \phi_n P(A)$.

The following result is well known.

LEMMA 6. Let $A_{n+m} \in \mathcal{F}_{n+m}^\infty$, $A_m \in \mathcal{F}_0^m$, and assume A3.5. Then as $n \rightarrow \infty$

$$|P\{A_{n+m}|\mathcal{F}_0^m\} - P\{A_{n+m}\}| \xrightarrow{P} 0,$$

$$|P\{A_m|\mathcal{F}_{n+m}^\infty\} - P\{A_m\}| \xrightarrow{P} 0,$$

$$E|P\{A_m|\mathcal{F}_{n+m}^\infty\} - P\{A_m\}|I_{A_{n+m}} \leq \phi_n P\{A_{n+m}\} \text{ and } \leq \phi_n P\{A_m\},$$

$$E|P\{A_{n+m}|\mathcal{F}_0^m\} - P\{A_{n+m}\}|I_{A_m} \leq \phi_n P\{A_{n+m}\} \text{ and } \leq \phi_n P\{A_m\}.$$

Let $|g_j| \leq 1$ with g_j being \mathcal{F}_j measurable. Then as $n \rightarrow \infty$

$$E[g_{n+m}|\mathcal{F}_0^m] - E g_{n+m} \xrightarrow{P} 0, \quad E[g_m|\mathcal{F}_{n+m}^\infty] - E g_m \xrightarrow{P} 0,$$

where the convergences are uniform in $\{g_j\}$, $\{A_m, A_{m+n}\}$ and m .

THEOREM 7. Assume A3.5. Let g_j be \mathcal{F}_j measurable with $|g_j| \leq 1$, and let $j > n$. Then $E[g_j|\mathcal{F}_0^n \cup \mathcal{F}_{j-1}] - E[g_j|\mathcal{F}_{j-1}] \rightarrow 0$ as $j - n \rightarrow \infty$, uniformly in j and $\{g_j\}$.

Proof. Let $G_{j-1} \in \mathcal{F}_{j-1}$ and $G_{0,n} \in \mathcal{F}_0^n$. The $v_{n,j}^i$ below are uniformly bounded and $E|v_{n,j}^i|I_{G_{j-1}} \leq 2\phi_{j-n}P\{G_{j-1}\}$, by Lemma 6. We have

$$\begin{aligned} \int_{G_{j-1} \cap G_{0,n}} E[g_j|\mathcal{F}_0^n \cup \mathcal{F}_{j-1}] dP &= \int_{G_{j-1}} g_j I_{G_{0,n}} dP \\ &= \int_{G_{j-1}} E[g_j I_{G_{0,n}}|\mathcal{F}_{j-1}] dP \\ &= \int_{G_{j-1}} E\{g_j E[I_{G_{0,n}}|\mathcal{F}_{j-1}^j]|\mathcal{F}_{j-1}\} dP \\ (3.3) \quad &= \int_{G_{j-1}} E[g_j|\mathcal{F}_{j-1}](P\{G_{0,n}\} + v_{n,j}^1) dP \\ &= \int_{G_{j-1}} E[g_j|\mathcal{F}_{j-1}](I_{G_{0,n}} + v_{n,j}^2) dP \\ &= \int_{G_{j-1} \cap G_{0,n}} E[g_j|\mathcal{F}_{j-1}] dP + \int_{G_{j-1}} v_{n,j}^3 dP. \end{aligned}$$

Now, suppose that the theorem is false. Then there is a sequence of sets $H_{j,n} \in \mathcal{F}_0^n \cup \mathcal{F}_{j-1}$ and an $\varepsilon > 0$ such that for some sequence $\{g_j\}$ and $j - n \rightarrow \infty$,

$$(3.4) \quad \int_{H_{j,n}} \{E[g_j|\mathcal{F}_0^n \cup \mathcal{F}_{j-1}] - E[g_j|\mathcal{F}_{j-1}]\} dP \geq \varepsilon$$

(and/or $\leq -\varepsilon$, we use (3.4) only for simplicity). For each $\delta > 0$, there are sets $G_{j-1}^i \in \mathcal{F}_{j-1}$ and $G_{0,n}^i \in \mathcal{F}_0^n$ with $\{G_{j-1}^i, i = 1, 2, \dots\}$ disjoint and $P\{H_{j,n}^\sigma \Delta H_{j,n}\} \leq \delta$, where $H_{j,n}^\sigma = \bigcup_i [G_{j-1}^i \cap G_{0,n}^i]$. Now, re-do the calculation (3.3) with G_{j-1} and $G_{0,n}$ superscripted

by i , and the integrals summed over i . For small enough $\delta > 0$, this yields a contradiction to (3.4), since $\sum_i E|v_{n,j}^3|I_{G^j_{j-1}} \rightarrow 0$ as $j - n \rightarrow \infty$. \square

THEOREM 8. Assume A3.1–A3.5. Let $\{\xi_j\}$ not depend on $\{X_j\}$; i.e., for all j

$$P\{d\xi_j|\xi_{u-1}, X_u, u \leq j\} = P\{d\xi_j|\xi_{u-1}, u \leq j\}.$$

Suppose that there is a measure $P(\xi, 1, \cdot)$ on Borel sets of S such that $P(\cdot, 1, B)$ is measurable for each Borel $B \subset S$. For each bounded, real-valued and continuous $g(\cdot)$, let

$$(3.5) \quad \begin{aligned} \int g(\xi_j)P(d\xi_j|\xi_{j-1} = \xi) &\rightarrow \int g(\xi')P(\xi, 1, d\xi') = G(\xi) \quad \text{for all } \xi, \\ \int f(x, \xi_j)P(d\xi_j|\xi_{j-1} = \xi) &\rightarrow \int f(x, \xi')P(\xi, 1, d\xi') = F(x, \xi) \quad \text{for each } x \text{ and } \xi, \end{aligned}$$

where $F(\cdot, \cdot)$ and $G(\cdot)$ are continuous. Then A3.2 and A3.3 hold.

The proof follows from Theorem 7, by letting g_j be either $g(\xi_j)$ or $f(x, \xi_j)$.

4. Examples.

4a. Application to a routing problem. To illustrate the power of the method, we consider the automata routing example described in [8, § 3]. Calls arrive at a transmitting or switching terminal at random, at discrete time instants $n = 0, 1, 2, \dots$, with $P\{\text{one call arrives at } n\text{th instant}\} = \mu$, $\mu \in (0, 1)$, $P\{>1 \text{ call arrives at } n\text{th instant}\} = 0$. From the terminal, there are two possible routings to the destination, route 1 and route 2; the i th route has N_i independent lines and can thus handle up to N_i calls simultaneously. Let $[n, n+1)$ denote the n th interval of time. The duration of each call is a random variable with a geometric distribution: $P\{\text{call completed in the } (n+1)\text{st interval} | \text{uncompleted at end of } n\text{th interval, route } i \text{ used}\} = \lambda_i$, $\lambda_i \in (0, 1)$. The members of the double sequence of the interarrival times and call durations are mutually independent. In [8], the “gain” per step was a constant, and a detailed study was made of the rate of convergence. Here, we do a stochastic approximation version; i.e., $a_n \rightarrow 0$. But the case where $a_n \equiv a > 0$ is handled in the same way. Let $\{y_n\}$ denote a sequence of random variables with values in $[0, 1]$. To get an unambiguous formulation, suppose that calls terminating in the n th interval actually terminate at $n + \frac{1}{2}$, and arrivals and route assignments are at the instants $0, 1, 2, \dots$. Define $\xi_n = (\xi_n^1, \xi_n^2) =$ route occupancy process (called X_n^e in [8]), where $\xi_n^i =$ number of lines of route i occupied at time n^+ . If a call arrives at instant $n+1$, the automaton “flips a coin,” choosing route 1 with probability y_n and route 2 with probability $(1 - y_n)$. If all lines of the chosen route i are occupied at instant $(n+1)^-$, then the call is switched to route j ($j \neq i$). If all lines of route j are also occupied at instant $(n+1)^-$, then the call is rejected, and disappears from the system. The model can be generalized considerably, both in the number of lines and switching nodes, and in the input and call length statistics. Let J_{in} denote the indicator of the event $\{\text{call arrives at } n+1, \text{ is assigned first to route } i \text{ and is accepted by route } i\}$. The algorithm is (4.1), where $0 < \alpha < \beta < 1$ are truncation points, and $y_0 \in (\alpha, \beta)$. The bar $|\alpha$ denotes truncation.

$$(4.1) \quad y_{n+1} = [y_n + a_n(1 - y_n)J_{1n} - a_n y_n J_{2n}]|\alpha.$$

Here, $P\{\xi_{n+1} = \xi' | y_n = y, \xi_n = \xi\}$ is a continuous function of (y, ξ) . The Markov chain is $\{y_n, \xi_n\}$ not $\{X_n, \xi_{n-1}\}$. For each fixed $y \in [a, \beta]$, $\{\xi_n(y), n \geq 0\}$ has a unique invariant measure $P^y(\cdot)$, and $E[J_{in} | y, \xi, l \leq n] = F_i(y, \xi_n)$, where $F_i(\cdot, \cdot)$ is a continuous function

of y for each (discrete) ξ . Define $y^n(\cdot)$ as $x^n(\cdot)$ was defined. By Theorem 1 or 3 and the extension cited after the statement of the theorem we immediately get the correct ODE (which must be satisfied by all the weak limits of $\{y^n(\cdot)\}$)

$$(4.2) \quad \dot{y} = [(1-y)E^y J_{1n} - yE^y J_{2n}] \quad \text{for } y \in (\alpha, \beta),$$

$y(\cdot)$ stops on first hitting α or β .

Simple! No analysis of rates of convergence of n -step transition functions, etc. is required. Also, no analysis of the x -dependence of the $\{\xi_n\}$ or $\{\xi_n(x)\}$ is required. The model [8] upon which the analysis was based appeared in [9].

4b. An adaptive quantizer. Efficient quantization of signals in telecommunications systems is of considerable current interest (e.g., of voice signals in telephone systems). Let the signal process $z(\cdot)$ be sampled at instants $n\Delta$, $n = 0, 1, \dots$, and let the samples $\{z(n\Delta)\}$ be quantized and then transmitted. Adaptive quantizers have been studied as a means to more efficient quantization. The quantization scale for "large" signals, should be different from that for "small" signals. An adaptive quantizer studied in [10], [11] takes roughly the following form. We use $a_n \equiv \varepsilon$, a constant. Let $0 = \psi_0 < \psi_1 < \dots < \psi_{L-1} < \psi_L = \infty$, $0 = \eta_1 < \eta_2 < \dots < \eta_L$, where the ψ_i , η_i are real numbers. For a scaling parameter $y > 0$, define the quantization function $q(\cdot)$. For $z(n\Delta) > 0$, set $q(z(n\Delta)) = y\eta_i$ if $z(n\Delta) \in [y\psi_{i-1}, y\psi_i)$ and set $q(-z) = -q(z)$. The parameter y should vary with the signal power. To get the adaptive quantizer of concern, fix real numbers $0 < M_1^\varepsilon < M_2^\varepsilon < \dots < M_L^\varepsilon < \infty$, where $M_1^\varepsilon < 1$, $M_L^\varepsilon > 1$, and set $\beta \in (0, 1]$. Let $0 < y_l < y_u < \infty$. Then we adapt the scale y according to

$$(4.3) \quad y_{n+1}^\varepsilon = (y_n^\varepsilon)^\beta B_n^\varepsilon|_{y_l}^{y_u}, \quad \text{where } B_n^\varepsilon = M_i^\varepsilon \text{ if } |z(n\Delta)| \in [y_n^\varepsilon \psi_{i-1}, y_n^\varepsilon \psi_i).$$

We do an asymptotic analysis of the sequence $y^\varepsilon(\cdot)$, defined as the piecewise linear interpolation of the function with values y_n^ε at time $n\varepsilon$. Let $y_0^\varepsilon = y_0 \in [y_l, y_u]$.

Now define $l_1 < l_2 < \dots < l_L$, $l_1 < 0$, $l_L > 0$, and $\alpha > 0$ such that $\varepsilon\alpha < 1$. Then set $M_i^\varepsilon = (1 + \varepsilon l_i)$, $\beta = 1 - \varepsilon\alpha$. Then using $y^{1-\varepsilon\alpha} = y[1 - \varepsilon\alpha \log y] + O(\varepsilon^2)$, and $(1 + \varepsilon b_n^2) = B_n^\varepsilon$,

$$(4.4) \quad y_{n+1}^\varepsilon = [y_n^\varepsilon(1 + \varepsilon b_n^\varepsilon) - \varepsilon\alpha y_n^\varepsilon \log y_n^\varepsilon + O(\varepsilon^2)]|_{y_l}^{y_u}$$

$$= [y_n^\varepsilon + \varepsilon F(y_n^\varepsilon, z(n\Delta)) + O(\varepsilon^2)]|_{y_l}^{y_u}.$$

Assume further that $Z(\cdot)$ is a stationary (finite order) Gaussian Markov process with $\text{Cov } Z(t) > 0$ and let $z(t) = h'Z(t)$, for some vector $h \neq 0$. In this example, the noise does not depend on the state and so the analysis is quite simple, even though $z(\cdot)$ is not a bounded process. Define $EF(y, z(0)) = \bar{F}(y)$. Then $\bar{F}(y)$ has a unique zero \bar{y} on $(0, \infty)$, and $\bar{F}(y)$ is positive for $y < \bar{y}$ and negative for $y > \bar{y}$ [8, § 7]. In [8, §§ 7-9], there is a detailed investigation of the limit of $[y_n^\varepsilon - \bar{y}]\sqrt{\varepsilon}$. Here, we are only concerned with the simpler question of the limit of $y^\varepsilon(\cdot)$. For some $c \geq 0$, $E_{Z(0)}F(y, z(n\Delta + \Delta + c\Delta))$ is continuous in $Z(0)$, y and tends to $\bar{F}(y)$ in the mean, uniformly in $y \in [y_l, y_u]$, as $c \rightarrow \infty$. This fact and the method of proof of Theorem 1 or of Theorem 3 and the extension cited after the theorems implies immediately that the weak limit of $\{y^\varepsilon(\cdot)\}$ satisfies

$$(4.5) \quad \dot{y} = \bar{F}(y), \quad y(0) = y_0 \quad \text{if } \bar{y} \in [y_l, y_u],$$

and if $\bar{y} \notin [y_l, y_u]$, $y(\cdot)$ stops on first hitting y_l or y_u .

REFERENCES

- [1] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 551–575.
- [2] H. J. KUSHNER AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Applied Math. Science Series 26, Springer, Berlin, 1978.
- [3] A. BENVENISTE, M. GOURSAT AND G. RUGET, *Analysis of stochastic approximation schemes with discontinuous and dependent forcing terms, with applications to data communication*, IEEE Trans. Automatic Control, AC-25 (1980), pp. 1042–1057.
- [4] H. J. KUSHNER, *Stochastic approximation with discontinuous dynamics and state dependent noise; w.p.1 and weak convergence*, Proc. Conference on Decision and Control, Albuquerque, NM, 1980.
- [5] ———, *On averaging methods for stochastic approximations with discontinuous dynamics and state dependent noise*, Chernoff Festschrift, J. Rustogi, ed., Academic Press, New York, 1983.
- [6] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1969.
- [7] A. V. SKOROKHOD, *Limit theorems for stochastic processes*, Theory Prob. and Appl., 1 (1956), pp. 262–290.
- [8] H. J. KUSHNER AND HAI HUANG, *Averaging methods for the asymptotic analysis of learning and adaptive systems with small adjustment rate*, this Journal, 19 (1981), pp. 635–650.
- [9] K. S. NARENDRA AND M. A. L. THATHACHAR, *On the behavior of a learning automaton in a changing environment with application to telephone traffic routing*, preprint, Dept. Engineering Yale Univ., New Haven, CT, 1979.
- [10] D. J. GOODMAN AND A. GERSHO, *Theory of an adaptive quantizer*, IEEE Trans. Comm., COM-22 (1974), pp. 1037–1045.
- [11] D. MITRA, *A generalized adaptive quantization system with a new reconstruction method for noisy transmission*, IEEE Trans. Comm., COM-27 (1979), pp. 1681–1689.
- [12] A. SHWARTZ, *Convergence of stochastic approximations: the invariant measure approach*, Thesis, Division of Engineering, Brown Univ., Providence, RI, 1982.

DECOUPLING OF MULTIVARIABLE CONTROL SYSTEMS OVER UNIQUE FACTORIZATION DOMAINS*

K. B. DATTA[†] AND M. L. J. HAUTUS[‡]

Abstract. Necessary and sufficient conditions are established for the existence of a state variable feedback decoupling of an m -input, m -output time invariant linear control system over a unique factorization domain. An explicit computation is provided for the feedback and the feedforward gain matrix. Also necessary and sufficient conditions for the existence of a stability-preserving state feedback decoupling are given. The results are illustrated by some examples.

Key words. decoupling, delay systems, systems over rings, multivariable systems, state feedback

1. Introduction. The design and synthesis of noninteracting control in multivariable control systems by state-variable feedback were initiated by Morgan (1964) and definitive results in this direction by establishing necessary and sufficient conditions for the existence of a decoupling feedback, as well as an explicit construction, were first given by Falb and Wolovich (1967). Their results were formulated for systems with real coefficients but they are easily seen to be extendible to systems over arbitrary fields. The extension of these results to systems over rings, however, is less obvious. On the other hand, systems over rings have shown to possess a wide range of potential applications such as delay systems, 2-D systems, parametrized systems, discrete time distributed systems, systems with integer coefficients, etc. We refer to the survey papers E. D. Sontag (1976), (1981), E. W. Kamen (1978), and the references therein.

This abundance of control systems which can conveniently be modelled as systems over rings is a motivation for a systematic investigation of systems over rings. This investigation was started with the thesis Rouchaleau (1972) and the paper Rouchaleau, Wyman and Kalman (1972) and it has received much attention recently.

The purpose of this paper is to formulate necessary and sufficient conditions for the existence of a decoupling state feedback for a linear time-invariant system over a unique factorization domain. This particular class of ring is wide enough to encompass almost all the models arising from the applications mentioned before and on the other hand, it allows a complete solution of the problem. The conditions which will be obtained reduce to the Falb-Wolovich conditions when applied to systems over a field. The method of proof, however, is completely different from the proof in Falb and Wolovich (1967). It can be regarded a generalization to systems over rings of the type of proof given in Hautus and Heymann (1980), (1983) and it is based on a characterization of feedback transformations given in Hautus and Heymann (1978).

It is possible to axiomatize the concept of stability for systems over rings in such a way that in each particular specification and application (delay systems, 2-D systems) the notion of stability customary in that field accommodates conveniently in the general framework. An example will be given in § 2. The treatment is based on what we have called "denominator set". This concept was introduced for systems over a field in

* Received by the editors May 26, 1981, and in revised form March 8, 1982. This research was partially supported by the National Science Foundation under grant ECS-7908673 and by the National Science Council, Taiwan under grant VE 80003.

[†] Department of Electrical Engineering, 11T, Kharagpur-721302, India. Formerly of the Department of Applied Physics, Calcutta University.

[‡] Department of Mathematics, University of Technology, Eindhoven, The Netherlands. This paper was completed while this author was on leave as TRW Visiting Lecturer at the Department of Electrical Engineering Systems, University of Southern California, Los Angeles, California 90007.

and Khargonekar and Sontag (1981) (where the name Hurwitz set is used). In the general framework of stability thus provided we give necessary and sufficient conditions for the existence of a stability preserving decoupling state feedback.

The problem formulation and the main results are given in § 3. Examples, illustrating the results, are given in § 4, and § 5 is devoted to the proof of our main result.

2. Stability of systems over rings. Throughout the paper \mathcal{R} will denote a unique factorization domain (= UFD) or factorial ring (see Samuel (1963), Barshay (1969)). We use the notations $\mathcal{R}[z]$ and $\mathcal{R}(z)$ to denote the rings of polynomials and rational functions over \mathcal{R} , respectively. A polynomial q is called *monic* if its leading coefficient equals 1. A rational function is called *causal* (or *proper*) if it has a representation of the form p/q , where q is a monic polynomial and $\deg p \leq \deg q$.

A *denominator set* is a subset \mathcal{D} of $\mathcal{R}[z]$ satisfying the following conditions:

- (i) \mathcal{D} is *multiplicative*, i.e. $1 \in \mathcal{D}$ and if $p, q \in \mathcal{D}$ then $pq \in \mathcal{D}$.
- (ii) Each polynomial $p \in \mathcal{D}$ is *monic* (in particular $0 \notin \mathcal{D}$).
- (iii) \mathcal{D} is *saturated*, i.e. if $p \in \mathcal{D}$ and q is monic and divides p then $q \in \mathcal{D}$.
- (iv) There exists $a \in \mathcal{R}$ such that $z - a \in \mathcal{D}$.

Since a denominator set is multiplicative, it is possible to associate with it a *ring of fractions* to be denoted by $\mathcal{R}_{\mathcal{D}}[z]$ (see Barshay (1969, Chap. 3)). Specifically $\mathcal{R}_{\mathcal{D}}[z]$ is the set of rational functions having a representation of the form p/q , where p and q are polynomials and $q \in \mathcal{D}$. It is well known and easily seen that $\mathcal{R}_{\mathcal{D}}[z]$ is a ring, even a UFD (see Samuel (1963, Thm. 4, p. 29)). In addition, we introduce the set of causal fractions in $\mathcal{R}_{\mathcal{D}}[z]$, i.e. elements of $\mathcal{R}_{\mathcal{D}}[z]$ that are causal rational functions. This set is denoted by $\mathcal{P}_{\mathcal{D}}[z]$, or, if the denominator set does not have to be specified, by \mathcal{P} .

LEMMA 2.1. \mathcal{P} is a UFD.

For a proof, see § 5. The set of all monic polynomials, which is denoted \mathcal{D}_0 , is an example of a denominator set. The corresponding set of causal fractions is denoted \mathcal{P}_0 .

A (free) linear system is identified by a quadruple (A, B, C, D) of matrices over \mathcal{R} of such dimensions that the following equations are well defined

$$(2.2) \quad x_{t+1} = Ax_t + Bu_t, \quad y_t = Cx_t + Du_t,$$

where

$$x_t \in \mathcal{X} := \mathcal{R}^n, \quad u_t \in \mathcal{U} := \mathcal{R}^m, \quad y_t \in \mathcal{Y} := \mathcal{R}^r.$$

Equations (2.2) give a discrete time interpretation of the system $\Sigma := (A, B, C, D)$. The system is called *reachable* if the columns of the matrix $[B, AB, \dots, A^{n-1}B]$ span the total state space \mathcal{R}^n . (See Sontag (1976) for details.) To the system Σ a *transfer function*

$$(2.3) \quad W(z) := W_{\Sigma}(z) := C(zI - A)^{-1}B + D$$

is associated. This is a matrix whose entries are causal rational functions. For a given transfer function $W(z)$, $\Sigma = (A, B, C, D)$ is called a *realization* if (2.3) holds.

Other interpretations of Σ can be given. Systems over rings can be used to model systems with parameters, systems with delays, 2-D systems, neutral systems (see Eising (1980), Hautus and Sontag (1981), E. W. Kamen (1978), Rouchaleau (1972), Sontag (1976), (1981)). We will give an example below. By a suitable choice of \mathcal{D} (and sometimes of \mathcal{R} , see Eising (1980, § 4.3)) one can accommodate various stability conditions one wants to impose on the transfer function. Once the ring \mathcal{R} and the

denominator set \mathcal{D} have been chosen, we call a rational function *stable* if it is in $\mathcal{R}_{\mathcal{D}}[z]$. A (single variable) stable transfer function is an element of $\hat{\mathcal{R}}_{\mathcal{D}}[z]$. An $n \times n$ matrix A is called a *stability matrix* if $\det(zI - A) \in \mathcal{D}$. Obviously, $W(z)$ is stable if A is a stability matrix. The converse is not always true, however, for any stable transfer function matrix, there exists a (free) realization Σ which is stable, i.e., for which A is a stability matrix (see Sontag (1976)).

Let us give some examples of interpretations of systems over rings and particular choices of denominator sets.

Example 2.4. In the case that $\mathcal{R} = \mathbb{R}$ (the field of real numbers) stability often is formulated in terms of pole location. Specifically, a set $\mathbb{C}^- \subseteq \mathbb{C}$ is given and a monic denominator $q(z)$ is in \mathcal{D} iff it has no zeros outside \mathbb{C}^- . It is easily seen that \mathcal{D} , thus defined, is a denominator set provided $\mathbb{C}^- \cap \mathbb{R} \neq \emptyset$. \square

Example 2.5. One can model a *delay system* with delays all multiple of a given positive real number τ by a system over the ring $\mathcal{R} = \mathbb{R}[\sigma]$ of polynomials in σ , where σ stands for the delay operator

$$\sigma x(t) = x(t - \tau).$$

The system then will be of the form

$$(2.6) \quad \dot{x} = A(\sigma)x + B(\sigma)u, \quad y = C(\sigma)x + D(\sigma)u$$

where A, B, C, D are polynomial matrices. The systemic significance of the transfer function

$$(2.7) \quad W(s, \sigma) = D(\sigma) + C(\sigma)(sI - A(\sigma))^{-1}B(\sigma)$$

is described in Sontag (1976). In particular, applying a Laplace transform to (2.6) yields

$$(2.8) \quad \hat{y}(s) = W(s, e^{-\tau s})\hat{u}(s).$$

It is well known (see Hale (1977, § 7.4)) that $\Sigma = (A, B, C, D)$ is (externally) stable iff $W(s, e^{-\tau s})$ has no pole in $\text{Re } s \geq 0$. Thus, here we define

$$(2.9) \quad \mathcal{D} := \{p \in \mathbb{R}[s, \sigma] \mid p \text{ is monic with respect to } s \text{ and } p(s, e^{-\tau s}) \neq 0 \text{ for } \text{Re } s \geq 0\}.$$

When saying p is monic with respect to s we mean that p is of the form

$$p(s, \sigma) = s^n + p_1(\sigma)s^{n-1} + \cdots + p_n(\sigma)$$

where $p_1, \dots, p_n \in \mathbb{R}[\sigma]$. It is easily seen that \mathcal{D} is a denominator set. In order that the system be internally stable one must require that $\det(sI - A(\sigma))$ be in \mathcal{D} . \square

Further examples demonstrating the generality of the stability concept described here can be given. Compare Datta and Hautus (1981), Eising (1980), Hautus and Sontag (1981), Kamen (1980).

3. Problem formulation and statement of the main results. First, we give a general formulation of the decoupling problem. We introduce the i/s -map corresponding to system Σ (i.e., system (2.2)) by (see Fig. 3.1)

$$(3.1) \quad W_s(z) := (zI - A)^{-1}B,$$

so that $W = CW_s + D$. Let F and G be dynamical systems with dimensions such that the formula

$$(3.2) \quad u = -F(z)x + G(z)y$$

is well defined. Then this formula defines a (combined) compensator, which transforms

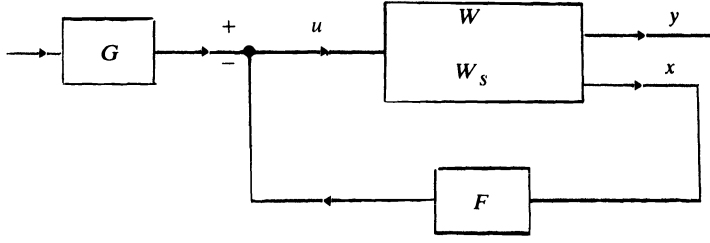


FIG. 3.1

Σ into a system $\Sigma_{F,G}$ with transfer matrix

$$(3.3) \quad W_{F,G}(z) = W(z)L_{F,G}(z),$$

where

$$(3.4) \quad L_{F,G}(z) = (I + F(z)W_s(z))^{-1}G(z).$$

We notice that the same transfer matrix is obtained if one replaces (F, G) by $(0, L_{F,G})$. A compensator in which $F = 0$ is called a *precompensator*. If G is *static* (no dynamics in the precompensator part) then we say that (F, G) is *pure (dynamic) feedback* and if, in addition, F is static then (F, G) is called a *static state feedback*. Our objective is to find a compensator of a specified class (precompensator, pure dynamic feedback, static feedback) such that the resulting transfer matrix $W_{F,G}$ is diagonal, in which case we call the resulting system *decoupled*. In order to guarantee that each output can effectively be controlled, we require in addition that the diagonal elements of $W_{F,G}$ be nonzero. This is equivalent to requiring G to be nonsingular. Sometimes one wants to impose stronger conditions, such as the invertibility (over \mathcal{R}) of G (compare Datta and Hautus (1981)). Finally, assuming that the original system is internally stable, we try to find (F, G) such that the resulting system is internally stable. Such a compensator will be called *stability preserving*. Notice that we do not attempt to stabilize and to decouple the system simultaneously. Rather, we try to decouple it while maintaining its stability. If the system is not stable at the outset, it has to be stabilized first and afterwards one has to design the decoupling compensator. It will follow from the results of this paper, that one cannot destroy the existence of such a decoupling compensator when applying the stabilizing feedback. We assume that $\mathcal{U} = \mathcal{Y}$, i.e., the number of input and output variables are equal.

It turns out that the problem of decoupling by precompensation or combined compensation (i.e., no restrictions on F, G) is very simple even if \mathcal{R} is an arbitrary integral domain.

THEOREM 3.5. *In the situation described above, the following statements are equivalent.*

- (i) *There exists a (stability preserving) decoupling combined compensator (F, G) .*
- (ii) *There exists a (stability preserving) decoupling precompensator $(0, G)$.*
- (iii) *W is nonsingular, i.e., $\det W$ is not identically zero.*

Proof. (ii) \Rightarrow (i) is trivial.

(i) \Rightarrow (iii). According to (3.3) we have

$$\det W \cdot \det L_{F,G} = \det W_{F,G} \neq 0,$$

since the diagonal elements of $W_{F,G}$ are nonzero.

(iii) \Rightarrow (ii). Let $\text{adj } W$ denote the adjoint of W (occurring in Cramér's rule). Then

$$W \text{ adj } W = (\det W)I.$$

Choose $a \in \mathcal{R}$ such that $z - a \in \mathcal{D}$. Then, for sufficiently high k , $(z - a)^{-k} \text{adj } W = G$ is causal and stable since the entries of W are stable. Hence, by

$$W \cdot G = (z - a)^{-k} (\det W) \cdot I,$$

G is a stable decoupling compensator, which has an internally stable realization. If G is internally stable, then the total realization is internally stable. \square

The condition for the existence of pure feedback decoupling compensators is more involved. To formulate it we need some notation. We write

$$(3.6) \quad W(z) = \begin{bmatrix} w_1(z) \\ \vdots \\ w_m(z) \end{bmatrix},$$

where $w_i(z)$ denotes the i th row of W . Let $d_i(z)$ denote a GCD over \mathcal{P} of the entries of $w_i(z)$. Such a GCD exists because \mathcal{P} is a UFD. An explicit construction of such a GCD is given in Lemma 3.11. We can write $w_i = d_i w_i^*$ for suitable w_i^* with entries in \mathcal{P} . Hence

$$(3.7) \quad W(z) = \Delta(z) W^*(z),$$

where $\Delta(z) = \text{diag}(d_1, \dots, d_m)$ and W^* is the matrix consisting of the rows w_1^*, \dots, w_m^* .

Now we are in the position to formulate the main result of this paper.

THEOREM 3.8. *Let Σ be a reachable, internally stable system with respect to the denominator set \mathcal{D} . Then the following statements are equivalent:*

- (i) Σ can be decoupled by a stability preserving static state feedback with G invertible over \mathcal{R} .
- (ii) Σ can be decoupled by a stability preserving, stable dynamic state feedback with G invertible over \mathcal{R} (and $F(z)$ stable).
- (iii) Σ can be decoupled by a stable precompensator L which is invertible over \mathcal{P} .
- (iv) W^* , as given in (3.7), is invertible over \mathcal{P} .

The proof of this result will be given in § 5.

In the theorem it is assumed that the gain matrix G is invertible, although this is not necessary in the original problem formulation: G nonsingular would do. It is possible to generalize the theorem to this more general case, but the formulation becomes more involved. There are two remarkable consequences of Theorem 3.8, already noted for systems over fields in Hautus and Heymann (1980), (1983). In the first place, if decoupling is possible by dynamic state feedback, it is also possible by static feedback. In the second place, the condition for the existence of a decoupling state feedback does not depend on the realization, provided the realization is reachable (for systems over fields this latter restriction is not necessary).

The GCD's used in defining W^* are not unique and consequently so is not W^* itself. However, the invertibility of W^* is independent of the particular choice of the GCD's, as easily can be seen. The condition on W^* can be checked by computing $w(z) = \det W^*$ and checking whether $(w(z))^{-1} \in \mathcal{P}$. Whether this condition can be verified effectively depends on \mathcal{D} . In the particular case that $\mathcal{D} = \mathcal{D}_0$, the set of all monic polynomials, the condition $(w(z))^{-1} \in \mathcal{P}_0$ can be checked very easily. To this

extent, expand $w(z)$ in powers of z^{-1} ,

$$w(z) = w_0 + w_1 z^{-1} + w_2 z^{-2} + \dots,$$

and $(w(z))^{-1} \in \mathcal{P}_0$ iff w_0 is invertible over \mathcal{R} . It is even not necessary to compute $w(z)$.

COROLLARY 3.9. *If $\mathcal{D} = \mathcal{D}_0$ in Theorem 3.8 and if W^* is expanded as*

$$W^*(z) = W_0^* + W_1^* z^{-1} + \dots$$

then the condition (iv) of Theorem 3.8 may be replaced by: W_0^ is invertible over \mathcal{R} .*

When restricted to the case that \mathcal{R} is a field, this is exactly the condition given in Falb and Wolovich (1967).

Although the particular choice of the GCD's used in the definition of W^* is of no relevance for condition (iv) of Theorem 3.8, it will turn out that for the actual construction of a decoupling feedback it is imperative that the GCD's satisfy an additional condition. To express this condition we use the notation $(p, q | \mathcal{R}[z])$ for the GCD's of elements in $\mathcal{R}[z]$.

DEFINITION 3.10. Let $w_1, \dots, w_m \in \mathcal{P}$. A GCD d of w_1, \dots, w_m is called *admissible* if there exist polynomials q and p such that pd and qw_i/pd , $i = 1, \dots, m$ are polynomials, and $(p, q | \mathcal{R}[z]) = 1$.

The existence and the construction of an admissible GCD is guaranteed by the following result.

LEMMA 3.11. *Let $w_1, \dots, w_m \in \mathcal{P}$ and let q be a least common denominator of w_1, \dots, w_m . Define $\tilde{p}_i := qw_i$ and let $v := (\tilde{p}_1, \dots, \tilde{p}_m | \mathcal{R}[z])$. Finally, write $p_i := \tilde{p}_i/v$. Choose $a \in \mathcal{R}$ such that $z - a \in \mathcal{D}$ and define $\mu := \min \{\deg q - \deg p_i | i = 1, \dots, m\}$. Then*

$$d := v \cdot (z - a)^{-\mu}$$

is an admissible GCD of w_1, \dots, w_m .

For a proof see § 5. The lemma provides us with an actual construction of an admissible GCD, at least if we have an algorithm for computing GCD's in $\mathcal{R}[z]$. This is for instance the case when $\mathcal{R} = \mathbb{R}[\sigma]$, see Bose (1976).

Now we can specify Theorem 3.8:

PROPOSITION 3.12. *If the elements d_i used in the construction of W^* are admissible GCD's of w_i then there exists a static feedback (F, G) such that $L_{F,G} = W^{*-1}$.*

For a proof see § 5.

Once it has been proved that a given precompensator L can be implemented by a (static) feedback (F, G) , the actual computation of F and G is straightforward. Let us start from L^{-1} , rather than L (recall that $L^{-1} = W^*$). The relation $L^{-1} = L_{F,G}^{-1}$ reads

$$G^{-1}(I + FW_s) = L^{-1} = M_0 + M_1 z^{-1} + \dots,$$

where we have expanded L^{-1} into powers of z^{-1} . Invertibility of L^{-1} over \mathcal{P} implies that M_0 is invertible. Since W_s is strictly causal we must have

$$(3.13) \quad G = M_0^{-1}.$$

Then it follows that $GL^{-1} = I =: V(z)$ is strictly causal and the map F has to be computed from

$$FW_s(z) = V(z).$$

On substitution of $W_s = (zI - A)^{-1}B$ into this equation and expanding both sides as a series in z^{-1} , one obtains

$$(3.14) \quad F[B, AB, \dots, A^{n-1}B] = [V_1, V_2, \dots, V_n],$$

where we have used the expansion $V(z) = V_1 z^{-1} + V_2 z^{-2} + \dots$. Because of reachability, the map $[B, AB, \dots, A^{n-1}B]$ has a right inverse and hence F can be solved uniquely from (3.14).

Notice that the maps F and G are uniquely determined by L . In particular, the stability-preservation property of (F, G) is automatically guaranteed by the invertibility of W^* over \mathcal{P} .

4. Examples.

Example 4.1. Consider the system $\Sigma = (A, B, C, D)$ over $\mathcal{R} = \mathbb{R}[\sigma]$, where

$$A = \begin{bmatrix} -1 & 1 \\ 1 & \sigma \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 1 & \sigma \end{bmatrix}, \quad C = \begin{bmatrix} -\sigma+1 & 1 \\ -\sigma^2+\sigma+1 & \sigma \end{bmatrix}, \quad D = 0.$$

Suppose that $\mathcal{D} = \mathcal{D}_0$. It is easily seen that Σ is reachable. The transfer matrix has the following expansion

$$\begin{aligned} W(z) &= CBz^{-1} + CABz^{-2} + \dots \\ &= \begin{bmatrix} 1 & 1 \\ \sigma & \sigma+1 \end{bmatrix} z^{-1} + \begin{bmatrix} 1 & 2\sigma \\ \sigma+1 & 2\sigma^2+\sigma-1 \end{bmatrix} z^{-2} \\ &\quad + \begin{bmatrix} 2\sigma & 2\sigma^2-\sigma+1 \\ 2\sigma^2+\sigma-1 & 2\sigma^3+2 \end{bmatrix} z^{-3} + \dots \end{aligned}$$

The system can be decoupled by state feedback since $W_0^* = \begin{bmatrix} 1 & 1 \\ \sigma & \sigma+1 \end{bmatrix}$ is invertible. According to (3.13), we have

$$G = W_0^{*-1} = \begin{bmatrix} \sigma+1 & -1 \\ -\sigma & 1 \end{bmatrix}.$$

Furthermore, F has to be computed from (3.14), where V_i is the coefficient of z^{-i-1} in the expansion of $W_0^{*-1}W(z)$. (Notice that $W^*(z) = z^{-1}W(z)$.) It follows that

$$F \begin{bmatrix} 0 & 1 & 1 & \sigma-1 \\ 1 & \sigma & \sigma & \sigma^2+1 \end{bmatrix} = \begin{bmatrix} 0 & \sigma+1 & \sigma+1 & \sigma^2-1 \\ 1 & \sigma-1 & \sigma-1 & \sigma^2-\sigma+2 \end{bmatrix}.$$

Equating the first two columns we obtain

$$F = \begin{bmatrix} \sigma+1 & 0 \\ -1 & 1 \end{bmatrix}.$$

The transfer function of the resulting system is $z^{-1}I$.

Example 4.2. The purpose of this example is to show that the reachability condition in Theorem 3.8 is essential. Let $\mathcal{R} = \mathbb{R}[\sigma]$, $\mathcal{D} = \mathcal{D}_0$

$$A = \begin{bmatrix} \sigma & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} \sigma & 0 \\ 0 & \sigma \end{bmatrix}, \quad C = I, \quad D = 0.$$

The system is not reachable, since

$$[B, AB] = \begin{bmatrix} \sigma & 0 & \sigma^2 & \sigma \\ 0 & \sigma & 0 & 0 \end{bmatrix}$$

and hence all reachable vectors have to be divisible by σ . The transfer function is

$$W = W_s = \frac{\sigma}{z(z-\sigma)} \begin{bmatrix} z & 1 \\ 0 & z-\sigma \end{bmatrix}.$$

Following the procedure described at the beginning of this section we obtain

$$W^* = \begin{bmatrix} \frac{z}{z-\sigma} & \frac{1}{z-\sigma} \\ 0 & 1 \end{bmatrix} = W_0^* + \dots$$

where

$$W_0^* = I$$

is invertible. We show that, nevertheless, decoupling by dynamic state feedback is impossible. Suppose that (F, G) is a dynamic state feedback decoupling the system. Then $W(I + FW)^{-1}G = \Delta$ for some diagonal matrix Δ (recall that $W = W_s$). Equivalently,

$$(4.3) \quad W = \Delta G^{-1}(I + FW).$$

Expanding the matrices D, W, F in powers of z^{-1} we have

$$\begin{aligned} W &= W_1 z^{-1} + W_2 z^{-2} + \dots, \\ \Delta &= D_0 + D_1 z^{-1} + D_2 z^{-2} + \dots, \\ F &= F_0 + F_1 z^{-1} + \dots, \end{aligned}$$

where

$$W_1 = B = \sigma I, \quad W_2 = AB = \begin{bmatrix} \sigma & \sigma^2 \\ 0 & 0 \end{bmatrix}.$$

Substitution into (4.3) yields

$$W_1 z^{-1} + W_2 z^{-2} + \dots = (D_0 + D_1 z^{-1} + \dots) G^{-1} (I + (F_0 + \dots)(W_1 z^{-1} + \dots)).$$

The coefficients of z_0 yields: $D_0 = 0$, and of z^{-1} : $W_1 = D_1 G^{-1}$ hence $\sigma G = D_1$. In particular, G is diagonal. Finally, equating the coefficients of z^{-2} we obtain

$$\begin{bmatrix} \sigma^2 & \sigma \\ 0 & 0 \end{bmatrix} = W_2 = D_1 G^{-1} F_0 W_1 + D_2 G^{-1} = \sigma^2 F_0 + D_2 G^{-1}.$$

The matrix $D_2 G^{-1}$ is diagonal and hence $\sigma^2 F_{0,12} = \sigma$, which has no solution in $\mathbb{R}[\sigma]$. Further examples can be found in Datta and Hautus (1981).

5. Proofs.

Proof of Lemma 2.1. It is well known that $\mathcal{R}_{\mathcal{D}}[z]$ is a UFD if \mathcal{D} is a multiplicative set (see Samuel (1969, Thm. 4, p. 29)). We define an isomorphism which maps \mathcal{P} onto $\mathcal{R}_{\mathcal{D}_1}[z]$ for some \mathcal{D}_1 . The isomorphism is

$$\varphi: r(z) \mapsto \hat{r}(z) := r\left(a + \frac{1}{z}\right)$$

defined on $\mathcal{R}(z)$, where $a \in \mathcal{R}$ is chosen such that $z - a \in \mathcal{D}$, and

$$\mathcal{D}_1 := \{z^n \hat{p}(z) \mid p \in \mathcal{D}, n = \deg p\}.$$

\mathcal{D}_1 is easily seen to be a multiplicative set in $\mathcal{R}[z]$. The map φ is invertible, and $\varphi^{-1}r(z) = r((z - a)^{-1})$. The homomorphism properties are readily verified. It remains to be shown that φ maps \mathcal{P} into $\mathcal{R}_{\mathcal{D}_1}[z]$ and φ^{-1} maps $\mathcal{R}_{\mathcal{D}_1}[z]$ into \mathcal{P} .

Let $r = p/q \in \mathcal{P}$. Then

$$\varphi r = \hat{r} = \frac{\hat{p}}{\hat{q}} = \frac{z^n \hat{p}(z)}{z^n \hat{q}(z)} \in \mathcal{R}_{\mathcal{D}_1}[z],$$

where $n = \deg q$. Notice that because of the causality of r , the numerator $z^n \hat{p}(z)$ is a polynomial. Conversely, let $r \in \mathcal{R}_{\mathcal{D}_1}[z]$, say

$$r(z) = \frac{p(z)}{z^n \hat{q}(z)}$$

for some $q \in \mathcal{D}$ with $\deg q = n$. Then

$$\varphi^{-1} r(z) = \frac{(z-a)^n}{q(z)} p\left(\frac{1}{z-a}\right) \in \mathcal{P}$$

since $(z-a)^{-1} \in \mathcal{P}$ and hence $p((z-a)^{-1}) \in \mathcal{P}$, and $q \in \mathcal{D}$ and hence $(z-a)^n/q(z) \in \mathcal{P}$. Since $\mathcal{R}_{\mathcal{D}_1}[z]$ is a UFD and isomorphic to \mathcal{P} , it follows that \mathcal{P} is a UFD. \square

Remark. \mathcal{D}_1 need not be a denominator set, in particular, the elements of \mathcal{D}_1 need not be monic.

Remark. The isomorphism φ is a standard device for transferring properties known about quotient rings to rings of causal quotients (Eising (1980, § 4.2), Hautus and Sontag (1980)).

Proof of Lemma 3.11. Because of the assumption that \mathcal{R} is a unique factorization domain, the ring $\mathcal{R}[z]$ is also a unique factorization domain (see Barshay (1969, Thm. 4.7)). Let us denote by $\tilde{\mathcal{D}}$ the set of polynomials v of the form $v = uw$, where u is a unit and $w \in \mathcal{D}$. If v is any polynomial in $\mathcal{R}[z]$, we can factorize it into primes $v = p_1 \cdots p_m$. As a consequence, we can decompose v into $v = v^+ v^-$, where v^- is the product of the prime factors of v which are in $\tilde{\mathcal{D}}$ and v^+ consists of the other factors. Except for unit factors this decomposition is unique. If we insist on a unique decomposition we can achieve this by requiring that v^- be monic. We call v^- the \mathcal{D} -part and v^+ the non- \mathcal{D} part of v . The following results follow easily from this definition.

PROPOSITION 5.1.

- (i) $(ab)^+ = a^+ b^+$, $(ab)^- = a^- b^-$.
- (ii) $p|q$ (in $\mathcal{R}[z]$) iff $p^+|q^+$ and $p^-|q^-$.
- (iii) $(\text{GCD}(v_1, \dots, v_n))^+ = \text{GCD}(v_1^+, \dots, v_n^+)$.
- (iv) If $p \in \tilde{\mathcal{D}}$, $q|p$ then $q \in \tilde{\mathcal{D}}$.

In view of the saturation condition imposed on \mathcal{D} , property (iv) follows from the fact that divisors of elements of $\tilde{\mathcal{D}}$ are monic up to a unit factor.

After this preparation we are in the position to prove that $d := v(z-a)^{-\mu}$ is an admissible GCD of w_1, \dots, w_m over \mathcal{P} . In the first place d is a divisor of w_1, \dots, w_m because (recall that $w_i = p_i v/q$)

$$\frac{w_i}{d} = \frac{p_i v}{q v} (z-a)^\mu = \frac{p_i (z-a)^\mu}{q} \in \mathcal{P},$$

since $q \in \tilde{\mathcal{D}}$ and $\mu + \deg p_i \leq \deg q$. Now let d_1 be also a divisor of w_1, \dots, w_m . We have to show that $d/d_1 \in \mathcal{P}$. Let $d_1 = \alpha/\beta$, $w_i/d_1 =: a_i = b_i/c_i$, with $\beta, c_i \in \mathcal{D}$.

We notice that

$$w_i = a_i d_1 = \frac{\alpha b_i}{\beta c_i} = \frac{\tilde{p}_i}{q},$$

and hence

$$q\alpha b_i = \tilde{p}_i \beta c_i.$$

Taking the non- \mathcal{D} part we obtain $\alpha^+ b_i^+ = \tilde{p}_i^+$. This shows that $\alpha^+ | \tilde{p}_i^+$ in $\mathcal{R}[z]$. Observing that v is a GCD of the \tilde{p}_i 's and hence v^+ is a GCD of the \tilde{p}_i^+ 's, we conclude that $\alpha^+ | v^+$, say $v^+ = \gamma \alpha^+$ with $\gamma \in \mathcal{R}[z]$. It follows that

$$\frac{d}{d_1} = \frac{\beta v}{\alpha(z-a)^\mu} = \frac{\beta \gamma v^-}{\alpha^-(z-a)^\mu} \in \mathcal{R}_{\mathcal{D}}[z]$$

since the denominator is in \mathcal{D} . It remains to be shown that d/d_1 is causal. Choose i such that $\deg q = \deg p_i + \mu$ (recall the definition of μ). Then we have that

$$\frac{d}{d_1} = \frac{d}{w_i} \cdot \frac{w_i}{d_1} = \frac{v}{(z-a)^\mu} \cdot \frac{q}{p_i v} \cdot a_i$$

is causal, since a_i is. It follows that d is a GCD of w_1, \dots, w_m . Finally, we show that d is admissible. We choose $p = (z-a)^\mu$, and q as already defined. Then $pd = v$ is a polynomial and pd and q are coprime. In addition, $qw_i/pd = p_i$ is also a polynomial. This completes the proof. \square

Proof of Theorem 3.8 (and Proposition 3.12).

(i) \Rightarrow (ii) is evident.

(ii) \Rightarrow (iii). If (F, G) is a stability preserving, stable dynamic state feedback and if G is invertible, then, according to (3.3), W is decoupled also by the precompensator $L_{F,G}$ defined by (3.4). It remains to be shown that the entries of $L_{F,G}$ are in \mathcal{P} and that $L_{F,G}$ is invertible over \mathcal{P} . It is easily seen that $L_{F,G}$ is invertible as a rational matrix and that $L_{F,G}^{-1} = G^{-1}(I + FW_s)$ has entries in \mathcal{P} . So, only the stability of $L_{F,G}$ itself has to be shown. By assumption, the resulting system is internally stable. This implies that the matrix $V = W_s L_{F,G}$ is stable. Since

$$(I + FW_s)^{-1} = I - FW_s(I + FW_s)^{-1}$$

it follows that

$$L_{F,G} = G - FV$$

is stable.

(iii) \Rightarrow (iv). Suppose that for some nonsingular diagonal matrix $E = \text{diag}(e_1, \dots, e_m)$ we have $W = EL$, where L is a matrix invertible over \mathcal{P} . The first row of this matrix equation reads

$$[w_{11}, \dots, w_{1m}] = e_1[l_{11}, \dots, l_{1m}],$$

which implies that e_1 is a divisor of $w_1 = [w_{11}, \dots, w_{1m}]$. Since d_i is a GCD of w_1 it follows that e_1 divides d_1 , i.e., $d_1 = h_1 e_1$ for some $h_1 \in \mathcal{P}$. Similar results hold for the other rows, so that we can write $\Delta = EH$ where $H = \text{diag}(h_1, \dots, h_m)$. It follows that

$$EL = W = \Delta W^* = EHW^*$$

and hence $L = HW^*$. Therefore $W^{*-1} = HL^{-1}$ is causal and stable. (The nonsingularity of the matrices involved is obvious).

(iv) \Rightarrow (i). If W^* is invertible we define $L = W^{*-1}$, and formula (3.7) reads

$$WL = \Delta.$$

Hence, the system is decoupled by the precompensator L which is a matrix over \mathcal{P} , hence causal and stable. We have to show that F and G exist such that $L = L_{F,G}$ (see (3.4)). To this extent, we formulate a generalization to systems over \mathcal{R} of a result given for systems over a field in Hautus and Heymann (1978). The result in question is:

THEOREM 5.2. *Let (A, B, C, D) denote a reachable system and L be a bicausal isomorphism (i.e., causal and with a causal inverse). Then there exists a static state feedback compensator (F, G) , with G invertible, and $L = L_{F,G}$ iff for each polynomial $u \in \mathcal{R}^m[z]$ we have: If $W_s u$ is polynomial then $L^{-1}u$ is polynomial.*

The proof of this result is completely analogous to the proof in the field case, so it will not be repeated here. Contrary to the field case, however, the reachability of the system is essential. (A counterexample in the nonreachable case can be deduced from Example 4.2.) In order to apply the theorem to our L , we have to prove that $L^{-1}u = W^*u$ is polynomial whenever u and $W_s u$ are polynomial. We show that the following stronger statement: *u and Wu are polynomial implies W^*u is polynomial* holds, provided W^* is constructed via admissible GCD's. (This will also prove Proposition 3.12.) Let us assume that Wu is polynomial. The first entry of this vector is $w_1 u$, which is a polynomial. Let d_1 be an admissible GCD of w_1 , and let $w_1^* = d_1^{-1} w_1$. We have to show that $w_1^* u$ is polynomial. According to Definition 3.10, there exist polynomials p and q such that $a := p d_1$ and v_1/a , where $v_1 := q w_1$ are polynomials and $(a, q | \mathcal{R}[z]) = 1$. Since $w_1 u = q^{-1} v_1 u$ is polynomial we have $q | v_1 u$. Also, $a | v_1 u$, since u and v_1/a are polynomial. Hence $q a | v_1 u$, q and a being coprime. But this means that $w_1^* u = d_1^{-1} w_1 u = (q a)^{-1} v_1 u p$ is polynomial. The same argument applies to each row. This shows that L can be realized by state feedback. It remains to be shown that the resulting system is internally stable. Since $W_{s,F,G}$, the i/s -map which results after feedback is applied, is equal to $W_s L_{F,G}$ and hence stable, the desired result follows from:

LEMMA 5.3. *Let $W_s = (zI - A)^{-1} B$ be a reachable i/s -map. Then W_s is stable iff $\det(zI - A) \in \mathcal{D}$.*

Proof. Since $(zI - A)^{-1} = \text{adj}(zI - A) / \det(zI - A)$, the “if” part is obvious. To prove the “only-if” part we note that because of the reachability of (A, B) , there exist polynomial matrices $P(z)$ and $Q(z)$ such that

$$(zI - A)P(z) + BQ(z) = I$$

(see Khargonekar and Sontag (1981, Lemma 3.2)). It follows that

$$(zI - A)^{-1} = P(z) + W_s(z)Q(z)$$

is stable whenever $W_s(z)$ is. But then also $\det(zI - A)^{-1} = 1 / \det(zI - A)$ is stable, and hence $\det(zI - A) \in \mathcal{D}$ (since \mathcal{D} is saturated). \square

Acknowledgment. K. B. Datta is grateful to the Eindhoven University of Technology, Department of Mathematics for offering him a Research Fellowship during the tenure of which he worked on the problem reported here.

REFERENCES

- J. BARSHAY (1969), *Topics in Ring Theory*, W. A. Benjamin, New York.
- N. K. BOSE (1976), *An algorithm for GCF extraction from two multivariable polynomials*, Proc. IEEE, pp. 185–186.
- K. B. DATTA AND M. L. J. HAUTUS (1981), *Decoupling of systems over a unique factorization domain*, in Proc. International Symposium on Mathematical Theory of Networks and Systems, N. Levan, ed., Santa Monica, CA, pp. 35–39.

- F. EISING (1980), *2-D systems, an algebraic approach*, Ph.D. Dissertation, Dept. of Mathematics, Univ. of Technology, Eindhoven.
- P. L. FALB AND W. A. WOLOVICH (1967), *Decoupling in the design and synthesis of multivariable control systems*, IEEE Trans. Aut. Contr., AC-12, pp. 651-659.
- J. HALE (1977), *Theory of Functional Differential Equations*, Springer, New York.
- M. L. J. HAUTUS AND M. HEYMANN (1978), *Linear feedback—an algebraic approach*, this Journal, 16, pp. 83-105.
- (1980), *New results in linear feedback decoupling*, in Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Control and Information Sciences 28, Springer, New York, pp. 562-577.
- (1983), *Linear feedback decoupling-transfer function analysis*, IEEE Trans. Aut. Control, to appear.
- M. L. J. HAUTUS AND E. D. SONTAG (1981), *An approach to detectability and observers*, in Algebraic and Geometric Methods in Linear System Theory, C. Byrnes and C. Martin, eds., Reidel, Dordrecht.
- E. W. KAMEN (1978), *Lectures on Algebraic System Theory: Linear systems over rings*, NASA Contractor Report 3016.
- (1980), *On the relationship between zero criteria for two-variable polynomials and asymptotic stability of delay differential equations*, IEEE Trans. Aut. Control, AC-25, pp. 983-984.
- P. P. KHARGONEKAR AND E. D. SONTAG (1981), *On the relation between stable matrix fraction factorizations and regulable realizations of linear systems over rings*, IEEE Trans. Aut. Control, AC-27, pp. 627-638.
- B. S. MORGAN, JR. (1964), *The synthesis of linear multivariable systems by state-variable feedback*, IEEE Trans. Aut. Control, AC-9, pp. 405-411.
- A. S. MORSE (1976), *System invariants under feedback and cascade control*, in Mathematical Systems Theory, G. Marchesini and S. K. Mitter, eds., Lecture Notes in Economics and Mathematical Systems, 131, Springer, Berlin, pp. 61-74.
- Y. ROUCHALEAU (1972), *Linear, discrete-time, finite-dimensional dynamical systems over some classes of commutative rings*, Ph.D. Dissertation, Stanford Univ., Stanford, CA.
- Y. ROUCHALEAU, B. F. WYMAN AND R. E. KALMAN (1972), *Algebraic structure of linear dynamical systems, III. Realization theory over a commutative ring*, Proc. Nat. Acad. Sci., 69, pp. 3404-3406.
- P. SAMUEL (1963), *Anneaux factoriels*, Redaction de Artibano Micali, Sociedade de Matemática de São Paulo.
- E. D. SONTAG (1976), *Linear systems over commutative rings: a survey*, Ricerche di Automatica, 7, pp. 1-34.
- (1981), *Linear systems over commutative rings: a (partial) updated survey*, Proc. 8th World Congress of IFAC, Kyoto, August 1981, Pergamon Press, Oxford, p. 325.

SOLVABLE APPROXIMATIONS TO CONTROL SYSTEMS*

P. E. CROUCH†

Abstract. This paper is concerned with extending certain results of Rothschild and Stein [Acta. Math., 137 (1976), pp. 247–320] and Goodman [Lecture Notes in Mathematics 562, Springer-Verlag, New York, 1976], in which a finite set of vector fields g_1, \dots, g_m are lifted to vector fields approximated by generators of a free nilpotent Lie algebra. We wish to add a vector field f , to the set g_1, \dots, g_m , and lift these vector fields to ones on a finite dimensional vector space V , approximated by generators F, G_1, \dots, G_m of a solvable Lie algebra, in which $\text{ad}^j F(G_i)$ generate a nilpotent, but not free, ideal. This procedure is accomplished in the context of a nonlinear control system, with outputs, in which f vanishes at the initial state, and in such a way that the output functions lift to the state space V , to define a system whose input output map is the same as the original system. The approximating system is obtained from a suitable realization of a truncation of the Volterra series expansion of the input output map. Such systems, with finite Volterra series, naturally exhibit the required Lie algebra structure.

Key words. approximation, graded Lie algebra, nilpotent, solvable, control systems, nonlinear

1. Introduction.

1.1. There have been several accounts in recent years of lifting a finite set of vector fields on a manifold M to a manifold $M \times \mathbb{R}^N$, in which the new vector fields are approximated in a precise sense by generators of a free, and hence graded, nilpotent Lie algebra N , whose underlying vector space is diffeomorphic, on some neighborhood of $0 \in N$, to some neighborhood of $p \times 0 \in M \times \mathbb{R}^N$ (see Rothschild and Stein [1] and Goodman [2]). These ideas reinforced similar ideas expressed earlier in Krener [12], in the context of a nonlinear control system

$$(1) \quad \dot{x} = f(x) + \sum_{i=1}^m u_i g_i(x), \quad x(0) = x_0, \quad x \in M,$$

where f, g_1, \dots, g_m are vector fields on a manifold M and all data is either smooth or real analytic. However, in Krener [12], and later Hermes [3] and Sussmann [4], approximating vector fields generating a nilpotent Lie algebra are constructed, and the lifting procedure is ignored. Another lifting procedure will be described in Krener [9].

The work in Hermes [3] and Sussmann [4] is in part devoted to applying these approximation procedures to obtain sufficient conditions for local controllability about a state x_0 for which $f(x_0) = 0$. It seems that in this and other applications such as optimal control, vector fields in the set $S = \{\text{ad}^j f(g_i) : j \geq 0, 1 \leq i \leq m\}$ ($\text{ad} f(g) = [f, g]$ being the Lie bracket) play an equal role. The conditions for local controllability in [3] and [4] are expressed in terms of S and its brackets. On the other hand f plays a role distinct from the vector fields in S , and can be viewed as the endomorphism $\text{ad} f : S \rightarrow S$. The papers [12], [4] and [3] all ignore this distinction, and the methods of Rothschild, Stein and Goodman do not directly apply to infinite sets of vector fields such as S .

The goal of this paper is to present a method by which we may extend the lifting procedure in Rothschild, Stein and Goodman to the infinite set S , and at the same time clarify the distinct roles played by f and the set S . This is accomplished by introducing a wider class of approximating systems, having a solvable but not

* Received by the editors August 16, 1982, and in revised form February 10, 1983. This work was supported in part by the Office of Naval Research under grant N00014-75-C-0648, while the author was on leave in the Division of Applied Sciences, Harvard University, Cambridge, Massachusetts.

† Department of Engineering, University of Warwick, Coventry, England CV4 7AL.

necessarily nilpotent Lie algebra, and demonstrating the lifting procedure in the context of an input-output system, (i.e. a system (1) with a finite set of outputs $y_i = h_i(x)$), as in the original paper by Krener [12]. We pay special attention to the process by which the approximating system is constructed and related to the original system. This enables the dimension of the state space for the approximating system to be reduced, and allows us to define the lifted system on the same state space. Although this does not enable the conditions for local controllability given in [3] and [4] to be improved, it should be useful in providing insight into future problems requiring the lifting and approximation techniques.

To be more specific we shall denote the lifted system by

$$(2) \quad \dot{z} = F'(z) + \sum_{i=1}^m u_i G'_i(z), \quad z \in V, \quad z(0) = 0,$$

where V is a finite dimensional vector space. The requirements for a lifted system, as proposed by the author, besides the existence of a map $\psi: V \rightarrow M$ locally carrying trajectories of (2) onto corresponding trajectories of (1), are that

$$F' = F + \tilde{F}, \quad G'_i = G_i + \tilde{G}_i, \quad 1 \leq i \leq m$$

where F and G_i are approximations to F' and G'_i such that the Lie algebra N generated by $ad^j F(G_i), j \geq 0, 1 \leq i \leq m$ is nilpotent, but not necessarily free, and transitive on V . Further, both V and N should have compatible graded structures, so that the gradation of N

$$N = \sum_{i=1}^K \oplus N_i$$

is defined by $N_1 = \text{Span}\{ad^j F(G_i), j \geq 0, 1 \leq i \leq m\}$, $N_i = \text{Span}\{i-1 \text{ iterated Lie brackets of elements in } N_1\}$, and is also described by $N_i = \{X \in N; X \text{ is a homogeneous vector field with polynomial coefficients on } V, \text{ of degree } i, \text{ with respect to the graded structure on } V\}$. (The precise definitions will be given in § 1.2.) This latter property ensures that in the case $f(x_0) = 0, F(0) = 0$, we may express the approximating system

$$\dot{z} = F(z) + \sum_{i=1}^m u_i G_i(z), \quad z \in V$$

by the system of equations

$$(3) \quad \begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \vdots \\ \dot{z}_K \end{bmatrix} = \begin{bmatrix} A_1 z_1 \\ A_2 z_2 + a_2(z_1) \\ \vdots \\ A_K z_K + a_K(z_1 \cdots z_{K-1}) \end{bmatrix} + \sum_{i=1}^m u_i \begin{bmatrix} b_{i1} \\ b_{i2}(z_1) \\ \vdots \\ b_{iK}(z_1 \cdots z_{K-1}) \end{bmatrix}$$

where the gradation of $V = \sum_{i=1}^K \oplus V_i$ is expressed in the given basis, by

$$z^t = (z_1^t \cdots z_K^t), \quad z \in V, \quad z_i \in V_i$$

and the homogeneity conditions give for $0 < t \in \mathbb{R}$

$$(4) \quad \begin{aligned} t^r a_r(z_1 \cdots z_{r-1}) &= a_r(tz_1 \cdots t^{r-1} z_{r-1}), \\ t^{r-1} b_{ir}(z_1 \cdots z_{r-1}) &= b_{ir}(tz_1 \cdots t^{r-1} z_{r-1}). \end{aligned}$$

We now make some observations on these requirements. The transitivity and nilpotent structure of N , ensure that N is a finite dimensional Lie algebra before any polynomial

structure is assumed, see [5, Lemma 3.3]; indeed, the polynomial structure of equations (3) can be deduced from these facts, as in Crouch [5]. The Lie algebra L generated by F, G_1, \dots, G_m , is clearly solvable, since N is nilpotent. Thus, whereas Rothschild, Stein and Goodman, lift a finite set of vector fields to ones approximated by generators of a free nilpotent Lie algebra, here f, g_1, \dots, g_m , are lifted to vector fields approximated by a solvable Lie algebra, with a codimension 1, graded, but not necessarily free, nilpotent ideal N .

The process by which one creates such a finite dimensional nilpotent Lie algebra N from an abstract, infinite dimensional, free nilpotent Lie algebra N^∞ , of step length $K < \infty$, requires the introduction of linear relations, which at the same time preserve the natural gradation on N^∞ . In this paper we will introduce these relations by constructing particular realizations of the truncated Volterra series, obtained from the input-output system (1), which have the desired Lie algebra structure. Clearly we are also interested in minimizing the state space dimension of these realizations, since this determines the dimension of the state space of the lifted system, and hence our ability to understand the complexity of the lifting map ψ .

The most important observation concerns the fact that we do not require the underlying vector space of N to be isomorphic to V , as is the case in Rothschild, Stein and Goodman. That is, if \mathcal{N} is the Lie group associated with N , and its action on V , then V may be regarded as a homogeneous space of \mathcal{N} , or more precisely, the tangent space to the homogeneous space, at the point at which \mathcal{N} acts. Such actions, and associated gradations on V , have been studied in the context of realizations of finite Volterra series in Crouch [5].

We consider here an example where we realize a Lie algebra in three inequivalent ways as vector fields, on particular vector spaces, demonstrating the possibilities and complexity of the situation just described for the approximating system. Assume that a Lie algebra consists of the independent vectors

$$f, \quad g, \quad [f, g], \quad [[f, g], g],$$

where all other brackets generated by f and g are zero. Consider the following three systems:

- (a) $\dot{z}_1 = u, \quad f = z_1^2 \frac{\partial}{\partial z_3}, \quad g = \frac{\partial}{\partial z_1};$
 $\dot{z}_3 = z_1^2.$
- (b) $\dot{z}_1 = u, \quad f = (z_1^2 + z_1) \frac{\partial}{\partial z_4}, \quad g = \frac{\partial}{\partial z_1};$
 $\dot{z}_4 = z_1^2 + z_1,$
- (c) $\dot{z}_1 = u, \quad f = z_1 \frac{\partial}{\partial z_2} + z_1^2 \frac{\partial}{\partial z_3}, \quad g = \frac{\partial}{\partial z_1};$
 $\dot{z}_2 = z_1,$
 $\dot{z}_3 = z_1^2.$

For all three systems the Lie algebra generated by f , and g is isomorphic to the one described above. The state space of each system has a gradation, agreeing with the structure defined in (3) and (4) only in cases (a) and (c). The Lie algebra N generated by $g, [f, g], [[f, g], g]$ is transitive on the state space in each case; in cases (a) and (b)

the state space is isomorphic to the tangent space of the homogeneous space obtained from the action of the Lie transformation group \mathcal{N} corresponding to N ; and only in case (c) is the state space isomorphic to the tangent space of \mathcal{N} , i.e. N .

Note also that the maps

$$\begin{aligned}(z_1, z_2, z_3) &\mapsto (z_1, z_4), & z_4 &= z_2 + z_3, \\ (z_1, z_2, z_3) &\mapsto (z_1, z_3)\end{aligned}$$

enable systems (b) and (a) to be lifted to system (c), respectively. It is also apparent from Krener [6] that the two systems (a) and (b) are not equivalent by diffeomorphism of the state space.

Finally, in this paper we shall restrict attention to the case where $f(x_0) = 0$, in (1), although the methods used are capable of generalization in the case $f(x_0) \neq 0$, under suitably strong hypotheses.

1.2. Preliminaries. In this section we introduce some terminology concerning graded structures; for more details see [1] or [2].

Let $V = \sum_{i=1}^K \oplus V_i$ be a finite dimensional graded vector space, and define a dilation $\delta_t: V \rightarrow V$, $t > 0$ by

$$\delta_t(v_1, \dots, v_K) = (tv_1, \dots, t^K v_K)$$

where $v^t = (v_1^t, \dots, v_K^t)$ is the graded decomposition of $v \in V$.

Let H_j be the space of homogeneous polynomials of order j on V ,

$$H_j = \{P, t^j P = P \circ \delta_t\}, \quad H_j = H_0, \quad j < 0.$$

Let C_j be the space of smooth functions on V which vanish to homogeneous order j in some neighborhood of $0 \in V$,

$$C_j = \{f, f(v) = O(|v|^j)\},$$

where $|\cdot|$ is a homogeneous norm on V , $|\delta_t(v)| = t|v|$, $C_j = C_0$ for $j < 0$.

A differential operator and, in particular, a vector field X on V , is said to be of (local) degree $\leq m$ if $X(h) \in C_{j-m}$ for each $h \in C_j$, $\forall j \geq 0$.

A vector field X , with polynomial coefficients, on V is said to be homogeneous of degree m , if $X(h) \in H_{j-m}$ for each $h \in H_j$, $\forall j \geq 0$, or in other words

$$X(h \circ \delta_t) = t^m X(h) \circ \delta_t.$$

Let L_m denote the vector space of vector fields of degree $\leq m$, and let Q_m denote the vector space of vector fields of homogeneous degree m . Then

$$(5) \quad [L_n, L_m] \subset L_{n+m}, \quad [Q_n, Q_m] \subset Q_{n+m}, \quad L_n = Q_n \oplus L_{n-1}.$$

We denote by a commutator in vector fields X_1, \dots, X_n , $\infty \geq n \geq 0$, a finite sequence of iterated Lie brackets of vector X_1, \dots, X_n , in some order.

Frequently we will want to associate a pseudo degree to a collection of differential operators, which reduces to degree for operators on a graded vector space. We do this by introducing the notion of weight. In this paper the vector fields $f(F)$ will be given the weight 0, and the vector fields $g_i(G_i)$ will be given the weight 1.

A differential operator $d(D)$ formed from a finite composition of vector fields $f(F)$, $g_i(G_i)$ will have weight $w(d)(w(D))$ equal to the sum of the weights of the constituent vector fields, and a linear sum of such operators, each of weight m , will be said to have weight m . In particular, a commutator will have weight equal to the sum of the weights of the constituent vector fields. We say a differential operator

$d(D)$ has weight $w(d) \leq m$ ($w(D) \leq m$) in X_1, \dots, X_m , if it is a linear sum of differential operators each of weight $\leq m$, formed by composing the vector fields X_1, \dots, X_m .

In this paper we shall consider the following nonlinear input-output system:

$$(6) \quad \begin{aligned} \dot{x} &= f(x) + \sum_{i=1}^m u_i g_i(x), & x(0) &= x_0, & x &\in \mathbb{R}^n = M, \\ y_i &= h_i(x), & 1 &\leq i \leq p. \end{aligned}$$

We shall also assume $f(x_0) = 0$, and that the data is smooth, or real analytic. Since we are dealing with an input-output system, and local questions only, there is no harm in assuming $M = \mathbb{R}^n$; however, some of our constructions do depend on fixing a system of coordinates for the system (unlike the lifting procedures in [1] and [2]).

The Volterra series representation of the input-output map defined by (6) is given by the series

$$(7) \quad \begin{aligned} y_i(t) &= w_0^i(t) + \int_0^t \sum_j w_1^{ij}(t, \sigma_1) u_j(\sigma_1) + \dots \\ &+ \int_0^t \int_0^{\sigma_1} \dots \int_0^{\sigma_{K-1}} \sum_{j_1, \dots, j_K} w_K^{ij_1 \dots j_K}(t, \sigma_1 \dots \sigma_K) \\ &\quad u_{j_1}(\sigma_1) \dots u_{j_K}(\sigma_K) d\sigma_1 \dots d\sigma_K + \dots \end{aligned}$$

For analytic data, the series converges uniformly and absolutely on arbitrary time intervals for the L_1 norm of u_i suitably small, and such that the solution of the uncontrolled system exists; for smooth data the series may be viewed as a Taylor series, cf. Krener and Lesiak [7] and Brockett [10].

We denote by the K th order truncated Volterra series, the Volterra series obtained by truncating the series (7), after terms involving K iterated integrals of the controls. We set $(t, x) \mapsto \gamma_f(t)(x)$, to be the flow of f (restricted to a suitable domain where it is defined) and set $g_i(\sigma)(x) = \gamma_f(-\sigma)_* g_i(\gamma_f(\sigma)(x))$. From [7] we obtain the following structure formulas for the Volterra kernels:

$$(8) \quad \begin{aligned} w_0^i(t) &= h_i \circ \gamma_f(t)(x_0), \\ w_K^{ij_1 \dots j_K}(t, \sigma_1, \dots, \sigma_K) &= g_{j_K}(\sigma_K - t)(\gamma_f(t)(x_0))(\dots g_{j_1}(\sigma_1 - t)(h_i) \dots) \\ &= g_{j_K}(\sigma_K)(x_0)(\dots g_{j_1}(\sigma_1)(h_i \circ \gamma_f(t)) \dots). \end{aligned}$$

1.3. Results. In § 2 we will use the methods of Crouch [5], [8], to construct realizations of truncated Volterra series which have the required properties as expressed by the following theorem.

THEOREM 1. *Given a smooth (analytic) system (6) with $f(x_0) = 0$ and commutators of weight $\leq K$ in f, g_1, \dots, g_m spanning $T_{x_0}M$, then there exists a system on a graded vector space $V = \sum_{i=1}^K \oplus V_i$, denoted by*

$$(9) \quad \begin{aligned} \dot{z} &= F(z) + \sum_{i=1}^m u_i G_i(z), & z &\in V, & z(0) &= 0, & F(0) &= 0, \\ \bar{y}_i &= H_i(z), \end{aligned}$$

and a smooth (analytic) submersion $\psi: \tilde{W} \rightarrow U$ where \tilde{W} and U are neighborhoods of 0, and x_0 in V and M respectively, satisfying $\psi(0) = x_0$, and

(a) *The input-output map defined by (9) is the K th order truncated Volterra series obtained from the Volterra series of system (6).*

(b) $ad^j F(G_i)j \geq 0, 1 \leq i \leq m$, generates a graded nilpotent algebra, $N = \sum_{i=1}^K \oplus N_i$, F being homogeneous of degree 0 and G_1, \dots, G_m being homogeneous of degree 1 with respect to the graded structure on V . N_j is spanned by the elements of N which are homogeneous of degree j .

(c) N acts transitively on V . If the underlying vector space of N is not isomorphic to V , so that $\dim(N) > \dim(V)$, then we may replace system (9) by a system with the same properties, and such that the underlying vector space of N is isomorphic to V .

$$(d) \quad \begin{aligned} D(h \circ \psi)|_{z=0} &= d(h) \circ \psi|_{z=0}, \\ DF(h \circ \psi)|_{z=0} &= df(h) \circ \psi|_{z=0} \end{aligned}$$

for smooth h , and a differential operator d of weight $\leq K$ in $ad^j f(g_i)$, $j \geq 0, 1 \leq i \leq m$, and D the corresponding operator obtained by replacing f, g_1, \dots, g_m by F, G_1, \dots, G_m , respectively.

(e) In particular, $\psi_*: T_0 V \rightarrow T_{x_0} M$, maps commutators (of weight $\leq K$) in F, G_1, \dots, G_m onto corresponding commutators in f, g_1, \dots, g_m .

In § 3 we follow some arguments of Goodman [2] to construct the required lifted vector fields, replacing the "partial homomorphism" of Goodman, by the map ψ of Theorem 1 and obtain the following result.

THEOREM 2. *Under the conditions and notation of Theorem 1, there exists a smooth (analytic) system defined on a possibly smaller neighborhood $W \subset \tilde{W}$, and represented by*

$$(10) \quad \begin{aligned} \dot{z} &= F'(z) + \sum_{i=1}^m u_i G'_i(z), \quad z \in W \subset V, \quad z(0) = 0, \quad F'(0) = 0, \\ y_i &= H'_i(z), \end{aligned}$$

satisfying

(a) The input-output map defined by system (10) is the same as that of system (6).

$$(b) \quad \begin{aligned} \psi_* F' &= f \circ \psi, \quad \psi_* G'_i = g_i \circ \psi, \quad 1 \leq i \leq m, \\ H'_i &= h_i \circ \psi, \quad 1 \leq i \leq p, \quad \text{on } W. \end{aligned}$$

$$(c) \quad \begin{aligned} F' &= F + \tilde{F}, \quad G'_i = G_i + \tilde{G}_i, \quad 1 \leq i \leq m, \\ H'_i &= H_i + \tilde{H}_i, \quad 1 \leq i \leq p, \quad \text{on } W, \end{aligned}$$

where \tilde{F} has local degree ≤ -1 , $\tilde{G}_1, \dots, \tilde{G}_m$ have local degree ≤ 0 , and \tilde{H}_i vanishes to homogeneous order $K+1$, with respect to the graded structure on V .

(d) If X'_i is the commutator of weight $m \leq K$ in F', G'_1, \dots, G'_m , and X_i is the corresponding commutator obtained by replacing F', G'_1, \dots, G'_m by F, G_1, \dots, G_m , respectively, then $X'_i = X_i \bmod$ vector fields of local degrees $\leq m-1$.

(e) In particular, commutators of weight $\leq K$ in F', G'_1, \dots, G'_m span $T_0 V$.

2. Theorem 1.

2.1. Graded realizations of truncated Volterra series. In this section we use the results of Crouch [8] to construct the initial realization of the K th order truncated Volterra series on a graded vector space.

We let $e_i: \mathbb{R}^n = M \rightarrow \mathbb{R}$, be the i th coordinate function, and apply these outputs successively to system (6), to obtain a Volterra series expansion of the state of

system (6). In particular, we set

$$(11) \quad e_i(z_n(t)(x)) = \int_0^t \int_0^{\sigma_1} \cdots \int_0^{\sigma_{n-1}} g_{i_n}(\sigma_n - t)(x) (\cdots g_{i_1}(\sigma_1 - t)(e_i) \cdots) \\ \cdot u_{i_n}(\sigma_n) \cdots u_{i_1}(\sigma_1) d\sigma_1 \cdots d\sigma_n$$

and obtain time varying vector fields $z_i(t)(x) \in T_x M$, $1 \leq i \leq K$. As shown in Crouch [8], we may simultaneously realize $(z_i(t)(x))$ as solutions of the partial differential equations

$$(12) \quad \begin{bmatrix} \partial z_1 / \partial t \\ \partial z_2 / \partial t \\ \vdots \\ \partial z_K / \partial t \end{bmatrix} = \begin{bmatrix} -\text{adf}(z_1) \\ -\text{adf}(z_2) + F_2(x)(z_1) \\ \vdots \\ -\text{adf}(z_K) + F_K(x)(z_1 \cdots z_{K-1}) \end{bmatrix} + \sum_{i=1}^m u_i \begin{bmatrix} g_i(x) \\ G_{2i}(x)(z_1) \\ \vdots \\ G_{Ki}(x)(z_1 \cdots z_{K-1}) \end{bmatrix},$$

$$z_i(0)(x) = 0, \quad 1 \leq i \leq K.$$

Here $F_i(x)$, $G_{ij}(x)$ are multilinear maps formed from derivatives of f , and g_i , and evaluated at x . Moreover, there exist multilinear maps $H_i(x): \mathbb{R}^{nK} \rightarrow \mathbb{R}$, formed from derivatives of h_i at x , such that the output functions

$$(13) \quad \bar{y}_i(t) = H_i(\gamma_f(t)(x_0))(z_1(t)(\gamma_f(t)(x_0)) \cdots z_k(t)(\gamma_f(t)(x_0)))$$

together with the system of equations (12) realize the k th order truncated Volterra series obtained from system (6).

We may turn the equations (12) into a set of ordinary differential equations depending on x by setting $z_i = \sum_j \alpha_{ij}(t) X_j(x)$, where X_i is a basis for $T_x M$ in some neighborhood of x_0 . Setting $x = \gamma_f(t)(x_0)$ we obtain a set of time varying ordinary differential equations, which with a suitable definition of output function realize the K th order truncated Volterra series. However, in this paper we restrict attention to the case where $f(x_0) = 0$, so $\gamma_f(t)(x_0) \equiv x_0$, and $-\text{adf}z_i(t)(x_0) = (df/dx)(x_0)z_i(t)(x_0)$. We may now consider (12) as ordinary differential equations in the state x_2

$$\dot{x}_2^t = (z_1^t, \cdots, z_K^t), \quad z_i(t) = z_i(t)(x_0).$$

We shall write the resulting autonomous equations in the form

$$(14) \quad \dot{x}_2 = f_2(x_2) + \sum_{i=1}^m u_i g_{2i}(x_2), \quad x_2(0) = 0 = x_2^0, \quad x_2 \in M_2,$$

$$\bar{y}_i = h_{2i}(x_2),$$

where $M_2 = \mathbb{R}^{nK}$ and $f_2(0) = 0$, and introduce a map $\psi_1: M_2 \rightarrow M$, defined by

$$\psi_1(x_2) = x_0 + z_1 + \cdots + z_K.$$

Clearly $\psi_1(x_2(t))$ is the K th order truncated Volterra series for the state of system (6). If we note the alternative descriptions of the Volterra kernels given in (8), it follows that for $1 \leq n \leq K$, $1 \leq j_i \leq m$, $\sigma_i, t \in \mathbb{R}$.

$$(15) \quad g_{2i_n}(\sigma_n)(g_{2i_{n-1}}(\sigma_{n-1}) \cdots g_{2i_1}(\sigma_1)(e_i \circ \psi_1 \circ \gamma_{f_2}(t)) \cdots)|_{x_2^0} \\ = g_{i_n}(\sigma_n)(g_{i_{n-1}}(\sigma_{n-1}) \cdots g_{i_1}(\sigma_1)(e_i \circ \gamma_f(t)) \cdots) \circ \psi_1|_{x_2^0},$$

where $g_{2j}(\sigma)(x_2)$ have the same definition as $g_j(\sigma)(x)$. Clearly we may replace e_i by arbitrary smooth functions on M_2 .

Using the dilation $\delta_s: M_2 \rightarrow M_2$ defined by $\delta_s(z_1, \dots, z_K) = (sz_1, \dots, s^K z_K)$, $s > 0$, on the graded space $M_2 = \sum_{i=1}^K \oplus \mathbb{R}^n$, equation (11) shows that system (14) satisfies

$$\frac{d}{dt} \delta_s x_2 = f_2(\delta_s x_2) + \sum_{i=1}^m g_{2i}(\delta_s x_2)(s u_i), \quad \delta_s x_2^0 = x_2^0 = 0.$$

It therefore follows that for smooth h

$$f_2(h) \circ \delta_s = f_2(h \circ \delta_s), \quad s g_{2i}(h) \circ \delta_s = g_{2i}(h \circ \delta_s).$$

In particular, with respect to the graded structure on M_2 , $f_2, g_{21}, \dots, g_{2m}$, are vector fields of homogeneous degree 0 and 1 respectively. It now follows that the Lie algebra L_2 generated by $f_2, g_{21}, \dots, g_{2m}$ is solvable and contains a finite dimensional nilpotent ideal N_2 which has a graded structure, defined by the degree of the vector fields. This completes the first part of the procedure for obtaining the approximating system. Although the realization has a Lie algebra with the required structure, it may not be transitive on its state space.

2.2. Proof of Theorem 1. In this section we restrict the system (14) to its orbit through the initial state, to obtain a system on a homogeneous space, and then further lift the system to one on a Lie group. We then express both systems in canonical coordinates of the second kind using techniques of Crouch [5] to complete the proof of Theorem 1.

Clearly we may obtain a system from that in (14), in which the Lie algebra L of the system is transitive on the state space, by restricting the system (14) to the maximal connected integral submanifold M_3 , through x_2^0 defined by L_2 , since L_2 is comprised of analytic vector fields. Let $\psi_2: M_3 \rightarrow M_2$ be the inclusion map. The induced system on M_3 will be written as

$$\begin{aligned} \dot{x}_3 &= f_3(x_3) + \sum_{i=1}^m u_i g_{3i}(x_3), & x_3(0) &= x_3^0, & f(x_3^0) &= 0, \\ (16) \quad \bar{y}_i &= h_{2i} \circ \psi_2(x_3), & x_3 &\in M_3. \end{aligned}$$

Clearly the Lie algebra L generated by f_3 and g_{31}, \dots, g_{3m} is a homomorphic image of L_2 and so has a codimension one ideal N generated by $ad^i f_{3g_{3i}}, j \geq 0, 1 \leq i \leq m$. However, it is not clear that N has the graded structure of N_2 .

LEMMA 1. *N is graded with gradation defined by the weight of commutators.*

Proof. Let $N^j, 1 \leq j \leq K$ be the subspace of N spanned by commutators of weight equal to j . To say that N is not graded by the weight of commutators is equivalent to postulating the existence of an element $X \in N^l$ such that X is a linear combination of elements in subspaces $N^j, j \neq l$. We note that $N^j(x_3^0) \cap N^i(x_3^0) = \{0\}$ for $j \neq i$, where $N^j(x_3^0)$ is the subspace of $T_{x_3^0} M_3$ spanned by elements of N^j ; since viewed as subspaces of $T_{x_3^0} M_2$, $N^j(x_3^0)$ coincides with the span at x_3^0 , of commutators of weight j , and hence degree j , in f, g_{21}, \dots, g_{2m} .

We define a filtration on N . Let N_0 be the subalgebra of N which vanishes at x_3^0 , and define N_{-1} by

$$N_{-1} = \{X; X \in N_0, [X, N] \in N_0\}$$

and inductively define N_{-i} by

$$N_{-i} = \{X; X \in N_{-i+1}, [\underbrace{\dots [X, N] \dots N}_{i \text{ brackets}}] \in N_0\}$$

The Jacobi identity ensures that each N_{-i} is a subalgebra of N , and the finite dimensionality of N ensures that the sequence is finite

$$N \supset N_0 \supset N_{-1} \supset \cdots \supset N_{-r} \supset N_{-r-1}.$$

The sequence may end in either of two ways. Either $N_{-r-1} = \{0\}$, or $[N_{-r-1}, N] \subset N_{-r-1}$ so that $N_{-r-2} = N_{-r-1}$. In the latter case N_{-r-1} is an ideal in N contained in N_0 , and so N_{-r-1} consists of zero vector fields on M_3 . Thus we may assume that $N_{-r-1} = \{0\}$. If $N_0 = N_{-1}$ then N_0 is an ideal in N and M_3 is a connected Lie group with Lie algebra N . Note also that $[N, N_{-i}] \subset N_{-i+1}$ $i = 1, \dots, r$.

Denote the subspace of N spanned by N^l_j $j \neq l$ by \hat{N}^l . We now have a sequence of subspaces for each l .

$$\hat{N}^l \cap N_{-r} \subset \hat{N}^l \cap N_{-r+1} \subset \cdots \subset \hat{N}^l \cap N_0 \subset \hat{N}^l.$$

Let X_i^r , $i = 1, \dots, n_r$ be a basis of $\hat{N}^l \cap V_{-n}$ and complete this with elements X_i^{r-1} , $i = 1, \dots, n_{r-1}$ to a basis for $\hat{N}^l \cap N_{-r+1}$, and proceed by induction to define a basis for $\hat{N}^l \cap N_0$ by completing a basis for $\hat{N}^l \cap N_{-j}$ to a basis for $\hat{N}^l \cap N_{-j+1}$ with elements X_i^{j-1} , $i = 1, \dots, n_{j-1}$. Finally, complete this basis to a basis for \hat{N}^l by elements X_i , $i = 1, \dots, n$.

By assumption we may write

$$X = \sum_i \alpha_i X_i + \sum_i \alpha_i^0 X_i^0 + \cdots + \sum_i \alpha_i^r X_i^r$$

for suitable α_i^l , $\alpha_i \in \mathbb{R}$. Evaluating this expression at x_3^0 gives $X(x_3^0) = \sum_i \alpha_i X_i(x_3^0)$, since all other terms vanish at x_3^0 . If $X(x_3^0) \neq 0$, we obtain a contradiction

$$N^l(x_3^0) \ni X(x_3^0) = \sum_i \alpha_i X_i(x_3^0) \in \hat{N}^l(x_3^0).$$

Thus $X(x_3^0) = 0$, and since X_i are linearly independent in $\hat{N}^l \setminus \hat{N}^l \cap N_0$ we deduce $\alpha_i = 0$, $i = 1, \dots, n$. We now take the Lie bracket of the remaining expression for X with an element $Y \in N^m$, to obtain

$$[X, Y] = \sum_i \alpha_i^0 [X_i^0, Y] + \cdots + \sum_i \alpha_i^r [X_i^r, Y].$$

Evaluating this expression at x_3^0 we obtain

$$[X, Y](x_3^0) = \sum_i \alpha_i^0 [X_i^0, Y](x_3^0).$$

If $[X, Y](x_3^0) \neq 0$ then we obtain a contradiction

$$N^{l+m}(x_3^0) \ni [X, Y](x_3^0) = \sum_i \alpha_i^0 [X_i^0, Y](x_3^0) \in \hat{N}^{l+m}(x_3^0).$$

Thus $\sum_i \alpha_i^0 X_i^0 \in N_{-1} \cap \hat{N}^l$ since $Y \in N^m$ $m = 1, \dots, k$, $\text{span } N$. However, X_i^0 are linearly independent in $\hat{N}^l \cap N_0 \setminus \hat{N}^l \cap N_{-1}$ so $\alpha_i^0 = 0$, $i = 1, \dots, n_0$. It follows that

$$X = \sum_i \alpha_i^1 X_i^1 + \cdots + \sum_i \alpha_i^r X_i^r.$$

Proceeding by a simple induction argument in which X is assumed to be a sum of terms of weight $\geq K$ and using the same arguments as above we deduce $X = 0$, and that N is graded by the weight of commutators. \square

The Lie algebra L of system (16) now has the required graded Lie algebra structure. We now show that we may view the state space as a graded vector space, with f_3 and g_{31}, \dots, g_{3m} being represented as vector fields of degree 0 and 1 respec-

tively. First observe that if \hat{e}_i are the coordinate functions on the vector space M_2 , then system (14), with outputs $y_i = \hat{e}_i(x_2)$, defines an observable system with a finite Volterra series. Therefore systems (16), with outputs $\hat{y}_i = \hat{e}_i \circ \psi_2(x_3)$, defines an observable and strongly accessible system on M_3 , with finite Volterra series. By Crouch [5, Lemma 4.1, Thm. 3.7] we see that the state space M_3 is diffeomorphic to a vector space V , and in fact V may be viewed as the tangent space $T_{x_3^0}M_3$.

We now outline the method originally contained in [11] but given in [5, Thms. 4.3, 4.10], in which the system is expressed in canonical coordinates of the second kind, taking care to ensure that only the same number of coordinates as the dimension of M_3 are used. These coordinates define the vector space V . This procedure is also given, in a slightly different setting, in Hermes [3].

We construct by induction a linearly independent set of vectors which completes any basis of N_0 (definitions as in Lemma 1), to a basis of N expressed as

$$(17) \quad X_1, \dots, X_{n_1}, X_{n_1+1}, \dots, X_{s_1}, X_{s_1+1}, \dots, X_{n_K}, X_{n_K+1}, \dots, X_{s_K}.$$

Let $X_{s_{K-1}+1}, \dots, X_{n_K}$ be linearly independent elements in N^K which complete any basis, $X_{n_K+1}, \dots, X_{s_K}$, of $N^K \cap N_0$ to a basis of N^K , and let $X_{s_{j-1}+1}, \dots, X_{n_j}$ be linearly independent elements of N^j , which complete any basis $X_{n_j+1}, \dots, X_{s_j}$ of $N^j \cap N_0$ to a basis of N^j for $1 \leq j \leq K$, $s_0 = 0$. By Lemma 1, the set of elements constructed in this way gives a basis for N ; since $N_0 = \sum_{j=1}^K \oplus (N^j \cap N_0)$. Note further that all elements $X_{s_{j-1}+1}, \dots, X_{s_j}$ have weight j , $1 \leq j \leq K$.

Let $(t, x) \mapsto \gamma_i(t)(x)$ be the flow of the (complete) vector fields X_i in (17), and define a map $\psi_3: V \rightarrow M_3$ by

$$\psi_3(z_1, \dots, z_{n_1}, z_{s_1+1}, \dots, z_{n_K}) = \gamma_1(z_1) \circ \dots \circ \gamma_{n_1}(z_{n_1}) \circ \dots \circ \gamma_{n_K}(z_{n_K})(x_3^0)$$

where $V = \sum_{i=1}^K \oplus V_i$ is the grading of the vector space V , with $z_{s_{i-1}+1}, \dots, z_{n_i}$ defining a set of coordinates for V_i . [5, Lemma 4.1] shows that ψ_3 defines a diffeomorphism of V onto M_3 . Further, using the basis (17) in [5, Thms. 4.3, 4.10] shows that system (16), expressed in the coordinates $z^t = (z_1, \dots, z_{n_1}, z_{s_1+1}, \dots, z_{n_K})$, takes the form of those in (3); and so the resulting vector fields F, G_1, \dots, G_m and systems denoted by (9) satisfy the conditions in paragraph (b) of Theorem 1. Moreover, if we define $H_i = h_{2i} \circ \psi_2 \circ \psi_3$, then system (9) with the outputs $\bar{y}_i = H_i(z)$ satisfies the requirements of paragraph (a) of Theorem 1.

By construction we have

$$(18) \quad (\psi_2 \circ \psi_3)_* F = f_2 \circ (\psi_2 \circ \psi_3), \quad (\psi_2 \circ \psi_3)_* G_i = g_{2i} \circ (\psi_2 \circ \psi_3).$$

Defining $\psi: V \rightarrow M$ by $\psi = \psi_1 \circ \psi_2 \circ \psi_3$ we now see that the relation (15) yields for $0 \leq n \leq K$, $1 \leq j_i \leq m$, and h any smooth function,

$$(19) \quad \begin{aligned} & G_{j_n}(\sigma_n) G_{j_{n-1}}(\sigma_{n-1}) \dots G_{j_1}(\sigma_1) (h \circ \psi \circ \gamma_F(t))|_{z=0} \\ &= g_{j_n}(\sigma_n) g_{j_{n-1}}(\sigma_{n-1}) \dots g_{j_1}(\sigma_1) (h \circ \gamma_f(t)) \circ \psi|_{z=0}. \end{aligned}$$

By differentiating these expressions with respect to the parameters σ_i and t , we now obtain the results of paragraphs (d) and (e) of Theorem 1.

It remains to prove paragraph (c). The connected Lie transformation group \mathcal{N} , corresponding to the Lie algebra N of system (16), acts transitively on M_3 , giving M_3 the structure of a homogeneous space. By [5, Thm 3.9], \mathcal{N} is simply connected and homeomorphic to a Cartesian space. We may now consider the system

$$(20) \quad \dot{x}_4 = f_4(x_4) + \sum_{i=1}^m u_i g_{4i}(x_4), \quad x_4(0) = e, \quad x_4 \in \mathcal{L},$$

where \mathcal{L} is the connected Lie transformation group corresponding to the Lie algebra L of system (16), and $f_4, g_{41}, \dots, g_{4m}$, are the generators of L , considered as left invariant vector fields on \mathcal{L} , e is the identity element in \mathcal{L} . If we let $\Phi: \mathcal{L} \times M_3 \rightarrow M_3$ be the smooth (analytic) map defining the action of \mathcal{L} on M_3 , $(x_4, x_3) \mapsto \Phi(x_4)x_3$, it follows that $x_4 \mapsto \Phi(x_4)(x_3^0) = \Phi'(x_4)$ satisfies $\Phi'(e) = x_3^0$,

$$\Phi'_* f_4 = f_3 \circ \Phi', \quad \Phi'_* g_{4i} = g_{3i} \circ \Phi', \quad 1 \leq i \leq m.$$

We now repeat the procedure described above, in order to express the system (20) in canonical coordinates of the second kind. In this case, however, the vector fields in (17) can be any graded basis of N , since $N_0 = \{0\}$. Although the vector field f_4 does not vanish at $e \in \mathcal{L}$, $\Phi'_* f_4(e) = 0$, so we reduce the dimension of the state space by one in the process, and hence obtain a system on a graded vector space V , isomorphic to the Lie algebra N . This completes the proof of Theorem 1.

We note that a process for obtaining realizations of finite Volterra series, with the structure given in paragraphs (a) and (b) of Theorem 1 and the state space V isomorphic to N , is given in [5, § 4.4]. Basically one constructs an arbitrary accessible realization of the finite Volterra series with Lie algebra L so that the Lie algebra N is nilpotent and finite dimensional but the subspaces N^i do not necessarily satisfy $N^i \cap N^j = \{0\}$, $i \neq j$. One can construct an abstract Lie algebra L' with the required graded property by discarding linear relations between the spaces N^i . This does not destroy the Lie algebra structure since the Jacobi identity relates commutators of equal weights. There is clearly a homomorphism of Lie algebras which maps L' onto L , enabling one to construct the desired realization from the system lifted to the Lie group \mathcal{L} , corresponding to L , as above.

The process described in this paper has two advantages; one does not have to go to a group structure in order to find a realization with the required properties, and one obtains without effort the submersion ψ , mapping the realization onto the system, from which the finite (truncated) Volterra series was obtained. One further notes that it is far from clear how to obtain realizations on homogeneous spaces from realizations on Lie groups, if one requires a graded structure, as illustrated by the example in the introduction.

3. Theorem 2.

3.1. In this section we prove a result which shows how closely vector fields defined by the approximate system (9), are related to the corresponding vector fields with the same weight for system (6), by calculating the order of vanishing of a suitable difference vector.

We shall use the graded basis X_i of N constructed in § 2.2, viewed as vector fields on V . We partition the set of indices I appearing in the basis by $I = I_1 \amalg \dots \amalg I_k$, such that $X_i \in N^j$ if $i \in I_j$, and denote by v_i the vector field on M , obtained from X_i by replacing F, G_1, \dots, G_m , in system (9) by f, g_1, \dots, g_m in system (6) respectively. By the assumption of Theorem 1 we may select a subset $J \subset I$ such that the vector fields $\{v_i, i \in J\}$ form a basis for $T_x M$, for all $x \in U$, a neighborhood of x_0 in M .

It now follows that on \tilde{W} , a neighborhood of $0 \in V$, $\psi: \tilde{W} \rightarrow U$,

$$\psi_* X_i = v_i \circ \psi + \sum_{j \in J} \theta_{ij} v_j \circ \psi, \quad i \in I, \quad (21)$$

$$\psi_* F = f \circ \psi + \sum_{j \in J} \eta_j v_j \circ \psi,$$

$$\psi_* G_i = g_i \circ \psi + \sum_{j \in J} \omega_{ij} v_j \circ \psi, \quad 1 \leq i \leq m \quad (22)$$

for suitable smooth (analytic) functions $\theta_{ij}, \eta_j, \omega_{ij}$ on \tilde{W} . By paragraph (e) of Theorem 1, and the independence of the set $v_j, j \in J$, we see that all functions $\theta_{ij}, \eta_j, \omega_{ij}$ vanish at $0 \in V$.

In this section $D_{(\alpha)}$ will always be a differential operator obtained by taking linear combinations and composing vector fields $ad^K F(G_i), K \geq 0, 1 \leq i \leq m$, and d will be the corresponding operator obtained by replacing F, G_1, \dots, G_m , by f, g_1, \dots, g_m , respectively.

LEMMA 2.

(a) $D\theta_{ij}|_{z=0} = 0$ for $i \in I_s$ and $w(D) \leq K - s$;

(b) $D\omega_{ij}|_{z=0} = 0$ for $w(D) \leq K - 1$;

(c) $D\eta_j|_{z=0} = 0$ for $w(D) \leq K$.

Proof. Clearly (b) is just a special case of (a). Let h be a smooth function on then from (21) and (22) we have

$$(23) \quad X_i(h \circ \psi) = v_i(h) \circ \psi + \sum_{j \in J} \theta_{ij} v_j(h) \circ \psi,$$

$$(24) \quad F(h \circ \psi) = f(h) \circ \psi + \sum_{j \in J} \eta_j v_j(h) \circ \psi.$$

We first prove by induction on the weight that if D is an arbitrary differential operator, under the restriction above,

$$(25) \quad DX_i(h \circ \psi) = dv_i(h) \circ \psi + \sum_{j \in J} (D\theta_{ij}) v_j(h) \circ \psi + \sum_{\alpha\beta\gamma} (D_\alpha \theta_{\beta\gamma}) \rho_\gamma$$

where $w(D_\alpha) + r < w(D) + s, i \in I_s, \beta \in I_r, w(D_\alpha) < w(D)$ for some differential operators D_α , and smooth functions ρ_γ on \tilde{W} .

Clearly (25) reduces to (23) when D is the identity operator. Assume (25) is true for operators D of weight $l < K$, and let X_p be a basis vector field of weight 1, i.e. $p \in I_1$. Applying X_p to (25) we obtain

$$\begin{aligned} X_p DX_i(h \circ \psi) &= X_p(dv_i(h) \circ \psi) + \sum_{j \in J} (X_p D\theta_{ij}) v_j(h) \circ \psi \\ &\quad + \sum (X_p D_\alpha \theta_{\beta\gamma}) \rho_\gamma + \sum (D\theta_{ij}) X_p(v_j(h) \circ \psi) + \sum (D_\alpha \theta_{\beta\gamma}) X_p \rho_\gamma. \end{aligned}$$

By (23) $X_p(dv_i(h) \circ \psi) = v_p dv_i(h) \circ \psi + \sum_{j \in J} \theta_{pj} v_j dv_i(h) \circ \psi$ so $X_p DX_i(h \circ \psi) = v_p dv_i(h) \circ \psi + \sum_{j \in J} (X_p D\theta_{ij}) v_j(h) \circ \psi + \text{remainder terms}$.

One checks easily that all the remainder terms may be written in the form $\sum_{\alpha'\beta'\gamma'} (D_{\alpha'} \theta_{\beta'\gamma'}) \rho_{\gamma'}$ where $w(D_{\alpha'}) + r' < w(X_p D) + s, \beta' \in I_r, w(D_{\alpha'}) < w(X_p D)$. Thus $X_p D$ satisfies (25) with $w(X_p D) \leq l + 1$. Since the operators of the form $X_p D$ generate all operators D' with $w(D') \leq l + 1$, the induction is complete.

We now prove by induction that

$$(26) \quad DF(h \circ \psi) = df(h) \circ \psi + \sum_{j \in J} (D\eta_j) v_j(h) \circ \psi + \sum_{\alpha\beta\gamma} (D_\alpha \theta_{\beta\gamma}) \rho_\gamma + \sum_{\delta i \epsilon} (D_\delta \eta_i) \mu_\epsilon,$$

where $w(D_\alpha) + r \leq w(D), \beta \in I_r$, and $w(D_\delta) < w(D)$. Clearly (26) reduces to (24) when D is the identity operator. Assume (26) is true for operators D with weight $l < K$, and let X_p be a basis vector field of weight one, as before. Applying X_p to (26) and making use of (23) we obtain

$$\begin{aligned} X_p DF(h \circ \psi) &= v_p df(h) \circ \psi + \sum_{j \in J} (X_p D\eta_j) v_j(h) \circ \psi \\ &\quad + \sum \theta_{pj} (v_j df(h)) \circ \psi + \sum (X_p D_\alpha \theta_{\beta\gamma}) \rho_\gamma + \sum (X_p D_\delta \eta_i) \mu_\epsilon \\ &\quad + \sum (D\eta_j) X_p(v_j(h) \circ \psi) + \sum (D_\alpha \theta_{\beta\gamma}) X_p \rho_\gamma + \sum (D_\delta \eta_i) X_p \mu_\epsilon. \end{aligned}$$

The first line of this identity has the same form as the first line of (26), and it is easily seen that the remaining terms have the form

$$\sum_{\alpha'\beta'\gamma'} (D_{\alpha'}\theta_{\beta'\gamma'})\rho_{\gamma'} + \sum_{\delta'i'\epsilon'} (D_{\delta'}\eta_{i'})\mu_{\epsilon'}$$

$$w(D_{\alpha'}) + r' \leq w(X_p D), \quad w(D_{\delta'}) < w(X_p D), \quad \beta' \in I_r.$$

Thus $X_p D$ satisfies (26) with $w(X_p D) \leq l + 1$, and hence as before for all operators D' with $w(D') \leq l + 1$. The induction is therefore complete.

We now prove (a) by induction. Assume $D\theta_{ij}|_{z=0} = 0$ for operators D , and $i \in I_s$ such that $w(D) + s \leq l < K$. Apply (25) to operators D and $i \in I_s$ such that $w(D) + s \leq l + 1$. By Theorem 1, paragraph (d), $DX_i(h \circ \psi)|_{z=0} = dv_i(h) \circ \psi|_{z=0}$, and by hypothesis $\sum (D_{\alpha}\theta_{\beta\gamma})\rho_{\gamma}$ vanishes at $z = 0$, because $w(D_{\alpha}) + r < w(D) + s$, $\beta \in I_r$. Hence, by the independence of the v_j , $j \in J$, $D\theta_{ij}$ all vanish at $z = 0$ for all operators D and $i \in I_s$ such that $w(D) + s \leq l + 1$. Clearly only $w(D) = 0, s = 1$ satisfies $w(D) + s \leq l + 1$, since $s \geq 1$. Since θ_{ij} all vanish at $z = 0$, the assertion is true for $l = 1$. The induction is complete.

We now prove (c) by induction. Assume $D\eta_j$ vanishes as $z = 0$ for operators D such that $w(D) \leq l < K$, and apply (26) to operators D such that $w(D) \leq l + 1$. By Theorem 1, paragraph (d), $DF(h \circ \psi)|_{z=0} = df(h) \circ \psi|_{z=0}$. By hypothesis, and part (a) of this lemma, $\sum (D_{\alpha}\theta_{\beta\gamma})\rho_{\gamma}$ and $\sum (D_{\delta}\eta_i)\mu_{\epsilon}$ vanish at $z = 0$, since $w(D_{\alpha}) + r \leq w(D) \leq K$, $\beta \in I_r$, and $w(D_{\delta}) < w(D)$. Thus, by the independence of the v_j , $j \in J$, $D\eta_j$ vanish at $z = 0$ for all operators D satisfying $w(D) \leq l + 1$. Clearly the induction step is true for $l = 0$, since η_j all vanish at $z = 0$. The induction argument is therefore complete. \square

We may gain insight into this result by using expressions (21) and (22) in (9). In particular, if we consider the pair of equations

$$\dot{z} = F(z) + \sum_{i=1}^m u_i G_i(z), \quad z(0) = 0, \quad z \in V,$$

$$\dot{x} = f(x) + \sum_{i=1}^m u_i g_i(x) + \sum_j \eta_j(z) v_j(x) + \sum_{ij} u_i \omega_{ij}(z) v_j(x), \quad x(0) = x_0, \quad x \in M,$$

we see that the map ψ satisfies $\psi(z(t)) = x(t)$ for all controls $u_i(\cdot)$. However, since we know that $\psi(z(t))$ is the K th order truncated Volterra series for the state of system (6), and that these terms are obtained entirely from the contribution of f, g_1, \dots, g_m in the second equation above we deduce that the net effect of the extra terms is to contribute only terms in the Volterra series expansion of the state $x(t)$ containing more than K iterated integrals of the controls. By expressing $\eta_j(z(t))$ and $\omega_{ij}(z(t))$ as Volterra series themselves, we see that the results (b) and (c) of Lemma 2 are compatible with this requirement.

We note that the vanishing properties established in Lemma 2 for the functions θ_{ij} and the expression (21) are identical to those obtained by Goodman [2]; however, the method of proof used is entirely different, and our expansion (21) is not just a formal power series identity, but is an identity between smooth, or analytic vector fields.

3.2. Proof of Theorem 2. In this section we prove Theorem 2, using a technique of Goodman [2], which we include here due to the different context. The proof is based on (21). Define a square matrix valued function on \tilde{W} , $z \mapsto S(z)$ such that $S_{ij} = \theta_{ij}$ for $j \in J$, and $S_{ij} = 0$ for $j \in I \setminus J$ for all $i \in I$. The matrix $S(z)$ is smooth (analytic) and vanishes at $z = 0$. Consequently,

$$z \mapsto \sum_{n \geq 1} (-S(z))^n = T(z)$$

is a smooth (analytic) matrix valued function on some smaller neighborhood $W \subset \tilde{W}$, of $0 \in V$. Clearly, $(I + T(z))(I + S(z)) = I$ on W , where I is the identity operator. Introduce the vector notation $\tilde{X} = (X_1, \dots, X_{s_K})$, $\tilde{v} = (v_1, \dots, v_{s_K})$, rewrite (21) in the form

$$\psi_* \tilde{X} = (I + S)\tilde{v} \circ \psi,$$

and operate by $I + T$ to obtain

$$\tilde{v} \circ \psi = \psi_* \tilde{X} + T\psi_* \tilde{X},$$

which may be written in coordinates as

$$v_i \circ \psi = \psi_* X_i + \sum_{j \in I} T_{ij} \psi_* X_j.$$

We define the smooth (analytic) vector fields Z_i on W by $Z_i = \sum_{j \in I} T_{ij} X_j$, to obtain the identities

$$(27) \quad \psi_* X_i + \psi_* Z_i = v_i \circ \psi.$$

We now set

$$(28) \quad \tilde{F} = - \sum_{j \in J} \eta_j (X_j + Z_j), \quad \tilde{G}_i = - \sum_{j \in J} \omega_{ij} (X_j + Z_j), \quad 1 \leq i \leq m,$$

and obtain from (21) and (22)

$$\psi_* (F + \tilde{F}) = f \circ \psi, \quad \psi_* (\tilde{G}_i + G_i) = g_i \circ \psi, \quad 1 \leq i \leq m.$$

Further, if we set $\tilde{H}_i = h_i \circ \psi - H_i$ and make the definitions

$$F' = F + \tilde{F}, \quad G'_i = \tilde{G}_i + G_i, \quad 1 \leq i \leq m, \quad H'_i = \tilde{H}_i + H_i, \quad 1 \leq i \leq p$$

we have proved paragraphs (a) and (b) of Theorem 2.

We now demonstrate the vanishing properties given in paragraph (c) of Theorem 2. Using the map $\psi_3: V \rightarrow M_3$ constructed in § 2.2, or an analogous map in the case of systems lifted to the Lie group, any function θ on M_3 is written as a function $\tilde{\theta}$ on V by the composition $\theta \circ \psi_3(z_1 \dots z_{n_1} z_{s_1+1} \dots z_{n_K}) = \tilde{\theta}(z_1 \dots z_{n_1} \dots z_{n_K})$. Consequently for $r_1 + r_2 + \dots + r_{n_K} = l$.

$$\frac{\partial^l \tilde{\theta}}{\partial z_1^{r_1} \dots \partial z_{n_K}^{r_{n_K}}} \Big|_{z=0} = X_1^{r_{n_K}} \dots X_1^{r_1} \tilde{\theta} \Big|_{z=0},$$

where X_i are the elements of the graded basis viewed as vector fields on V , and X_i^r is X_i composed with itself, as a differential operator, r times. We may now use Lemma 2; and this expression to see that ω_i vanish to homogeneous order K , and η_j vanish to homogeneous order $K+1$. The relations (28) now show that \tilde{F} is a vector field of degree ≤ -1 , and $\tilde{G}_1, \dots, \tilde{G}_m$ are vector fields of degree ≤ 0 .

Now by construction system (9), with outputs $\tilde{y}_i = H_i(z)$, has the K th order truncated Volterra series of system (6), and by (19) the same is true for system (9) with outputs $\tilde{y}_i = h_i \circ \psi(z)$. It follows that $D(h_i \circ \psi)|_{z=0} = D(H_i)|_{z=0}$ for all operators D of degree $\leq K$ in $ad^l F(G_i)$, $j \geq 0$, $1 \leq i \leq m$. Thus $\tilde{H}_i = h_i \circ \psi - H_i$ vanishes to homogeneous order $K+1$. This completes the proof of paragraph (c) of Theorem 2.

It remains to prove paragraphs (d) and (e). Note that $F' = F + \tilde{F}$ where F has degree ≤ 0 and \tilde{F} has degree ≤ -1 , $G'_i = G_i + \tilde{G}_i$ where \tilde{G}_i has degree ≤ 0 and G_i has degree ≤ 1 , and hence by (5) $[F', G'_i] = [F, G_i] + Z$ where $[F, G_i]$ has degree ≤ 1 , and Z has degree ≤ 0 . In general, one now easily shows by induction that $X'_i = X_i + Z_i$,

where X_i is any basis element constructed in § 2.2 of degree r , X'_i is the corresponding vector field of degree r obtained by replacing F, G_1, \dots, G_m by F', G'_1, \dots, G'_m and Z_i has degree $\leq r-1$. This demonstrates paragraph (d). Noting that vector fields of degree ≤ 0 vanish at $z=0$, and by (5) $\text{Span}\{X_i|_{z=0}; i \in I_1 \amalg \dots \amalg I_r\} = \text{Span}\{X|_{z=0}; X = \text{vector field degree } \leq r\}$, we see that

$$X'_i|_{z=0} = X_i|_{z=0} + Z_i|_{z=0}, \quad i \in I_r, \quad 1 \leq r \leq K$$

forms a triangular set of equations, showing that commutators of weight $\leq K$ in F', G'_1, \dots, G'_m span T_0V , proving paragraph (e). \square

It is instructive to write system (10) explicitly in coordinates, completing the coordinate version of the system (9) given by the equations (3), by using the relations (28), and the vanishing properties of η_j, ω_{ij} and \tilde{H}_i . Here z_i will represent the vector comprising the coordinates $z_{s_{i-1}+1}, \dots, z_{n_i}$

$$\begin{aligned} \begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \vdots \\ \dot{z}_K \end{bmatrix} &= \begin{bmatrix} A_1 z_1 \\ A_2 z_2 + a_2(z_1) \\ \vdots \\ A_K z_K + a_K(z_1 \dots z_{K-1}) \end{bmatrix} + \sum_{i=1}^m u_i \begin{bmatrix} b_{i1} \\ b_{i2}(z_1) \\ \vdots \\ b_{iK}(z_1 \dots z_{K-1}) \end{bmatrix} + \begin{bmatrix} O(|z|^{K+1}) \\ O(|z|^{K+1}) \\ \vdots \\ O(|z|^{K+1}) \end{bmatrix} \\ &\quad + \sum_{i=1}^m u_i \begin{bmatrix} O_i(|z|^K) \\ O_i(|z|^K) \\ \vdots \\ O_i(|z|^K) \end{bmatrix}, \end{aligned}$$

$$y_i = H_{i0} + H_{i1}(z_1) + \dots + H_{iK}(z_1, \dots, z_K) + O(|z|^{K+1}),$$

where $|z|$ is a homogeneous norm for the dilation $\delta_t(z_1, \dots, z_K) = (tz_1, \dots, t^K z_K)$, and

$$a_j \circ \delta_t = t^j a_j, \quad b_{ij} \circ \delta_t = t^{j-1} b_{ij}, \quad 1 \leq i \leq m,$$

$$H_{ij} \circ \delta_t = t^j H_{ij}, \quad 1 \leq i \leq p.$$

REFERENCES

- [1] L. P. ROTHSCILD AND E. M. STEIN, *Hypoelliptic differential operators and nilpotent groups*, Acta Math. 137 (1976), pp. 247–320.
- [2] R. W. GOODMAN, *Nilpotent Lie Groups, Structure and Applications to Analysis*, Lecture Notes in Mathematics, 562, Springer-Verlag, New York, 1976.
- [3] H. HERMES, *Control systems which generate decomposable Lie algebras*, Differential Equations, 44 (1982), pp. 166–187.
- [4] H. J. SUSSMANN, *Lie brackets and local controllability: a sufficient condition for scalar input systems*, this Journal, 21 (1983), pp. 686–713.
- [5] P. E. CROUCH, *Dynamical realizations of finite Volterra series*, this Journal, 19 (1981), pp. 177–202.
- [6] A. J. KRENER, *On the equivalence of control systems and the linearization of nonlinear systems*, this Journal, 15 (1977), pp. 813–829.
- [7] A. J. KRENER AND C. M. LESIAK, *The existence and uniqueness of Volterra series for nonlinear systems*, Trans. IEEE, AC 23 (1978), pp. 1090–1095.
- [8] P. E. CROUCH, *Realizations of a single Volterra kernel*, Analyse des Systèmes, Bordeaux, 11–16 Sept. 1978; Asterisque 75–76 (1980), pp. 77–85.
- [9] A. J. KRENER, *Lifting techniques in optimal control*, to appear.
- [10] R. W. BROCKETT, *Volterra series and geometric control theory*, Automatica, 12 (1976), pp. 167–176.
- [11] P. E. CROUCH, *Dynamical realizations of finite Volterra series*, Ph.D. thesis, Harvard University, Cambridge, MA, 1977.
- [12] A. J. KRENER, *Bilinear and nonlinear realizations of input-output maps*, this Journal, 13 (1975), pp. 827–834.

OPTIMAL CONSUMPTION FOR GENERAL DIFFUSIONS WITH ABSORBING AND REFLECTING BARRIERS*

S. E. SHREVE,[†] J. P. LEHOCZKY[‡] AND D. P. GAVER[§]

Abstract. Two stochastic control problems of the storage or inventory type are considered for general diffusion processes. In the absorption problem, a diffusion process is controlled by subtracting a nondecreasing withdrawal process. The controlled process is absorbed if it reaches zero. The objective function to be maximized is the expected discounted value of the withdrawals plus a discounted penalty for absorption. In the reflection problem, the process can be controlled by subtracting a withdrawal process and adding a deposit process, and the controlled process must be nonnegative. One seeks to maximize an expected discounted weighted difference in withdrawals and deposits. The value function is computed, and a necessary and sufficient condition for the existence of an optimal policy is given. When they exist, optimal policies are found to be the minimal processes which keep the controlled process inside an interval.

Key words. optimal consumption, diffusion process, reflection, absorption

1. Introduction. We consider two stochastic control problems of the storage or inventory type. In the “absorption problem,” denoted (AP), a diffusion process can be controlled by subtracting off a nondecreasing “withdrawal” process. If the controlled process reaches zero, it is absorbed. The objective to be maximized is the expected discounted value of all withdrawals plus a penalty, discounted back to the initial time, for absorption if it occurs. It is shown that an optimal policy, if it exists, is characterized by an upper reflecting barrier U^* . The optimal policy is to make the minimal amount of withdrawals necessary to keep the controlled process below U^* . Thus the controlled process is a diffusion reflected at U^* and absorbed at zero; the optimal policy is a constant multiple of the local time of the controlled process at U^* .

In the “reflection problem” (RP), a diffusion process can be controlled by subtracting out a “withdrawal” process and adding in a “deposit” process. The utility gained from a withdrawal is no greater and might be less than the utility forfeited by a deposit of the same size. Sufficient deposits must be made to keep the controlled process nonnegative. Thus there is reflection rather than absorption at zero. In this model, the optimal policy, if it exists, is generally characterized by two barriers, L^* and U^* . After an initial deposit or withdrawal to bring the wealth process to the nearest boundary of $[L^*, U^*]$, deposits should be made only when the controlled process drops to L^* , and then only in sufficient quantity to prevent dropping below L^* . Withdrawals should be made only at U^* . We will give assumptions which guarantee $L^* = 0$. In both models, we will give a necessary and sufficient condition for an optimal policy to exist and will determine the value function even when an optimal policy fails to exist.

Applications of diffusion models of this kind to inventory/production control and control of dams are discussed in [8], [1], [2], [5], [6], [19], [20], [17], [16]. In the case of constant drift and diffusion, our model (RP) is very close to models studied by Harrison et al. [8], [10]. Harrison and Taylor [8] admit a linear holding cost, but an

* Received by the editors June 28, 1982, and in revised form January 21, 1983.

[†] Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213. The research of this author was partially supported by the Donors of The Petroleum Research Fund, administered by the American Chemical Society.

[‡] Department of Statistics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213. The research of this author was supported in part by the National Science Foundation under grant ECS 8101576.

[§] Department of Operations Research, Naval Postgraduate School, Monterey, California 93940. The research of this author was supported in part by the Office of Naval Research.

integration by parts will reduce that model to our (RP). Harrison and Taksar [10] admit a convex, extended real-valued, holding cost. They show the optimal policy is characterized by an interval $[L^*, U^*]$ in the same way described earlier for our (RP). Despite this similarity, there is no apparent way to reformulate one of these models into the other. It may well be that they are both special cases of a still more general model. Harrison, Sellke and Taylor [9] treat a model in which there is a positive cost for each deposit and withdrawal. The optimal deposit and withdrawal functions are, consequently, step functions rather than local times.

Our absorption problem (AP) is not in the tradition of these works. It can be used to analyze a wide range of problems which are not terminated when the controlled process reaches zero. In a nonterminated model, there is a value associated with the state zero, and if we set the (AP) penalty equal to this value, the value function for the nonterminated problem will agree with that of (AP). For example, we could postulate a model in which, upon reaching zero, the state is set to a positive value at a positive cost. The value function for this problem can be obtained as a value function for (AP) with proper choice of the penalty. It is also the case that if the L^* corresponding to an optimal policy in (RP) is zero, there is an equivalent (AP) problem which can be constructed in this way.

Excluding any initial jumps, the optimal withdrawal and/or deposit processes in our models are constant multiples of local times for the controlled processes. Control problems in which optimal policies involve reflection and local times have been studied by [3], [4], [13], [14].

In § 2, we compute for later use expectations of some random variables involving hitting times and local times. In § 3 we formulate our models and discuss the relevant Hamilton-Jacobi-Bellman conditions. In § 4 we determine the value functions and give necessary and sufficient conditions for existence of optimal policies. Section 5 is devoted to examples.

2. Integrals with respect to local times. In this section we compute the expected values of some random variables defined in terms of hitting times and/or local times for diffusion processes on an interval. The diffusion process may be either absorbed or reflected at barriers. These computations will be used in the analysis in §§ 3 and 4 of the control problem.

Let $a(\cdot)$ and $\sigma(\cdot)$ be real-valued, Lipschitz continuous functions defined on an interval $[a, b]$. We assume σ is nonvanishing. For initial $x \in [a, b]$, we wish to define processes $\{x_i(t), t \geq 0\}$, $i = 1, 2, 3$, which satisfy

$$(2.1) \quad dx_i(t) = a(x_i(t)) dt + \sigma(x_i(t)) dw(t)$$

when $a < x_i(t) < b$,

$$(2.2) \quad x_i(0) = x,$$

and x_1 is absorbed at a and b , x_2 is reflected at a and absorbed at b , and x_3 is reflected at a and b . The process $\{w(t), t \geq 0\}$ appearing in (2.1) is a standard Brownian motion.

Corresponding to a given Brownian motion, the $x_1(\cdot)$ process exists and is (pathwise) unique [7, § 6]. The situation for $x_2(\cdot)$ and $x_3(\cdot)$ is somewhat more complex. There exists a Brownian motion $\{w(t), t \geq 0\}$ and a filtration $\{\mathcal{F}(t), t \geq 0\}$ such that $w(\cdot)$ is adapted to $\mathcal{F}(\cdot)$, $\{w(s), s \geq t\}$ is independent of $\mathcal{F}(t)$, and there exists a corresponding $x_2(\cdot)$ process also adapted to $\mathcal{F}(\cdot)$ such that up to the time of absorption at b ,

$$(2.3) \quad x_2(t) = x + \int_0^t a(x_2(s)) ds + \int_0^t \sigma(x_2(s)) dw(s) + \zeta_2^a(s).$$

The process $\zeta_2^a(\cdot)$ is adapted to $\mathcal{F}(\cdot)$, nondecreasing, continuous, satisfies $\zeta_2^a(0) = 0$, and for fixed ω , is constant on any interval where $x_2(\cdot) > a$ [7, § 23]. The question of uniqueness is dealt with by Watanabe [18, Theorem 1], who showed that all pairs of processes $(x_2(\cdot), \zeta_2^a(\cdot))$ satisfying the above conditions, even those defined on different probability spaces relative to different Brownian motions, induce the same distribution on the space of paths.

In regard to the doubly reflected process, there exists a Brownian motion $\{w(t), t \geq 0\}$ and a filtration $\{\mathcal{F}(t), t \geq 0\}$ as before, and there exists a corresponding $x_3(\cdot)$ process adapted to $\mathcal{F}(\cdot)$ such that

$$(2.4) \quad x_3(t) = x + \int_0^t a(x_3(s)) ds + \int_0^t \sigma(x_3(s)) dw(s) + \zeta_3^a(t) - \zeta_3^b(t).$$

The processes $\zeta_3^a(\cdot)$ and $\zeta_3^b(\cdot)$ are adapted to $\mathcal{F}(\cdot)$, nondecreasing, continuous, and zero at the origin. On any interval where $x_3(\cdot) > a$, $\zeta_3^a(\cdot)$ is constant. On any interval where $x_3(\cdot) < b$, $\zeta_3^b(\cdot)$ is constant. The triple $(x_3(\cdot), \zeta_3^a(\cdot), \zeta_3^b(\cdot))$ is unique in law. See [7, § 23] and [12, Chapter IV, § 7].

We define stopping times

$$(2.5) \quad \tau_i^y = \inf \{t \geq 0: x_i(t) = y\},$$

where $\tau_i^y = \infty$ if $x_i(t)$ never reaches y . Choose $\beta > 0$. In this section we characterize the four functions on $[a, b]$,

$$(2.6) \quad \varphi_1(x) = E_x e^{-\beta \tau_1^a},$$

$$(2.7) \quad \varphi_2(x) = E_x e^{-\beta \tau_2^b},$$

$$(2.8) \quad \psi_2(x) = E_x \int_0^{\tau_2^b} e^{-\beta t} d\zeta_2^a(t),$$

$$(2.9) \quad \psi_3(x) = E_x \int_0^\infty e^{-\beta t} d\zeta_3^a(t).$$

All these functions will be found to satisfy the differential equation

$$(2.10) \quad \beta f(x) = a(x)f'(x) + \frac{1}{2}\sigma^2(x)f''(x).$$

They are thus characterized by their boundary behavior, which is given as a corollary to the following lemma.

LEMMA 2.1. *Let f be a solution of (2.10). For $a \leq x \leq b$,*

$$(2.11) \quad f(x) = E_x e^{-\beta(\tau_1^a \wedge \tau_1^b)} f(x_1(\tau_1^a \wedge \tau_1^b)),$$

$$(2.12) \quad f(x) = E_x e^{-\beta \tau_2^b} f(b) - f'(a) E_x \int_0^{\tau_2^b} e^{-\beta t} d\zeta_2^a(t),$$

$$(2.13) \quad f(x) = f'(b) E_x \int_0^\infty e^{-\beta t} d\zeta_3^b(t) - f'(a) E_x \int_0^\infty e^{-\beta t} d\zeta_3^a(t).$$

Proof. We prove (2.11) and (2.13). The proof of (2.12) can be given by combining techniques used to prove the others.

Apply Ito's lemma to $y_1(t) := e^{-\beta t} f(x_1(t))$ to obtain

$$dy_1(t) = e^{-\beta t} f'(x_1(t)) \sigma(x_1(t)) dw(t).$$

Integrating from 0 to $\tau_1^a \wedge \tau_2^b$ and noting that the integrand of the stochastic integral is bounded for $a \leq x_1(t) \leq b$, so the expectation of the stochastic integral is zero, we obtain (2.11).

Now apply the generalization of Ito's lemma for semimartingales [15, p. 278 or p. 301] to $y_3(t) := e^{-\beta t} f(x_3(t))$ to obtain

$$(2.14) \quad dy_3(t) = e^{-\beta t} f'(x_3(t)) [\sigma(x_3(t)) dw(t) + d\zeta_3^a(t) - d\zeta_3^b(t)].$$

The measure induced on $[0, \infty)$ by $\zeta_3^a(t)$ assigns zero measure to the set $\{t: x_3(t) \neq a\}$. Therefore,

$$f'(x_3(t)) d\zeta_3^a(t) = f'(a) d\zeta_3^a(t).$$

Likewise,

$$f'(x_3(t)) d\zeta_3^b(t) = f'(b) d\zeta_3^b(t).$$

From (2.14) we have

$$(2.15) \quad E_x e^{-\beta T} f(x_3(T)) - f(x) = E_x \int_0^T e^{-\beta t} [f'(a) d\zeta_3^a(t) - f'(b) d\zeta_3^b(t)].$$

We can let $T \rightarrow \infty$ and separate the terms on the right side of (2.15) to obtain (2.13) if we have

$$(2.16) \quad E_x \int_0^\infty e^{-\beta t} d\zeta_3^a(t) < \infty,$$

$$(2.17) \quad E_x \int_0^\infty e^{-\beta t} d\zeta_3^b(t) < \infty.$$

To prove (2.16), let g be a solution of (2.10) with $g'(a) = -1$, $g'(b) = 0$. In place of (2.15), we have

$$E_x e^{-\beta T} g(x_3(T)) - g(x) = -E_x \int_0^T e^{-\beta t} d\zeta_3^a(t).$$

Letting $T \rightarrow \infty$, we have

$$E_x \int_0^\infty e^{-\beta t} d\zeta_3^a(t) = g(x) < \infty.$$

The proof of (2.17) is similar. \square

COROLLARY 2.2. *The functions $\varphi_1, \varphi_2, \psi_2$ and ψ_3 are solutions to the differential equation (2.10) satisfying the boundary conditions*

$$\begin{aligned} \varphi_1(a) &= 0, & \varphi_1(b) &= 1, \\ \varphi_2'(a) &= 0, & \varphi_2(b) &= 1, \\ \psi_2'(a) &= -1, & \psi_2(b) &= 0, \\ \psi_3'(a) &= -1, & \psi_3(b) &= 0. \end{aligned}$$

Proof. In each case, let f be a solution to (2.10) satisfying the specified boundary conditions and use Lemma 2.1 to show that f satisfies the appropriate definition (2.6)–(2.9). \square

3. Model formulation. Let $a(\cdot)$ and $\sigma(\cdot)$ be real-valued functions defined on $[0, \infty)$ which have Lipschitz continuous derivatives, and assume $a(\cdot)$ and $\sigma(\cdot)$ grow at most linearly and $\sigma(\cdot)$ is nonvanishing.

In the *absorption problem* (AP) we are given P , a penalty for absorption. For each $x \geq 0$, we select a Brownian motion $\{w(t); t \geq 0\}$ relative to a right-continuous filtration $\{\mathcal{F}(t); t \geq 0\}$ and an adapted, nondecreasing, left-continuous, process $\xi(t)$ with $\xi(0) = 0$. We define $x(\cdot)$ by the stochastic integral equation

$$(3.1) \quad x(t) = x + \int_0^t a(x(s)) ds + \int_0^t \sigma(x(s)) dw(s) - \xi(t),$$

and define stopping times

$$(3.2) \quad \tau_y = \inf \{t \geq 0: x(t) = y\}.$$

It is known that (3.1) has a unique solution for $0 \leq t \leq \tau_0$ [7, p. 51]. Define

$$(3.3) \quad V_\xi(x) = E_x \left\{ \int_0^{\tau_0} e^{-\beta t} d\xi(t) + e^{-\beta \tau_0} P \right\},$$

where $\beta > 0$ is a discount factor. We denote by $V^*(x)$ the supremum of $V_\xi(x)$ over all $\xi(\cdot)$ obtained as above.

In the *reflection problem* (RP) we are given a constant $k \geq 1$, and for $x \geq 0$ we must choose a Brownian motion adapted to a right-continuous filtration $\{w(t), \mathcal{F}(t); t \geq 0\}$ and two nondecreasing, left-continuous processes $\{\xi^-(t), t \geq 0\}, \{\xi^+(t), t \geq 0\}$ adapted to $\{\mathcal{F}(t), t \geq 0\}$ and satisfying $\xi^-(0) = \xi^+(0) = 0$. We wish to define a solution to the integral equation

$$(3.4) \quad x(t) = x + \int_0^t a(x(s)) ds + \int_0^t \sigma(x(s)) dw(s) - \xi^-(t) + \xi^+(t).$$

Because we have placed no bounds on $\xi^-(t)$ and $\xi^+(t)$, we must do this somewhat indirectly. Define

$$(3.5) \quad \nu_n = \inf \{t \geq 0: \xi^-(t) \vee \xi^+(t) \geq n\},$$

$$(3.6) \quad \xi_n^-(t) = \xi^-(t \wedge \nu_n),$$

$$(3.7) \quad \xi_n^+(t) = \xi^+(t \wedge \nu_n).$$

For each n , there is a unique solution ([7], p. 51) to

$$(3.8) \quad x_n(t) = x + \int_0^t a(x_n(s)) ds + \int_0^t \sigma(x_n(s)) dw(s) - \xi_n^-(t) + \xi_n^+(t).$$

This solution has the property that on $\{\nu_n \geq t\}$, $x_n(t) = x_m(t)$ a.s. for $m \geq n$. It is thus possible to choose a left-continuous process $\{x(t), t \geq 0\}$ such that $x(t) = x_n(t)$ a.s. on $\{\nu_n \geq t\}$. This process satisfies

$$(3.9) \quad x(t \wedge \nu_n) = x + \int_0^{t \wedge \nu_n} a(x(s)) ds + \int_0^{t \wedge \nu_n} \sigma(x(s)) dw(s) - \xi^-(t \wedge \nu_n) + \xi^+(t \wedge \nu_n)$$

for every n . Since $\nu_n \rightarrow \infty$ a.s. as $n \rightarrow \infty$, we can interpret $x(t)$ as a solution to (3.4). We will say the pair $\{\xi^-(t), \xi^+(t), t \geq 0\}$ is admissible if $x(t) \geq 0$ for all $t \geq 0$. Admissible pairs do exist. Indeed, if we let $\{\xi^+(t), t \geq 0\}$ correspond to reflection at zero as discussed in § 2 and take $\xi^-(t)$ to be identically zero, the pair $\{\xi^-(t), \xi^+(t), t \geq 0\}$ will be

admissible. For an admissible pair, we define

$$(3.10) \quad V_{\xi^-, \xi^+}(x) = \overline{\lim}_{n \rightarrow \infty} E_x \int_0^{\nu_n} e^{-\beta t} [d\xi^-(t) - k d\xi^+(t)].$$

The reflection problem is to find a Brownian motion and admissible $\{\xi^-(t), \xi^+(t), t \geq 0\}$ such that $V_{\xi^-, \xi^+}(x)$ is maximized. Again, we denote the value function by V^* .

LEMMA 3.1. *Consider the model (AP). Let $F: [0, \infty) \rightarrow \mathbb{R}$ be continuous, have two continuous derivatives, and satisfy*

$$(3.11) \quad F(0) \geq P,$$

$$(3.12) \quad F'(x) \geq 1, \quad x \geq 0,$$

$$(3.13) \quad F''(x) \leq 0, \quad x \geq 0,$$

$$(3.14) \quad \beta F(x) \geq a(x)F'(x) + \frac{1}{2}\sigma^2(x)F''(x), \quad x \geq 0.$$

Fix $x \geq 0$ and let $x(\cdot), \xi(\cdot)$ be as in (3.1). Then $F(x) \geq V_\xi(x)$, and consequently,

$$(3.15) \quad F(x) \geq V^*(x), \quad x \geq 0.$$

If $F(x) = V^*(x)$ and $\xi(\cdot)$ is optimal at x , then

$$(3.16) \quad F(x) = E_x \left\{ \int_0^{\tau \wedge \tau_0} e^{-\beta t} d\xi(t) + e^{-\beta(\tau \wedge \tau_0)} F(x(\tau \wedge \tau_0)) \right\}$$

for any almost surely bounded stopping time τ . If $F = V^*$ and

$$(3.17) \quad F'(x) > 1, \quad x \geq 0,$$

then for $x > 0$ there is no optimal policy.

Proof. We show first that for any $x(\cdot), \xi(\cdot)$ pair,

$$W(t) := \int_0^{t \wedge \tau_0} e^{-\beta s} d\xi(s) + e^{-\beta(t \wedge \tau_0)} F(x(t \wedge \tau_0))$$

is a supermartingale relative to $\{\mathcal{F}(t); t \geq 0\}$. In fact, we will show directly the optional sampling result that if τ_1 and τ_2 are bounded, $\{\mathcal{F}(t)\}$ -stopping times with $\tau_1 \leq \tau_2$ a.s., then $E\{W(\tau_2)|\mathcal{F}(\tau_1)\} \leq W_{\tau_1}$. We subsequently show that, as would be expected, when $\xi(\cdot)$ is optimal, the $\{W_n, \mathcal{F}_t\}$ process is a martingale.

Let $x_0(t) = x(t \wedge \tau_0)$ and $\xi_0(t) = \xi(t \wedge \tau_0)$. The process $\{x_0(t); t \geq 0\}$ is left-continuous with limits from the right, contrary to the usual convention. However, the filtration $\{\mathcal{F}(t), t \geq 0\}$ is right-continuous, so the right-continuous process $\{x_0(t+), t \geq 0\}$ is also adapted to this filtration. Moreover,

$$x(t_2 \wedge \tau_0) = \lim_{t \uparrow t_2} x_0(t+), \quad t_2 > 0.$$

For $0 \leq t_1 \leq t$, the differentiation formula for semi-martingales [15, p. 278 or p. 301] implies

$$\begin{aligned} e^{-\beta(t \wedge \tau_0)} F(x_0(t+)) &= e^{-\beta(t_1 \wedge \tau_0)} F(x_0(t_1+)) \\ &+ \int_{t_1 \wedge \tau_0}^{t \wedge \tau_0} e^{-\beta s} [-\beta F(x(s)) + a(x(s))F'(x(s)) + \frac{1}{2}\sigma^2(x(s))F''(x(s))] ds \\ &- \int_{(t_1 \wedge \tau_0, t \wedge \tau_0]} e^{-\beta s} F'(x(s)) d\xi_0(s+) + \int_{t_1 \wedge \tau_0}^{t \wedge \tau_0} F'(x(s))\sigma(x(s)) dw(s) \\ &+ \sum_{t_1 \wedge \tau_0 < s \leq t \wedge \tau_0} [F(x_0(s+)) - F(x_0(s)) - F'(x_0(s))(x_0(s+) - x_0(s))]. \end{aligned}$$

In light of (3.13) and (3.14), this leads to

$$e^{-\beta(t \wedge \tau_0)} F(x_0(t+)) \leq e^{-\beta(t_1 \wedge \tau_0)} F(x_0(t_1+)) \\ - \int_{(t_1 \wedge \tau_0, t \wedge \tau_0]} e^{-\beta s} F'(x(s)) d\xi_0(s+) + \int_{t_1 \wedge \tau_0}^{t \wedge \tau_0} F'(x(s)) \sigma(x(s)) dw(s).$$

The concavity of F also implies

$$0 \leq e^{-\beta(t_1 \wedge \tau_0)} \{F(x_0(t_1)) - F(x_0(t_1+)) - F'(x_0(t_1))[\xi_0(t_1+) - \xi_0(t_1)]\},$$

and adding this to the previous inequality, we obtain

$$e^{-\beta(t \wedge \tau_0)} F(x_0(t+)) \\ \leq e^{-\beta(t_1 \wedge \tau_0)} F(x_0(t_1)) - \int_{[(t_1 \wedge \tau_0), t \wedge \tau_0]} e^{-\beta s} F'(x(s)) d\xi_0(s+) + \int_{t_1 \wedge \tau_0}^{t \wedge \tau_0} F'(x(s)) \sigma(x(s)) dw(s).$$

If $0 \leq t_1 < t_2$, we can let $t \uparrow t_2$ to obtain

$$e^{-\beta(t_2 \wedge \tau_0)} F(x(t_2 \wedge \tau_0)) \\ \leq e^{-\beta(t_1 \wedge \tau_0)} F(x(t_1 \wedge \tau_0)) - \int_{[t_1 \wedge \tau_0, t_2 \wedge \tau_0)} e^{-\beta s} F'(x(s)) d\xi(s) + \int_{t_1 \wedge \tau_0}^{t_2 \wedge \tau_0} F'(x(s)) \sigma(x(s)) dw(s).$$

In other words,

$$(3.18) \quad W(t_2) \leq W(t_1) + \int_{[t_1 \wedge \tau_0, t_2 \wedge \tau_0)} e^{-\beta s} [1 - F'(x(s))] d\xi(s) \\ + \int_{t_1 \wedge \tau_0}^{t_2 \wedge \tau_0} F'(x(s)) \sigma(x(s)) dw(s).$$

Observe that because (3.18) is a pathwise inequality, we can replace t_1 and t_2 by bounded stopping times τ_1 and τ_2 which satisfy $\tau_1 \leq \tau_2 < \infty$ a.s. Making such a replacement and taking conditional expectations, we obtain

$$E_x\{W(\tau_2)|\mathcal{F}(\tau_1)\} \leq W(\tau_1) + E_x\left\{\int_{[\tau_1 \wedge \tau_0, \tau_2 \wedge \tau_0)} e^{-\beta s} [1 - F'(x(s))] d\xi(s) | \mathcal{F}(\tau_1)\right\} \\ + E_x\left\{\int_{\tau_1 \wedge \tau_0}^{\tau_2 \wedge \tau_0} F'(x(s)) \sigma(x(s)) dw(s) | \mathcal{F}(\tau_1)\right\}.$$

By assumption, F' is bounded between 1 and $F'(0)$, and there is a constant K such that $|\sigma(x)| \leq K(x+1)$, $x \geq 0$. Let $x_1(t)$ be the solution of

$$x_1(t) = x + \int_0^t a(x_1(s)) ds + \int_0^t \sigma(x_1(s)) dw(s).$$

By a straightforward generalization of the comparison theorem of Ikeda–Watanabe [11], $x_1(t) \geq x(t)$, $0 \leq t \leq \tau_0$. Therefore,

$$\int_{\tau_1 \wedge \tau_0}^{\tau_2 \wedge \tau_0} [F'(x(s)) \sigma(x(s))]^2 ds \leq [F'(0)K]^2 \int_{\tau_1 \wedge \tau_0}^{\tau_2 \wedge \tau_0} (x_1(s) + 1)^2 ds,$$

which has finite expectation [7, p. 40]. It follows that

$$E_x \left\{ \int_{\tau_1 \wedge \tau_0}^{\tau_2 \wedge \tau_0} F'(x(s)) \sigma(x(s)) dw(s) | \mathcal{F}(\tau_1) \right\} = 0,$$

and so for any bounded stopping times $\tau_1 \leq \tau_2$ a.s.,

$$(3.19) \quad \begin{aligned} E_x \{W(\tau_2) | \mathcal{F}(\tau_1)\} &\leq W(\tau_1) + E_x \left\{ \int_{[\tau_1 \wedge \tau_0, \tau_2 \wedge \tau_0)} e^{-\beta s} [1 - F'(x(s))] d\xi(s) | \mathcal{F}(\tau_1) \right\} \\ &\leq W(\tau_1). \end{aligned}$$

By definition, $W(0) = F(x)$, and since $W(t)$ is a supermartingale and $F \geq P$,

$$F(x) \geq E_x W(t) \geq E_x \left\{ \int_0^{t \wedge \tau_0} e^{-\beta s} d\xi(s) + e^{-\beta(t \wedge \tau_0)} P \right\}.$$

Letting $t \rightarrow \infty$ we obtain $F(x) \geq V_\xi(x)$, and (3.15) follows.

If $F = V^*$ and $\xi(\cdot)$ is optimal, then for any bounded stopping time τ ,

$$\begin{aligned} F(x) = W(0) &\geq E_x W(\tau) \geq \lim_{t \rightarrow \infty} E_x W(\tau + t) \\ &\geq \lim_{t \rightarrow \infty} E_x \left\{ \int_0^{(\tau+t) \wedge \tau_0} e^{-\beta s} d\xi(s) + e^{-\beta((\tau+t) \wedge \tau_0)} P \right\} \\ &= V_\xi(x) = V^*(x) = F(x), \end{aligned}$$

and consequently, $F(x) = E_x W(\tau)$. This proves (3.16) and also shows that $\{W(t), \mathcal{F}(t); t \geq 0\}$ is a martingale.

We now assume $F = V^*$ and (3.17) holds. Note first of all that for all $x \geq 0$, $V^*(x) \geq P + x$, because $P + x$ is the return associated with an initial jump to zero. Thus, for some \bar{x} ,

$$V^*(x) > E_x P e^{-\beta \tau_0}, \quad x \geq \bar{x},$$

where τ_0 is the passage time to zero of the process

$$(3.20) \quad x(t) = x + \int_0^t a(x(s)) ds + \int_0^t \sigma(x(s)) dw(s).$$

(If $P \geq 0$, we may take \bar{x} to be any positive number, but if $P < 0$, \bar{x} must be sufficiently large.) Now suppose for some $x > 0$, there is an optimal $\xi(\cdot)$. Assumption (3.17) and inequality (3.19) imply ξ is identically zero almost surely. Thus, $V_\xi(x) = E_x P e^{-\beta \tau_0}$, and so $0 < x < \bar{x}$. The process given by (3.20) has positive probability of arrival at \bar{x} before absorption at 0, so (3.16), Fatou's lemma, and the strong Markov property imply

$$\begin{aligned} E_x P e^{-\beta \tau_0} &= V^*(x) \\ &= \lim_{t \rightarrow \infty} E_x e^{-\beta(t \wedge \tau_{\bar{x}} \wedge \tau_0)} V^*(x(t \wedge \tau_{\bar{x}} \wedge \tau_0)) \geq E_x e^{-\beta(\tau_{\bar{x}} \wedge \tau_0)} V^*(x(\tau_{\bar{x}} \wedge \tau_0)) \\ &= E_x \{ \chi_{\{\tau_0 < \tau_{\bar{x}}\}} e^{-\beta \tau_0} P \} + E_x \{ \chi_{\{\tau_{\bar{x}} < \tau_0\}} e^{-\beta \tau_{\bar{x}}} V^*(\bar{x}) \} \\ &> E_x \{ \chi_{\{\tau_0 < \tau_{\bar{x}}\}} e^{-\beta \tau_0} P \} + E_x \{ \chi_{\{\tau_{\bar{x}} < \tau_0\}} e^{-\beta \tau_{\bar{x}}} E_{\bar{x}} \{ P e^{-\beta \tau_0} \} \} \\ &= E_x \{ \chi_{\{\tau_0 < \tau_{\bar{x}}\}} e^{-\beta \tau_0} P \} + E_x \{ \chi_{\{\tau_{\bar{x}} < \tau_0\}} E_x [P e^{-\beta \tau_0} | \mathcal{F}(\tau_{\bar{x}})] \} \\ &= E_x \{ \chi_{\{\tau_0 < \tau_{\bar{x}}\}} e^{-\beta \tau_0} P \} + E_x \{ \chi_{\{\tau_{\bar{x}} < \tau_0\}} e^{-\beta \tau_0} P \} \\ &= E_x P e^{-\beta \tau_0}. \end{aligned}$$

This contradiction shows there can be no optimal policy at x . \square

For some of the later analysis, we will need to refer to auxiliary control problems (APU) and (RPU). Let U be a positive “upper barrier.” The problems (APU) and (RPU) are like (AP) and (RP), respectively, except that in the modified problems, the control policies must be chosen so that the $x(\cdot)$ process never exceeds U . The value function V_U^* for these problems is defined only on $[0, U]$. The proof of Lemma 3.1 is also a proof of the following lemma.

LEMMA 3.2. *Consider the model (APU). Let $F: [0, \infty) \rightarrow \mathbb{R}$ have two continuous derivatives and satisfy (3.12)–(3.14) for $0 \leq x \leq U$ and also satisfy (3.11). Then $F(x) \geq V_U^*(x)$, $0 \leq x \leq U$. If $F = V_U^*$ and $\xi(\cdot)$ is optimal, then (3.16) holds for any almost surely bounded stopping time.*

LEMMA 3.3. *Consider the model (RP). Let $F: [0, \infty) \rightarrow \mathbb{R}$ have two continuous derivatives and satisfy*

$$(3.21) \quad 1 \leq F'(x) \leq k, \quad x \geq 0,$$

$$(3.22) \quad F''(x) \leq 0, \quad x \geq 0,$$

$$(3.23) \quad \beta F(x) \geq a(x)F'(x) + \frac{1}{2}\sigma^2(x)F''(x), \quad x \geq 0.$$

Fix $x \geq 0$ and let $x(\cdot)$, $\xi^-(\cdot)$, $\xi^+(\cdot)$ be as in (3.4). Then $F(x) \geq V_{\xi^-, \xi^+}(x)$, and consequently,

$$(3.24) \quad F(x) \geq V^*(x), \quad x \geq 0.$$

If $F(x) = V^*(x)$, and $(\xi^-(\cdot), \xi^+(\cdot))$ is optimal at x , and $\rho_n = \inf\{t \geq 0: x(t) \vee \xi^-(t) \vee \xi^+(t) \geq n\}$, then

$$(3.25) \quad F(x) = E_x \left\{ \int_0^{\tau \wedge \rho_n} e^{-\beta t} [d\xi^-(t) - k d\xi^+(t)] + e^{-\beta(\tau \wedge \rho_n)} F(x(\tau \wedge \rho_n)) \right\},$$

for any almost surely bounded stopping time τ . If $F = V^*$ and

$$(3.26) \quad F'(x) > 1, \quad x \geq 0,$$

then for $x \geq 0$, there is no optimal policy.

Proof. Define

$$W_n(t) = \int_0^{t \wedge \rho_n} e^{-\beta s} [d\xi^-(s) - k d\xi^+(s)] + e^{-\beta(t \wedge \rho_n)} F(x(t \wedge \rho_n)).$$

By an argument similar to the proof of Lemma 3.1, we can show that for any almost surely bounded stopping times τ_1 and τ_2 satisfying $\tau_1 \leq \tau_2$ a.s.,

$$\begin{aligned} E_x \{W_n(\tau_2) | \mathcal{F}(\tau_1)\} \\ \leq W_n(\tau_1) + E_x \left\{ \int_{\tau_1 \wedge \rho_n}^{\tau_2 \wedge \rho_n} e^{-\beta s} [(1 - F'(x(s))) d\xi^-(s) - (k - F'(x(s))) d\xi^+(s)] | \mathcal{F}(\tau_1) \right\} \\ + E_x \left\{ \int_{\tau_1 \wedge \rho_n}^{\tau_2 \wedge \rho_n} F'(x(s)) \sigma(x(s)) dw(s) | \mathcal{F}(\tau_1) \right\}. \end{aligned}$$

Since $x(s)$ is bounded on $[\tau_1 \wedge \rho_n, \tau_2 \wedge \rho_n]$, the Itô integral term has conditional expectation zero. Using this fact and (3.21), we obtain

$$(3.27) \quad \begin{aligned} E_x\{W_n(\tau_2)|\mathcal{F}(\tau_1)\} &\leq W_n(\tau_1) + E_x\left\{\int_{\tau_1 \wedge \rho_n}^{\tau_2 \wedge \rho_n} e^{-\beta s} [(1 - F'(x(s))) d\xi^-(s) \right. \\ &\quad \left. - (k - F'(x(s))) d\xi^+(s)] | \mathcal{F}(\tau_1)\right\} \\ &\leq W_n(\tau_1), \end{aligned}$$

so $\{W_n(t), \mathcal{F}(t)\}$ is a supermartingale.

Let τ be an almost surely bounded stopping time. From (3.27) we have

$$E_x W_{n+q}(0) \geq E_x W_{n+q}(\tau \wedge \rho_n) \geq E_x W_{n+q}((\tau + t) \wedge \nu_m), \quad m \geq n \geq 0, \quad t \geq 0, \quad q \geq 0.$$

In other words,

$$\begin{aligned} F(x) &\geq E_x \left\{ \int_0^{\tau \wedge \rho_n} e^{-\beta s} [d\xi^-(s) - k d\xi^+(s)] + e^{-\beta(\tau \wedge \rho_n)} F(x(\tau \wedge \rho_n)) \right\} \\ &\geq E_x \left\{ \int_0^{(\tau+t) \wedge \nu_m \wedge \rho_{n+q}} e^{-\beta s} [d\xi^-(s) - k d\xi^+(s)] \right. \\ &\quad \left. + e^{-\beta((\tau+t) \wedge \nu_m \wedge \rho_{n+q})} F(x((\tau+t) \wedge \nu_m \wedge \rho_{n+q})) \right\}. \end{aligned}$$

For fixed m , $\int_0^{(\tau+t) \wedge \nu_m \wedge \rho_{n+q}} e^{-\beta s} [d\xi^-(s) - k d\xi^+(s)]$ is uniformly bounded in t, q and ω . Thus, we can let $t \rightarrow \infty, q \rightarrow \infty$ and use the bounded convergence theorem on this term and Fatou's lemma on the other term in the last expectation to obtain

$$\begin{aligned} F(x) &\geq E_x \left\{ \int_0^{\tau \wedge \rho_n} e^{-\beta s} [d\xi^-(s) - k d\xi^+(s)] + e^{-\beta(\tau \wedge \rho_n)} F(x(\tau \wedge \rho_n)) \right\} \\ &\geq E_x \left\{ \int_0^{\nu_m} e^{-\beta s} [d\xi^-(s) - k d\xi^+(s)] + e^{-\beta \nu_m} F(x(\nu_m)) \right\} \\ &\geq E_x \left\{ \int_0^{\nu_m} e^{-\beta s} [d\xi^-(s) - k \xi^+(s)] + e^{-\beta \nu_m} F(0) \right\}. \end{aligned}$$

If we now take $\lim_{m \rightarrow \infty}$ of the last term, we see that $F(x) \geq V_{\xi^-, \xi^+}(x)$ and (3.24) follows. Moreover, if $F(x) = V^*(x) = V_{\xi^-, \xi^+}(x)$, then (3.25) must also hold.

Suppose now $F = V^*$. If initial wealth is x , the policy of immediately jumping to zero and then following a nearly (within $\varepsilon > 0$) optimal policy returns reward at least $x + V^*(0) - \varepsilon$. Since this is positive for large x , we can choose \bar{x} such that $V^*(x) > 0$ for $x \geq \bar{x}$. Suppose that (3.26) holds and for some $x \geq 0$, there is an optimal policy $(\xi^-(\cdot), \xi^+(\cdot))$. From (3.25) we see that $E_x W_n(0) = E_x W_n(t), n \geq 0, t \geq 0$. In order to achieve this equality in (3.27), we must have $\xi^- \equiv 0$ a.s. Therefore, the corresponding $x(\cdot)$ process satisfies

$$(3.28) \quad x(t) = x + \int_0^t a(x(s)) ds + \int_0^t \sigma(x(s)) dw(s) + \xi^+(t),$$

and

$$(3.29) \quad V^*(x) = E_x \left\{ -k \int_0^\infty e^{-\beta t} d\xi^+(t) \right\} \leq 0.$$

This implies $x < \bar{x}$. Extend $a(\cdot)$ and $\sigma(\cdot)$ to all of $[-1, \infty)$ so that they remain Lipschitz continuous and $\sigma(\cdot)$ does not vanish. Let $x_1(\cdot)$ be the process on $(-\infty, \infty)$ absorbed at -1 and \bar{x} and satisfying

$$x_1(t) = x + \int_0^t a(x_1(s)) ds + \int_0^t \sigma(x_1(s)) dw(s).$$

Define stopping times

$$\tau^{\bar{x}} = \inf \{t \geq 0: x(t) = \bar{x}\}, \quad \tau_1^{\bar{x}} = \inf \{t \geq 0: x_1(t) = \bar{x}\}.$$

Corollary (2.2) implies that

$$\varphi_1(x) := E_x e^{-\beta \tau^{\bar{x}}}$$

is a nonzero solution to (2.10) for $-1 \leq x \leq \bar{x}$, and $\varphi_1(-1) = 0$, $\varphi_1(\bar{x}) = 1$. Since φ_1 is nonnegative, any zero of φ_1 in $[0, \bar{x})$ would also be a zero of φ_1' , in which case φ_1 would vanish identically. Therefore,

$$E_x e^{-\beta \tau^{\bar{x}}} > 0, \quad 0 \leq x \leq \bar{x}.$$

By a straightforward generalization of the comparison theorem of Ikeda and Watanabe [11], almost surely

$$x(t) \geq x_1(t), \quad t \geq 0,$$

so

$$(3.30) \quad E_x e^{-\beta \tau^{\bar{x}}} \geq E_x e^{-\beta \tau_1^{\bar{x}}} > 0, \quad 0 \leq x \leq \bar{x}.$$

From (3.25) we have (setting $\tau = t \wedge \tau^{\bar{x}}$ and then letting $t \rightarrow \infty$, $n \rightarrow \infty$)

$$(3.31) \quad V^*(x) = E_x \left\{ -k \int_0^{\tau^{\bar{x}}} e^{-\beta t} d\xi^+(t) + e^{-\beta \tau^{\bar{x}}} V^*(\bar{x}) \right\}.$$

A comparison of (3.29) and (3.31) shows

$$E_x \left\{ -k \int_{\tau^{\bar{x}}}^{\infty} e^{-\beta t} d\xi^+(t) \right\} = E_x \{ e^{-\beta \tau^{\bar{x}}} V^*(\bar{x}) \},$$

which contradicts (3.30) and the fact that $V^*(\bar{x}) > 0$. \square

The proof of Lemma 3.3 is also a proof of the following results concerning the model (RPU) defined just before Lemma 3.2.

LEMMA 3.4. *Consider the model (RPU). Let $F: [0, \infty) \rightarrow \mathcal{R}$ have two continuous derivatives and satisfy (3.21)–(3.23) for $0 \leq x \leq U$. Then $F(x) \geq V_U^*(x)$, $0 \leq x \leq U$. If $F = V_U^*$ and $(\xi^-(\cdot), \xi^+(\cdot))$ is optimal, then (3.25) holds for any almost surely bounded stopping time τ .*

4. Determination of the value function. In this section, we determine the value function for (AP) and (RP) under the assumption

$$(4.1) \quad a'(x) \leq \beta, \quad x \geq 0.$$

In § 5 we give examples, one of which indicates the complexity of the behavior which can occur when this assumption is violated. We begin with some lemmas concerning the behavior of solutions to (2.10).

LEMMA 4.1. *Let k be a solution, not identically zero, to the differential equation*

$$k''(x) = \gamma(x)k(x) + \delta(x)k'(x)$$

on some interval $[a, b]$. Assume $\gamma(\cdot)$ is Lipschitz continuous and nonnegative. If, for some $\bar{x} \in [a, b]$, $k(\bar{x}) > 0$ and $k'(\bar{x}) \leq 0$, then $k'(x) \leq 0$ for $a \leq x \leq \bar{x}$. If k has a zero in $[a, b]$, then k' has no zero in $[a, b]$. If $\gamma(x) > 0$ for all x and for some \bar{x} , $k'(\bar{x}) = 0$, then $(x - \bar{x})k(x)k'(x) > 0$ for $x \in [a, b]$, $x \neq \bar{x}$.

Proof. We consider first the case $\delta \equiv 0$. Under this condition k is convex (strictly convex when $\gamma(x) > 0$) on any interval where it is positive and concave on any interval where it is negative. Suppose for some $\bar{x} \in [a, b]$, $k(\bar{x}) > 0$. Let

$$\bar{y} = \inf \{y \in [a, b] : k(x) \geq 0 \text{ for } y \leq x \leq b\}.$$

Since k is nonnegative on $[\bar{y}, \bar{x}]$, it is convex there. If, in addition, $k'(\bar{x}) \leq 0$, then $k' \leq 0$ on $[\bar{y}, \bar{x}]$, which implies $k(\bar{y}) \geq k(\bar{x}) > 0$. According to the definition of \bar{y} and the continuity of k , we must have $\bar{y} = a$, i.e., $k'(x) \leq 0$ for $a \leq x \leq \bar{x}$.

Suppose now $k(\bar{z}) = 0$ for some $\bar{z} \in [a, b]$. We cannot also have $k'(\bar{z}) = 0$, since the only solution to the differential equation

$$k''(x) = \gamma(x)k(x)$$

satisfying $k(\bar{z}) = k'(\bar{z}) = 0$ is the identically zero solution. Assume without loss of generality that $k'(\bar{z}) > 0$. Define

$$\bar{w} = \sup \{w \in [\bar{z}, b] : k'(x) > 0 \text{ for } \bar{z} \leq x < w\}.$$

Since $k'(w) > 0$ on $[\bar{z}, \bar{w})$, $k(w) > 0$ on (\bar{z}, \bar{w}) , and therefore k is convex on $[\bar{z}, \bar{w}]$. This implies $k'(\bar{w}) \geq k'(\bar{z}) > 0$, and so $\bar{w} = b$. Therefore, k' has no zero in $[\bar{z}, b]$. A similar argument shows k' has no zero in $[a, \bar{z}]$.

When δ is not identically zero, we introduce a change of variable $\varphi(x) = \int_0^x [\exp \int_0^u \delta(v) dv] du$, and define $l(y) = k\varphi^{-1}(y)$. Then

$$l''(y) = [\varphi'(\varphi^{-1}(y))]^{-2} \gamma(\varphi^{-1}(y))l(y),$$

and since $l'(\varphi(x))$ has the same sign as $k'(x)$, k inherits the desired properties from l .

For the final assertion of the lemma, observe that we have already proved under the assumption $k(\bar{x}) > 0$ and $k'(\bar{x}) = 0$ that $k'(x) \leq 0$ for $a \leq x \leq \bar{x}$. When $\gamma(x) > 0$ for $a \leq x \leq b$, we obtain the stronger conclusion $k'(x) < 0$ and so

$$(x - \bar{x})k(x)k'(x) > 0, \quad a \leq x < \bar{x}.$$

When $k'(\bar{x}) = 0$, we have seen that $k(\bar{x}) \neq 0$. In the case that $k(\bar{x}) < 0$, an analogous argument leads to the conclusion that $k'(x) > 0$ and $k(x) < 0$, $a \leq x < \bar{x}$. Analogous arguments also show that

$$(x - \bar{x})k(x)k'(x) > 0, \quad \bar{x} < x \leq b. \quad \square$$

Recall our assumption that $a'(\cdot)$ and $\sigma'(\cdot)$ are Lipschitz continuous.

LEMMA 4.2. Assume (4.1). Let f be a solution, not identically constant, to (2.10) on some interval $[a, b]$.

- (a) If f has a zero in $[a, b]$, then f' has no zero in $[a, b]$.
- (b) If $f'(\bar{x}) = 0$, then $(x - \bar{x})f(x)f'(x) > 0$ for $x \in [a, b]$, $x \neq \bar{x}$.
- (c) If $f'(\bar{x}) > 0$ and $f''(\bar{x}) \leq 0$, then $f''(x) \leq 0$ for $a \leq x \leq \bar{x}$.

Proof. Parts (a) and (b) follow directly from Lemma 4.1. To prove (c), differentiate (2.10) to obtain

$$f'''(x) = \frac{2[\beta - a'(x)]}{\sigma^2(x)} f'(x) - \frac{2}{\sigma^2(x)} [a(x) + \sigma(x)\sigma'(x)] f''(x).$$

Inequality (4.1) allows us to apply Lemma 4.1 to f' . \square

We now investigate (AP). Given an “upper barrier” $U > 0$ and $0 \leq x \leq U$, there is a Brownian motion on a probability space such that there is a diffusion $x_U(\cdot)$ reflected downward at U , absorbed at zero, and satisfying

$$(4.2) \quad x_U(t) = x + \int_0^t a(s) ds + \int_0^t \sigma(s) dw(s) - \xi_U(t), \quad 0 \leq t \leq \tau_U^0,$$

where

$$\tau_U^0 = \min \{t \geq 0: x_U(t) = 0\}.$$

Such reflected processes were discussed in § 2. The reward associated with this policy ζ_U of downward reflection at U is

$$(4.3) \quad V_U(x) := \begin{cases} E_x \left\{ \int_0^{\tau_U^0} e^{-\beta t} d\zeta_U(t) + e^{-\beta \tau_U^0} P \right\}, & 0 \leq x \leq U, \\ x - U + V_U(U), & x \geq U. \end{cases}$$

The definition of $V_U(x)$ for $x > U$ reflects the fact that if $x > U$, we understand ζ_U to have a jump of size $U - x$ at $t = 0$, and after this initial jump, $x_U(\cdot)$ is a diffusion reflected at U and absorbed at zero. For $0 \leq x \leq U$, $V_U(x)$ is the sum of the functions

$$E_x \left\{ \int_0^{\tau_U^0} e^{-\beta t} d\zeta_U(t) \right\} \quad \text{and} \quad PE_x e^{-\beta \tau_U^0}.$$

These fall into the framework of ψ_2 and φ_2 , respectively, Corollary 2.2, except that here reflection occurs at the right endpoint of $[0, U]$ rather than the left endpoint of $[a, b]$. We conclude that on $[0, U]$, V_U is a solution of (2.10) with boundary conditions

$$(4.4) \quad V_U(0) = P, \quad V_U'(U) = 1.$$

Note that while V_U' exists and is equal to one, $V_U''(U)$ may not exist. The right-hand second derivative is zero, but the left-hand second derivative may not be. To avoid complicating the notation, we will use $V_U''(U)$ to denote the left-hand second derivative at U .

THEOREM 4.3. *Consider the model (AP) and assume (4.1) holds. If $a(0) \leq \beta P$, then*

$$(4.5) \quad V^*(x) = P + x, \quad x \geq 0,$$

and the policy of initially jumping to zero ($\xi(0^+) = x$) is optimal. If $a(0) > \beta P$ and there is a $U^ > 0$ such that*

$$(4.6) \quad V_{U^*}''(U^*) = 0,$$

then $V^ = V_{U^*}$ and ζ_{U^*} is optimal. If no such U^* exists, then*

$$(4.7) \quad V^*(x) = \lim_{U \rightarrow \infty} V_U(x), \quad x \geq 0,$$

which limit exists and is finite for every $x \geq 0$, and there does not exist an optimal policy for $x > 0$.

Proof. If $a(0) \leq \beta P$, Lemma 3.1 implies $F(x) := P + x$ majorizes $V^*(x)$. Since the policy of initially jumping to zero gives return $P + x$, this policy is optimal and (4.5) holds.

Suppose $a(0) > \beta P$ and (4.6) holds. From (4.4), (4.6) and Lemma 4.2(c), we see that $V_{U^*}''(x) \leq 0$, $0 \leq x \leq U^*$. According to Lemma 3.1, $V_{U^*} \geq V^*$, so $V_{U^*} = V^*$ and ζ_{U^*} is optimal.

Suppose now that $a(0) > \beta P$ and no $U^* > 0$ satisfying (4.6) exists. Let g and h be linearly independent solutions of (2.10), so for $0 \leq x \leq U$,

$$(4.8) \quad V_U(x) = \frac{g(0)h(x) - h(0)g(x) + P[h'(U)g(x) - g'(U)h(x)]}{g(0)h'(U) - h(0)g'(U)}.$$

Note that $g(x)h'(U) - h(x)g'(U)$ is a solution to (2.10), and since its first derivative has a zero at $x = U$, Lemma 4.2 guarantees $g(0)h'(U) - h(0)g'(U) \neq 0$. Our assumption of nonexistence of U^* satisfying (4.6) is equivalent to assuming

$$\psi(U) := g(0)h''(U) - h(0)g''(U) + P[h'(U)g''(U) - g'(U)h''(U)]$$

has no zero in $(0, \infty)$. Using the fact that g and h solve (2.10), we can write

$$\psi(0) = \frac{2[a(0) - \beta P][h(0)g'(0) - g(0)h'(0)]}{\sigma^2(0)}.$$

The function $h(0)g(x) - g(0)h(x)$ is a solution to (2.10) which has a zero at $x = 0$, so its derivative has no zero. Consequently,

$$(4.9) \quad \frac{h(0)g'(0) - g(0)h'(0)}{g(0)h'(U) - h(0)g'(U)} < 0,$$

and so $\psi(0)$ has the opposite sign of $g(0)h'(U) - h(0)g'(U)$. Since ψ has no zero in $(0, \infty)$, $\psi(U)$ has the same sign as $\psi(0)$, from which we conclude

$$V_U''(U) = \frac{\psi(U)}{g(0)h'(U) - h(0)g'(U)} < 0.$$

Lemma 4.2(c) implies that $V_U''(x) \leq 0$, $0 \leq x \leq U$. Lemma 3.2 implies V_U dominates the value function for (APU), and since ζ_U is feasible in this model, V_U is the value function for (APU). For $x \leq U_1$ and $U_1 \leq U_2$, every policy feasible in (APU_1) is also feasible in (APU_2) , so $V_{U_1}(x) \leq V_{U_2}(x)$. Thus we can define

$$V(x) = \lim_{U \rightarrow \infty} V_U(x), \quad x \geq 0.$$

It is clear that $V \leq V^*$. Choose a policy $\xi(\cdot)$ for (AP), let $x(\cdot)$ be the state process satisfying (3.1), and let τ_U be as in (3.2). As in (3.19), the differentiation formula for semimartingales applied to $e^{-\beta t} V_U(x(t))$ implies that for $t \geq 0$,

$$\begin{aligned} V_U(x) &\geq E_x \left\{ \int_0^{\tau_0 \wedge \tau_U \wedge t} e^{-\beta s} d\xi(s) + e^{-\beta(\tau_0 \wedge \tau_U \wedge t)} V_U(x(\tau_0 \wedge \tau_U \wedge t)) \right\} \\ &\geq E_x \left\{ \int_0^{\tau_0 \wedge \tau_U \wedge t} e^{-\beta s} d\xi(s) + e^{-\beta(\tau_0 \wedge \tau_U \wedge t)} P \right\}. \end{aligned}$$

Letting first $t \rightarrow \infty$ and then $U \rightarrow \infty$, we obtain $V(x) \geq V_\xi(x)$. Equation (4.7) follows.

To see that V^* is finite, we pose a modified control problem in which P is replaced by $a(0)/\beta$. Let V_1^* be the associated value function. It is clear that $V^* \leq V_1^*$, and the part of the theorem already proved states that $V_1^*(x) = a(0)/\beta + x$.

We now show that when $a(0) > \beta P$ and (4.6) does not have a solution, then for $x > 0$, there is no optimal policy. Let $x(\cdot)$ be a process absorbed at zero and satisfying

$$x(t) = x + \int_0^t a(x(s)) ds + \int_0^t \sigma(x(s)) dw(s), \quad 0 \leq t \leq \tau_0,$$

where τ_y is given by (3.2). For $0 < U \leq \hat{U}$, the process $x(t)$ induces the same distribution on sample paths on $[0, \tau_U \wedge \tau_0]$ as does a process absorbed at zero and reflected at \hat{U} . Lemma 3.2 implies (setting $\xi = \xi_U$, $\tau = \tau_U \wedge t$ and letting $t \rightarrow \infty$)

$$V_{\hat{U}}(x) = E_x e^{-\beta(\tau_U \wedge \tau_0)} V_{\hat{U}}(x(\tau_U \wedge \tau_0)).$$

Letting $\hat{U} \rightarrow \infty$, we have

$$V^*(x) = PE_x 1_{\{\tau_0 < \tau_U\}} e^{-\beta\tau_0} + V^*(U) E_x 1_{\{\tau_U < \tau_0\}} e^{-\beta\tau_U}, \quad 0 \leq x \leq U.$$

According to Corollary 2.2, $V^*(x)$ is thus twice continuously differentiable and satisfies (2.10). Since each V_U is concave, so is V^* . It follows that $(d^2/dx^2)V^*(x) \leq 0$, $x \leq 0$. Choose $x \geq 0$, $h \geq 0$. Concavity implies

$$\begin{aligned} \frac{d}{dx} V^*(x) &\geq \frac{1}{h} [V^*(x+h) - V^*(x)] = \lim_{U \rightarrow \infty} \frac{1}{h} [V_U(x+h) - V_U(x)] \\ &\geq \lim_{U \rightarrow \infty} V'_U(x+h) \geq 1. \end{aligned}$$

If, for any U , $(d/dx)V^*(U) = 1$, then V^* is linear for $x \geq U$. We would have $V^*(x) = V_U(x)$ and $V''_U(U) = 0$, contrary to our assumption. It follows that $(d/dx)V^*(x) > 1$ for $x \geq 0$. Lemma 3.1 implies there exists no optimal policy. \square

Remark. The fact that $V^* = V_{U^*}$, where U^* is chosen to make V_{U^*} twice continuously differentiable is a manifestation of the “heuristic principle of smooth fit” advanced in [4] and further expounded in [13], [14]. In this model, satisfying this principle of smooth fit is equivalent to maximizing over U the expression for $V_U(x)$, $0 \leq x \leq U$, given in (4.8).

We now turn our attention to (RP). Given a “lower barrier” $L \geq 0$ and an “upper barrier” $U > L$, there is a diffusion $x_{L,U}(\cdot)$ reflected upward at L and downward at U and satisfying

$$x_{L,U}(t) = x + \int_0^t a(x(s)) ds + \int_0^t \sigma(x(s)) dw(s) - \xi_U^-(t) + \xi_L^+(t), \quad t \geq 0.$$

The reward associated with (ξ_U^-, ξ_L^+) is

$$V_{L,U}(x) := \begin{cases} k(x-L) + V_{L,U}(L), & 0 \leq x \leq L, \\ E_x \left\{ \int_0^\infty e^{-\beta t} [d\xi_U^-(t) - k d\xi_L^+(t)] \right\}, & L \leq x \leq U, \\ x - U + V_{L,U}(U), & x \geq U. \end{cases}$$

The definition of $V_{L,U}$ for $x < L$ and $x > U$ reflects the fact that if $x \notin [L, U]$, then (ξ_U^-, ξ_L^+) causes an immediate jump to the nearest endpoint of $[L, U]$. According to Corollary 2.2, $V_{L,U}$ is a solution of (2.10) for $L \leq x \leq U$, $V'_{L,U}(L) = k$, $V'(U) = 1$. At L and U , $V''_{L,U}$ may not exist. We denote by $V''_{L,U}(L)$ the right-hand second derivative at L and by $V''_{L,U}(U)$ the left-hand second derivative at U .

In § 5 an example will be given in which $V^* = V_{L^*,U^*}$ and $L^* > 0$. However, we will show that under assumption (4.1), we can have only $L^* = 0$. When $k = 1$, the problem becomes trivial, so we dispatch that case first.

THEOREM 4.4. *Consider the model (RP) and assume $k = 1$ and (4.1) holds. Then*

$$V^*(x) = x + \frac{a(0)}{\beta}, \quad x \geq 0,$$

and an optimal policy is given by

$$(4.10) \quad \xi^+(t) = \sigma(0)w^-(t) + ta^-(0)$$

$$(4.11) \quad \xi^-(t) = \begin{cases} 0, & t = 0, \\ x + \sigma(0)w^+(t) + ta^+(0), & t > 0, \end{cases}$$

where $w^\pm(t) = \max\{0, \pm w(t)\}$, $a^\pm(0) = \max\{0, \pm a(0)\}$.

Proof. Lemma 3.3 implies that $F(x) := x + a(0)/\beta \geq V^*(x)$, $x \geq 0$. Let $\xi^+(\cdot)$ and $\xi^-(\cdot)$ be given by (4.10) and (4.11). It is easily verified that

$$x(t) := \begin{cases} x, & t = 0, \\ 0, & t > 0, \end{cases}$$

satisfies (3.9), so $\{\xi^-(\cdot), \xi^+(\cdot), t \geq 0\}$ is feasible. Moreover,

$$\begin{aligned} V_{\xi^-, \xi^+}(x) &= \overline{\lim}_{n \rightarrow \infty} \left\{ x + E_x \int_0^{\nu_n} e^{-\beta t} [\sigma(0) dw(t) + a(0) dt] \right\} \\ &= x + \frac{a(0)}{\beta}. \end{aligned} \quad \square$$

THEOREM 4.5. *Consider the model (RP) and assume $k > 1$ and (4.1) hold. If there is a $U^* > 0$ which satisfies*

$$(4.12) \quad V''_{0,U^*}(U^*) = 0,$$

then $V^ = V_{0,U^*}$ and $(\xi_{U^*}^-, \xi_0^+)$ is optimal. If no such U^* exists, then*

$$(4.13) \quad V^*(x) = \lim_{U \rightarrow \infty} V_{0,U}(x), \quad x \geq 0,$$

which limit exists and is finite for every $x \geq 0$, and there does not exist an optimal policy for $x \geq 0$.

Proof. If U^* satisfying (4.12) exists, then Lemma 4.2(c) implies $V''_{0,U^*}(x) \leq 0$, $0 \leq x \leq U^*$. Lemma 3.3 implies $V_{0,U^*} \geq V^*$, so $V^* = V_{0,U^*}$ and $(\xi_{U^*}^-, \xi_0^+)$ is optimal.

Suppose now there does not exist a $U^* > 0$ satisfying (4.12). As in the proof of Theorem 4.3, we let g and h be linearly independent solutions of (2.10), so for $U > 0$ and $0 \leq x \leq U$,

$$(4.14) \quad V_{0,U}(x) = \frac{g'(0)h(x) - h'(0)g(x) + k[g(x)h'(U) - h(x)g'(U)]}{g'(0)h'(U) - h'(0)g'(U)}.$$

Define

$$\psi(U) = g'(0)h''(U) - h'(0)g''(U) + k[g''(U)h'(U) - h''(U)g'(U)],$$

which by assumption has no zero in $(0, \infty)$. Application of (2.10) results in

$$\psi(0) = \frac{2\beta(k-1)}{\sigma^2(0)} [g(0)h'(0) - h(0)g'(0)],$$

from which we conclude that $\psi(0)$ has the same sign as $g(0)h'(0) - h(0)g'(0)$. We know from (4.9) that this latter quantity is nonzero and has the same sign as $f(0)$, where

$$f(x) := g(x)h'(U) - h(x)g'(U).$$

Since f solves (2.10) and $f'(U) = 0$, Lemma 4.2(b) guarantees $f(0)$ and $f'(0)$ have opposite signs. Therefore, $\psi(0)$ and $f'(0)$ have opposite signs, and since $\psi(U)$ is nonzero

in $(0, \infty)$ and $\psi(0)$ is also nonzero

$$V''_{0,U}(U) = \frac{\psi(U)}{g'(0)h'(U) - h'(0)g'(U)} < 0.$$

Lemmas 4.2(c) and 3.4 imply that $V_{0,U}$ is the value function for (RPU) and (ξ_U^-, ξ_0^+) is optimal in the model. Since the class of admissible policies in (RPU) enlarges with increasing U , $V_{0,U}(x)$ is nondecreasing in U , and we can define

$$V(x) = \lim_{U \rightarrow \infty} V_{0,U}(x), \quad x \geq 0.$$

We have $V \leq V^*$.

The proof that $V = V^*$ is slightly more involved than in Theorem 4.3. We begin by showing $V^*(x) < \infty$, $x \geq 0$. Let $x \geq 0$, $x(\cdot)$, $\xi^-(\cdot)$, $\xi^+(\cdot)$ be as in (3.9), ν_n as in (3.5), so that

$$E_x x(t \wedge \nu_n) = x + E_x \int_0^{t \wedge \nu_n} a(x(s)) ds - E_x [\xi^-(t \wedge \nu_n) - \xi^+(t \wedge \nu_n)], \quad t \geq 0.$$

Let $y_n(t)$ satisfy the deterministic integral equation

$$(4.15) \quad y_n(t) = x + \int_0^t [a^+(0) + \beta y_n(s)] ds - E_x [\xi^-(t \wedge \nu_n) - \xi^+(t \wedge \nu_n)], \quad t \geq 0,$$

where $a^+(0) = \max\{a(0), 0\}$. Inequality (4.1) implies

$$0 \leq E_x x(t \wedge \nu_n) \leq y_n(t), \quad t \geq 0.$$

The solution to (4.15) is

$$y_n(t) = e^{\beta t} \left[x + \frac{a^+(0)}{\beta} (1 - e^{-\beta t}) - \int_0^t e^{-\beta s} dE_x (\xi^-(s \wedge \nu_n) - \xi^+(s \wedge \nu_n)) \right],$$

and so

$$\int_0^t e^{-\beta s} dE_x [\xi^-(s \wedge \nu_n) - \xi^+(s \wedge \nu_n)] \leq x + \frac{a^+(0)}{\beta}, \quad t \geq 0.$$

From definition (3.10), we see that

$$V_{\xi^-, \xi^+}(x) \leq x + \frac{a^+(0)}{\beta},$$

so

$$V(x) \leq V^*(x) \leq x + \frac{a^+(0)}{\beta}, \quad x \geq 0.$$

Now let $x(\cdot)$ be a process with drift $a(\cdot)$ and diffusion $\sigma(\cdot)$ which is reflected upward at zero. For $0 \leq x \leq U \leq \hat{U}$, Lemma 3.4 implies (setting $\tau = \tau_U \wedge t$ and letting $t \rightarrow \infty$, $n \rightarrow \infty$)

$$V_{0,\hat{U}}(x) = E_x \left\{ -k \int_0^{\tau_U} e^{-\beta t} d\xi_0^+(t) + e^{-\beta \tau_U} V_{0,\hat{U}}(U) \right\},$$

where $\tau_U = \inf\{t \geq 0: x(t) = U\}$. Letting $\hat{U} \rightarrow \infty$, we obtain

$$V(x) = E_x \left\{ -k \int_0^{\tau_U} e^{-\beta t} d\xi_0^+(t) + e^{-\beta \tau_U} V(U) \right\}.$$

From Corollary 2.2 we conclude that V is twice continuously differentiable, is a solution to (2.10), and $V'(0) = k$. Since each $V_{0,U}$ is concave and has first derivative at least 1, V also has these properties. We conclude from Lemma 3.3 that $V \geq V^*$, so $V = V^*$. If $V'(U) = 1$ for any U , then V would be linear for $x \geq U$ and this U would satisfy (4.12). Therefore, $V'(x) > 1$ for $x \geq 0$, and Lemma 3.3 implies that no optimal policy can exist. \square

Remark. As in the model (AP), condition (4.12) is the principle of smooth fit. If U^* satisfies this condition, it maximizes $V_{0,U}$ (cf. (4.14)) over U .

5. Examples. In this section, we give several examples to illustrate the results of § 4. The first two examples deal with ordinary Brownian motion and a general logarithmic Brownian motion. Both satisfy the condition $\beta > a'(x)$. A third example is designed to illustrate the many possibilities which can occur when the condition $\beta \geq a'(x)$ is violated.

Example 1. Brownian motion. We take $a(x) = \mu$ and $\sigma(x) = \sigma$, $\beta > 0 = a'(x)$. The general solution of (2.10) is of the form

$$(5.1) \quad f(x) = c_1 e^{r_1 x} + c_2 e^{r_2 x},$$

where $r_1 > 0 > r_2$ are the roots of

$$(5.2) \quad \sigma^2 r^2 + 2\mu r - 2\beta = 0.$$

We note that $r_2^2 > r_1^2$ if and only if $\mu > 0$.

For (AP) we must select c_1, c_2 , and U^* such that

$$(5.3) \quad f(0) = P, \quad f'(U^*) = 1, \quad f''(U^*) = 0.$$

If $P < \mu/\beta$, the resulting value function will be given by

$$(5.4) \quad V^*(x) = \begin{cases} c_1 e^{r_1 x} + c_2 e^{r_2 x}, & 0 \leq x \leq U^*, \\ V^*(U^*) + x - U^*, & x \geq U^*. \end{cases}$$

If $P \geq \mu/\beta$, the value function will be given by

$$(5.5) \quad V^*(x) = x + P.$$

Conditions (5.3) applied to (5.4) give rise to transcendental equations. A simple special case occurs when $P = 0$ and $\mu > 0$. In this case

$$(5.6) \quad U^* = \frac{1}{r_1 - r_2} \log \frac{r_2^2}{r_1^2},$$

and

$$(5.7) \quad V^*(x) = \begin{cases} \frac{e^{r_1 x} - e^{r_2 x}}{r_1 e^{r_1 U^*} - r_2 e^{r_2 U^*}}, & 0 \leq x \leq U^*, \\ V^*(U^*) + x - U^*, & x > U^*. \end{cases}$$

If $P \leq \mu/\beta$, then U^* is the unique positive root of

$$(5.8) \quad P = \left(1 - \frac{r_1^2}{r_2^2} e^{(r_1 - r_2)U^*}\right) / \left(\frac{r_1}{r_2}(r_2 - r_1) e^{r_1 U^*}\right).$$

In the case of (RP), the conditions analogous to (5.3) are

$$(5.9) \quad f'(0) = k, \quad f'(U^*) = 1, \quad f''(U^*) = 0.$$

The value function will be given by (5.4) with c_1 , c_2 , and U^* appropriately chosen to satisfy (5.9). Again transcendental equations arise. It is, however, interesting to relate the two control problems (AP) and (RP). A simple example is afforded by the case $\mu = 0$ and $P \leq 0$. In this case $r_1 = -r_2 = \sqrt{2\beta/\sigma}$ and

$$(5.10) \quad U^* = \frac{1}{r_1} \log(-r_1 P + \sqrt{r_1^2 P^2 + 1}).$$

The resulting value function for (AP) is identical to that for (RP) with k taken to be

$$(5.11) \quad k = \sqrt{r_1^2 P^2 + 1} \geq 1.$$

The results are in general agreement with those of Harrison and Taylor [8].

Example 2. General logarithmic Brownian motion. In this example, we let $a(x) = ax + b$, $\sigma(x) = \theta(ax + b)$, with $\beta > a = a'(x) > 0$ and $b > 0$. Solutions of (2.10) are of the form

$$(5.12) \quad f(x) = c_1(ax + b)^{r_1} + c_2(ax + b)^{r_2},$$

where $r_1 > 0 > r_2$ are the roots of

$$(5.13) \quad \frac{1}{2}\theta^2 a^2 r^2 + r(a - \frac{1}{2}\theta^2 a^2) - \beta = 0.$$

For (AP) one must select c_1 , c_2 , and U^* to satisfy (5.3). The resulting value function will be of the form

$$(5.14) \quad V^*(x) = \begin{cases} c_1(ax + b)^{r_1} + c_2(ax + b)^{r_2}, & 0 \leq x \leq U^*, \\ V^*(U^*) + x - U^*, & x \geq U^*, \end{cases}$$

if $P < b/\beta$, while it is given by (5.5) if $P \geq b/\beta$.

By way of illustration, when $P = 0$ we find

$$(5.15) \quad \begin{aligned} c_2 &= -c_1 b^{r_1 - r_2}, \\ U^* &= \frac{b}{a} \left(\left(\frac{r_2(r_2 - 1)}{r_1(r_1 - 1)} \right)^{1/(r_1 - r_2)} - 1 \right), \\ c_1 &= (r_1 - 1) / ((r_2 - r_1)r_2 a (aU^* + b)^{r_2 - 1} b^{r_1 - r_2}). \end{aligned}$$

The condition $\beta > a$ ensures $r_1 > 1$.

When $\beta = a$, there is no $U^* < \infty$ such that V_{U^*} satisfies (5.3) or (5.9). The value function is finite and given by (4.7) or (4.13). For example in (AP) with $P = 0$

$$V^*(x) = x + \frac{b}{a} - \left(\frac{b}{a} \right)^{1-r_2} \left(x + \frac{b}{a} \right)^{r_2}.$$

There is no optimal policy.

The reflecting problem (RP) involves solutions of the form (5.14) with boundary conditions (5.9). Again transcendental equations arise. As an illustration, the value function specified by (5.15) for $P = 0$ arises in (RP) when k is taken to be

$$(5.16) \quad k = c_1 a b^{r_1 - 1} (r_1 - r_2).$$

Example 3. This example is designed to illustrate the possible behavior which can occur if we remove the condition $\beta \geq a'(x)$. In the (RP) formulation we specify $k > 1$. In the (AP) formulation we specify $P = -((k - 1)/2) \cdot \pi/2$.

Let

$$(5.17) \quad a(x) = \frac{\beta((k+1)/2)x - (\beta+1)((k-1)/2) \cos x}{(k+1)/2 + ((k-1)/2) \sin x}$$

$$\sigma(x) = \sqrt{2}.$$

The value function for (AP) and (RP) is given by

$$(5.18) \quad V^*(x) = \begin{cases} k\left(x - \frac{\pi}{2}\right) + \frac{k+1}{2} \frac{\pi}{2}, & 0 \leq x \leq \frac{\pi}{2}, \\ \frac{k+1}{2} x - \left(\frac{k-1}{2}\right) \cos x, & x \geq \frac{\pi}{2}. \end{cases}$$

The value function corresponds to the following optimal policy. If $0 < x < \pi/2$, an immediate jump up to $\pi/2$ is made. The region $[\pi/2, \infty)$ is divided into intervals of length π . If $x \in [\pi/2 + 2\pi n, 3\pi/2 + 2\pi n]$ for some $n \geq 0$, then ξ^+ and ξ^- are taken to be the minimal processes which keep the x process inside this interval. If $x \in [3\pi/2 + 2\pi n_0, 5\pi/2 + 2\pi n_0]$ for some $n_0 \geq 0$, then ξ^+ and ξ^- are taken to be 0 until the first time x reaches $3\pi/2 + 2\pi n_0$ or $5\pi/2 + 2\pi n_0$. At this time, ξ^+ and ξ^- are the minimal processes which keep the x process in the adjacent interval. It is easy to check that V satisfies (3.14) (with equality for $x \in [\pi/2, \infty)$),

$$V(0) = P, \quad V'(0) = k, \quad V'(\pi/2 + 2\pi n) = k, \quad V''(\pi/2 + 2\pi n) = 0,$$

$$V'(3\pi/2 + 2\pi n) = 1, \quad \text{and} \quad V''(3\pi/2 + 2\pi n) = 0.$$

Acknowledgments. The approach represented by Lemmas 4.1 and 4.2 was suggested by Charles Coffman. The proof of Lemma 2.1 is due to Ioannis Karatzas. We wish to express our gratitude to them.

REFERENCES

- [1] J. A. BATHER, *A continuous time inventory model*, J. Appl. Prob., 3 (1966), pp. 538–549.
- [2] ———, *A diffusion model for control of a dam*, J. Appl. Prob., 5 (1968), pp. 55–71.
- [3] J. A. BATHER AND H. CHERNOFF, *Sequential decisions in the control of a spaceship*, Proc. Fifth Berkeley Symposium on Math. Statistics and Probability, 3, 1967, Univ. California Press, Berkeley, pp. 181–207.
- [4] V. E. BENEŠ, L. A. SHEPP AND H. S. WITSENHAUSEN, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 39–83.
- [5] M. J. FADDY, *Optimal control of finite dams: Discrete (2-stage) output procedure*, J. Appl. Prob., 11 (1974), pp. 111–121.
- [6] ———, *Optimal control of finite dams*, Adv. Appl. Prob., 6 (1974), pp. 689–710.
- [7] I. I. GIKHMAN AND A. V. SKOROKHOD, *Stochastic Differential Equations*, Springer, New York, 1972.
- [8] J. M. HARRISON AND A. J. TAYLOR, *Optimal control of a Brownian storage system*, Stoch. Proc. Appl., 6 (1978), pp. 179–194.
- [9] J. M. HARRISON, T. M. SELLKE AND A. J. TAYLOR, *Impulse control of Brownian motion*, Tech. Rep., Dept. Operations Research, Stanford Univ., Stanford, CA, 1981.
- [10] J. M. HARRISON AND M. I. TAKSAR, *Instantaneous control of Brownian motion*, Tech. Rep., Dept. Operations Research, Stanford Univ., Stanford, CA, 1981.
- [11] N. IKEDA AND S. WATANABE, *A comparison theorem for solutions of stochastic differential equations and its applications*, Osaka J. Math., 14 (1977), pp. 619–633.
- [12] ———, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, New York, 1981.
- [13] I. KARATZAS, *The monotone follower problem in stochastic decision theory*, Appl. Math. Optim., 7 (1981), pp. 175–189.
- [14] ———, *A class of singular stochastic control problems*, Adv. Appl. Probability, 15 (1983), to appear.

- [15] P. A. MEYER, ed., *Lecture Notes in Mathematics* 511, Séminaire de Probabilités X, Université de Strasbourg, Springer, New York, 1976.
- [16] S. R. PLISKA, *A diffusion model for the optimal operation of a reservoir system*, J. Appl. Prob., 12 (1975), pp. 859–863.
- [17] M. L. PUTERMAN, *A diffusion process model for a storage system*, TIMS Studies in Management Sciences, 1 (1975), pp. 859–863.
- [18] S. WATANABE, *On stochastic differential equations for multidimensional diffusion processes with boundary conditions*, J. Math. Kyoto Univ., 11 (1971), pp. 169–180.
- [19] W. WHITT, *Diffusion models for inventory and production systems*, Tech. Rep., Dept. Administrative Sciences, Yale Univ., New Haven, CT, 1973.
- [20] ———, *The diffusion inventory model with discounting*, Tech. Rep., Dept. Administrative Sciences, Yale Univ., New Haven, CT, 1973.

RICCATI EQUATION ARISING IN A BOUNDARY CONTROL PROBLEM WITH DISTRIBUTED PARAMETERS*

FRANCO FLANDOLI†

Abstract. Global existence is proved for the solution of a Riccati differential equation connected with the synthesis of a boundary control problem governed by parabolic partial differential equations.

Key words. Riccati equation, optimal control, parabolic systems, Dirichlet boundary control

1. Introduction. In this work we present a direct study of a Riccati operator equation which is connected with the synthesis of several boundary control problems governed by parabolic partial differential equations. It may be found in this case that the optimal control may be represented by means of an operator valued function $P(\cdot)$ which verifies formally the following differential equation

$$(1.1) \quad \begin{aligned} P'(t) + A^*P(t) + P(t)A + P(t)ABR(t)B^*A^*P(t) &= M(t), \quad t \in [0, T] \\ P(0) &= P_0. \end{aligned}$$

Here $(-A)$ is the infinitesimal generator of an analytic semigroup of bounded linear operators in a Hilbert space E , $M(t)$ and P_0 are bounded linear operators in E , B is a bounded linear operator mapping an Hilbert space U (the control space) into E , $R(t)$ is a bounded linear operator in U , P' denotes the derivative of P with respect to t , A^* and B^* denote the adjoint operators respectively of A and B .

Note that in (1.1) the unbounded operator A also appears in the nonlinear term; this is a new feature in comparison with the Riccati equation connected with distributed control problems.

Equation (1.1) arises in the problem of minimizing

$$(1.2) \quad \int_0^T \{ \langle M(T-s)y(s), y(s) \rangle_E + \langle R^{-1}(T-s)u(s), u(s) \rangle_U \} ds + \langle P_0 y(T), y(T) \rangle_E$$

under the state equation

$$y'(t) + A(y(t) - Bu(t)) = 0, \quad y(0) = y_0$$

(see for example [1] and Proposition 3.3) where we have denoted by $\langle \cdot, \cdot \rangle_E$ and $\langle \cdot, \cdot \rangle_U$ the scalar products in E and U respectively. The problem of the boundary control of a parabolic system through a Dirichlet condition or a Newman condition (with distributed observation) may be studied in this way (see [1] and [6]). Up to now the synthesis of these problems has been studied using variational methods; for the case when $P_0 = 0$, A. V. Balakrishnan (see [1]) derived for the decoupling operator P the equation

$$(1.3) \quad \begin{aligned} \frac{d}{dt} \langle P(t)x, y \rangle_E + \langle P(t)x, Ay \rangle_E + \langle Ax, P(t)y \rangle_E \\ + \langle R(t)B^*A^*P(t)x, B^*A^*P(t)y \rangle_U = \langle M(t)x, y \rangle_E, \\ P(0) = P_0 \end{aligned}$$

* Received by the editors June 15, 1982, and in revised form December 27, 1982.

† Classe di Scienze, Scuola Normale Superiore, Pisa, Italy.

where $t \in [0, T]$, and x and y belong to the domain of A : in [9] and [10], M. Sorine has studied a class of stationary and differential Riccati equations similar to (1.1), considering also some initial values $P_0 \neq 0$ (with a regularity condition a little stronger than (2.3)): in his works the existence of the decoupling operator P is proved using J. L. Lions' techniques (see [7, Chap. 3]), while a new technique is employed to derive a Riccati equation for P and to prove the unicity of the solution.

The class of problems covered by [1] and [10] is similar to ours (apart from the hypotheses on P_0): in particular, regularity conditions on M are not required.

We present here a direct approach to equation (1.1), without using any of the results of the control problem. In this way we are able to solve the synthesis of problem (1.2) using the dynamic programming method. This is believed to be the first direct study of (1.1); for the case when A does not appear in the nonlinear term, direct studies of infinite dimensional Riccati equations may be found in the papers of G. Da Prato [3], L. Tartar [12], and the book of R. Curtain and A. Pritchard [2]. An application of our techniques to stochastic problems may be found in [4].

The contents of this paper are outlined below.

In § 2 the contraction principle is used to prove the local existence of a mild solution of (1.1): in § 3 we prove the "a priori" estimation which leads to the main result of global existence and unicity. Moreover we show in § 3 that (1.3) is verified, and that some regularity results hold: finally we use the dynamic programming technique to solve the synthesis of problem (1.2).

We conclude this section by listing some notation.

Let U and E be two complex Hilbert spaces, with scalar products $\langle \cdot, \cdot \rangle_U$ and $\langle \cdot, \cdot \rangle_E$ respectively. We shall denote by $L(U, E)$ the Banach space of bounded linear operators from U to E , and we shall set $L(E) = L(E, E)$. We shall denote by $H(E)$ the Banach space of hermitian linear operators in E : we shall set $H^+(E) = \{T \in H(E), \text{ such that } \langle Tx, x \rangle_E \geq 0 \forall x \in E\}$. If $T \in L(U, E)$ we shall denote by T^* its dual operator from E to U .

If X is a Banach space (with norm $\|\cdot\|_X$) and $P \in [1, +\infty)$, $a, b \in \mathbb{R}$, $a < b$, then we shall denote by $L^P(a, b; X)$ the Banach space of functions $f: [a, b] \rightarrow X$, Bochner measurable, such that $\int_a^b \|f(t)\|_X^P dt < +\infty$. We shall denote by $C(a, b; X)$, the Banach space of all continuous functions from $[a, b]$ to X .

We shall denote by $C_s(a, b; L(U, E))$ the set of all mappings $T(\cdot): [a, b] \rightarrow L(U, E)$ such that $T(\cdot)x$ is continuous for any $x \in U$. Moreover we shall set $C_s(a, b; H(E)) = \{T \in C_s(a, b; L(E)), \text{ such that } T(t) \in H(E) \text{ for any } t \in [a, b]\}$, and similarly for $C_s(a, b; H^+(E))$. We shall set $C_s^1(a, b; H^+(E)) = \{T \in C_s(a, b; H^+(E)), \text{ such that } T(\cdot)x \text{ is continuously differentiable on } [a, b] \text{ for any } x \in E\}$.

Let $(-A)$ be the infinitesimal generator of an analytic semigroup in E : we shall denote by $D(A)$ the domain of A , by $\rho(A)$ the resolvent set of A , by A^* the adjoint of A , by e^{-tA} and e^{-tA^*} , $t \geq 0$, the semigroups generated by $(-A)$ and $(-A^*)$ respectively. If $\alpha \in \mathbb{R}$, we shall denote by A^α the fractional powers of A (see, for example, [11]) and by $D(A^\alpha)$ the domain of A^α .

2. Local solution. We study (1.1) under the following hypotheses:

- (2.1) $(-A)$ is the infinitesimal generator of an analytic semigroup in the Hilbert space E ; moreover we assume for simplicity that $0 \in \rho(A)$;
- (2.2) B is a bounded linear operator from U to E , and there exists $\alpha \in (0, 1)$ such that $B \in L(U, D(A^\alpha))$.

These assumptions are verified in several boundary control problems (see [6]).

We assume further that

$$(2.3) \quad M \in C_s(0, T; H(E)), \quad R \in C_s(0, T; H(U)), \quad P_0 \in H(E) \cap L(E, D(A^{*1-\alpha})).$$

Equation (1.1) is not well defined because of the unbounded operator A . In order to study (1.1) we set $Q(t) = (A^*)^{1-\alpha} P(t)$ and $Q_0 = (A^*)^{1-\alpha} P_0$: then formally Q verifies the following integral equation:

$$(2.4) \quad \begin{aligned} Q(t)x = & e^{-tA^*} Q_0 e^{-tA} x + \int_0^t (A^*)^{1-\alpha} e^{-(t-s)A^*} \\ & \cdot (M(s) - Q^*(s) A^\alpha B R(s) (A^\alpha B)^* Q(s)) e^{-(t-s)A} x \, ds, \end{aligned}$$

$x \in E$, $t \in [0, T]$. If $Q \in C_s(0, T_1; L(E))$ is a solution of (2.4) on $[0, T_1]$, then we say that $P(\cdot) = (A^*)^{\alpha-1} Q(\cdot)$ is a mild solution of (1.1) on $[0, T_1]$.

We remark that the integral in (2.4) is meaningful because from assumption (2.1) it follows that

$$(2.5) \quad \|(A^*)^{1-\alpha} e^{-tA^*}\|_{L(E)} \leq \frac{c_0}{t^{1-\alpha}} \quad \forall t \in (0, T),$$

for a suitable constant $c_0 > 0$ (see for instance [11]).

It is useful to introduce a family of equations, with bounded operators, which "approximates" (1.1). Let us set $A_n = nA(A+n)^{-1}$, $A_n^* = nA^*(A^*+n)^{-1}$, for any $n \in N$; since A_n and A_n^* are bounded linear operators, we may consider the analytic functions e^{-tA_n} and $e^{-tA_n^*}$, $t \in \mathbb{C}$. We consider the equation

$$(2.6) \quad \begin{aligned} P_n(t) + A_n^* P_n(t) + P_n(t) A_n + P_n(t) A_n B R(t) B^* A_n^* P_n(t) &= M(t), \quad t \in [0, T], \\ P(0) &= P_0. \end{aligned}$$

Setting $Q_n(t) = (A^*)^{1-\alpha} n(A^*+n)^{-1} P_n(t)$, and $Q_{0,n} = (A^*)^{1-\alpha} n(A^*+n)^{-1} P_0$, we find that Q_n verifies the following integral equation:

$$(2.7) \quad \begin{aligned} Q_n(t) = & e^{-tA_n^*} Q_{0,n} e^{-tA_n} + \int_0^t (A^*)^{1-\alpha} n(A^*+n)^{-1} e^{-(t-s)A_n^*} \\ & \cdot (M(s) - Q_n^*(s) A^\alpha B R(s) (A^\alpha B)^* Q_n(s)) e^{-(t-s)A_n} \, ds. \end{aligned}$$

It is easy to see that there exists a constant $c_1 > 0$ such that

$$(2.8) \quad \|(A^*)^{1-\alpha} n(A^*+n)^{-1} e^{-tA_n^*}\| \leq \frac{c_1}{t^{1-\alpha}} \quad \forall t \in (0, T),$$

$\forall n \in N$; using the contraction principle, and properties (2.5) and (2.8), we can prove the following proposition on the local solutions of (2.4) and (2.7).

PROPOSITION 2.1. *Suppose that (2.1), (2.2), (2.3) hold. Let $Q_0, Q_{0,n} \in L(E)$ be such that $Q_{0,n}x \rightarrow Q_0x \, \forall x \in E$, and let $m_0 \geq \sup_{n \in N} \|Q_{0,n}\|_{L(E)}$. Let*

$$c_2 = \sup_{n \in N, t \in [0, T]} \{\|e^{-tA}\|_{L(E)}, \|e^{-tA_n}\|_{L(E)}\}.$$

Then for any $\rho > c_2^2 m_0$ there exists $T_\rho > 0$ such that (2.4) has a unique solution $Q \in C_s(0, T_\rho; L(E))$ on $[0, T_\rho]$, and (2.7) has a unique solution $Q_n \in C_s(0, T_\rho; L(E))$ on $[0, T_\rho] \, \forall n \in N$. Moreover, $Q_n \rightarrow Q$ in $C_s(0, T; L(E))$. Finally, T_ρ depends only on the numbers ρ, m_0, c_2, c_0, c_1 (see (2.5) and (2.8)), $\alpha, \|A^\alpha B\|_{L(U,E)}$,

$$\sup_{t \in [0, T]} \|M(t)\|_{L(E)}, \quad \sup_{t \in [0, T]} \|R(t)\|_{L(U)}.$$

Proof. We denote by $B(0, T_1; L(E))$, $T_1 \in [0, T]$, the Banach space of applications $U: [0, T_1] \rightarrow L(E)$ such that $U, U^* \in C_s(0, T_1; L(E))$, endowed with the norm $\|U\|_{B(0, T_1)} = \sup_{t \in [0, T_1]} \|U(t)\|_{L(E)}$. We set $B_{\rho, T_1} = \{U \in B(0, T_1; L(E)), \text{ such that } \|U\|_{B(0, T_1)} \leq \rho\}$. When $\rho > c_2^2 m_0$, using (2.5) and (2.8), we may find $T_\rho > 0$ such that the contraction principle applies in B_{ρ, T_ρ} to equations (2.4) and (2.7). The convergence of Q_n to Q follows by the contraction principle with the same techniques of [3]. \square

By Proposition 2.1, this corollary follows immediately.

COROLLARY 2.1. *Suppose that (2.1), (2.2), (2.3) hold. Then there exists $T_1 > 0$ such that (1.1) and (2.6) have unique mild solutions, P and P_n respectively, on $[0, T_1]$. Moreover, $P \in C_s(0, T_1; L(E, D(A^{*1-\alpha})))$, $P_n \rightarrow P$ in $C_s(0, T_1; H(E))$ and $B^* A_n^* P_n \rightarrow (A^\alpha B) A^{*1-\alpha} P$ in $C_s(0, T_1; L(E, U))$.*

In the next section we shall prove a global existence result for (1.1) using the following lemma.

LEMMA 2.1. *Suppose that the hypotheses of Proposition 2.1 hold, suppose that there exists a unique global solution $Q_n \in C_s(0, T; L(E))$ of (2.7) $\forall n \in \mathbb{N}$, and suppose that there exists a constant $c > 0$ such that*

$$(2.9) \quad \|Q_n(t)\|_{L(E)} \leq c \quad \forall t \in [0, T], \quad \forall n \in \mathbb{N}.$$

Then there exists a unique global solution $Q \in C_s(0, T; L(E))$ of (2.4) and $Q_n \rightarrow Q$ in $C_s(0, T; L(E))$.

Proof. Let $T_\rho > 0$ be given by Proposition 2.1. We may write (2.4) in the form

$$\begin{aligned} Q(t)x &= e^{-(t-T_\rho)A^*} Q(T_\rho) e^{-(t-T_\rho)A} x \\ &+ \int_{T_\rho}^t A^{*1-\alpha} e^{-(t-s)A^*} (M(s) - Q^*(s) A^\alpha B R(s) (A^\alpha B)^* Q(s)) e^{-(t-s)A} x ds; \end{aligned}$$

and by Proposition 2.1, we obtain a solution on $[T_\rho, 2T_\rho]$. We conclude by a finite number of steps. \square

3. Global solution. We prove now the “a priori” estimate (2.9) assuming (2.1), (2.2), (2.3), and moreover that

$$(3.1) \quad P_0 \in H^+(E), M(t) \in H^+(E), R(t) \in H^+(E) \quad \forall t \in [0, T], \text{ and there exists } \nu > 0 \text{ such that } \langle R(t)x, x \rangle_U \geq \nu \|x\|_U^2 \quad \forall x \in U, \forall t \in [0, T].$$

It is well known that under these hypotheses there exists a unique solution $P_n \in C_s^1(0, T; H^+(E))$ of (2.6), $\forall n \in \mathbb{N}$. Note that equation (2.6) may be written in the form

$$(3.2) \quad \begin{aligned} P_n'(t) + A_n^* P_n(t) + P_n(t) (A_n + A_n B R(t) B^* A_n^* P_n(t)) &= M(t), \quad t \in [0, T], \\ P_n(0) &= P_0, \end{aligned}$$

or even in the form

$$(3.3) \quad \begin{aligned} P_n'(t) + (A_n^* + P_n(t) A_n B R(t) B^* A_n^*) P_n(t) \\ + P_n(t) (A_n - A_n B R(t) B^* A_n^* P_n(t)) &= M(t) + P_n(t) A_n B R(t) B^* A_n^* P_n(t), \\ P_n(0) &= P_0. \end{aligned}$$

Denote by $U_n(t, s)$ the evolution operator associated to the family $\{-A_n^* - P_n A_n BR(t) B^* A_n^*\}_{t \in [0, T]}$. We have

$$\begin{aligned} \frac{\partial}{\partial t} U_n(t, s) &= [-A_n^* - P_n(t) A_n BR(t) B^* A_n^*] U_n(t, s), \\ \frac{\partial}{\partial s} U_n(t, s) &= -U_n(t, s) [-A_n^* - P_n(s) A_n BR(s) B^* A_n^*], \end{aligned}$$

for $0 \leq s \leq t \leq T$, and by transposition we have

$$(3.4) \quad \frac{\partial}{\partial s} U_n^*(t, s) = -[-A_n - A_n BR(s) B^* A_n^* P_n(s)] U_n^*(t, s).$$

From (2.6) it follows that

$$(3.5) \quad \begin{aligned} P_n(t) &= e^{-tA_n^*} P_0 e^{-tA_n} \\ &+ \int_0^t e^{-(t-s)A_n^*} (M(s) - P_n(s) A_n BR(s) B^* A_n^* P_n(s)) e^{-(t-s)A_n} ds. \end{aligned}$$

From (3.3) it follows that

$$(3.6) \quad \begin{aligned} P_n(t) &= U_n(t, 0) P_0 U_n^*(t, 0) \\ &+ \int_0^t U_n(t, s) (M(s) + P_n(s) A_n BR(s) B^* A_n^* P_n(s)) U_n^*(t, s) ds, \end{aligned}$$

and from (3.2) it follows that

$$(3.7) \quad P_n(t) = e^{-tA_n^*} P_0 U_n^*(t, 0) + \int_0^t e^{-(t-s)A_n^*} M(s) U_n^*(t, s) ds.$$

By multiplication on the right by $U_n^*(\tau, t)$ in (3.7) we obtain

$$(3.8) \quad \begin{aligned} P_n(t) U_n^*(\tau, t) &= e^{-tA_n^*} P_0 U_n^*(\tau, 0) \\ &+ \int_0^t e^{-(t-s)A_n^*} M(s) U_n^*(\tau, s) ds \quad \forall \tau \geq t. \end{aligned}$$

From (3.4) we have

$$(3.9) \quad U_n^*(\tau, t) = e^{-(\tau-t)A_n} + \int_t^\tau e^{-(s-t)A_n} A_n BR(s) B^* A_n^* P_n(s) U_n^*(\tau, s) ds.$$

Finally we set

$$(3.10) \quad \varphi_n(t, s) = B^* A_n^* P_n(s) U_n^*(t, s).$$

Hence, from (3.9) and (3.8) we have respectively

$$(3.11) \quad U_n^*(\tau, t) = e^{-(\tau-t)A_n} + \int_t^\tau A^{1-\alpha} n(A+n)^{-1} e^{-(s-t)A_n} A^\alpha BR(s) \varphi_n(\tau, s) ds,$$

$$(3.12) \quad \begin{aligned} \varphi_n(\tau, t) &= (A^\alpha B)^* (A^*)^{1-\alpha} n(A^*+n)^{-1} e^{-tA_n^*} P_0 U_n^*(\tau, 0) \\ &+ \int_0^t (A^\alpha B)^* (A^*)^{1-\alpha} n(A^*+n)^{-1} e^{-(t-s)A_n^*} M(s) U_n^*(\tau, s) ds. \end{aligned}$$

LEMMA 3.1. *Let φ_n be defined by (3.10). Then there exists a constant $c > 0$ such that*

$$\int_0^t \|\varphi_n(t, s)\|_{L(E, U)}^2 ds \leq c \quad \forall t \in [0, T], \quad \forall n \in N.$$

Proof. From (3.6) we have that $\forall x \in E$

$$\int_0^t \langle \varphi_n(t, s)x, R(s)\varphi_n(t, s)x \rangle_U ds \leq \langle P_n(t)x, x \rangle_E.$$

By (3.5) there exists a constant $c_1 > 0$ such that $\langle P_n(t)x, x \rangle_E \leq c_1 \forall t \in [0, T], \forall n \in N, \forall x \in E$. The thesis follows from (3.1). \square

LEMMA 3.2. *Suppose $\alpha < \frac{1}{2}$. Then for any $\varepsilon > 0$ sufficiently small there exists a constant $c_1(\varepsilon)$ such that*

$$\|U_n^*(\tau, \cdot)\|_{L^{2/(1-2\alpha)-\varepsilon}(0, \tau; L(E))} \leq c_1(\varepsilon) \quad \forall n \in N, \quad \forall \tau \in [0, T].$$

Proof. From (3.11) we have

$$(3.13) \quad \|U_n^*(\tau, t)\|_{L(E)} < c + c \int_t^\tau \frac{\|\varphi_n(\tau, s)\|_{L(E, U)}}{(t-s)^{1-\alpha}} ds$$

(we have used an inequality for $A^{1-\alpha}n(A+n)^{-1}e^{-tA_n}$ similar to (2.8), and we have denoted by c a suitable constant independent of $t \in [0, T]$ and $n \in N$). From Lemma 3.1 and from the Young inequality (see [5, p. 290], the thesis follows. \square

LEMMA 3.3. *Suppose $P_0 \in H(E) \cap L(E, D((A^*)^{1-\alpha}))$. Let $\beta \in (0, 1-\alpha)$ and $\varepsilon \in (0, \beta)$. Then, there exists a bounded linear operator in E which extends $(A^*)^{1-\alpha-\beta}P_0A^{\beta-\varepsilon}$.*

Proof. Set $Q_0 = (A^*)^{1-\alpha}P_0$, and $R_0 = Q_0 + Q_0^*$. Denote by $e^{-tA^{1-\alpha}}$ and $e^{-t(A^*)^{1-\alpha}}$, $t \geq 0$, the analytic semigroups generated by $A^{1-\alpha}$ and $(A^*)^{1-\alpha}$ respectively (see for instance [11]). Since $0 \in \rho(A^{1-\alpha})$ we can easily check that

$$P_0 = \int_0^{+\infty} e^{-t(A^*)^{1-\alpha}} R_0 e^{-tA^{1-\alpha}} dt.$$

Due to the analyticity of $e^{-tA^{1-\alpha}}$ and $e^{-t(A^*)^{1-\alpha}}$, the integral

$$\int_0^{+\infty} (A^*)^{1-\alpha-\beta} e^{-t(A^*)^{1-\alpha}} R_0 A^{\beta-\varepsilon} e^{-tA^{1-\alpha}} dt$$

is meaningful (it may be verified for example using [11, Thm. 3.3.3]), and clearly it coincides with the closure of $(A^*)^{1-\alpha-\beta}P_0A^{\beta-\varepsilon}$. \square

LEMMA 3.4. *Suppose $\alpha < \frac{1}{2}$. Then for any $\varepsilon > 0$ sufficiently small there exists a constant $c_2(\varepsilon)$ such that*

$$\|\varphi_n(\tau, \cdot)\|_{L^{2/(1-2\alpha)-\varepsilon}(0, \tau; L(E, U))} \leq c_2(\varepsilon) \quad \forall n \in N, \quad \forall \tau \in [0, T].$$

Proof. Take a small $\tilde{\varepsilon} > 0$. From (3.12) we have

$$\begin{aligned} \|\varphi_n(\tau, t)\|_{L(E, U)} &< c \|(A^*)^{1-\alpha-1/2+\tilde{\varepsilon}} n(A^*+n)^{-1} e^{-tA_n^*} (A^*)^{1/2-\tilde{\varepsilon}} \\ &\quad \cdot P_0 U_n^*(\tau, 0)\|_{L(E)} + c \int_0^t \frac{\|U_n(\tau, s)\|}{(t-s)^{1-\alpha}} ds, \end{aligned}$$

where $c > 0$ is a suitable constant. The second term is bounded in virtue of the Young inequality and Lemma 3.2; consider the first term. By Lemma 3.3 (with $\beta = \frac{1}{2} - \alpha - \tilde{\varepsilon}$)

there exists an operator $L \in L(E)$ which extends $(A^*)^{1/2-\tilde{\varepsilon}} P_0 A^{1-\alpha-1/2+\tilde{\varepsilon}/2}$. Using (3.11) we have

$$\begin{aligned} & \| (A^*)^{1-\alpha-1/2+\tilde{\varepsilon}} n(A^*+n)^{-1} e^{-tA_n^*} (A^*)^{1/2-\tilde{\varepsilon}} P_0 U_n^*(\tau, 0) \| \\ & \leq \frac{c_1}{t^{1-\alpha-1/2+\tilde{\varepsilon}}} \|L\|_{L(E)} \cdot \left\{ c_2 + \int_0^\tau \|A^{1/2-\tilde{\varepsilon}/2} n(A+n)^{-1} e^{-(s-t)A_n} A^\alpha BR(s) \varphi_n(\tau, s)\| ds \right\} \\ & \leq c_3 t^{-(1-\alpha-1/2+\tilde{\varepsilon})} \end{aligned}$$

by Lemma 3.1 and the Hölder inequality (we have denoted by c_1, c_2, c_3 suitable constants independent of t and n). Since $(1-\alpha-1/2+\tilde{\varepsilon})(2/(1-2\alpha)-\varepsilon) < 1$ for $\tilde{\varepsilon}$ sufficiently small, the claim is proved.

LEMMA 3.5. *There exists a constant $c > 0$ such that*

$$\|U_n^*(\tau, t)\|_{L(E)} \leq c \quad \forall n \in N, \quad \forall \tau, t \in [0, T], \quad \text{with } \tau \geq t.$$

Proof. If $\alpha > \frac{1}{4}$ the conclusion follows from (1.13), the Hölder inequality, and Lemma 3.4. If $\alpha < \frac{1}{4}$ we use Lemma 3.4 and applying the Young inequality (see [5, p. 290] in (3.13), we obtain that $\|U_n^*(\tau, \cdot)\|_{L^{2/(1-4\alpha)-\varepsilon}(0, \tau; L(E))} < c(\varepsilon)$. Then we may prove the claim of Lemma 3.4 with 4α in place of 2α . Iterating Lemmas 3.2 and 3.4, with the new exponents, for a suitable number of times, we prove the claim. \square

We apply now Lemma 3.5 and (2.8) to (3.7). Recalling that $Q_n(t) = (A^*)^{1-\alpha} n(A^*+n)^{-1} P_n(t)$, we have finally proved the “a priori” estimate (2.9). Hence, by Lemma (2.1), this proposition follows.

PROPOSITION 3.1. *Suppose that (2.1), (2.2), (2.3), (3.1) hold, and set $Q_0 = (A^*)^{1-\alpha} P_0$, $Q_{0,n} = (A^*)^{1-\alpha} n(A^*+n)^{-1} P_0$. Then there exists a unique solution $Q \in C_s(0, T; L(E))$ of (2.4), and $Q_n \rightarrow Q$ in $C_s(0, T; L(E))$. Moreover there exists a unique mild solution $P \in C_s(0, T; H^+(E))$ of (1.1), and $P_n \rightarrow P$ in $C_s(0, T; H(E))$.*

We conclude with a regularity result connected with the interpretation of (1.3). If $a, b \in \mathbb{R}$, $a < b$, and $\beta \in (0, 1)$, we set

$$C_s^\beta(a, b; L(E)) = \{U \in C_s(a, b; L(E)), \text{ such that } U(\cdot)x \text{ is } \beta\text{-Hölder}, \forall x \in E\}.$$

LEMMA 3.6. *Assume that the hypotheses of Proposition 3.1 hold. Then for any $\varepsilon \in (0, T)$ and for any $\beta \in (0, \alpha)$ we have $Q \in C_s^\beta(\varepsilon, T; L(E))$.*

Proof. Consider (2.4). Set

$$(3.14) \quad F(s) = M(s) - Q^*(s) A^\alpha BR(s) (A^\alpha B)^* Q(s), \quad s \in [0, T].$$

By Proposition 3.1 we have $F \in C_s(0, T; L(E))$ (note that $Q^* \in C_s(0, T; L(E))$ by the proof of Proposition 2.1). For any $t > \tau$ we have

$$\begin{aligned} & \left\| \int_0^t (A^*)^{1-\alpha} e^{-(t-s)A^*} F(s) e^{-(t-s)A} ds - \int_0^\tau (A^*)^{1-\alpha} e^{-(\tau-s)A^*} F(s) e^{-(\tau-s)A} ds \right\| \\ & \leq \int_\tau^t \| (A^*)^{1-\alpha} e^{-(t-s)A^*} F(s) e^{-(t-s)A} \| ds \\ & \quad + \int_0^\tau \{ \| (A^*)^{1-\alpha} (e^{-(t-s)A^*} - e^{-(\tau-s)A^*}) F(s) e^{-(t-s)A} \| \\ & \quad + \| (A^*)^{1-\alpha} e^{-(\tau-s)A^*} F(s) (e^{-(t-s)A} - e^{-(\tau-s)A}) \| \} ds \\ & \leq \int_\tau^t \frac{c}{(t-s)^{1-\alpha}} ds + \int_0^\tau \left\{ \int_{\tau-s}^{t-s} \frac{cd\sigma}{\sigma^{2-\alpha}} + \frac{c}{(\tau-s)^{1-\alpha}} \right\} ds \leq c_1(t-\tau)^\beta. \end{aligned}$$

We have used the equality

$$e^{-(t-s)A^*} - e^{-(\tau-s)A^*} = \int_{\tau-s}^{t-s} -A^* e^{-\sigma A^*} d\sigma,$$

and the estimate [11, 3.3.3]. Since $t \rightarrow e^{-tA^*} Q_0 e^{-tA}$ is analytic for $t > 0$ we conclude by way of (2.4). \square

Let $\mathcal{O} \in (0, 1)$. Consider the real interpolation spaces $D_A(\mathcal{O}, \infty)$ in E defined by [8]. We have

$$(3.15) \quad D_A(\mathcal{O}, \infty) = \left\{ x \in E, \frac{\|e^{-tA}x - x\|_E}{t^{\mathcal{O}}} \leq c \quad \forall t \in (0, T) \right\}.$$

LEMMA 3.7. *Suppose that (2.1), (2.2), (2.3), (3.1) hold. Suppose moreover that $P_0 \in L(E, D(A^*))$, and that there exists $\gamma \in (0, \alpha)$ such that $M \in C_s^\gamma(0, T; L(E))$, $R \in C_s^\gamma(0, T; L(U))$. Then $\forall \mathcal{O} \in (0, 1)$, $\forall x \in D_A(\mathcal{O}, \infty)$, $\forall t \in [0, T]$ we have $P(t)x \in D(A^*)$.*

Proof. From (2.4) it follows that

$$P(t)x = e^{-tA^*} P_0 e^{-tA} x + \int_0^t e^{-(t-s)A^*} F(s) e^{-(t-s)A} x ds,$$

where $F(s)$ is given by (3.14). The first term belongs to $D(A^*)$. Moreover,

$$\begin{aligned} & \int_0^t e^{-(t-s)A^*} F(s) e^{-(t-s)A} x ds \\ &= \int_0^t e^{-(t-s)A^*} F(s) (e^{-(t-s)A} x - x) ds + \int_0^t e^{-(t-s)A^*} F(s) x ds \\ & \quad + \int_0^{t/2} e^{-(t-s)A^*} (F(s) - F(t)) x ds + \int_{t/2}^t e^{-(t-s)A^*} (F(s) - F(t)) x ds. \end{aligned}$$

It is easy to verify that each term belongs to $D(A^*)$, using (3.15), and the equality

$$\int_0^t -A^* e^{-(t-s)A^*} F(s) x ds = (e^{-tA^*} - 1) F(t) x. \quad \square$$

PROPOSITION 3.2. *Assume that the hypotheses of Proposition 3.1 hold, and let P be the mild solution of (1.1). Then $\forall x, y \in D(A)$, $\forall t \in [0, T]$, we have*

$$\begin{aligned} & \frac{d}{dt} \langle P(t)x, y \rangle_E + \langle P(t)x, Ay \rangle_E + \langle Ax, P(t)y \rangle_E \\ & \quad + \langle R(t)(A^\alpha B)^*(A^*)^{1-\alpha} P(t)x, (A^\alpha B)^*(A^*)^{1-\alpha} P(t)y \rangle_U = \langle M(t)x, y \rangle_E, \\ & P(0) = P_0. \end{aligned}$$

Moreover assume that the hypotheses of Lemma 3.7 hold. Then P verifies (1.3). Moreover $\forall x, y \in D_A(\mathcal{O}, \infty)$, with $\mathcal{O} \in (0, 1)$, and $\forall t \in [0, T]$ we have

$$\begin{aligned} & \frac{d}{dt} \langle P(t)x, y \rangle_E + \langle A^* P(t)x, y \rangle_E + \langle x, A^* P(t)y \rangle_E \\ & \quad + \langle R(t)B^* A^* P(t)x, B^* A^* P(t)y \rangle_U = \langle M(t)x, y \rangle_E, \\ & P(0) = P_0. \end{aligned}$$

Proof. From (2.6) it follows that

$$\begin{aligned} & \langle M(t)x, y \rangle_E - \langle P_n(t)x, A_n y \rangle_E - \langle A_n x, P_n(t)y \rangle_E \\ & - \langle R(t)B^*A_n^*P_n(t)x, B^*A_n^*P_n(t)y \rangle_U = \frac{d}{dt} \langle P_n(t)x, y \rangle_E. \end{aligned}$$

By Proposition 3.1 the second term converges uniformly on $[0, T]$; hence we have the first part. The second part is a consequence of Lemma 3.7. \square

Using Proposition 3.1, we solve now the synthesis of the optimal control problem of minimizing

$$(3.16) \quad J(u) = \int_0^T \{ \langle M(T-t)y(t), y(t) \rangle + \langle R^{-1}(T-t)u(t), u(t) \rangle \} dt + \varphi(y(T))$$

over all controls $u \in L^2(0, T; U)$; where the state y is defined by the formula

$$(3.17) \quad y(t) = e^{-tA}y_0 + \int_0^t A^{1-\alpha} e^{-(t-s)A} A^\alpha B u(s) ds \quad \text{for a.e. } t \in [0, T],$$

with $y_0 \in E$, and where the final cost $\varphi(y(T))$ is defined by (3.18) below. Formula (3.17) defines a function $y \in L^2(0, T; E)$ which may be considered as a mild solution of some parabolic equations with boundary control (see for instance [6]). We consider now the case when $\alpha < \frac{1}{2}$, because it covers the largest class of boundary conditions (in the Dirichlet case we have $\alpha = \frac{1}{4} - \varepsilon$). When $\alpha < \frac{1}{2}$, y is not always continuous, and $y(T)$ is not always well defined. In order to define the final cost $\varphi(y(T))$ we write (3.17) in the form

$$y(t) = A^{1/2-\alpha/2} z(t),$$

where

$$z(t) = A^{\alpha/2-1/2} e^{-tA} y_0 + \int_0^t A^{1/2-\alpha/2} e^{-(t-s)A} A^\alpha B u(s) ds.$$

Now we note that if $u \in L^2(0, T; U)$ then $z \in C(0, T; E)$ and if P_0 satisfies (2.3), then by Lemma 3.3 there exists $\tilde{P}_0 \in L(E)$ which extends $(A^*)^{1/2-\alpha/2} P_0 A^{1/2-\alpha/2}$. Hence we may define

$$(3.18) \quad \varphi(y(T)) = \langle \tilde{P}_0 z(T), z(T) \rangle_E.$$

We note that if u is so regular that the corresponding trajectory y belongs to $C(0, T; E)$, then we have $\varphi(y(T)) = \langle P_0 y(T), y(T) \rangle_E$.

The synthesis is solved by the following proposition.

PROPOSITION 3.3. *Suppose that the hypotheses of Proposition 3.1 hold. Then there exists a unique optimal control u for the problem (3.16), (3.17), and we have*

$$(3.19) \quad u(t) = -R(T-t)(A^\alpha B)^* A^{*1-\alpha} P(T-t)y(t), \quad t \in [0, T],$$

where y is the unique solution of the equation

$$(3.20) \quad y(t) = e^{-tA} y_0 - \int_0^t A e^{-(t-s)A} B R(T-s)(A^\alpha B)^* A^{*1-\alpha} P(T-t)y(s) ds.$$

Moreover the optimal cost is $\langle P(T)y_0, y_0 \rangle_E$.

Proof. Let P_n be the solution of (2.6). Let $u \in L^2(0, T; U)$ and y_n be given by

$$(3.21) \quad y_n'(t) + A_n y_n(t) = A_n B u(t), \quad y_n(0) = y_0.$$

We set

$$f_n(t) = \langle P_n(T-t)y_n(t), y_n(t) \rangle_E - \int_t^T \{ \langle M(T-t)y_n(s), y_n(s) \rangle_E + \langle R^{-1}(T-s)u(s), u(s) \rangle_U \} ds.$$

Deriving f_n in t , and using (3.21) and (2.6), we find easily

$$f'_n(t) = \|R^{1/2}(T-t)B^*A_n^*P_n(T-t)y_n(t) + R^{-1/2}(T-t)u(t)\|^2.$$

Integrating on $[0, T]$, we obtain

$$(3.22) \quad \int_0^T \{ \langle M(T-s)y_n(s), y_n(s) \rangle_E + \langle R^{-1}(T-s)u(s), u(s) \rangle_U \} ds + \langle P_0y_n(T), y_n(T) \rangle_E \\ = \langle P_n(T)y_0, y_0 \rangle_E + \int_0^T \|R^{1/2}(T-s)B^*A_n^*P_n(T-s)y_n(s) + R^{-1/2}(T-s)u(s)\|_U^2 ds.$$

Now, by standard arguments it may be proved that $y_n \rightarrow y$ in $L^2(0, T; E)$ and that $\langle P_0y_n(T), y_n(T) \rangle_E \rightarrow \langle \tilde{P}_0z(T), z(T) \rangle_E$; hence, taking the limit in (3.22), we obtain

$$(3.23) \quad J(u) = \langle P(T)y_0, y_0 \rangle_E + \int_0^T \|R^{1/2}(T-s)(A^\alpha B)^*(A^*)^{1-\alpha}P(T-s)y(s) + R^{-1/2}(T-s)u(s)\|_U^2 ds,$$

from which we deduce that $J(u) \geq \langle P(T)y_0, y_0 \rangle_E$. Using the contraction principle for the local existence and an iteration technique for the global existence, it may be proved immediately that equation (3.20) has a unique solution y in $L^2(0, T; E)$ and also in $C(0, T; E)$. Then, taking u as in (3.19), we deduce by (3.23) that $J(u) = \langle P(T)y_0, y_0 \rangle_E$, and that (3.19) defines an optimal control. Conversely, if $v \in L^2(0, T; U)$ is an optimal control, then by (3.23) it is $v(t) = -R(T-t)(A^\alpha B)^*(A^*)^{1-\alpha}P(T-t)y(t)$ for a.e. $t \in [0, T]$, where y is the state which correspond to v . But then y is the solution in $L^2(0, T; E)$ of (2.20), which is unique; hence v is equal to u given by (3.19). \square

Remark. It may be proved that under the hypotheses of Lemma 3.7 we may write in (3.19) and (3.10) $B^*A^*P(T-t)y(t)$ in place of $(A^\alpha B)^*(A^*)^{1-\alpha}P(T-t)y(t)$.

REFERENCES

- [1] A. V. BALAKRISHNAN, *Boundary control of parabolic equations: L-Q-R theory*, in Theory of Nonlinear Operators, Proc. Fifth Intern. Summer School, Berlin, 1977.
- [2] R. CURTAIN AND A. PRITCHARD, *Infinite Dimensional Linear System Theory*, Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, 1978.
- [3] G. DA PRATO, *Quelques resultats d'existence, unicité et régularité pour un problème de la théorie du contrôle*, J. Math. Pures et Appl., 52 (1973), pp. 353-375.
- [4] F. FLANDOLI, *Riccati equation arising in a stochastic optimal control problem with boundary control*, to appear in Bollettino Unione Matematica Italiana, 1982.
- [5] G. H. HARDY, J. E. LITTLEWOOD AND G. POLYA, *Inequalities*, Cambridge University Press, Cambridge, 1934, p. 290.
- [6] I. LASIECKA, *Unified theory for abstract parabolic boundary problems—a semigroup approach*, Appl. Math. and Optim., 6 (1980), pp. 287-334.
- [7] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.
- [8] J. L. LIONS AND J. PEETRE, *Sur une classe d'espaces d'interpolation*, Publ. de l'I.S.H.E., Paris, 1964.
- [9] M. SORINE, *Une resultat d'existence et unicité pour l'équation de Riccati stationnaire*, Rapport INRIA no. 55, 1981.

- [10] M. SORINE, *Sur le semigroupe non linéaire associé à l'équation de Riccati*, Rapport du CRMA no. 1055, Université de Montréal, 1981.
- [11] H. TANABE, *Equation of Evolution*, Monographs and Studies in Mathematics, Pitman, London, 1979.
- [12] L. TARTAR, *Sur l'étude directe d'équations non linéaires intervenant en théorie du contrôle optimal*, J. Functional Analysis, 6 (1974), pp. 1–47.

CONTROL OF A DIFFUSION BY SWITCHING BETWEEN TWO DRIFT-DIFFUSION COEFFICIENT PAIRS*

J. M. McNAMARA†

Abstract. We consider the control of a particular one-dimensional diffusion process over a finite time interval $[0, T]$. Two drift-diffusion pairs (μ_1, σ_1) and (μ_2, σ_2) are given and the process is controlled by switching between these pairs. The objective is to maximize the probability that the process lies in the half line $[0, \infty)$ at final time T .

The case where $\mu_1 = \mu_2 = 0$ is considered first. Let the control σ_0 be given by the rule: choose the smaller diffusion coefficient if and only if the current state is nonnegative. A result is proved which, loosely speaking, says that σ_0 is optimal in this special case, and remains optimal even if we know the final value of the driving Brownian motion in advance.

The general problem (with drift) is then solved by an application of this result and the Girsanov transformation.

Key words. optimal stochastic control, risk

1. Introduction. We consider the control of a one-dimensional diffusion process over a finite time interval $[0, T]$. The process is controlled by simultaneously varying the drift and diffusion coefficients. Two pairs of real numbers (μ_1, σ_1) and (μ_2, σ_2) are given with $0 < \sigma_1 < \sigma_2$, and the idea is that at any time during $[0, T]$ we may either choose μ_1 as the drift coefficient and σ_1^2 as the diffusion coefficient or choose μ_2 as the drift and σ_2^2 as the diffusion coefficient. The objective is to maximize the expected value of a function of the state at final time T .

Formally, let (Ω, \mathcal{F}, P) be a probability space with filtration $\{\mathcal{F}_t\}_{0 \leq t \leq T}$. Let $\{B(t): 0 \leq t \leq T\}$ be a Brownian motion which is adapted to the filtration and whose increments after time t are independent of \mathcal{F}_t for each $t \in [0, T]$. We further assume $B(0) = 0$. An admissible control is defined to be a measurable process $\{\sigma(t): 0 \leq t \leq T\}$, which is adapted to the filtration $\{\mathcal{F}_t\}_{0 \leq t \leq T}$ and takes values in the two element set $\{\sigma_1, \sigma_2\}$. We denote the set of all admissible controls by \mathcal{A} . For each $x \in \mathbb{R}$ an admissible control σ determines a process $\{\xi_\sigma(t): 0 \leq t \leq T\}$ according to the equation

$$(1) \quad \xi_\sigma(t) = x + \int_0^t \mu(\sigma(\lambda)) d\lambda + \int_0^t \sigma(\lambda) dB(\lambda), \quad 0 \leq t \leq T,$$

where $\mu(\cdot)$ denotes the function which maps σ_i to μ_i , $i = 1, 2$.

The terminal reward function R is taken to be the step function

$$R(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0. \end{cases}$$

For $\sigma \in \mathcal{A}$ the corresponding expected future reward function is the map $f_\sigma: \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$f_\sigma(x) = E_x^\sigma \{R(\xi_\sigma(T))\},$$

where E_x^σ denotes the expectation given that the process $\{\xi_\sigma(t): 0 \leq t \leq T\}$ satisfies equation (1). In this paper we find a control in \mathcal{A} which maximizes $f_\sigma(x)$ for all $x \in \mathbb{R}$.

We outline two situations which provide motivation for this mathematical model. Consider a game in which a given fixed set of lotteries (plays) are available at each

* Received by the editors June 15, 1982, and in revised form November 18, 1982.

† School of Mathematics, University of Bristol, Bristol, England B28 1TW.

stage, and each play yields a number of points. A single player has to choose lotteries sequentially from this set until a nonrandom number n have been chosen. The current score is defined as the sum of the points gained on previous plays, and the player wins the game if and only if his score after n plays exceeds some threshold level e . If we define the state after k plays, X_k , as the current score minus e , then the objective of the player is to maximize $E\{R(X_n)\}$.

We have in mind here a game played by a single person against nature; but with slight modification we can consider a game played against an opponent who employs a fixed strategy. In this latter case one takes the state after k plays, X_k , to be the difference in scores. Thus a player maximizes the probability of having the larger score after n plays by maximizing $E\{R(X_n)\}$.

In such games the number of points gained in a play will usually be random with a variety of distributions being available. Many games involve a tradeoff between mean gain and risk: plays with a high mean gain often having a high variance (although, as we shall see, a high variance can be advantageous). Our model is intended to investigate this type of situation, but for simplicity (and tractability) restricts attention to the case where just two lotteries are available.

As a second example consider a small bird foraging for food in winter. During cold spells birds such as goldcrests forage during almost the whole daylight period. During the night considerable amounts of energy are needed to keep the bird warm and unless reserves at dusk are sufficiently high the bird will be dead by morning. Thus, as a rough approximation, a bird will maximize its probability of survival by maximizing the probability that energy reserves at dusk exceed some critical level e .

It is clear that food sources in the wild differ in their variability as well as their mean food yield, and several authors have been concerned with the effect of variability on foraging strategy [1], [2], [4], [7], [9]. Our problem can be seen to give a simple model of foraging in which a bird can choose between two such food sources and must maximize its chances of overnight survival.

In § 2 we consider the special case $\mu_1 = \mu_2 = 0$, and, loosely speaking consider optimal control conditional on the final state $B(T)$ of the driving Brownian motion. We exhibit a control which is optimal uniformly over all values of $B(T)$. Then in § 3 we reduce the general case to that considered in § 2 using the Girsanov transformation, and hence solve the general problem.

2. Uniform optimality in a special case. In this section we consider the case $\mu_1 = \mu_2 = 0$. For this special case the control problem outlined in the introduction can be shown to have the simple bang-bang optimal control given by the rule: if the current state is positive use σ_1 , if negative use σ_2 (McNamara [5]; the result is also outlined in McNamara [6]). We show that this control remains optimal “if we know the final value of the driving Brownian motion in advance”.

To see that this rule gives an admissible control in our sense let $a_0: \mathbb{R} \rightarrow \{\sigma_1, \sigma_2\}$ be defined by

$$(2) \quad a_0(x) = \begin{cases} \sigma_1 & \text{if } x \geq 0, \\ \sigma_2 & \text{if } x < 0, \end{cases}$$

and let $x \in \mathbb{R}$. Then by Nakao [8], the stochastic differential equation

$$\xi(0) = x, \quad d\xi(t) = a_0(\xi(t)) dB(t), \quad 0 \leq t \leq T$$

has a unique strong solution. Thus we can define the process $\sigma_0 \in \mathcal{A}$ by $\sigma_0(t) = a_0(\xi(t))$ for $0 \leq t \leq T$. McNamara [5] then shows that σ_0 maximizes $f_\sigma(x)$. Of course in our

terminology there is a different optimal control σ_0 for every initial point x . However, since they all arise from the same feedback control a_0 we will sometimes just loosely refer to σ_0 as the optimal control without reference to the state at time 0.

To make the idea of control conditional on $B(T)$ rigorous we consider the control of a two-dimensional process over $[0, T]$. Let $(x, y) \in \mathbb{R}^2$ and let $\sigma \in \mathcal{A}$. Then the vector process $\{(\xi_\sigma(t), \eta(t)): 0 \leq t \leq T\}$ is defined by

$$(3) \quad \xi_\sigma(t) = x + \int_0^t \sigma(\lambda) dB(\lambda)$$

and

$$(4) \quad \eta(t) = y + \int_0^t dB(\lambda).$$

Let $\phi: \mathbb{R} \rightarrow [0, \infty)$ satisfy an exponential growth condition:

$$\phi(x) \leq K e^{k|x|}, \quad x \in \mathbb{R},$$

for constants K and k . In this section we consider the problem of maximizing the product $R(\xi_\sigma(T))\phi(\eta(T))$: i.e. of maximizing

$$g_\sigma(x, y) = E_{x,y}^\sigma \{R(\xi_\sigma(T))\phi(\eta(T))\},$$

where $E_{x,y}^\sigma$ denotes the expectation given that the process $\{(\xi_\sigma(t), \eta(t)): 0 \leq t \leq T\}$ satisfies (3) and (4). We investigate optimality by using the familiar Bellman equations. Let $\psi: \mathbb{R}^2 \times [0, T] \rightarrow \mathbb{R}$, then we say that ψ satisfies the dynamic programming equation if the following five conditions are satisfied.

(D1) ψ is continuous on $\mathbb{R}^2 \times [0, T] - \{(0, y, T): y \in \mathbb{R}\}$.

(D2) ψ satisfies an exponential growth condition in x and y uniformly over $t \in [0, T]$; i.e. there exist constants B and b such that

$$|\psi(x, y, t)| \leq B e^{b(|x|+|y|)} \quad \forall (x, y, t) \in \mathbb{R}^2 \times [0, T].$$

(D3) The derivatives $\partial\psi/\partial t$, $\partial^2\psi/\partial x^2$, $\partial^2\psi/\partial x \partial y$ and $\partial^2\psi/\partial y^2$ exist and are continuous on $\mathbb{R}^2 \times (0, T)$.

(D4) $\partial\psi/\partial t + \frac{1}{2} \max_{a=\sigma_1, \sigma_2} G_a \psi = 0$ on $\mathbb{R}^2 \times (0, T)$, where G_a denotes the differential operator $a^2 \partial^2/\partial x^2 + 2a \partial^2/\partial x \partial y + \partial^2/\partial y^2$.

(D5) $\psi(x, y, T) = R(x)\phi(y) \quad \forall (x, y) \in \mathbb{R}^2 - \{(0, y): y \in \mathbb{R}\}$.

The following lemma gives a sufficient condition for σ_0 to be optimal.

VERIFICATION LEMMA. *Suppose there exists a solution, ψ , of the dynamic programming equation which satisfies*

$$(5) \quad (G_{\sigma_1} \psi)(x, y, t) \geq (G_{\sigma_2} \psi)(x, y, t), \quad x > 0$$

and

$$(6) \quad (G_{\sigma_1} \psi)(x, y, t) \leq (G_{\sigma_2} \psi)(x, y, t), \quad x < 0$$

for all $(y, t) \in \mathbb{R} \times (0, T)$. Then for each $(x, y) \in \mathbb{R}^2$

$$g_{\sigma_0}(x, y) = \max_{\sigma \in \mathcal{A}} g_\sigma(x, y) = \psi(x, y, 0)$$

where σ_0 denotes the control defined for initial state x as prescribed above.

The proof is of standard type and is hence omitted. For general results of this type see Fleming and Rishel [3]. Although these authors assume a continuous terminal reward function their results remain valid in our case since $P^\sigma(\xi_\sigma(T) = 0) = 0$ for all $(x, y) \in \mathbb{R}^2$ and all $\sigma \in \mathcal{A}$.

In the theorem given below we merely write down a solution of the dynamic programming equation which satisfies the condition of the lemma. However, before doing this we motivate the formula used by considering the case where $\phi = \delta$, the delta function. Our discussion will be heuristic and we make no attempt to give a rigorous justification of the remarks made in it.

Consider the (x, y) -plane shown in Fig. 1. When ϕ is the delta function the objective is to steer the process so that the state at time T is on the half line OX extended. Let P lie on the half line $\sigma_1 y = x$, $x \geq 0$ and let Q lie on $\sigma_2 y = x$, $x \leq 0$, as shown. We denote the open region above QOP by C_1 and the open region below QOP by C_2 . Whilst the diffusion coefficient has value σ_1^2 motion is restricted to a line of slope σ_1^{-1} , i.e., a line parallel to OP . Similarly while the diffusion coefficient has value σ_2^2 motion is restricted to a line parallel to OQ . By using a combination of σ_1^2 and σ_2^2 any point in the plane is accessible from any other point.

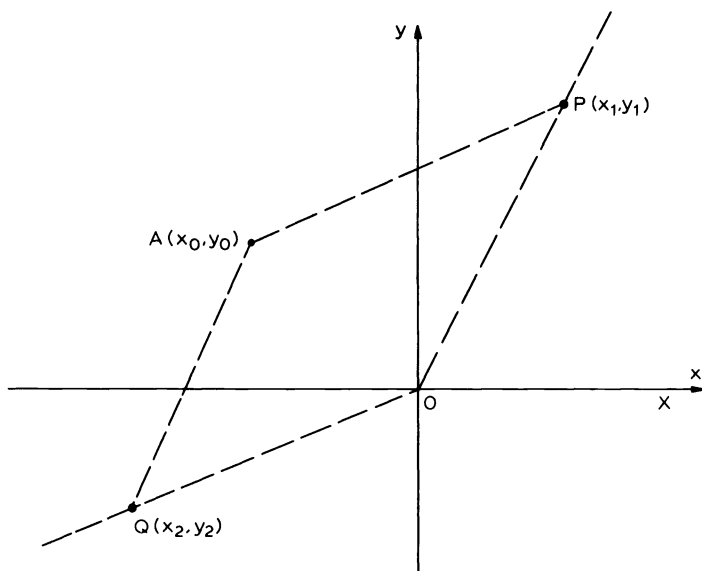


FIG. 1

First of all suppose that the process starts at point P at time 0. Since one cannot control the y -component of the process (i.e., η) the probability of being "at" 0 at final time cannot exceed the density of $\eta(T)$ at 0, namely $(2\pi T)^{-1/2} \exp\{-y_1^2/2T\}$. But by using the control $\sigma \equiv \sigma_1$ one can achieve this maximum. Thus starting at P it is optimal to use σ_1 for the remaining time. Similarly starting at Q it is optimal to use σ_2 .

If the initial state (x, y) is in C_2 it can be seen that one can either use σ_1 or σ_2 , or alternate between these alternatives, provided control σ_1 is used while on OP and control σ_2 is used while on OQ . In this way the process remains below or on QOP , and lies on OX at time T if and only if $\eta(T) = 0$.

Finally suppose that the process is at point A in C_1 at time 0. If control σ_2 is used until the process hits P , and then σ_1 is used from then on, the probability of being "at" 0 at time T is $(2\pi T)^{-1/2} \exp\{-v^2/2T\}$ where $v = (y_1 - y_0) + y_1 = (\sigma_2 - \sigma_1)^{-1}[(\sigma_1 + \sigma_2)y_0 - 2x_0]$. Similarly if σ_1 is used until the process hits Q and then σ_2 is used from then on the probability of being "at" 0 at time T is $(2\pi T)^{-1/2} \exp\{-v'^2/2T\}$

where $v' = (y_0 - y_2) + (-y_2)$. But since $OPAQ$ is a parallelogram, $v = v'$. This suggests that one can either use σ_1 or σ_2 or alternate between these while in C_1 provided the appropriate control is used if the process hits QOP .

Thus we have surmised that any control which used σ_1 on OP and σ_2 on OQ is optimal. Furthermore, since one can express this problem in terms of time to go rather than forward time one can immediately extend this result to processes which start at (x, y) at time $t \in [0, T)$ rather than time 0. For such processes the expected reward has a maximum of

$$(7) \quad g(x, y, t) = (2\pi(T-t))^{-1/2} \exp \left\{ -\frac{v^2(x, y)}{2(T-t)} \right\},$$

where

$$v(x, y) = \begin{cases} \frac{(\sigma_1 + \sigma_2)y - 2x}{\sigma_2 - \sigma_1}, & (x, y) \in C_1, \\ |y| & \text{otherwise.} \end{cases}$$

Note that g is continuous on $\mathbb{R}^2 \times [0, T)$ since v is continuous on \mathbb{R}^2 . The conjecture that either of σ_1 or σ_2 is optimal while not on QOP is strengthened by the fact that

$$(8) \quad \frac{\partial g}{\partial t} + \frac{1}{2} G_a g = 0$$

in $C_1 \cup C_2$ for both $a = \sigma_1$ and $a = \sigma_2$. Of course g is not differentiable on QOP .

Now consider the policy of using σ_1 for $x \geq 0$ and using σ_2 for $x < 0$. This is one of many optimal policies when $\phi = \delta$, but possesses the additional property that it is optimal for all ϕ of the form $\phi(x) = \delta(x - x_0)$, $x_0 \in \mathbb{R}$. Thus one expects this policy to be optimal for all nonnegative ϕ . This motivates the study of σ_0 and we now return to the rigorous exposition.

THEOREM. *Let ϕ be continuous and nonnegative. Then σ_0 is an optimal control.*

Proof. Let $\psi: \mathbb{R}^2 \times [0, T] \rightarrow \mathbb{R}$ be given by

$$(9) \quad \psi(x, y, t) = \int_{-\infty}^{\infty} g(x, y - v, t) \phi(v) dv, \quad (x, y, t) \in \mathbb{R}^2 \times [0, T)$$

and

$$(10) \quad \psi(x, y, T) = R(x) \phi(y),$$

where g is given by (7). We show that ψ satisfies the conditions of the verification lemma. Conditions (D1), (D2) and (D5) of the dynamic programming equation can be verified directly from (9) and (10) and the expression for g . By expressing ψ as

$$\psi(x, y, t) = \int_{-\infty}^{y-x/a_0(x)} g(x, y-v, t) \phi(v) dv + \int_{y-x/a_0(x)}^{\infty} g(x, y-v, t) \phi(v) dv,$$

where a_0 is given by (2), and differentiating the two terms on the right-hand side of this equation separately, one can verify that (D3) is satisfied and that

$$(11) \quad \frac{\partial \psi}{\partial t}(x, y, t) = \int_{-\infty}^{\infty} \frac{\partial g}{\partial t}(x, y-v, t) \phi(v) dv$$

and

$$(12) \quad \begin{aligned} (G_a \psi)(x, y, t) &= \int_{-\infty}^{\infty} (G_a g)(x, y - v, t) \phi(v) dv \\ &\quad - \frac{2x \operatorname{sgn} x}{(\sigma_2 - \sigma_1)(T - t)} f\left(x, \frac{x}{a_0(x)}, t\right) \phi\left(y - \frac{x}{a_0(x)}\right) - \left(\frac{a}{a_0(x)} - 1\right)^2, \\ &\quad a = \sigma_1, \sigma_2. \end{aligned}$$

Here $\partial g / \partial t$ and $G_a g$ are the appropriate derivatives of g when these exist, and are taken to be zero when they do not exist. By (8) $G_{\sigma_1} g = G_{\sigma_2} g$; thus by (12)

$$(13) \quad \max_{a=\sigma_1, \sigma_2} (G_a \psi)(x, y, t) = G_{a_0(x)} \psi(x, y, t)$$

where

$$(14) \quad (G_{a_0(x)} \psi)(x, y, t) = \int_{-\infty}^{\infty} (G_a g)(x, y - v, t) \phi(v) dv, \quad a = \sigma_1, \sigma_2.$$

By (13) and the definition of a_0 , (5) and (6) hold. Finally by (11), (13) and (14)

$$\left[\frac{\partial \psi}{\partial t} + \frac{1}{2} \max_{a=\sigma_1, \sigma_2} G_a \psi \right] \Big|_{(x, y, t)} = \int_{-\infty}^{\infty} \left[\frac{\partial g}{\partial t} + \frac{1}{2} G_a g \right] \Big|_{(x, y - v, t)} \phi(v) dv,$$

which equals 0 by (8); thus condition (D4) holds.

This completes the proof.

Remark. Let A be a Borel subset of \mathbb{R} with positive measure. Then since $P(\xi_\sigma(T) \geq 0 \text{ and } \eta(T) \in A) = P(\xi_\sigma(T) \geq 0 | \eta(T) \in A) P(\eta(T) \in A)$, and since $P(\eta(T) \in A)$ cannot be controlled, we see that σ maximizes $P(\xi_\sigma(T) > 0 \text{ and } \eta(T) \in A)$ if and only if it maximizes $P(\xi_\sigma(T) \geq 0 | \eta(T) \in A)$. This is the motivation behind the intuitive remark that σ_0 is optimal conditional on $B(T)$.

3. The general problem. We now return to the general problem defined in § 1.

Note first of all that if we relabel coordinates as (x', t') where $x' = x - \alpha(T - t)$ and $t' = t$, then a drift of μ_i and a diffusion coefficient of σ_i^2 becomes a drift of $\mu'_i = \mu_i + \alpha$ and diffusion coefficient of $\sigma_i'^2 = \sigma_i^2$ in the new coordinates. Thus setting

$$(15) \quad \alpha = \frac{\mu_2 \sigma_1 - \mu_1 \sigma_2}{\sigma_2 - \sigma_1}$$

we have $\mu'_i = \kappa \sigma'_i$, $i = 1, 2$, where

$$\kappa = \frac{\mu_2 - \mu_1}{\sigma_2 - \sigma_1}.$$

It is therefore sufficient to consider the case

$$\frac{\mu_1}{\sigma_1} = \frac{\mu_2}{\sigma_2} \quad (= \kappa).$$

We will use Girsanov's results to reduce this case to that considered in § 2.

Let \tilde{P} denote the measure defined on (Ω, \mathcal{F}) by

$$(16) \quad \tilde{P}(A) = \int_A \exp \{-\kappa B(T) - \tfrac{1}{2} \kappa^2 T\} dP, \quad A \in \mathcal{F}.$$

Then \tilde{P} is a probability measure since $E\{\exp\{-\kappa B(T) - \frac{1}{2}\kappa^2 T\}\} = 1$. Let $\{\tilde{B}(t): 0 \leq t \leq T\}$ be defined by

$$(17) \quad \tilde{B}(t) = B(t) + \kappa t.$$

Then by the Cameron–Martin–Girsanov results \tilde{B} is a Brownian motion on $(\Omega, \mathcal{F}, \tilde{P})$ which is adapted to the filtration $\{\mathcal{F}_t\}_{0 \leq t \leq T}$. Now let $\sigma \in \mathcal{A}$ and suppose the $\{\xi(t): 0 \leq t \leq T\}$ is defined by

$$\xi_\sigma(t) = x + \kappa \int_0^t \sigma(\lambda) d\lambda + \int_0^t \sigma(\lambda) dB(\lambda).$$

Then by the Girsanov transformation formula

$$(18) \quad \xi_\sigma(t) = x + \int_0^t \sigma(\lambda) d\tilde{B}(\lambda), \quad 0 \leq t \leq T.$$

By (16) and (17)

$$f_\sigma(x) = \tilde{E}_x^\sigma \{R(\xi_\sigma(T)) \exp\{\kappa \tilde{B}(T) - \kappa T + \frac{1}{2}\kappa^2 T\}\},$$

where \tilde{E}_x^σ denotes the \tilde{P} expectation given that $\{\xi_\sigma(t): 0 \leq t \leq T\}$ satisfies equation (18). Thus we can express $f_\sigma(x)$ as

$$f_\sigma(x) = \tilde{E}_x^\sigma \{R(\xi_\sigma(T)) \phi(\tilde{\eta}(T))\},$$

where ϕ is given by

$$\phi(y) = \exp\{\kappa y - \kappa T + \frac{1}{2}\kappa^2 T\}$$

and the process $\{\tilde{\eta}(t): 0 \leq t \leq T\}$ is given by

$$\tilde{\eta}(t) = 0 + \int_0^t d\tilde{B}(\lambda) \equiv \tilde{B}(t).$$

It is thus natural to compare our present control problem with the two-dimensional control problem considered in § 2, modified so that \tilde{B} rather than B is the driving Brownian motion. Since \mathcal{A} is defined in terms of the filtration, and \tilde{B} and B are both adapted to this filtration and their increments after time t are independent of \mathcal{F}_t for each $t \in [0, T]$, the two problems share the same set of admissible controls. Also the analysis given in § 2 is valid for any Brownian motion and is hence valid for \tilde{B} . Thus, if we let \tilde{g}_σ denote the analogue of g_σ we have

$$f_\sigma(x) = \tilde{g}_\sigma(x, 0), \quad \sigma \in A, \quad x \in \mathbb{R},$$

and

$$f_{\sigma_0}(x) = \tilde{g}_{\sigma_0}(x, 0) = \sup_{\sigma \in \mathcal{A}} \tilde{g}_\sigma(x, 0) = \sup_{\sigma \in \mathcal{A}} f_\sigma(x) \quad \forall x \in \mathbb{R}.$$

That is σ_0 is optimal when (μ_1, σ_1) and (μ_2, σ_2) satisfy $\mu_1/\sigma_1 = \mu_2/\sigma_2$. Thus for general pairs (μ_1, σ_1) and (μ_2, σ_2) the optimal control is given by the rule:

$$\text{use } (\mu_1, \sigma_1) \text{ if } x \geq \alpha(T-t),$$

$$\text{use } (\mu_2, \sigma_2) \text{ if } x < \alpha(T-t)$$

where α is given by (15).

REFERENCES

- [1] T. CARACO, *On foraging time allocation in a stochastic environment*, Ecology, 61 (1980), pp. 119–128.
- [2] T. CARACO, S. MARTINDALE AND T. S. WHITTAM, *An empirical demonstration of risk sensitive foraging preferences*, Anim. Behav., 28 (1980), pp. 820–830.
- [3] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, 1975.
- [4] A. I. HOUSTON AND J. M. McNAMARA, *A sequential approach to risk taking*, Anim. Behav., 30 (1982), pp. 1260–1261.
- [5] J. M. McNAMARA, *Sequential choice between high risk and safe alternatives*, Univ. Bristol School of Mathematics Rep. No. S-82-04, 1982.
- [6] ———, *Control of the diffusion coefficient of a simple diffusion process*, Math. Oper. Res. (1983), to appear.
- [7] J. M. McNAMARA AND A. I. HOUSTON, *Short-term behaviour and life-time fitness*, in *Functional Ontogeny*, D. J. McFarland, ed., Pitman, London, 1982.
- [8] S. NAKAO, *On pathwise uniqueness of solutions of one-dimensional stochastic differential equations*, Osaka J. Math., 9 (1972), pp. 513–518.
- [9] D. W. STEPHENS, *The logic of risk sensitive foraging preferences*, Anim. Behav., 29 (1981), pp. 628–629.

DISCRETE APPROXIMATION METHODS FOR PARAMETER IDENTIFICATION IN DELAY SYSTEMS*

I. GARY ROSEN†

Abstract. We construct approximation schemes for parameter identification problems in which the governing state equation is a linear functional differential equation of retarded type. The basis of the schemes is the replacement of the parameter identification problem having an infinite dimensional state equation by a sequence of approximating parameter identification problems in which the states are given by finite dimensional discrete difference equations. The difference equations are constructed using linear semigroup theory and rational function approximations to the exponential. Sufficient conditions are given for the convergence of solutions to the approximating problems, which can be obtained using conventional methods, to solutions to the original parameter identification problem. Finite difference and spline based schemes using Padé rational function approximations to the exponential are constructed, and shown to satisfy the sufficient conditions for convergence. A discussion and analysis of numerical results obtained through the application of the schemes to several examples is included.

Key words. delay systems, parameter identification, approximation schemes, finite difference approximation, spline approximation

1. Introduction. The purpose of this paper is the investigation of approximation methods for the identification of parameters in control systems where the state equation is a linear retarded functional differential equation (LRFDE). The parameters which we are interested in being able to estimate include system coefficients, initial conditions and the delays themselves. The methods which we shall discuss are based upon the discrete approximation framework for the integration of LRFDE initial value problems developed in [27] and [28]. The approach we take is to first replace the LRFDE which governs the dynamics of the system by an equivalent abstract evolution equation set in an infinite dimensional Hilbert space. The abstract evolution equation is then approximated by a finite dimensional discrete difference equation. This in turn leads to a totally discrete finite dimensional approximating parameter identification problem which can then be solved using standard techniques (see for instance [30]) and software packages which are readily available. That the solutions to these approximating problems in some sense approximate solutions to the original parameter identification problem is the primary result discussed in this paper.

As is pointed out in [7] there is very little literature about the parameter identification problem for delay systems (PIDDS). This is especially true for the case in which the delays are among the parameters to be identified. More recently, however, research in this area has been undertaken. Banks, Burns and Cliff [7] have extended the approximation framework, which had previously been developed to solve optimal control problems governed by delay differential systems [1], [4], [5], [9], [12], [17], [22], to the PIDDS as well. Their approach relies upon semi-discrete methods (i.e. those methods in which the LRFDE state equation is replaced by an approximating ordinary differential equation) for the solution of the delay differential equation which governs the dynamics of the system. The convergence arguments given by the authors

* Received by the editors December 10, 1981, and in revised form September 20, 1982. This research was supported in part by the Air Force Office of Scientific Research under contract AFOSR 76-3092D, in part by the National Science Foundation under grant NSF-MCS 7905774-02, and in part by the U.S. Army Research Office under contract ARO-DAAG29-79-C-0161. Additional research was carried out while the author was in residence at the Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA under NASA Contract Nos. NAS1-15810 and NAS1-16394.

† Department of Mathematics, Bowdoin College, Brunswick, Maine 04011.

rely heavily upon an abstract formulation of the problem which permits the use of linear semigroup theory and the associated approximation results which have been developed for application in such a setting. Their approximation framework is applicable to an extremely wide class of problems, includes methods having an arbitrarily high order of convergence and is capable of identifying the delays which appear in the state equation.

Banks in [2] develops spline based semi-discrete approximation schemes which are applicable to PIDDS in which the delays are not among the parameters to be estimated and in which the state equation is a nonlinear delay system satisfying global Lipschitz and differentiability conditions. While an equivalent abstract formulation of the problem is employed, in [2] Banks has avoided the use of semigroup theory entirely. Instead, the convergence of the approximation schemes is argued via the dissipativeness of the nonlinear operators involved and the Gronwall inequality. In [8] and [14] the ideas discussed in [2] are further extended so as to be applicable to problems in which the delays are also to be identified.

In [3] and [7] in addition to the construction of approximation schemes, a discussion of modeling problems, arising in physiology, enzyme kinetics and unsteady aerodynamics, which involve parameter identification and control for delay systems, can be found.

The results which we discuss are closely related to the ideas contained in [7]. We treat essentially the same class of problems, rely upon the same abstract formulation, apply many of the same functional analytic techniques to argue convergence and, in fact, incorporate the same state approximations as those discussed in [7]. The primary difference between the two approaches, however, is that our methods result in a complete discretization of the problem and hence require no further approximation when implemented. The methods included in our framework are capable of identifying delays and the integration schemes which they rely upon may be constructed with an arbitrarily high order of convergence. In [11] Banks and Rosen provide a detailed comparison of the performance of the semi-discrete schemes developed in [7] and totally discrete schemes similar to those which will be discussed here, when applied to parameter identification problems in which the delays themselves are not among the parameters to be estimated.

An alternative treatment of the problem of approximating solutions to the PIDDS, which is also based upon discrete approximation via difference equations, is given by Burns and Hirsch in [13]. These authors have taken a more straightforward approach by studying a specific scheme (as opposed to developing an approximation framework as is done in [7] and will be done here) which can be applied to PIDDS in which the LRFDE contains a single discrete delay term only. (The schemes developed here and those discussed in [7] are capable of handling equations which contain multiple discrete delay terms as well as a distributed delay term.) The approximating difference equation is derived via a modification of standard numerical integration schemes for ordinary differential equations (i.e. Euler's method, fourth order Runge-Kutta, etc.) so as to be applicable to delay differential systems. The authors are able to argue first order convergence for the Euler based scheme directly and hence can avoid the necessity for a functional analytic formulation of the problem. Computational evidence supporting the feasibility of extending these ideas to higher order schemes is also provided. However, the authors point out that the possibility of extending the relevant convergence arguments is uncertain. The Burns and Hirsch paper also addresses the difficulties which can arise in the construction of approximation schemes for PIDDS in which the delays are to be identified due to the fact that solutions to delay differential systems

may not be smooth with respect to the delays. This can pose problems since most standard optimization packages require differentiability with respect to the parameters.

Although it does not concern itself with the PIDDS directly, the work of Banks and Kunisch [10] should also be included in this historical outline. In this paper the authors treat parameter identification problems in which the governing state equations are semi-linear parabolic or hyperbolic partial differential equations. The approach that they take is similar to the one that is taken in [7]. Indeed, the infinite dimensional identification problem is replaced by an equivalent abstract formulation which is then used to develop finite dimensional semi-discrete approximation schemes. In a similar manner, the totally discrete schemes which will be developed below could easily be modified so as to be applicable to parameter identification problems with partial differential state equations.

We conclude this section with a brief outline of our presentation. In § 2 we state the PIDDS and show that it can be reformulated as an equivalent parameter identification problem in which the state equation is an abstract evolution equation set in an infinite dimensional Hilbert space. In § 3 we first discuss approximation results for abstract evolution equations and then use these results to construct the approximating parameter identification problems and to show that under the appropriate hypotheses, solutions to the approximating problems converge to solutions to the PIDDS. In § 4 we construct actual approximation schemes which satisfy the hypotheses and conditions necessary for convergence, while in § 5 we discuss and analyze numerical results obtained through the application of these schemes to several examples.

The notation we employ is, for the most part, standard. The symbol $\mathcal{L}_{n \times n}$ is used to denote the space of n square matrices. We denote the space of functions defined on (a, b) with range in R^n and p continuous derivatives by $C_p^n(a, b)$. The space of piecewise continuous functions and the space of continuous functions on (a, b) with range in R^n are denoted by $PC^n(a, b)$ and $C^n(a, b)$ respectively. The Lebesgue spaces of R^n -valued functions on (a, b) are denoted by $L_p^n(a, b)$ while the Sobolev spaces of functions ϕ with $\phi^{(m-1)}$ absolutely continuous and $\phi^{(m)}$ in $L_p^n(a, b)$ are denoted by $W_{m,p}^n(a, b)$. For a function $\phi \in W_{1,2}^n(a, b)$ we shall use the notations $D\phi$ and $\dot{\phi}$ interchangeably to denote the derivative of ϕ . Finally for a linear operator \mathcal{T} , the symbols $D(\mathcal{T})$ and $R(\mathcal{T})$ are used to denote the domain of \mathcal{T} and the range of \mathcal{T} respectively.

2. The PIDDS and its abstract formulation. In this section we formulate the parameter identification problem for delay systems and show that it has an equivalent formulation, whereby the dynamics of the governing control system in the form of an LRFDE are replaced by an abstract evolution equation set in an infinite dimensional Hilbert space. Since the PIDDS and the associated approximation schemes which we shall develop here are closely related to the problem and schemes discussed by Banks, Burns and Cliff [7] the reader is instructed to note the similarities which exist between the material and notation to follow in this section and that which is contained in [7, §§ 2, 2.1] and [28, § 2].

We begin with the definition of the admissible initial data/parameter set and a formal statement of the PIDDS. Let $r > 0$ and Ω a compact convex subset of R^μ be given. Define the compact convex set $Q \subset R^{\mu+\nu}$ by $Q \equiv \Omega \times \mathcal{H}$ where

$$\mathcal{H} \equiv \{h = (r_1, r_2, \dots, r_\nu) \in R^\nu \mid 0 \leq r_i \leq r_{i+1} \leq r, i = 1, 2, \dots, \nu - 1\}.$$

In addition let \mathcal{S} be a compact convex subset of $R^n \times L_2^n(-r, 0)$ and define

$$\Gamma = \mathcal{S} \times Q = \mathcal{S} \times \Omega \times \mathcal{H}$$

to be the admissible initial data/parameter set. We further assume that we have been provided with an input/output pair $(u, \zeta) \in PC^m(0, T) \times C^l(0, T)$ for some $T > 0$. We refer to (u, ζ) as an input/output pair since it is assumed that if given input $u \in PC^m(0, T)$ the physical system to be identified produces output $\zeta \in C^l(0, T)$.

With the above definitions in hand, we can state the PIDDS:

PIDDS. Given an input/output pair $(u, \zeta) \in PC^m(0, T) \times C^l(0, T)$ for some $T > 0$, find $\gamma^* = (\eta^*, \phi^*, q^*) = (\eta^*, \phi^*, \alpha^*, h^*) \in \Gamma$ which minimizes

$$(2.1) \quad J(\gamma) = |y(0; \gamma, u) - \zeta(0)|_{w_1}^2 + |y(T; \gamma, u) - \zeta(T)|_{w_2}^2 + \int_0^T |y(t; \gamma, u) - \zeta(t)|_{w_3}^2 dt,$$

subject to

$$(2.2) \quad \dot{x}(t) = L(q)x_t + B(\alpha)u(t), \quad t \in [0, T],$$

$$(2.3) \quad x(0) = \eta, \quad x_0 = \phi,$$

$$(2.4) \quad y(t) = C(\alpha)x(t) + D(\alpha)u(t),$$

where for each $\alpha \in \Omega$, $B(\alpha)$, $C(\alpha)$ and $D(\alpha)$ are $n \times m$, $l \times n$ and $l \times m$ matrices respectively, $|\cdot|_{w_j}$, $j = 1, 2, 3$ represent appropriately weighted (application dependent) norms on R^l , x_t denotes the function $\theta \rightarrow x(t + \theta)$, $-r \leq \theta \leq 0$ and the notation $y(\cdot; \gamma, u)$ is employed in order to exhibit the explicit dependence of the output y of the theoretical system on the initial conditions and parameter values γ and the given input u . For each $q = (\alpha, h) = (\alpha, r_1, r_2, \dots, r_\nu) \in Q$ the operator $L(q): L_2^n(-r, 0) \rightarrow R^n$ is assumed to be of the form

$$L(q)\phi = \sum_{i=0}^{\nu} A_i(\alpha)\phi(-r_i) + \int_{-r_\nu}^0 K(\alpha, \theta)\phi(\theta) d\theta$$

with $r_0 \equiv 0$ and where for each $\alpha \in \Omega$, $A_i(\alpha)$, $i = 0, 1, 2, \dots, \nu$ are $n \times n$ matrices and $\theta \rightarrow K(\alpha, \theta)$ is an $n \times n$ matrix-valued function in $L_2((-r, 0), \mathcal{L}_{n \times n})$. It is assumed that $A_i(\alpha)$, $i = 0, 1, 2, \dots, \nu$, $B(\alpha)$, $C(\alpha)$, $D(\alpha)$, $K(\alpha, \cdot)$ are continuous in α .

Before we go on to discuss the parameter identification problem, let us take a moment to consider the LRFDE initial value problem given by (2.2)–(2.3). Given $\gamma = (\eta, \phi, q) \in \Gamma$, a solution to the initial value problem is a function $x: [-r, T] \rightarrow R^n$ such that $x \in W_{1,2}^n(0, T)$, x satisfies (2.2) almost everywhere on $[0, T]$, $x(0) = \eta$ and $x_0 = \phi$. Standard arguments [24] can be used to demonstrate that for each $\gamma \in \Gamma$ (2.2)–(2.3) has a unique solution which depends continuously upon $\gamma \in \Gamma$ and the nonhomogeneous term u (as an element of $L_2^m(0, T)$). The notation $x(t; \gamma, u)$ (and $x_t(\gamma, u)$) will be used to denote this unique solution of (2.2)–(2.3) (and its past history on $[t-r, t]$) corresponding to a particular choice of $\gamma \in \Gamma$ and $u \in L_2^m(0, T)$.

Remark. One might be tempted to question the validity of choosing a least squares payoff functional of the form given in (2.1) for the PIDDS since in actual practice it is usually the case that for a given input u , output can only be measured at discrete times $0 \leq t_0 < t_1 < \dots < t_m \leq T$. In this instance a more appropriate choice for a payoff functional would be the one used in [7] which is given by

$$J(\gamma) = \frac{1}{2} \sum_{j=0}^m |y(t_j; \gamma, u) - \zeta_j|^2,$$

where the $\{\zeta_j\}_{j=0}^m$ are the given discrete output observations obtained from the actual system which is to be identified. Oddly enough, it is the discrete nature of the approximation schemes to be discussed which necessitates the use of the distributed

payoff functional given by (2.1). However, this restriction can be circumvented via the use of an interpolation scheme applied either to the observational data provided in order to generate a continuous observation $\hat{\zeta}(\cdot) \in C^1(0, T)$ or to the discrete output generated by the difference-equation-based approximation schemes. The latter approach is the one which is employed in [13] in order to overcome this very same problem.

We next show that the PIDDS has an equivalent formulation as a parameter identification problem in which the governing state equation is given by an abstract evolution equation set in the Hilbert space Z

$$Z \equiv R^n \times L_2^n(-r, 0)$$

with inner product

$$\langle \cdot, \cdot \rangle_Z = \langle \cdot, \cdot \rangle_{R^n} + \langle \cdot, \cdot \rangle_{L_2}.$$

The quantity r which appears in the definition of the space Z is as previously defined. For $q = (\alpha, h) \in Q$ and $(\eta, \phi) \in Z$ we define the parameterized family of operators $S(t; q): Z \rightarrow Z$ for $t \geq 0$ by

$$S(t; q)(\eta, \phi) = (x(t; (\eta, \phi, q), 0), x_t((\eta, \phi, q), 0)),$$

where $x(\cdot, (\eta, \phi, q), 0)$ denotes the unique solution of (2.2), (2.3) corresponding to $q \in Q$, $(\eta, \phi) \in Z$ and $u \equiv 0$. In light of the existence, uniqueness and continuous dependence results for solutions to the initial value problem (2.2)–(2.3) discussed earlier, it is not difficult to show that for each $q \in Q$ the operators $\{S(t; q): t \geq 0\}$ form a \mathcal{C}_0 semigroup of bounded linear operators on Z . Furthermore, for each $q \in Q$ the infinitesimal generator $\mathcal{A}(q): D(\mathcal{A}(q)) \subset Z \rightarrow Z$ of the semigroup and its domain of definition (which is independent of q) can be calculated. They are given by

$$D(\mathcal{A}(q)) = D = \{(\eta, \phi) \in Z \mid \phi \in W_{1,2}^n(-r, 0), \eta = \phi(0)\},$$

$$\mathcal{A}(q)(\phi(0), \phi) = (L(q)\phi, D\phi).$$

Using the fact that $\mathcal{A}(q)$ satisfies a dissipative inequality with respect to a weighted inner product on Z (see [29]), it can be shown that

$$(2.5) \quad |S(t; q)| \leq \sqrt{\nu} e^{\omega(q)t},$$

where

$$(2.6) \quad \omega(q) = \frac{\nu+1}{2} + |A_0(\alpha)| + \frac{1}{2} \sum_{i=1}^{\nu} |A_i(\alpha)|^2 + \frac{1}{2} \int_{-r}^0 |K(\alpha, \theta)|^2 d\theta.$$

Turning our attention to the nonhomogeneous equation, for each $\alpha \in \Omega$ we define the operator $\hat{B}(\alpha): R^m \rightarrow Z$ by $\hat{B}(\alpha)u = (B(\alpha)u, 0)$ and consider

$$(2.7) \quad z(t; \gamma, u) = S(t; q)(\eta, \phi) + \int_0^t S(t-\sigma; q)\hat{B}(\alpha)u(\sigma) d\sigma, \quad 0 \leq t \leq T$$

for each $\gamma = (\eta, \phi, q) = (\eta, \phi, \alpha, h) \in \Gamma$ and $u \in L_2^m(0, T)$. Using standard results from linear semigroup theory [20], it is easily verified that the expression for z given in (2.7) is well defined and continuous in t . Furthermore, under the somewhat more restrictive conditions that $(\eta, \phi) \in D$ and $u \in C_1^m(0, T)$ we have that (2.7) is the unique strong solution to the initial value problem in Z given by

$$(2.8) \quad \dot{z}(t) = \mathcal{A}(q)z(t) + \hat{B}(\alpha)u(t),$$

$$(2.9) \quad z(0) = (\eta, \phi).$$

It can be shown (see [4], [5]) that under the same conditions

$$w(t) \equiv (x(t; \gamma, u), x_t(\gamma, u))$$

is a strong solution to the initial value problem given by (2.8)–(2.9), as well. It therefore must follow that $z(t)$ and $w(t)$ coincide for $0 \leq t \leq T$. By making use of standard density and continuous dependence arguments the equivalence of z and w can be extended so as to hold for all $(\eta, \phi) \in Z$ and $u \in L_2^m(0, T)$. We state this conclusion in the form of a theorem.

THEOREM 2.1. *Let $x(\cdot; \gamma, u)$ denote the unique solution of the LRFDE initial value problem (2.2)–(2.3) corresponding to $\gamma \in \Gamma$ and $u \in L_2^m(0, T)$. Then for $0 \leq t \leq T$ we have*

$$z(t; \gamma, u) = (x(t; \gamma, u), x_t(\gamma, u)).$$

In light of Theorem 2.1 above, the equivalence which exists between solutions of the LRFDE initial value problem (2.2)–(2.3) and the Z valued function given by expression (2.7) permits the reformulation of the PIDDS as an equivalent (1–1 correspondence between solutions) parameter identification problem in which the governing state equation is now given by (2.7). Indeed, if we define for each $\alpha \in \Omega$ the operator $\hat{C}(\alpha): Z \rightarrow R^l$ by $\hat{C}(\alpha)(\eta, \phi) = C(\alpha)\eta$ then the PIDDS is equivalent to the following abstract parameter identification problem.

(APIDDS): Given an input/output pair $(u, \xi) \in PC^m(0, T) \times C^l(0, T)$ for some $T > 0$, find $\gamma^* = (\eta^*, \phi^*, q^*) = (\eta^*, \phi^*, \alpha^*, h^*) \in \Gamma$ which minimizes $J(\gamma)$ given by (2.1) subject to

$$(2.10) \quad z(t) = S(t; q)(\eta, \phi) + \int_0^t S(t-\sigma; q) \hat{B}(\alpha) u(\sigma) d\sigma,$$

$$(2.11) \quad y(t) = \hat{C}(\alpha)z(t) + D(\alpha)u(t), \quad t \in [0, T].$$

The fact that there exists a 1–1 correspondence between solutions to the APIDDS above and solutions to the PIDDS forms the basis for the approximation schemes which we construct in the succeeding sections.

3. Approximation results for the PIDDS. Fundamental to our approximation schemes for the PIDDS is the construction of convergent finite dimensional discrete approximation schemes for the state equation given by (2.10). Indeed, the approach we take is to replace the APIDDS by a parameter identification problem in which the governing state equation is now a finite dimensional linear nonhomogeneous difference equation which depends upon $q \in Q$. For each $q \in Q$, the solutions to these difference equations will, in some sense, approximate the solutions of (2.10). If we let N represent the degree of approximation, and if the N th approximating problem is solved for γ_N^* , an element in the admissible initial data/parameter set Γ which minimizes a discrete least squares payoff functional which approximates (2.1), then we shall see that the compactness of Γ and the convergence of the state approximation are sufficient to guarantee the existence of a subsequence $\{\gamma_{N_k}^*\}$ of $\{\gamma_N^*\}$ and a $\gamma^* \in \Gamma$ such that $\gamma_{N_k}^* \rightarrow \gamma^*$ as $k \rightarrow \infty$ with γ^* a solution to the APIDDS. In light of the equivalence established in § 2, it necessarily follows that γ^* is a solution to the PIDDS as well.

The above ideas will be made precise once we have outlined the abstract approximation results which are fundamental to the construction of convergent approximations to the state equation. These approximation results can be considered to be discrete analogues of the well-known Trotter–Kato theorem [20], [23] which is frequently

employed to establish the convergence of semidiscrete approximations to semigroups of operators. While Theorem 3.1 below is quite similar to the result given in [28, Thm. 4.9], a direct application of [28] cannot establish the convergence of the state approximations for the PIDDS. This is a consequence of the fact that since the delays can appear among the unknown parameters, the largest delay, r_v , which plays a crucial role in our formulation, may no longer remain fixed with respect to the degree of approximation. With minor modifications the proof of [28, Thm. 4.9] can be employed to verify Theorem 3.1. Therefore, we simply state the result and omit its proof.

In what is to follow we shall employ the following conventions. For a rational function of a complex variable $r(z) = p(z)/q(z)$ we denote the degree of $r(z) = \text{degree of } p(z) - \text{degree of } q(z)$ by $\deg r(z)$. For $\mathcal{T}: D(\mathcal{T}) \subset H \rightarrow H$ a linear transformation on a Hilbert space H we say $\mathcal{T} \in G(M, \beta)$ if \mathcal{T} is the infinitesimal generator of a \mathcal{C}_0 semigroup of operators $\{T(t): t \geq 0\}$ satisfying $|T(t)| \leq M e^{\beta t}$. Furthermore, if $\rho(\mathcal{T})$ denotes the resolvent set of \mathcal{T} , we denote the resolvent of \mathcal{T} , $(\lambda I - \mathcal{T})^{-1}$ by $R_\lambda(\mathcal{T})$, $\lambda \in \rho(\mathcal{T})$.

We formulate our approximation framework in the same general setting used in [7]. Let $(Z, \langle \cdot, \cdot \rangle)$, $(Z_N, \langle \cdot, \cdot \rangle_N)$, $N = 1, 2, \dots$ be Hilbert spaces with norms $|\cdot|$ and $|\cdot|_N$ respectively. For each $N = 1, 2, \dots$ let X_N be a closed subspace of Z_N and let $\Pi_N: Z_N \rightarrow X_N$ be the orthogonal projection of Z_N onto X_N with respect to the $\langle \cdot, \cdot \rangle_N$ inner product. Let $\mathcal{J}_N: Z \rightarrow Z_N$ be a mapping which is onto Z_N and which satisfies $|\mathcal{J}_N z|_N \leq |z|$ for each $z \in Z$. Finally, we define $P_N: Z \rightarrow X_N$ by $P_N = \Pi_N \mathcal{J}_N$, and note that $|P_N z|_N \leq |z|$ for each $z \in Z$.

THEOREM 3.1. *Let Z , Z_N , X_N and P_N be as defined above and let T be a fixed positive real number. Suppose for some M, β we have that $\mathcal{A} \in G(M, \beta)$ is the infinitesimal generator of the \mathcal{C}_0 semigroup of operators on Z , $\{S(t): t \geq 0\}$ and $\mathcal{A}_N \in G(M, \beta)$ is the infinitesimal generator of the \mathcal{C}_0 semigroup of operators on X_N , $\{S_N(t): t \geq 0\}$, $N = 1, 2, \dots$. Suppose further that:*

(1) *There exists $\mathcal{D} \subset D(\mathcal{A})$, a dense subset of Z such that $R_\lambda(\mathcal{A})\mathcal{D} \subset \mathcal{D}$ for each $\lambda \in C$ with $\text{Re } \lambda > \beta$ and for each $z \in \mathcal{D}$ we have*

$$|\mathcal{A}_N P_N z - P_N \mathcal{A} z|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

(2) *$c(z)$ is a rational function of the complex variable $z \in C$ such that*

(a) $|c(z) - e^z| = O(|z|^{m+1})$ as $z \rightarrow 0$ with $m > 0$;

(b) $\deg c(z) \leq m + 1$;

(c) $c(z)$ has no poles in $\{z \in C: \text{Re } z \leq 0\}$.

(3) $\{r_v^N\}_{N=1}^\infty$ *is a sequence of positive real numbers satisfying* $0 < r_v^N \leq r \leq T$, $N = 1, 2, \dots$.

(4) $\{\rho^N\}_{N=1}^\infty$ *is a sequence of positive integers determined by the following relation:*

$$\frac{\rho^N r_v^N}{N} \leq T < (\rho^N + 1) \frac{r_v^N}{N}, \quad N = 1, 2, \dots$$

Then there exists an \bar{N} such that the operators on X_N given by $c((r_v^N/N)\mathcal{A}_N)$ exist for all $N > \bar{N}$ and moreover if the infinite collection of operators

$$\left\{ c\left(\frac{r_v^N}{N}\mathcal{A}_N\right)^k \right\}_{k=0}^{\rho^N}, \quad N > \bar{N}$$

are uniformly bounded, then given $\varepsilon > 0$, there exists an $\hat{N} > \bar{N}$ such that

$$\left| c\left(\frac{r_v^N}{N}\mathcal{A}_N\right)^k P_N z - P_N S\left(\frac{kr_v^N}{N}\right) z \right|_N < \varepsilon,$$

$k = 0, 1, 2, \dots, \rho^N$ for all $N > \hat{N}$ and each $z \in Z$. (Equivalently stated:

$$\left| c\left(\frac{r_\nu^N}{N}\mathcal{A}_N\right)^k P_{Nz} - P_N S\left(\frac{kr_\nu^N}{N}\right)z \right|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

uniformly in $k, k \in \{0, 1, 2, \dots, \rho^N\}$.)

Remark 3.1. Although in Theorem 3.1 the spaces and the transformations between them are unspecified, when actually realized for the purpose of developing approximation schemes for the PIDDS, the constructs appearing in the theorem take the following form. The space Z is of course $R^n \times L_2^n(-r, 0)$, $Z_N = R^n \times L_2^n(-r_\nu^N, 0)$, X_N is a finite dimensional subspace of Z_N such as the AVE or spline subspaces discussed in [28], \mathcal{J}_N is the operator that takes $(\eta, \phi) \in Z$ into $\tilde{z} = (\eta, \tilde{\phi})$ in Z_N where $\tilde{\phi}$ is the restriction of ϕ to $[-r_\nu^N, 0]$ and $c(z)$ might for example be chosen from among the Padé rational function approximations to the exponential (see [28], [31]). Once a basis for X_N has been chosen, \mathcal{A}_N can be represented by a matrix as can the operators $c((r_\nu^N/N)\mathcal{A}_N)^k$, $k = 0, 1, 2, \dots, \rho^N$. If the \mathcal{A}_N are constructed and $c(z)$ is chosen so as to comply with the hypotheses and conditions of Theorem 3.1, for $z_0 \in Z$ and $t_k^N = kr_\nu^N/N \in [0, T]$, $k = 0, 1, 2, \dots, \rho^N$ we have that $z(t_k^N) = S(t_k^N)z_0$ is approximated by $z_k^N = c((r_\nu^N/N)\mathcal{A}_N)^k P_{Nz_0}$. The construction of X_N and \mathcal{A}_N and the selection of $c(z)$ so as to lead to convergent approximation schemes is examined in detail in § 4.

Remark 3.2. Implicit in condition (1) in Theorem 3.1 is the assumption that $P_N \mathcal{D} \subset D(\mathcal{A}_N)$, $N = 1, 2, \dots$. However, in practice X_N is chosen to be finite dimensional in which case $\mathcal{A}_N: X_N \rightarrow X_N$ is a bounded operator with $D(\mathcal{A}_N) = X_N$.

Remark 3.3. In what is to follow we shall frequently refer to $\mathcal{A}_N \in G(M, \beta)$ as the spatial stability condition and the uniform boundedness of the operators $\{c((r_\nu^N/N)\mathcal{A}_N)^k\}_{k=0}^{\rho^N}$ as the temporal stability condition.

It is not surprising that an estimate of the rate of convergence in Theorem 3.1 would depend upon both the degree to which the \mathcal{A}_N approximate \mathcal{A} and the degree to which $c(z)$ approximates e^z . An application of [7, Thm. 3.2] and arguments similar to those used to verify [28, Thm. 4.17] can be used to establish that for all $z \in \mathcal{B}$, \mathcal{B} a subset of Z with sufficiently well behaved elements, for which there exists a $K = K(z)$ such that

$$|\mathcal{A}_N P_{Nz} - P_N \mathcal{A} z|_N \leq \frac{K(z)}{N^p},$$

there exist constants $K_1 = K_1(z)$ and $K_2 = K_2(z)$ such that

$$\left| c\left(\frac{r_\nu^N}{N}\mathcal{A}_N\right)^k P_{Nz} - P_N S\left(\frac{kr_\nu^N}{N}\right)z \right|_N \leq K_1 \left(\frac{r}{N}\right)^p + K_2 \left(\frac{r}{N}\right)^m,$$

$k = 0, 1, 2, \dots, \rho^N$, where m is as in Theorem 3.1 condition (2).

For $Z = R^n \times L_2^n(-r, 0)$ let $\pi^0: Z \rightarrow R^n$, $\pi^1: Z \rightarrow L_2^n(-r, 0)$ be the canonical coordinate projections of Z given by $\pi^0(\eta, \phi) = \eta$ and $\pi^1(\eta, \phi) = \phi$, respectively. For $\{q_N\}$ a sequence of elements in Q with

$$q_N = (\alpha_N, h_N) = (\alpha_N, (r_1^N, r_2^N, \dots, r_\nu^N))$$

let $Z_N = Z_N(q_N) = R^n \times L_2^n(-r_\nu^N, 0)$, let $X_N = X_N(q_N)$ be a closed subspace of Z_N , let $\Pi_N = \Pi_N(q_N)$ be the orthogonal projection of Z_N onto X_N with respect to the Z_N inner product $\langle \cdot, \cdot \rangle_N$ and let $\mathcal{J}_N = \mathcal{J}_N(q_N): Z \rightarrow Z_N$ be the mapping which takes $(\eta, \phi) \in Z$ into $(\eta, \tilde{\phi}) \in Z_N$ where $\tilde{\phi}$ denotes the restriction of ϕ to $[-r_\nu^N, 0]$. Define $P_N = P_N(q_N): Z \rightarrow X_N$ by $P_N(q_N) = \Pi_N(q_N)\mathcal{J}_N(q_N)$ and let $\mathcal{A}_N(q_N)$ be a linear transformation

defined on X_N with range contained in X_N . Finally let $c(z)$ and $d(z)$ be rational functions of the complex variable z and let θ be a fixed positive scalar with $0 \leq \theta \leq 1$. With these definitions in hand, the approximating parameter identification problems can be stated as follows.

NPIDDS. Given an input/output pair $(u, \zeta) \in PC^m(0, T) \times C^l(0, T)$ for some $T > 0$ find $\gamma_N^* = (\eta_N^*, \phi_N^*, q_N^*) = (\eta_N^*, \phi_N^*, \alpha_N^*, h_N^*) \in \Gamma$ which minimizes

$$J_N(\gamma) = |y_0^N(\gamma; u) - \zeta(0)|_{w_1}^2 + |y_{\rho^N}^N(\gamma; u) - \zeta(T)|_{w_2}^2 \\ + \frac{r_\nu}{N} \sum_{j=0}^{\rho^N-1} \left| y_j^N(\gamma; u) - \zeta\left(j \frac{r_\nu}{N}\right) \right|_{w_3}^2$$

subject to

$$(3.1) \quad z_j^N = c\left(\frac{r_\nu}{N} \mathcal{A}_N(q)\right)^j P_N(q)(\eta, \phi) \\ + \frac{r_\nu}{N} \sum_{l=1}^j c\left(\frac{r_\nu}{N} \mathcal{A}_N(q)\right)^{j-l} d\left(\theta \frac{r_\nu}{N} \mathcal{A}_N(q)\right) P_N(q) \hat{B}(\alpha) u\left(\frac{lr_\nu}{N}\right),$$

$$(3.2) \quad y_j^N = \hat{C}(\alpha) z_j^N + D(\alpha) u\left(\frac{jr_\nu}{N}\right), \quad j = 0, 1, 2, \dots, \rho^N,$$

where $\hat{B}(\alpha)$, $\hat{C}(\alpha)$, $D(\alpha)$ are as defined in § 2, $\alpha = (\eta, \phi, q) = (\eta, \phi, \alpha, h) = (\eta, \phi, \alpha, (r_1, r_2, \dots, r_\nu))$ and ρ^N is that positive integer for which $\rho^N (r_\nu/N) \leq T < (\rho^N + 1)(r_\nu/N)$.

Under reasonable continuity assumptions (which will be satisfied by the specific schemes we construct in § 4), for each N , the approximating parameter identification problem becomes the minimization of a continuous function over a compact set, and hence we are assured of the existence of a solution.

Remark 3.4. The inclusion of the operator $d((\theta r_\nu/N) \mathcal{A}_N(q))$ in the state equation is a consequence of the theory developed in [28, § 10]. In that paper it is shown that if $d(z)$ is chosen as a rational function approximation to the exponential for which $d((\theta r_\nu/N) \mathcal{A}_N(q))$ satisfies the hypotheses of Theorem 3.2 below then the convergence properties of the state approximation will be enhanced.

Remark 3.5. If for each $q \in Q$ and $N = 1, 2, \dots$, we define the operators $B^N(q) : R^m \rightarrow X_N$ and $A^N(q) : X_N \rightarrow X_N$ by $B^N(q)u = d((\theta r_\nu/N) \mathcal{A}_N(q)) P_N(q) \hat{B}(\alpha)u = d((\theta r_\nu/N) \mathcal{A}_N(q)) P_N(q) (B(\alpha)u, 0)$ and $A^N(q) = c((r_\nu/N) \mathcal{A}_N(q))$, respectively, and let $z_0^N(\gamma) = P_N(q)(\eta, \phi)$ and $u_j^N = u(jr_\nu/N)$, $j = 0, 1, 2, \dots, \rho^N$, then it is immediately clear that (3.1) is the classical variation of parameters solution to the linear non-homogeneous difference equation in X_N given by

$$(3.3) \quad z_j^N = A^N(q) z_{j-1}^N + B^N(q) u_{j-1}^N, \quad j = 1, 2, \dots, \rho^N$$

with initial condition $z_0^N = z_0^N(\gamma)$. Furthermore when the state equation is written in the form given by (3.3) with the exception of the fact that in its most general form the admissible initial data set is infinite dimensional, the approximating parameter identification problems are easily recognized to be in the standard form of a finite dimensional discrete linear-least squares parameter identification problem for which conventional numerical methods can be used to obtain solutions (see [30, Chapt. 4]). In practice the compact admissible initial data set \mathcal{S} (see § 2) is almost always finite dimensional. In fact, the set \mathcal{S} is usually chosen to be the span of a finite collection of elements $\{\hat{\psi}_1, \hat{\psi}_2, \dots, \hat{\psi}_L\} \subset Z$ over a bounded subset of R^L where the unknown parameters to be determined are the coefficients.

In the theorem which follows we apply Theorem 3.1 and ideas similar to those discussed in [28, §§ 8, 9] to show that for a sequence of solutions γ_N^* to the approximating parameter identification problems which converges to an element $\gamma^* \in Q$, the solutions to the corresponding state equations converge as well.

THEOREM 3.2. *Suppose $\{\gamma_N^*\} = \{(z_N^*, q_N^*)\} = \{(\eta_N^*, \phi_N^*, q_N^*)\} = \{(\eta_N^*, \phi_N^*, \alpha_N^*, (r_1^{N*}, r_2^{N*}, \dots, r_\nu^{N*}))\} \subset \Gamma$ is a sequence of solutions to the NPIDDs problems and there exists $\gamma^* = (z^{0*}, q^*) = (\eta^*, \phi^*, q^*) = (\eta^*, \phi^*, \alpha^*, (r_1^*, \dots, r_\nu^*)) \in \Gamma$ such that $\gamma_N^* \rightarrow \gamma^*$ in the sense that (a) $q_N^* \rightarrow q^*$ in $R^{\mu+\nu}$ and (b) $z_N^{0*} \rightarrow z^{0*}$ in Z as $N \rightarrow \infty$. Suppose further that $P_N = P_N(q_N^*): Z \rightarrow X_N(q_N^*)$, $\mathcal{A}_N = \mathcal{A}_N(q_N^*): X_N(q_N^*) \rightarrow X_N(q_N^*)$, $\mathcal{A} = \mathcal{A}(q^*): D \subset Z \rightarrow Z$, $c(z)$, $\{r_\nu^{N*}\} = \{(q_N^*)_{\mu+\nu}\}$, $\rho^{N*} = \rho^{N*}(r_\nu^{N*}) = \rho^{N*}(q_N^*)$ satisfy the conditions and hypotheses of Theorem 3.1 and that:*

(1) *The infinite collection of operators*

$$\left\{ c\left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*)\right)^k \right\}_{k=0}^{\rho^{N*}}$$

are uniformly bounded for all N sufficiently large.

(2) *For $\theta \in [0, 1]$ fixed and each $z \in Z$ we have that for the rational function $d(z)$ the operators $d(\theta(r_\nu^{N*}/N) \mathcal{A}_N(q_N^*))$ exist and satisfy the condition*

$$(3.4) \quad \left| d\left(\frac{\theta r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*)\right) P_N(q_N^*) z - P_N(q_N^*) z \right|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

Then

$$|P_N(q_N^*) z(t_k^N; \gamma^*, u) - z_k^N(\gamma^*; u)|_N \rightarrow 0$$

as $N \rightarrow \infty$ uniformly in k , $k \in \{0, 1, 2, \dots, \rho^{N}\}$ where z and z_k^N are given by (2.7) and (3.1) respectively and $t_k^N = kr_\nu^{N*}/N$, $k = 0, 1, 2, \dots, \rho^{N*}$.*

Proof. The existence of the operators $c((r_\nu^{N*}/N) \mathcal{A}_N(q_N^*))$ for all N sufficiently large is guaranteed by Theorem 3.1. Let M_0 be such that

$$\left| c\left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*)\right)^k \right|_N \leq M_0, \quad k = 0, 1, 2, \dots, \rho^{N*}$$

for all N sufficiently large. Then

$$\begin{aligned} & |P_N(q_N^*) z(t_k^N; \gamma^*, u) - z_k^N(\gamma^*; u)| \\ &= \left| P_N(q_N^*) S(t_k^N; q^*) z^{0*} + P_N(q_N^*) \int_0^{t_k^N} S(t_k^N - \sigma; q^*) \hat{B}(\alpha^*) u(\sigma) d\sigma \right. \\ &\quad \left. - c\left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*)\right)^k P_N(q_N^*) z_N^{0*} - \frac{r_\nu^{N*}}{N} \sum_{j=1}^k c\left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*)\right)^{k-j} \right. \\ &\quad \left. \cdot d\left(\theta \frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*)\right) P_N(q_N^*) \hat{B}(\alpha_N^*) u\left(\frac{j r_\nu^{N*}}{N}\right) \right|_N \\ &\leq \left| P_N(q_N^*) S(t_k^N; q^*) z^{0*} - c\left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*)\right)^k P_N(q_N^*) z^{0*} \right|_N \\ &\quad + \left| c\left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*)\right)^k P_N(q_N^*) (z^{0*} - z_N^{0*}) \right|_N \\ &\quad + \left| P_N(q_N^*) \int_0^{t_k^N} S(t_k^N - \sigma; q^*) \hat{B}(\alpha^*) u(\sigma) d\sigma \right. \\ &\quad \left. - \frac{r_\nu^{N*}}{N} \sum_{j=1}^k c\left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*)\right)^{k-j} d\left(\theta \frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*)\right) P_N(q_N^*) \hat{B}(\alpha_N^*) u\left(\frac{j r_\nu^{N*}}{N}\right) \right|_N \\ &\equiv T_1^N + T_2^N + T_3^N. \end{aligned}$$

The term T_1^N tends to 0 as $N \rightarrow \infty$ uniformly in $k, k \in \{0, 1, 2, \dots, \rho^{N*}\}$ by Theorem 3.1 while

$$T_2^N = \left| c \left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right)^k P_N(q_N^*)(z^{0*} - z_N^{0*}) \right|_N \leq M_0 |z^{0*} - z_N^{0*}| \rightarrow 0$$

as $N \rightarrow \infty$ uniformly in $k, k \in \{0, 1, 2, \dots, \rho^{N*}\}$. We next consider the term T_3^N .

$$\begin{aligned} T_3^N &\leq \left| \int_0^{t_k^N} P_N(q_N^*) S(t_k^N - \sigma; q^*) \hat{B}(\alpha^*) u(\sigma) d\sigma \right. \\ &\quad \left. - \int_0^{t_k^N} P_N(q_N^*) S(t_k^N - \sigma; q^*) \hat{B}(\alpha^*) u_N(\sigma) d\sigma \right|_N \\ &\quad + \left| \int_0^{t_k^N} P_N(q_N^*) S(t_k^N - \sigma; q^*) \hat{B}(\alpha^*) u_N(\sigma) d\sigma \right. \\ &\quad \left. - \frac{r_\nu^{N*}}{N} \sum_{j=1}^k P_N(q_N^*) S(t_{k-j}^N; q^*) \hat{B}(\alpha^*) u \left(\frac{j r_\nu^{N*}}{N} \right) \right|_N \\ &\quad + \left| \frac{r_\nu^{N*}}{N} \sum_{j=1}^k P_N(q_N^*) S(t_{k-j}^N; q^*) \hat{B}(\alpha^*) u \left(\frac{j r_\nu^{N*}}{N} \right) \right. \\ &\quad \left. - \frac{r_\nu^{N*}}{N} \sum_{j=1}^k c \left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right)^{k-j} P_N(q_N^*) \hat{B}(\alpha^*) u \left(\frac{j r_\nu^{N*}}{N} \right) \right|_N \\ &\quad + \left| \frac{r_\nu^{N*}}{N} \sum_{j=1}^k c \left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right)^{k-j} P_N(q_N^*) \hat{B}(\alpha^*) u \left(\frac{j r_\nu^{N*}}{N} \right) \right. \\ &\quad \left. - \frac{r_\nu^{N*}}{N} \sum_{j=1}^k c \left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right)^{k-j} d \left(\theta \frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right) P_N(q_N^*) \hat{B}(\alpha^*) u \left(\frac{j r_\nu^{N*}}{N} \right) \right|_N \\ &\quad + \left| \frac{r_\nu^{N*}}{N} \sum_{j=1}^k c \left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right)^{k-j} d \left(\theta \frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right) P_N(q_N^*) \hat{B}(\alpha^*) u \left(\frac{j r_\nu^{N*}}{N} \right) \right. \\ &\quad \left. - \frac{r_\nu^{N*}}{N} \sum_{j=1}^k c \left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right)^{k-j} d \left(\theta \frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right) P_N(q_N^*) \hat{B}(\alpha^*) u \left(\frac{j r_\nu^{N*}}{N} \right) \right|_N \\ &= \mathcal{T}_1^N + \mathcal{T}_2^N + \mathcal{T}_3^N + \mathcal{T}_4^N + \mathcal{T}_5^N, \end{aligned}$$

where $u_N \in PC^m(0, T)$ is defined by

$$u_N(\sigma) = \begin{cases} u(t_k^N), & \sigma \in [t_{k-1}^N, t_k^N), \quad k = 1, 2, \dots, \rho^{N*}, \\ u(T), & \sigma \in [t_{\rho^{N*}}^N, T]. \end{cases}$$

For each N sufficiently large and each $t \in [0, T]$ we define the following parameterized families of bounded linear operators with domain R^n and range in X_N . For $\eta \in R^n$ and $t \in [0, T]$, let

- (i) $\hat{T}_N(t)\eta = P_N(q_N^*) S(t, q^*)(\eta, 0)$,
- (ii) $\hat{S}_N(t)\eta = P_N(q_N^*) S(t_k^N; q^*)(\eta, 0), \quad t \in [t_k^N, t_{k+1}^N), \quad k = 0, 1, 2, \dots, \rho^{N*},$
- (iii) $\hat{c}_N(t)\eta = c \left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right)^k P_N(q_N^*)(\eta, 0), \quad t \in [t_k^N, t_{k+1}^N),$
 $k = 0, 1, 2, \dots, \rho^{N*},$
- (iv) $\hat{d}_N\eta = d \left(\theta \frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right) P_N(q_N^*)(\eta, 0)$,
- (v) $\hat{I}_N\eta = P_N(q_N^*)(\eta, 0)$.

Using the fact that $\{S(t, q^*): t \geq 0\}$ is a \mathcal{C}_0 semigroup of bounded linear operators on Z and Theorem 3.1 it is not difficult to show (see [29, Lemma 9.1]) that for each $t \in [0, T]$

$$(3.5) \quad \|\hat{T}_N(t) - \hat{S}_N(t)\| \rightarrow 0, \quad \text{as } N \rightarrow \infty,$$

$$(3.6) \quad \|\hat{S}_N(t) - \hat{c}_N(t)\| \rightarrow 0, \quad \text{as } N \rightarrow \infty,$$

$$(3.7) \quad \|\hat{d}_N - I_N\| \rightarrow 0 \quad \text{as } N \rightarrow \infty,$$

where the norm in (3.5), (3.6) and (3.7) above is the one which is induced by the uniform operator topology on $\mathcal{B}(R^n, Z_N)$, the space of all bounded linear operators with domain R^n and range in Z_N .

We now turn to the terms \mathcal{T}_j^N , $j = 1, 2, 3, 4, 5$ and treat each one separately and in turn. Since $u \in PC^m(0, T)$ it is therefore Riemann integrable on $[0, T]$ and hence

$$\begin{aligned} \mathcal{T}_1^N &= \left| \int_0^{t_k^N} P_N(q_N^*) S(t_k^N - \sigma; q^*) \hat{B}(\alpha^*) (u(\sigma) - u_N(\sigma)) d\sigma \right|_N \\ &= \left| \int_0^{t_k^N} \hat{T}_N(t_k^N - \sigma) B(\alpha^*) (u(\sigma) - u_N(\sigma)) d\sigma \right|_N \\ &\leq \int_0^{t_k^N} |\hat{T}_N(t_k^N - \sigma)| |B(\alpha^*)| |u(\sigma) - u_N(\sigma)| d\sigma \\ &\leq M e^{\beta T} |B(\alpha^*)| \int_0^T |u(\sigma) - u_N(\sigma)| d\sigma \rightarrow 0 \quad \text{as } N \rightarrow \infty \end{aligned}$$

uniformly in k , $k \in \{0, 1, 2, \dots, \rho^{N*}\}$.

Using (3.5) above we have that $\|\hat{T}_N(T - \sigma) - \hat{S}_N(T - \sigma)\|$ tends to zero for each $\sigma \in [0, T]$ as $N \rightarrow \infty$. Moreover $\|\hat{T}_N(T - \sigma) - \hat{S}_N(T - \sigma)\|$ is dominated by $g(\sigma) = 2M e^{\beta(T-\sigma)}$ which is integrable on $[0, T]$. Therefore, by the Lebesgue dominated convergence theorem we have

$$\begin{aligned} \mathcal{T}_2^N &= \left| \int_0^{t_k^N} \hat{T}_N(t_k^N - \sigma) B(\alpha^*) u_N(\sigma) d\sigma - \sum_{j=1}^k \int_{t_{j-1}^N}^{t_j^N} \hat{S}_N(t_k^N - \sigma) B(\alpha^*) u_N(\sigma) d\sigma \right|_N \\ &= \left| \int_0^{t_k^N} (\hat{T}_N(t_k^N - \sigma) - \hat{S}_N(t_k^N - \sigma)) B(\alpha^*) u_N(\sigma) d\sigma \right|_N \\ &\leq \int_0^T \|\hat{T}_N(T - \sigma) - \hat{S}_N(T - \sigma)\| |B(\alpha^*)| |u_N(\sigma)| d\sigma \\ &\leq |B(\alpha^*)| \|u\|_\infty \int_0^T \|\hat{T}_N(T - \sigma) - \hat{S}_N(T - \sigma)\| d\sigma \rightarrow 0 \end{aligned}$$

as $N \rightarrow \infty$ uniformly in k , $k \in \{0, 1, 2, \dots, \rho^{N*}\}$. Using (3.6), the fact that $g(\sigma) =$

$M e^{\beta(T-\sigma)} + M_0$ and reasoning similar to that used above we have

$$\begin{aligned} \mathcal{T}_3^N &= \left| \frac{r_\nu^{N*}}{N} \sum_{j=1}^k \left(P_N(q_N^*) S(t_{k-j}^N; q^*) - c \left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right)^{k-j} P_N(q_N^*) \right) \hat{B}(\alpha^*) u \left(\frac{j r_\nu^{N*}}{N} \right) \right|_N \\ &= \left| \sum_{j=1}^k \int_{t_{j-1}^N}^{t_j^N} (\hat{S}_N(t_k^N - \sigma) - \hat{c}_N(t_k^N - \sigma)) B(\alpha^*) u_N(\sigma) d\sigma \right|_N \\ &\leq \int_0^{t_k^N} \|\hat{S}_N(t_k^N - \sigma) - \hat{c}_N(t_k^N - \sigma)\| |B(\alpha^*)| |u_N(\sigma)| d\sigma \\ &\leq |B(\alpha^*)| |u|_\infty \int_0^T \|\hat{S}_N(T - \sigma) - \hat{c}_N(T - \sigma)\| d\sigma \rightarrow 0 \end{aligned}$$

as $N \rightarrow \infty$ uniformly in k , $k \in \{0, 1, 2, \dots, \rho^{N*}\}$. Using (3.7) we have

$$\begin{aligned} \mathcal{T}_4^N &= \left| \frac{r_\nu^{N*}}{N} \sum_{j=1}^k c \left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right)^{k-j} \left(P_N(q_N^*) I - d \left(\frac{\theta r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right) \right) P_N(q_N^*) \hat{B}(\alpha^*) u \left(\frac{j r_\nu^{N*}}{N} \right) \right|_N \\ &\leq M_0 \frac{r_\nu^{N*}}{N} \sum_{j=1}^k \left| (\hat{I}_N - \hat{d}_N) B(\alpha^*) u \left(\frac{j r_\nu^{N*}}{N} \right) \right|_N \\ &\leq M_0 \frac{r_\nu^{N*}}{N} \sum_{j=1}^k \|\hat{I}_N - \hat{d}_N\| |B(\alpha^*)| |u|_\infty \\ &\leq M_0 \|\hat{I}_N - \hat{d}_N\| |B(\alpha^*)| |u|_\infty \rho^{N*} \frac{r_\nu^{N*}}{N} \\ &\leq M_0 |B(\alpha^*)| |u|_\infty T \|\hat{I}_N - \hat{d}_N\| \rightarrow 0 \quad \text{as } N \rightarrow \infty \end{aligned}$$

uniformly in k , $k \in \{0, 1, 2, \dots, \rho^{N*}\}$. Finally, recalling that B has been assumed to depend continuously upon the parameters, we have

$$\begin{aligned} \mathcal{T}_5^N &= \left| \frac{r_\nu^{N*}}{N} \sum_{j=1}^k c \left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right)^{k-j} d \left(\frac{\theta r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right) P_N(q_N^*) (\hat{B}(\alpha^*) - \hat{B}(\alpha_N^*)) u \left(\frac{j r_\nu^{N*}}{N} \right) \right|_N \\ &\leq \frac{r_\nu^{N*}}{N} \sum_{j=1}^k \left| c \left(\frac{r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right)^{k-j} \right| \left| d \left(\frac{\theta r_\nu^{N*}}{N} \mathcal{A}_N(q_N^*) \right) \right| |B(\alpha^*) - B(\alpha_N^*)| \left| u \left(\frac{j r_\nu^{N*}}{N} \right) \right| \\ &\leq M_0 M_1 |B(\alpha^*) - B(\alpha_N^*)| |u|_\infty \rho^{N*} \frac{r_\nu^{N*}}{N} \\ &\leq M_0 M_1 |u|_\infty T |B(\alpha^*) - B(\alpha_N^*)| \rightarrow 0 \quad \text{as } N \rightarrow \infty \end{aligned}$$

uniformly in k , $k \in \{0, 1, 2, \dots, \rho^{N*}\}$, where M_1 is the uniform bound on the operators $d(\theta(r_\nu^{N*}/N)\mathcal{A}_N(q_N^*))$ guaranteed to exist by the strong convergence condition given in (3.4).

Therefore

$$T_3^N = \mathcal{T}_1^N + \mathcal{T}_2^N + \mathcal{T}_3^N + \mathcal{T}_4^N + \mathcal{T}_5^N \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

uniformly in k , $k \in \{0, 1, 2, \dots, \rho^{N*}\}$ and the theorem is proven.

LEMMA 3.1. *If, under the hypotheses and conditions of Theorem 3.2 we have*

$$(3.8) \quad \pi^0 P_N(q_N^*) z \rightarrow \pi^0 z$$

in R^n as $N \rightarrow \infty$ for each $z \in Z$. Then

$$|y_k^N(\gamma_N^*; u) - y(t_k^N; \gamma^*, u)| \rightarrow 0$$

as $N \rightarrow \infty$ uniformly in k , $k \in \{0, 1, 2, \dots, \rho^{N*}\}$ where for each $\gamma \in \Gamma$, $u \in PC^m(0, T)$, $k \in \{0, 1, 2, \dots, \rho^{N*}\}$ and all N sufficiently large $y(t_k^N; \gamma, u)$ is given by (2.11) and $y_k^N(\gamma; u)$ is given by (3.2).

Proof.

$$\begin{aligned}
 |y_k^N(\gamma_N^*; u) - y(t_k^N; \gamma^*, u)| &= \left| \hat{C}(\alpha_N^*) z_k^N(\gamma_N^*; u) + D(\alpha_N^*) u \left(\frac{kr_v^{N*}}{N} \right) \right. \\
 &\quad \left. - \hat{C}(\alpha^*) z(t_k^N; \gamma^*, u) - D(\alpha^*) u(t_k^N) \right| \\
 &\leq |(C(\alpha_N^*) - C(\alpha^*)) \pi^0 z_k^N(\gamma_N^*; u)| \\
 &\quad + |C(\alpha^*) (\pi^0 z_k^N(\gamma_N^*; u) - \pi^0 P_N(q_N^*) z(t_k^N; \gamma^*, u))| \\
 &\quad + |C(\alpha^*) (\pi^0 P_N(q_N^*) z(t_k^N; \gamma^*, u) - \pi^0 z(t_k^N; \gamma^*, u))| \\
 &\quad + |(D(\alpha_N^*) - D(\alpha^*)) u(t_k^N)| \\
 &\leq |C(\alpha_N^*) - C(\alpha^*)| |z_k^N(\gamma_N^*; u)|_N \\
 &\quad + |C(\alpha^*)| |z_k^N(\gamma_N^*; u) - P_N(q_N^*) z(t_k^N; \gamma^*, u)|_N \\
 &\quad + |C(\alpha^*)| |\pi^0 P_N(q_N^*) z(t_k^N; \gamma^*, u) - \pi^0 z(t_k^N; \gamma^*, u)| \\
 &\quad + |u|_\infty |D(\alpha_N^*) - D(\alpha^*)| \\
 &\equiv \mathcal{T}_1^N + \mathcal{T}_2^N + \mathcal{T}_3^N + \mathcal{T}_4^N.
 \end{aligned}$$

In light of the convergence guaranteed by Theorem 3.2, it is easily verified that $\{|z_k^N(\gamma_N^*; u)|_N\}_{k=0}^{\rho^{N*}}$ lie in a bounded subset of R which is independent of N for all N sufficiently large. Therefore using the assumptions that $\alpha_N^* \rightarrow \alpha^*$ in R^μ as $N \rightarrow \infty$ and that $C(\alpha)$ and $D(\alpha)$ depend continuously upon the parameters we have $\mathcal{T}_1^N \rightarrow 0$ and $\mathcal{T}_4^N \rightarrow 0$ as $N \rightarrow \infty$ uniformly in k , $k \in \{0, 1, 2, \dots, \rho^{N*}\}$. The term \mathcal{T}_2^N tends to zero as $N \rightarrow \infty$ uniformly in k , $k \in \{0, 1, 2, \dots, \rho^{N*}\}$ as a consequence of Theorem 3.2. Finally (3.8), the fact that the set $S \equiv \{z(t; \gamma^*, u) : t \in [0, T]\}$ is a compact subset of Z (being the continuous image of a compact subset of R) and the uniform boundedness of the operators $\pi^0 P_N(q_N^*)$ imply $\pi^0 P_N(q_N^*) \rightarrow \pi^0$ uniformly on S as $N \rightarrow \infty$ and hence $\mathcal{T}_3^N \rightarrow 0$ as $N \rightarrow \infty$ uniformly in k , $k \in \{0, 1, 2, \dots, \rho^{N*}\}$.

We can now state and prove the major result of this paper which is that in a certain sense (which will be made precise in the statement of Theorem 3.3 below) a solution γ_N^* to the N th approximating parameter identification problem is in fact an approximation of a solution γ^* of the PIDDS.

THEOREM 3.3. *Suppose $\{\gamma_N^*\} = \{(z_N^{0*}, q_N^*)\} \subset \Gamma$ is a sequence of solutions to the problems NPIDDS. Then there exist a $\gamma^* = (z^{0*}, q^*) \in \Gamma$ and a subsequence $\{\gamma_{N_k}^*\}$ of $\{\gamma_N^*\}$ such that $\gamma_{N_k}^* \rightarrow \gamma^*$ as $k \rightarrow \infty$ in the sense that (a) $q_{N_k}^* \rightarrow q^*$ in $R^{\mu+\nu}$ and (b) $z_{N_k}^{0*} \rightarrow z^{0*}$ in Z as $k \rightarrow \infty$. If in addition $P_N = P_N(q_N^*)$, $\mathcal{A}_N = \mathcal{A}_N(q_N^*)$, $\mathcal{A} = \mathcal{A}(q^*)$, $c(z)$, $d(z)$ and $\rho^{N*} = \rho^{N*}(q_N^*)$ satisfy the hypotheses and conditions of Theorem 3.2 and if $P_N(q_N^*)$ satisfies (3.8) then γ^* is a solution of the APIDDS (and therefore to the PIDDS as well).*

Proof. Since $\mathcal{S} \subset Z$ has been assumed compact, there exists a subsequence $\{z_{N_j}^{0*}\}$ of $\{z_N^{0*}\}$ such that $z_{N_j}^{0*} \rightarrow z^{0*} \in \mathcal{S}$ as $j \rightarrow \infty$. Similarly $Q \subset R^{\mu+\nu}$ compact implies the existence of a subsequence $\{q_{N_l}^*\}$ of $\{q_N^*\}$ such that $q_{N_l}^* \rightarrow q^* \in Q$ as $l \rightarrow \infty$. Letting $\gamma^* = (z^{0*}, q^*)$ and reindexing, we obtain a subsequence $\{\gamma_{N_k}^*\}$ of $\{\gamma_N^*\}$ such that $\gamma_{N_k}^* \rightarrow \gamma^* \in \Gamma$ as $k \rightarrow \infty$.

For each $\gamma = (z^0, q) = (z^0, \alpha, (r_1, r_2, \dots, r_\nu)) \in \Gamma$, $u \in PC^m(0, T)$, $\zeta \in C^l(0, T)$ and all N sufficiently large we define $y^N, \hat{y}^N, \zeta^N \in PC^l(0, T)$ by

$$\begin{aligned} y^N(\sigma) &= y^N(\sigma; \gamma, u) = y(t_k^N; \gamma, u), \\ \hat{y}^N(\sigma) &= \hat{y}^N(\sigma; \gamma, u) = y_k^N(\gamma; u), \\ \zeta^N(\sigma) &= \zeta^N(\sigma; \gamma) = \zeta(t_k^N), \end{aligned}$$

for $\sigma \in I_k^N$, $k = 0, 1, 2, \dots, \rho^N$, where $I_k^N = I_k^N(\gamma) = [t_k^N, t_{k+1}^N)$, $t_k^N = t_k^N(\gamma) = (kr_\nu/N)$, $k = 0, 1, 2, \dots, \rho^N$ and ρ^N is that integer for which $\rho^N(r_\nu/N) \leq T < (\rho^N + 1)(r_\nu/N)$ and where $y(\cdot; \gamma, u)$ and $y^N(\gamma; u)$ are given by (2.11) and (3.2) respectively.

If $\{\gamma_N\}$ is a sequence of elements in Γ for which $\gamma_N \rightarrow \gamma \in \Gamma$, then Lemma 3.1 implies

$$(3.9) \quad |\hat{y}^N(\sigma; \gamma_N, u) - y^N(\sigma; \gamma, u)| \rightarrow 0$$

as $N \rightarrow \infty$ uniformly in σ for $\sigma \in [0, T]$. Furthermore the continuity of $y(\cdot, \gamma, u)$ and $\zeta(\cdot)$ and the fact that $\text{length}(I_k^N) = r_\nu^N/N \leq r/N \rightarrow 0$ as $N \rightarrow \infty$ imply

$$(3.10) \quad |y^N(\sigma; \gamma, u) - y(\sigma; \gamma, u)| \rightarrow 0,$$

and

$$|\zeta^N(\sigma; \gamma_N) - \zeta(\sigma)| \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

for each $\sigma \in [0, T]$. The triangle inequality, (3.9) and (3.10) imply $\hat{y}^N(\sigma; \gamma_N, u) \rightarrow y(\sigma; \gamma, u)$ for each $\sigma \in [0, T]$ and hence by the Lebesgue dominated convergence theorem we have for any $\gamma \in \Gamma$

$$\begin{aligned} J(\gamma^*) &= |y(0; \gamma^*, u) - \zeta(0)|_{w_1}^2 + |y(T; \gamma^*, u) - \zeta(T)|_{w_2}^2 + \int_0^T |y(t; \gamma^*, u) - \zeta(t)|_{w_3}^2 dt \\ &= \lim_{k \rightarrow \infty} |\hat{y}^{N_k}(0; \gamma_{N_k}^*, u) - \zeta(0)|_{w_1}^2 + \lim_{k \rightarrow \infty} |\hat{y}^{N_k}(T; \gamma_{N_k}^*, u) - \zeta(T)|_{w_2}^2 \\ &\quad + \int_0^T \lim_{k \rightarrow \infty} |\hat{y}^{N_k}(t; \gamma_{N_k}^*, u) - \zeta^{N_k}(t; \gamma_{N_k}^*)|_{w_3}^2 dt \\ &= \lim_{k \rightarrow \infty} |y_0^{N_k}(\gamma_{N_k}^*; u) - \zeta(0)|_{w_1}^2 + \lim_{k \rightarrow \infty} |y_{\rho^{N_k}}^{N_k}(\gamma_{N_k}^*; u) - \zeta(T)|_{w_2}^2 \\ &\quad + \lim_{k \rightarrow \infty} \int_0^T |\hat{y}^{N_k}(t; \gamma_{N_k}^*, u) - \zeta^{N_k}(t; \gamma_{N_k}^*)|_{w_3}^2 dt \\ &= \lim_{k \rightarrow \infty} |y_0^{N_k}(\gamma_{N_k}^*; u) - \zeta(0)|_{w_1}^2 + \lim_{k \rightarrow \infty} |y_{\rho^{N_k}}^{N_k}(\gamma_{N_k}^*; u) - \zeta(T)|_{w_2}^2 \\ &\quad + \lim_{k \rightarrow \infty} \sum_{j=0}^{\rho^{N_k}-1} \int_{jr_\nu^*/N_k}^{(j+1)r_\nu^*/N_k} |\hat{y}^{N_k}(t; \gamma_{N_k}^*, u) - \zeta^{N_k}(t, \gamma_{N_k}^*)|_{w_3}^2 dt \\ &= \lim_{k \rightarrow \infty} \left[|y_0^{N_k}(\gamma_{N_k}^*; u) - \zeta(0)|_{w_1}^2 + |y_{\rho^{N_k}}^{N_k}(\gamma_{N_k}^*; u) - \zeta(T)|_{w_2}^2 \right. \\ &\quad \left. + \frac{r_\nu^{N_k}}{N_k} \sum_{j=0}^{\rho^{N_k}-1} \left| y_j^{N_k}(\gamma_{N_k}^*; u) - \zeta\left(\frac{jr_\nu^{N_k}}{N_k}\right) \right|_{w_3}^2 \right] \end{aligned}$$

$$\begin{aligned}
&= \lim_{k \rightarrow \infty} J_{N_k}(\gamma_{N_k}^*) \leq \lim_{k \rightarrow \infty} J_{N_k}(\gamma) \\
&= \lim_{k \rightarrow \infty} \left[|y_0^{N_k}(\gamma; u) - \zeta(0)|_{w_1}^2 + |y_{\rho}^{N_k}(\gamma; u) - \zeta(T)|_{w_2}^2 \right. \\
&\quad \left. + \frac{r_\nu}{N_k} \sum_{j=0}^{\rho N_k - 1} \left| y_j^{N_k}(\gamma; u) - \zeta\left(\frac{j r_\nu}{N_k}\right) \right|_{w_3}^2 \right] \\
&= \lim_{k \rightarrow \infty} \left[|y_0^{N_k}(\gamma; u) - \zeta(0)|_{w_1}^2 + |y_{\rho}^{N_k}(\gamma; u) - \zeta(T)|_{w_2}^2 \right. \\
&\quad \left. + \sum_{j=0}^{\rho N_k - 1} \int_{j r_\nu / N_k}^{(j+1)r_\nu / N_k} |\hat{y}^{N_k}(t; \gamma, u) - \zeta^{N_k}(t; \gamma)|_{w_3}^2 dt \right] \\
&= \lim_{k \rightarrow \infty} \left[|\hat{y}^{N_k}(0; \gamma, u) - \zeta(0)|_{w_1}^2 + |\hat{y}^{N_k}(T; \gamma, u) - \zeta(T)|_{w_2}^2 \right. \\
&\quad \left. + \int_0^T |\hat{y}^{N_k}(t; \gamma, u) - \zeta^{N_k}(t; \gamma)|_{w_3}^2 dt \right] \\
&= |y(0; \gamma, u) - \zeta(0)|_{w_1}^2 + |y(T; \gamma, u) - \zeta(T)|_{w_2}^2 \\
&\quad + \int_0^T \lim_{k \rightarrow \infty} |\hat{y}^{N_k}(t; \gamma, u) - \zeta^{N_k}(t, \gamma)|_{w_3}^2 dt \\
&= |y(0; \gamma, u) - \zeta(0)|_{w_1}^2 + |y(T; \gamma, u) - \zeta(T)|_{w_2}^2 + \int_0^T |y(t; \gamma, u) - \zeta(t)|_{w_3}^2 dt \\
&= J(\gamma).
\end{aligned}$$

Thus $J(\gamma^*) \leq J(\gamma)$ for any $\gamma \in \Gamma$ and γ^* is a solution to the APIDDS.

4. Examples of convergent approximation schemes for the PIDDS. In this section we construct specific examples of convergent approximation schemes for the PIDDS. That is, given a sequence $\{q_N\} \subset Q$ with $q_N \rightarrow \bar{q} \in Q$ as $N \rightarrow \infty$, for each $N = 1, 2, \dots$, we define X_N a closed subspace of $Z_N = \mathbb{R}^n \times L_2^n(-r_\nu^N, 0)$ with inner product $\langle \cdot, \cdot \rangle_N$, $\Pi_N: Z_N \rightarrow X_N$ the orthogonal projection of Z_N onto X_N with respect to $\langle \cdot, \cdot \rangle_N$, linear operators $\mathcal{A}_N(q_N): X_N \rightarrow X_N$, and choose rational functions $c(z)$ and $d(z)$ which satisfy the hypotheses and conditions of Theorem 3.3. We require that:

- (4.1) There exist constants M and β such that $\mathcal{A}_N(q_N) \in G(M, \beta)$ on X_N for all N sufficiently large and $\mathcal{A} = \mathcal{A}(\bar{q}) \in G(M, \beta)$ on Z .
- (4.2) There exists a dense subset of Z , $\mathcal{D} \subset D(\mathcal{A}(\bar{q}))$ such that $R_\lambda(\mathcal{A}(\bar{q}))\mathcal{D} \subset \mathcal{D}$ for each $\lambda \in C$ with $\operatorname{Re} \lambda > \beta$ and for each $z \in \mathcal{D}$ we have

$$|\mathcal{A}_N(q_N)P_{Nz} - P_N\mathcal{A}(\bar{q})z|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty, \quad \text{where } P_N = \Pi_N \mathcal{I}_N.$$

- (4.3) $\pi^0 P_{Nz} \rightarrow \pi^0 z$ for each $z \in Z$ where $\pi^0: \{Z_N \rightarrow \mathbb{R}^n$ is defined by $\pi^0(\eta, \phi) = \eta$.

- (4.4) $c(z)$ is a rational function approximation to the exponential for which
 - (a) $|c(z) - e^z| = O(|z|^{m+1})$ as $z \rightarrow 0$ with $m > 0$;
 - (b) $\deg c(z) \leq m + 1$;
 - (c) $c(z)$ has no poles in $\{z \in C: \operatorname{Re} z \leq 0\}$;

(d) there exists a constant M_0 such that

$$\left| c\left(\frac{r_\nu^N}{N} \mathcal{A}_N(q_N)\right)^k \right| \leq M_0, \quad k = 0, 1, 2, \dots, \rho^N$$

for all N sufficiently large where ρ^N is that positive integer for which

$$\rho^N \frac{r_\nu^N}{N} \leq T < (\rho^N + 1) \frac{r_\nu^N}{N}.$$

(4.5) $d(z)$ is a rational function approximation to the exponential for which

- (a) the operators $d((\theta r_\nu^N/N) \mathcal{A}_N(q_N))$ exist for all N sufficiently large
- (b) $|d((\theta r_\nu^N/N) \mathcal{A}_N(q_N)) P_N z - P_N I z|_N \rightarrow 0$ as $N \rightarrow \infty$, where $0 \leq \theta \leq 1$.

For a given choice of X_N , Π_N , $\mathcal{A}_N(q_N)$, $c(z)$ and $d(z)$, the triple $\{X_N, \Pi_N, \mathcal{A}_N(q_N)\}$ will be referred to as the state approximation, while the collection $\{X_N, \Pi_N, \mathcal{A}_N(q_N), c(z), d(z)\}$ itself will be referred to as an approximation scheme. We shall consider two particular families of state approximations which can be shown to satisfy conditions (4.1), (4.2) and (4.3) above. The first, and more primitive of the two is the averaging or AVE state approximation ([4], [5], [7], [28]) in which the functional component of the subspace X_N is chosen to be the span of a finite collection of piecewise constant functions defined on $[-r_\nu^N, 0]$. The second family of state approximations is spline based and is known as the SPL state approximation ([7], [9], [28]). In this case the subspace X_N is chosen to be the span of a finite collection of elements in Z_N having first or higher order spline functions as their functional component. We note that in both the AVE and SPL state approximations X_N is finite dimensional.

Once a state approximation has been set, rational functions $c(z)$ and $d(z)$ must be chosen in order to complete the construction of the approximation scheme. Although others are available, we shall restrict our attention to choices of $c(z)$ and $d(z)$ from the Padé table of rational function approximations to the exponential ([28], [31]). We shall demonstrate that for appropriate choices for $c(z)$ and $d(z)$, taken from the Padé table, the AVE and SPL state approximations generate approximation schemes which satisfy conditions (4.1) through (4.5) above and hence yield approximate solutions to the PIDDS.

All of the ideas discussed in this section have appeared elsewhere. In particular, since our discrete schemes are based upon the semi-discrete approximation schemes for the PIDDS developed by Banks, Burns and Cliff [7], the AVE and SPL state approximations are the same as those used in [7] for a similar purpose. Furthermore, since our schemes are also based upon the discrete approximation framework for the integration of LRFDE initial value problems developed in [28], the theory underlying the appropriate choice of the rational functions $c(z)$ and $d(z)$ can be found in [28]. Therefore, the construction of the state approximations, the choosing of the rational functions and the arguments used in the verification of conditions (4.1) through (4.5) for the particular schemes will only be outlined and summarized here. For a detailed explanation of the various constructs which we define and the verification of the many results which we state without proof, the interested reader is advised to consult [7], [28] and [29].

4.1. The AVE state approximation. Let $\{q_N\} = \{(\alpha_N, r_1^N, r_2^N, \dots, r_\nu^N)\} \subset Q$ be given with $q_N \rightarrow \bar{q} \in Q$. Define $\chi_j^N \in L_2^n(-r_\nu^N, 0)$ to be the characteristic function of the interval $[-jr_\nu^N/N, -(j-1)r_\nu^N/N]$, $j = 2, 3, \dots, N$ and χ_1^N to be the characteristic function of the interval $[-r_\nu^N/N, 0]$. Let X_N be the closed subspaces of $Z_N =$

$R^n \times L_2^n(-r_\nu^N, 0)$ given by

$$X_N = \left\{ (\eta, \phi) \in Z_N \mid \eta \in R^n, \phi = \sum_{j=1}^N v_j^N \chi_j^N, v_j^N \in R^n \right\},$$

and let $\langle \cdot, \cdot \rangle_N$ denote the standard inner product on Z_N . With X_N as above, the orthogonal projection Π_N of Z_N onto X_N with respect to $\langle \cdot, \cdot \rangle_N$ can be computed and is given by

$$\Pi_N(\eta, \phi) = \left(\eta, \sum_{j=1}^N \phi_j^N \chi_j^N \right), \quad \text{where } \phi_j^N = \frac{N}{r_\nu^N} \int_{-jr_\nu^N/N}^{-(j-1)r_\nu^N/N} \phi(\theta) d\theta, \quad j = 1, 2, \dots, N.$$

In order to define the operators $\mathcal{A}_N(q_N)$ we first define the operators $L_N(q_N): X_N \rightarrow R^n$ and $D_N(q_N): X_N \rightarrow L_2^n(-r_\nu^N, 0)$ by

$$L_N(q_N) \left(\eta, \sum_{j=1}^N v_j^N \chi_j^N \right) = A_0(\alpha_N) \eta + \sum_{i=1}^N \sum_{j=1}^N A_i(\alpha_N) v_j^N \chi_j^N (-r_i^N) + \frac{r_\nu^N}{N} \sum_{j=1}^N K_j^N(\alpha_N) v_j^N$$

where

$$K_j^N(\alpha) = \frac{N}{r_\nu^N} \int_{-jr_\nu^N/N}^{-(j-1)r_\nu^N/N} K(\alpha, \theta) d\theta, \quad j = 1, 2, \dots, N$$

and

$$D_N(q_N) \left(\eta, \sum_{j=1}^N v_j^N \chi_j^N \right) = \sum_{j=1}^N \frac{N}{r_\nu^N} (v_{j-1}^N - v_j^N) \chi_j^N,$$

where $v_0 \equiv \eta$ respectively. Let $\mathcal{A}_N(q_N): X_N \rightarrow X_N$ be given by

$$(4.6) \quad \mathcal{A}_N(q_N)(\eta, \phi) = (L_N(q_N)(\eta, \phi), D_N(q_N)(\eta, \phi)).$$

An elementary but rather tedious computation can be performed which demonstrates that the operators on Z_N , $\mathcal{A}_N(q_N)$ given by (4.6) satisfy

$$\langle \mathcal{A}_N(q_N)z, z \rangle_N \leq \omega(q_N) \|z\|_N^2,$$

where $\omega(q)$ is given by (2.6) and $\langle \cdot, \cdot \rangle_N$ and $\|\cdot\|_N$ denote an innerproduct on Z and its corresponding norm which satisfies

$$|\cdot|_N \leq \|\cdot\|_N \leq \sqrt{\nu} |\cdot|_N.$$

Since X_N is finite dimensional we have that $D(\mathcal{A}_N(q_N)) = X_N$ and therefore that the operators $\mathcal{A}_N(q_N) - \omega(q_N)I$ are maximal dissipative (see [21, Definition 4.1, p. 86]). This fact together with (2.5) yields that $\mathcal{A}(q), \mathcal{A}_N(q_N) \in G(\sqrt{\nu}, \beta)$, $q, q_N \in Q$, $N = 1, 2, \dots$, where $\beta = \max_{q \in Q} \omega(q)$ and hence that (4.1) is satisfied. In addition it can also be shown (see [25]) that

$$(4.7) \quad \left\| I + \frac{r_\nu^N}{N} \mathcal{A}_N(q_N) \right\|_N \leq 1 + K(q_N) \frac{r_\nu^N}{N} \leq 1 + \frac{\kappa}{N},$$

where κ is a constant independent of N and $q \in Q$. The bound given in (4.7) is a somewhat stronger result than dissipativeness in that (4.7) implies that the operators $\mathcal{A}_n(q_N) - \kappa I/2$ are dissipative (see [28, Lemma 5.15]). The importance of condition (4.7) will become clear when the choice of the rational function $c(z)$ is discussed in § 4.3.

If we let $\mathcal{D} = \{(\phi(0), \phi) \in Z \mid \phi \in C_1^n(-r, 0)\}$, then \mathcal{D} is a dense subset of Z , $\mathcal{D} \subset D = D(\mathcal{A}(\bar{q}))$ and for $\lambda \in C$ with $\operatorname{Re} \lambda > \beta$ we have $R_\lambda(\mathcal{A}(\bar{q}))\mathcal{D} \subset D(\mathcal{A}^2(\bar{q})) \subset \mathcal{D}$.

Moreover, it can be shown (see [7]) that for $z \in \mathcal{D}$

$$|\mathcal{A}_N(q_N)P_N z - P_N \mathcal{A}(\bar{q})z|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

and (4.2) is satisfied.

Finally, for $(\eta, \phi) \in Z$ we have

$$\pi^0 P_N z = \pi^0 \left(\eta, \sum_{j=1}^N \phi_j^N \chi_j^N \right) = \eta = \pi^0 z$$

and hence that conditions (4.1), (4.2) and (4.3) are satisfied for the AVE state approximation.

4.2. The SPL state approximation. In this subsection we describe spline based state approximations using first order or linear splines. All of the results stated below can be modified so as to be applicable to spline based state approximations employing higher order splines.

Once again we assume $\{q_N\} = \{(\alpha_N, r_1^N, r_2^N, \dots, r_\nu^N)\} \subset Q$ with $q_N \rightarrow \bar{q} \in Q$ as $N \rightarrow \infty$. We partition each of the subintervals $[-r_k^N, -r_{k-1}^N]$, $k = 1, 2, \dots, \nu$ into N equal subintervals to define the partition $\{\varepsilon_j^N\}_{j=1}^{\nu N}$ of $[-r_\nu^N, 0]$ where

$$\varepsilon_j^N = -(j - (k-1)N) \frac{r_k^N - r_{k-1}^N}{N} + r_{k+1}^N$$

$j = (k-1)N, \dots, kN$, $k = 1, 2, \dots, \nu$, and define the finite dimensional subspace X_N of Z_N by

$$X_N = \{(\phi(0), \phi) \in Z_N \mid \phi \text{ is a first order spline with knots at } \{\theta_j^N\}_{j=1}^{\nu N}\}.$$

Let $\langle \cdot, \cdot \rangle_N$ denote the weighted inner product on Z_N defined in § 5 of [29] and let Π_N be the orthogonal projection of Z_N onto X_N with respect to $\langle \cdot, \cdot \rangle_N$. Finally we let $\mathcal{A}_N(q_N) : X_N \rightarrow X_N$ be given by

$$(4.8) \quad \mathcal{A}_N(q_N) = \Pi_N \mathcal{A}(q_N) \Pi_N.$$

We note that

$$R(\Pi_N) = X_N \subset \{(\eta, \phi) \in Z_N \mid \eta = \phi(0), \phi \in W_{1,2}^n(-r_\nu^N, 0)\} = D(\mathcal{A}(q_N))$$

and hence that the expression for $\mathcal{A}_N(q_N)$ given by (4.8) is well defined.

Using the fact that Π_N is the orthogonal projection of Z_N onto X_N it can be shown that

$$\langle \mathcal{A}_N(q_N)z, z \rangle_N \leq \beta |z|_N$$

for each $z \in X_N$, where β is as defined above. Since X_N is finite dimensional with $D(\mathcal{A}_N(q_N)) = X_N$ we have therefore that $\mathcal{A}_N(q_N) \in G(\sqrt{\nu}, \beta)$, $N = 1, 2, \dots$, and condition (4.1) is satisfied.

Next, if we define $\mathcal{D} = D(\mathcal{A}^3(\bar{q}))$, we have that \mathcal{D} is a dense subset of Z (see [23]) and for $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda > \beta$, $R_\lambda(\mathcal{A}(\bar{q}))\mathcal{D} \subset \mathcal{D}$. Using the properties of interpolatory splines, and the fact that Π_N is an orthogonal projection (and hence has certain minimality properties) it can be demonstrated that

$$|\mathcal{A}_N(q_N)P_N z - P_N \mathcal{A}(\bar{q})z|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

for each $z \in \mathcal{D}$. Furthermore it can also be argued that for $\hat{\phi} \in \mathcal{D}$, $|\Pi_N \mathcal{J}_N \hat{\phi} - \mathcal{J}_N \hat{\phi}|_N \rightarrow 0$ as $N \rightarrow \infty$. However \mathcal{D} is a dense subset of Z and the operators $\{\Pi_N \mathcal{J}_N - \mathcal{J}_N\}$ are uniformly bounded. Recalling that $P_N = \Pi_N \mathcal{J}_N$ it follows therefore

that $|P_{NZ} - \mathcal{J}_{NZ}|_N \rightarrow 0$ as $N \rightarrow \infty$ for all $z \in Z$. This in turn implies that

$$\pi^0(P_{NZ}) \rightarrow \pi^0 z$$

for all $z \in Z$ and the SPL state approximation defined above satisfies conditions (4.1), (4.2) and (4.3).

4.3. Selecting the rational functions $c(z)$ and $d(z)$. Our primary objective in this subsection is to summarize the theory developed in [28] for the selection of rational functions $c(z)$ and $d(z)$ which satisfy conditions (4.4) and (4.5) respectively for a given state approximation triple $\{X_N, \Pi_N, \mathcal{A}_N(q_N)\}$. For a given approximation scheme $\{X_N, \Pi_N, \mathcal{A}_N(q_N), c(z), d(z)\}$ the most difficult condition to verify is the temporal stability condition (4.4)(d). As we shall soon see, it is the happy circumstance that the relatively easily verified spatial stability condition (4.1) (which we already know is satisfied by the AVE and SPL state approximations) is, under the appropriate hypotheses, sufficient to guarantee that (4.4)(d) holds as well.

Although there are many families of rational functions which satisfy the required exponential approximation property (4.4)(a), among the most widely studied are the well known Padé approximants. The Padé approximations, which can be arranged in a tableau $\{p_{jk}(z)\}$ commonly referred to as the Padé table, are defined by the following formulae

$$p_{jk}(z) = \frac{n_{jk}(z)}{d_{jk}(z)}, \quad j, k = 1, 2, \dots, \quad \text{where } n_{jk}(z) = \sum_{l=0}^k \frac{(j+k-l)! k!}{(j+k)! l! (k-l)!} z^l,$$

$$d_{jk}(z) = \sum_{l=0}^j \frac{(j+k-l)! j!}{(j+k)! l! (j-l)!} (-z)^l.$$

It is well known that

$$(4.9) \quad |p_{jk}(z) - e^z| = O(|z|^{j+k+1}) \quad z \rightarrow 0,$$

it is easily seen that

$$(4.10) \quad \deg p_{jk}(z) = k - j,$$

and Ehle [15] has demonstrated that

$$(4.11) \quad |p_{jk}(z)| \leq 1, \quad j = k, k+1, k+2, \quad z \in \{z \in C : \operatorname{Re} z \leq 0\}.$$

Thus from (4.9), (4.10) and (4.11) it is immediately clear that if the rational function $c(z)$ is chosen from among the entries on the diagonal or the first two subdiagonals of the Padé table the resulting approximation scheme will satisfy conditions (4.4)(a), (b) and (c). We note further that this will also be true if $c(z)$ is chosen from the top row of the Padé table, the Maclaurin polynomials for e^z .

We next turn our attention to the temporal stability condition (4.4)(d). The dissipativeness of the operators $\mathcal{A}_N(q_N) - \beta I$ discussed above (both the AVE and SPL constructions), von Neumann's theory of spectral sets [26], a result due to Hersh and Kato [18] and (4.11) can be used to demonstrate that for $j = k, k+1, k+2, k = 1, 2, \dots$

$$\left| p_{jk} \left(\frac{r_\nu^N}{N} \mathcal{A}_N(q_N) \right) \right|_N^l \leq \sqrt{\nu} \left(1 + \mathcal{K} \frac{r_\nu^N}{N} \right)^l \leq \sqrt{\nu} e^{\beta \mathcal{K} T},$$

where \mathcal{K} is a constant independent of N . Furthermore using the properties of the

Padé approximants and (4.7) we have

$$\left| p_{0k} \left(\frac{r_N^N}{N} \mathcal{A}_N(q_N) \right) \right|_N \leq \sqrt{\nu} e^{\beta \kappa T}, \quad k = 1, 2, \dots$$

for the AVE state approximation triple. Therefore if we define \mathfrak{A}_p and \mathfrak{E}_p to be the subclasses of the Padé approximants given by

$$\mathfrak{A}_p = \{p_{jk}(z)\}, \quad j = k, k+1, k+2, \quad k = 1, 2, \dots$$

and

$$\mathfrak{E}_p = \{p_{0k}(z)\}, \quad k = 1, 2, \dots,$$

respectively, (4.4) will be satisfied for the AVE state approximation triple if $c(z) \in \mathfrak{A}_p \cup \mathfrak{E}_p$ and for the SPL state approximation triple if $c(z) \in \mathfrak{A}_p$.

In [28] a heuristic argument is given in support of choosing $d(z)$ as a rational function approximation to the exponential. This argument is supported empirically by computational results. Indeed enhanced convergence rates are observed for schemes constructed with $d(z)$ chosen in this way. Therefore $d(z)$ should be chosen as a rational function approximation to the exponential for which condition (4.5) is satisfied. It is easily verified (see [28, Thm. 10.3]) that if $d(z)$ satisfies condition (4.4) it will satisfy condition (4.5) as well. For the AVE state approximation, therefore, $d(z)$ can be chosen from $\mathfrak{A}_p \cup \mathfrak{E}_p$, while for the SPL state approximation $d(z)$ can be chosen from \mathfrak{A}_p . However, it is shown in [28] that for the SPL state approximation $d(z)$ can actually be chosen from $\mathfrak{A}_p \cup \mathfrak{E}_p$ and still satisfy condition (4.5).

Finally, the results of this section can be summarized as follows: For an approximation scheme $\{X_N, \Pi_N, \mathcal{A}_N(q_N), c(z), d(z)\}$ constructed with the AVE state approximation and $c(z)$ and $d(z)$ chosen from $\mathfrak{A}_p \cup \mathfrak{E}_p$, conditions (4.1) through (4.5) will be satisfied and a sequence of solutions to the resulting sequence of approximating parameter identification problems will contain a subsequence converging to a solution of the PIDDS. A similar statement can be made for approximation schemes constructed with the SPL state approximation, $c(z)$ chosen from \mathfrak{A}_p and $d(z)$ chosen from $\mathfrak{A}_p \cup \mathfrak{E}_p$.

5. Numerical results. In this section we discuss and analyze numerical results obtained by applying the approximation schemes developed in the previous sections to actual parameter identification problems in which the governing control system is a linear functional differential equation of retarded type. All of the computations for the examples which follow were performed on an IBM 370/158 using software packages written in Fortran. We provide no information regarding storage requirements or computational efficiency in that our primary objective in performing these tests was to demonstrate the feasibility of our methods.

The approximating parameter identification problems given in § 3 were constructed using the AVE and SPL state approximations defined in § 4, $c(z) = p_{22}(z) \in \mathfrak{A}_p$, $d(z) = p_{02}(z) \in \mathfrak{E}_p$ and $\theta = .5$. The effect of variations in the choice of $c(z)$, $d(z)$ and θ were not tested here since this was studied extensively in [28]. We have assumed that we have been given observational data on the interval $[0, 2]$ which resulted from input $u = u_l \in PC^1(0, 2)$ where

$$u_l(t) = \begin{cases} 0, & t < l, \\ 1, & l \leq t. \end{cases}$$

The norms $|\cdot|_{w_1}, |\cdot|_{w_2}, |\cdot|_{w_3}$ which appear in (2.1) have all been taken to be the standard Euclidean norm on R^l . To obtain observational data ζ , for each example the state equation was integrated using the method of steps [16], a fourth order Runge–Kutta numerical integration scheme for ordinary differential equation initial value problems and a preselected set of true parameter values $\gamma^* = (\eta^*, \phi^*, \alpha^*, h^*)$. We emphasize that the integration method used to obtain the observational data was completely independent of the approximation schemes being tested and hence should not have contaminated our results.

The resulting finite dimensional approximating parameter identification problems were solved using a modified version of the integration package for LRFDE initial value problems developed in [28] and the IMSL [19] routine ZXSSQ, a finite difference Levenberg–Marquardt scheme for solving the problem of minimizing the sum of squares of M nonlinear functions in N -unknowns. The Levenberg–Marquardt algorithm is an iterative gradient projection scheme which must be provided with an initial estimate of the unknown parameters.

Since among the principal advantages of our approximation schemes is their ability to identify the delays, it is this feature which we are most interested in testing. The examples which have been included below, therefore, all have the delays in the problem among the parameters to be identified. A discussion of the performance of the schemes on examples in which the delays need not be identified can be found in [11].

Two of the four examples which appear below have also been included in [6] where they are used to test the semi-discrete schemes developed in [7]. A comparison of the performance of the two methods (based upon the two examples below, and others not included here) reveals that they exhibit similar behavior. The similarity becomes especially apparent for the cases $N = 16$ and 32 , at which point the r_v^N/N time step in the totally discrete schemes becomes comparable to the $1/32$ time step used in the integration of the resulting approximating ordinary differential equation in the semi-discrete schemes. In addition, as N increases, the number of observational data points, ρ^N used by the totally discrete schemes increases and becomes comparable to the 101 (N independent) data points used in the testing of the semi-discrete schemes in [6]. It is interesting to note that a reasonably good fit can be achieved using relatively few observations.

Example 5.1 (Banks, Burns, Cliff [6, Example S 2.2]). In this example we identify the time delay r in the scalar first order equation given by

$$(5.1) \quad \dot{x}(t) = .05x(t) - 4.0x(t-r) + u_{.1}(t)$$

with initial conditions

$$(5.2) \quad x(0) = 1.0, \quad x_0(s) = 1, \quad -r \leq s \leq 0$$

and output

$$(5.3) \quad y(t) = x(t).$$

Observational data was generated by using a true parameter value of $r^* = 1$. The initial estimate of the parameter was taken to be $r^{N,0} = .6$. In Table 5.1 for each N and each state approximation we give the final converged value for the parameter as returned by the routine ZXSSQ as a solution to the approximating parameter identification problem.

Based upon the numerical results discussed in [28], it is not surprising to find the performance of the SPL state approximation superior to that of the AVE.

TABLE 5.1

N	AVE	SPL
2	.976458	.982173
4	1.11242	.984818
8	1.08012	.984677
16	1.04227	.996628
32	1.10351	1.00126
$r^* = 1.0$		$r^* = 1.0$

Example 5.2. In this example we consider the state equation (5.1), initial data (5.2) and output (5.3) of Example 5.1

$$\begin{aligned}\dot{x}(t) &= .05x(t) - a_1x(t-r) + u_{.1}(t), \\ x(0) &= 1, \quad x_0(s) = 1, \quad -r \leq s \leq 0, \\ y(t) &= x(t),\end{aligned}$$

and identify the coefficient a_1 of the delay term and the delay r itself. The true values of the parameters were taken to be $a_1^* = 4.0$ and $r^* = 1$, respectively, with start-up values given by $a_1^{N,0} = 3.0$ and $r^{N,0} = .6$. Our results are summarized in Table 5.2.

TABLE 5.2

N	AVE		SPL	
2	did not converge		did not converge	
4	4.59759	1.20779	4.13681	.991267
8	did not converge		4.09309	.987206
16	4.17380	1.04557	4.02157	.996570
32	4.06641	1.02561	3.99287	1.00124
$a_1^* = 4.0$		$r^* = 1.0$	$a_1^* = 4.0$	$r^* = 1.0$

Example 5.3 (Banks, Burns, Cliff [6 Example O1.2]). In this example we identify the time delay r in the damped harmonic oscillator with delayed damping and delayed restoring force given by

$$(5.4) \quad \ddot{x}(t) + 36x(t) + 2.5\dot{x}(t-r) + 9.0x(t-r) = u_{.1}(t),$$

together with initial conditions and output given by

$$(5.5) \quad x(0) = 1, \quad \dot{x}(0) = 0,$$

$$(5.6) \quad x_0(s) = 1, \quad \dot{x}_0(s) = 0, \quad -r \leq s \leq 0,$$

and

$$(5.7) \quad y(t) = x(t),$$

respectively. The initial value problem (5.4), (5.5), (5.6), (5.7) can be written as an equivalent first order system:

$$\dot{X}(t) = \begin{bmatrix} 0 & 1 \\ -36 & 0 \end{bmatrix} X(t) + \begin{bmatrix} 0 & 0 \\ -9.0 & -2.5 \end{bmatrix} X(t-r) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_{.1}(t),$$

$$X(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad X_0(s) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad -r \leq s \leq 0,$$

$$y(t) = [1, 0]X(t), \quad \text{where } X(t) = \begin{bmatrix} x(t) \\ \dot{x}(t) \end{bmatrix}.$$

The true parameter value was taken to be $r^* = 1.0$ with start-up value given by $r^{N,0} = 1.2$. Our results for this example, which are given in Table 5.3 once again exhibit the fact that the SPL schemes are superior to the AVE.

TABLE 5.3

N	AVE	SPL
2	did not converge	1.05621
4	1.22407	1.18990
8	1.14306	.991904
16	1.03183	.998599
$r^* = 1.0$		$r^* = 1.0$

Example 5.4. Here we once again consider the state equation (5.4), initial conditions (5.5), (5.6) and output (5.7) and identify the coefficient of the restoring force term and the time delay. Written as an equivalent first order system, the state equation, initial conditions and output are given by

$$\dot{X}(t) = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \end{bmatrix} X(t) + \begin{bmatrix} 0 & 0 \\ -9.0 & -2.5 \end{bmatrix} X(t-r) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_{.1}(t),$$

$$X(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad X_0(s) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad -r \leq s \leq 0,$$

$$y(t) = [1, 0]X(t), \quad \text{where } X(t) = \begin{bmatrix} x(t) \\ \dot{x}(t) \end{bmatrix}.$$

The true parameter values were taken to be $\omega^* = 6.0$ and $r^* = 1.0$ with start-up values given by $\omega^{N,0} = 5.0$ and $r^{N,0} = 1.2$ respectively. Our results for this example are summarized in Table 5.4.

TABLE 5.4

N	AVE		SPL	
2	4.53647	1.16643	6.26975	1.00952
4	6.28624	.895982	6.34399	.921017
8	did not converge		6.05748	.985784
16	6.07952	1.04665	6.01031	.997449
$\omega^* = 6.0$		$r^* = 1.0$	$\omega^* = 6.0$	$r^* = 1.0$

In this example, as was the case in all multi-parameter, higher dimensional examples we studied, the SPL schemes performed far better than the AVE. In fact, even for large values of N , it was not uncommon for the SPL schemes to converge while the AVE schemes did not. In all examples studied, for N sufficiently large, the SPL based schemes would always produce a solution to the approximating parameter identification problem. Moreover, as N increased, the solutions to the approximating problems appeared to be converging to the true parameter values used to generate the observational data.

Acknowledgments. The author would like to thank the referees for their careful review of the manuscript and for the helpful comments and suggestions which they provided. The author would also like to thank Professor H. T. Banks for the time he found to discuss the results contained in this paper throughout their development.

REFERENCES

- [1] H. T. BANKS, *Approximation of nonlinear functional differential equation control systems*, J. Optim. Theory Appl., 29 (1979), pp. 383–408.
- [2] ———, *Identification of nonlinear delay systems using spline methods*, Proc. International Conference on Nonlinear Phenomena in the Mathematical Sciences, Univ. Texas, Arlington, June 16–20, 1980, Academic Press, to appear.
- [3] ———, *Parameter identification techniques for physiological control systems*, Lectures in Applied Math., 19, American Mathematical Society, Providence, RI, 1981.
- [4] H. T. BANKS AND J. A. BURNS, *An abstract framework for approximate solutions to optimal control problems governed by hereditary systems*, International Conference on Differential Equations, H. Antosiewicz, ed., Academic Press, New York, 1975, pp. 10–25.
- [5] ———, *Hereditary control problems: numerical methods based on averaging approximations*, this Journal, 16 (1978), pp. 169–208.
- [6] H. T. BANKS, J. A. BURNS AND E. M. CLIFF, *A comparison of numerical methods for identification and optimization problems involving control systems with delays*, Brown Univ. LCDS Tech. Rep. 79-7, 1979, Providence, RI.
- [7] ———, *Parameter estimation and identification for systems with delays*, this Journal, 19 (1981), pp. 791–828.
- [8] H. T. BANKS AND P. K. DANIEL LAMM, *Estimation of delays and other parameters in nonlinear functional differential equations*, this Journal, 21 (1983), pp. 895–915.
- [9] H. T. BANKS AND F. KAPPEL, *Spline approximations for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496–522.
- [10] H. T. BANKS AND K. KUNISCH, *An approximation theory for nonlinear partial differential equations with applications to identification and control*, Institute for Computer Applications in Science and Engineering Rep. 81-16, 1981, Hampton, VA, this Journal, 20 (1982), pp. 815–849.
- [11] H. T. BANKS AND I. G. ROSEN, *Approximation techniques for parameter estimation in hereditary control systems*, Proc. IEEE Conference on Decision and Control, Albuquerque, NM, December, 1980, pp. 741–743.
- [12] J. A. BURNS AND E. M. CLIFF, *Methods for approximating solutions to linear hereditary quadratic optimal control problems*, IEEE Trans. Automatic Control, 23 (1978), pp. 21–36.
- [13] J. A. BURNS AND P. D. HIRSCH, *A difference equation approach to parameter estimation for differential-delay equations*, Appl. Math. Comp., 7 (1980), pp. 281–311.
- [14] P. L. DANIEL, *Spline-based approximation methods for the identification and control of nonlinear functional differential equations*, Ph.D. thesis, Brown Univ., Providence, RI, June, 1981.
- [15] R. L. EHLE, *A-stable methods and Padé approximations to the exponential*, SIAM J. Math. Anal., 4 (1973), pp. 671–680.
- [16] L. E. EL'SGOL'TS, *Introduction to the Theory of Differential Equations with Deviating Arguments*, Holden-Day, San Francisco, 1966.
- [17] J. S. GIBSON, *Linear quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximations*, this Journal, 21 (1983), pp. 95–139.
- [18] R. HERSH AND T. KATO, *High-accuracy stable difference schemes for well-posed initial-value problems*, SIAM J. Numer. Anal., 16 (1979), pp. 670–682.

- [19] *International Mathematical and Statistical Libraries, Library 1*, Edition 6, 1977, ZXSSQ-1-6.
- [20] T. KATO, *Perturbation Theory for Linear Operators*, second edition, Springer-Verlag, New York, 1976.
- [21] S. G. KREIN, *Linear differential equations in Banach space*, Trans. Math. Monographs, 29, American Mathematical Society, Providence, RI, 1971.
- [22] K. KUNISCH, *Approximation schemes for the linear quadratic optimal control problem associated with delay equations*, this Journal, to appear.
- [23] A. PAZY, *Semigroups of linear operators and applications to partial differential equations*, Lecture Notes, 10, Mathematics Dept., Univ. Maryland, College Park, 1974.
- [24] D. REBER, *Approximation and optimal control of linear hereditary systems*, Ph.D. thesis, Brown Univ., Providence, RI, November, 1977.
- [25] ———, *A finite difference technique for solving optimization problems governed by linear functional differential equations*, J. Differential Equations, 32 (1979), pp. 192–232.
- [26] F. REISZ AND B. SZ.-NAGY, *Functional Analysis*, Ungar, New York, 1975.
- [27] I. G. ROSEN, *A discrete approximation framework for hereditary systems*, Ph.D. thesis, Brown Univ., Providence, RI, June 1980.
- [28] ———, *A discrete approximation framework for hereditary systems*, J. Differential Equations, 40 (1981), pp. 377–449.
- [29] ———, *Discrete approximation methods for parameter identification in delay systems*, Institute for Computer Applications in Science and Engineering, Rep. 81-36, 1981, Hampton, VA.
- [30] A. P. SAGE AND J. L. MELSA, *System Identification*, Academic Press, New York, 1971.
- [31] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962, Chapt. 8.3.

OPTIMAL STOCHASTIC SCHEDULING OF POWER GENERATION SYSTEMS WITH SCHEDULING DELAYS AND LARGE COST DIFFERENTIALS*

G. L. BLANKENSHIP† AND J.-L. MENALDI‡

Abstract. The optimal scheduling or unit commitment of power generation systems to meet a random demand involves the solution of a class of dynamic programming inequalities for the optimal cost and control law. We study the behavior of this optimality system in terms of two parameters: (i) a scheduling delay, e.g., the startup time of a generation unit; and (ii) the relative magnitudes of the costs (operating or starting) of different units. In the first case we show that under reasonable assumptions the optimality system has a solution for all values of the delay, and, as the delay approaches zero, that the solutions converge uniformly to those of the corresponding system with no delays. In the second case we show that as the cost of operating or starting a given machine increases relative to the costs of the other machines, there is a point beyond which the expensive machine is not used, except in extreme situations. We give a formula for the relative costs that characterize this point. Moreover, we show that as the relative cost of the expensive machine goes to infinity the optimal cost of the system including the expensive machine approaches the optimal cost of the system without the machine.

1. Introduction. Optimal scheduling of continuously evolving stochastic dynamical systems admitting costly, discrete state transitions as control actions involves the analysis of partial differential inequalities which constitute the dynamic programming optimality conditions for the problem. These are the “quasi-variational inequalities” (QVI’s) introduced for such problems by A. Bensoussan and J. L. Lions [1] [2]. While there is an extensive analytical theory for the existence, uniqueness, and regularity properties of the solutions of QVI’s, it is very difficult to describe the solutions and the associated optimal scheduling rules, i.e., the control laws, in any but the simplest cases. For this reason it is useful to examine the behavior of the solutions to QVI’s as a function of various parameters which have simple interpretations in specific settings.

In this paper we consider the problem of scheduling a collection of power generation machines to meet a random demand for power, that is, the “unit commitment” problem. There are positive startup and operating costs associated with each machine, and the scheduling problem is to commit the units and operate them (set their power output levels) to meet the demand at minimum cost. The “demand” is modeled here as a diffusion process. In § 3 we study the problem including scheduling delays in unit starting. (In power systems operations such delays correspond to the times for boiler reheating in steam turbine generators or crew travel times in manual start units [3].) In § 4 we consider the scheduling problem when some machines are much more expensive to start and/or operate than any of the other machines.

Under reasonable assumptions on the demand dynamics and the cost functions we show that the optimality system (the QVI’s) has a well-defined solution, cost and control policy, for all values of the scheduling delay, and, as the delay approaches zero, that the optimal cost converges uniformly to that of the corresponding system

* Received by the editors July 7, 1982, and in revised form August 15, 1982.

† Electrical Engineering Department, University of Maryland, College Park, Maryland 20742. The research of this author was supported in part by the U.S. Department of Energy under contract DE-AC01-79ET29244.

‡ Department of Mathematics, Wayne State University, Detroit, Michigan 48202. The research of this author was supported in part during a visit at the University of Maryland by the U.S. Department of Energy under contracts DE-AC01-79ET29244 and 01-80RA-50154; and in part by the Office of Sponsored Research Programs, Wayne State University.

with no delay. The results of M. Robin [4] and J.-L. Menaldi [5] [6] form the basis for our arguments. In § 4 we show that as the cost of starting and/or operating a designated machine increases relative to the costs of the other machines there is a point beyond which the expensive machine is not used, except in extreme situations. We give an inequality on the relative costs that characterizes this point. Moreover, we show that as the relative operating cost of the expensive machine goes to infinity the optimal cost of the system including the expensive machine approaches the optimal cost of the system excluding the machine.

Related work on the asymptotic analysis of QVI's in general and optimal scheduling problems in particular may be found in the papers [7]–[10] (among others). For the most part these are concerned with the asymptotic behavior as the noise intensity approaches zero, i.e., as the system dynamics reduce from stochastic to deterministic. The QVI's are, in such cases, singularly perturbed. The problems treated here are of a different type, although the case of large cost differentials has an order reduction effect in the asymptotic limit.

In [11] a result is given (Thm. 1.2, p. 192) which characterizes the optimal switching among alternatives in terms of a simple inequality on the costs. However, the problems considered in [11] do not include explicit costs for switching, and the methods used are quite different.

2. Problem statement and an existence result. Let (Ω, \mathcal{F}, P) be a probability space, $\{\mathcal{F}_t, t \geq 0\}$ a nondecreasing, right-continuous family of completed sub- σ -fields of \mathcal{F} , and let $w(t)$, $t \geq 0$, be a standard R^N -valued Brownian motion with respect to \mathcal{F}_t , $t \geq 0$.

Let $m \geq 1$ be the *number of machines*. Let $A = \{0, 1\}^m$ be the *set of schedules*. If $\mathbf{a} \in A$, and a_i is the i th element of \mathbf{a} , then $a_i = 0$ means machine i is down, and $a_i = 1$ means it is up. Let $\{\theta_j, j = 1, 2, \dots\}$ be an increasing sequence of stopping times with respect to \mathcal{F}_t which are convergent to infinity and which satisfy $\theta_{j+1} \geq \theta_j + h$, for each j and some $h \geq 0$, the *scheduling delay*. A *scheduling policy* $\mathbf{a}(t)$, $t \geq 0$, is an A -valued random process starting at $\mathbf{a} \in A$ and adapted to \mathcal{F}_t satisfying

$$(2.1) \quad \mathbf{a}(t) = \begin{cases} \mathbf{a}, & 0 \leq t < \theta_1, \\ \mathbf{a}^j, & \theta_j \leq t < \theta_{j+1}, \quad j = 1, 2, \dots \end{cases}$$

Let $\mathcal{A}_{h,\mathbf{a}}$ be the set of all *scheduling policies* starting at \mathbf{a} with delay h . These are the discrete controls for our system. The components $a_i(t)$, $i = 1, 2, \dots, m$, of $\mathbf{a}(t)$ are the unit commitment schedules for the individual machines. Let $[\check{p}_i, \hat{p}_i] \subset (0, \infty)$, $i = 1, \dots, m$, be the *output capacities* of the machines when up, and let $P = [\check{p}_1, \hat{p}_1] \times \dots \times [\check{p}_m, \hat{p}_m] \subset R^m$. Then $\mathbf{p} \in P$ is the vector of *power generations* from the ensemble of machines. The system control—the *power production*—is

$$(2.2) \quad \mathbf{v}(t) = \mathbf{a}(t) \circ \mathbf{p}(t) = \{a_j(t)p_j(t), j = 1, \dots, m\}$$

for $\mathbf{a}(t) \in \mathcal{A}_{h,\mathbf{a}}$, $p(\cdot): [0, \infty) \rightarrow P$. We have used the Schur product notation in (2.2). Let $P_0 = [0, \check{p}_1] \times \dots \times [0, \check{p}_m]$ be the output powers of the ensemble of machines including the possibility of shutdowns.

Now let $g(x, \mathbf{a})$, $\sigma(x, \mathbf{a})$ be two given functions on $R^N \times A$ into R^N and $R^N \otimes R^N$, respectively, which are Lipschitz continuous in x for each fixed \mathbf{a} , $g = (g_i)$, $\sigma = (\sigma_{ij})$,

$$(2.3) \quad \frac{\partial g_i}{\partial x_k}, \frac{\partial \sigma_{ij}}{\partial x_k} \in B(R^N), \quad i, j, k = 1, \dots, N \quad \forall \mathbf{a} \in A$$

where $B(R^N)$ is the set of R -valued, Borel measurable, bounded functions on R^N .

The R^N -valued diffusion $y(t) = y_x(t, \mathbf{a}(\cdot))$ with drift g and diffusion $\sigma\sigma^T$ characterizes the *demand* on the system. We permit the demand to depend on the schedule. Let $\{\mathbf{a}(t), t \geq 0\} \in \mathcal{A}_{h,\mathbf{a}}$ and

$$(2.4) \quad dy(t) = g[y(t), \mathbf{a}(t)] dt + \sigma[y(t), \mathbf{a}(t)] dw(t), \quad y(0) = x \in R^N, \quad t \geq 0.$$

The process y has continuous paths almost surely.

Let \mathcal{O} be a bounded subset of R^N and let $\bar{\mathcal{O}}$ be its closure. We denote by $\tau = \tau_x(\mathbf{a}(\cdot))$ the first *exit time* of $y_x(t, \mathbf{a}(\cdot))$ from \mathcal{O} . That is,

$$(2.5) \quad \tau_x(\mathbf{a}(\cdot)) = \inf \{t \geq 0: y_x(t, \mathbf{a}(\cdot)) \notin \bar{\mathcal{O}}\}$$

for each $\mathbf{a} \in A$ and $\mathbf{a}(\cdot) \in \mathcal{A}_{h,\mathbf{a}}$. (Recall $\mathbf{a} \in A$ is $\mathbf{a}(0)$ for $\mathbf{a}(\cdot) \in \mathcal{A}_{h,\mathbf{a}}$.) Let $\Gamma_{\mathbf{a}}$ be the set of *regular points of $\partial\mathcal{O}$ from \mathbf{a}* (cf. [12])

$$(2.6) \quad \Gamma_{\mathbf{a}} = \{x \in \partial\mathcal{O}: P(\tau_{x,\mathbf{a}} > 0) = 0\}$$

where \mathbf{a} is a constant scheduling policy. If $\{\theta_i, i = 1, 2, \dots\}$ is a sequence of stopping times, $\mathbf{a} \in A$, $\mathbf{a}(\cdot) \in \mathcal{A}_{h,\mathbf{a}}$, and $\{\mathbf{p}(t), t \geq 0\}$ is a P -valued process adapted to \mathcal{F}_t with right continuous trajectories (having left-hand limits), then $\{\mathbf{v}(t) = \mathbf{a}(t) \circ \mathbf{p}(t), t \geq 0\}$ is called an *admissible control*. We have

$$(2.7) \quad y(\tau, \mathbf{a}) \in \Gamma_{\mathbf{a}} \quad \text{a.s. on } \{\tau < \infty\} \quad \forall \mathbf{a} \in A.$$

The cost functional for the problem is defined as follows: let $f: R^N \times P_0 \rightarrow (0, \infty)$ be continuous; f is the *operating cost rate*. The *switching cost* $k: A \times A \rightarrow [0, \infty)$ is

$$(2.8) \quad k(\mathbf{a}, \mathbf{b}) = \sum_{j=1}^m k_j [b_j - a_j]^+, \quad \mathbf{a}, \mathbf{b} \in A$$

where $k_j \geq k_0 > 0$ for $j = 1, \dots, m$. The *cost* is

$$(2.9) \quad J_{x\mathbf{a}}(\mathbf{v}) = E_{x\mathbf{a}} \left\{ \int_0^\tau f[y(t, \mathbf{a}(\cdot)), \mathbf{v}(t)] e^{-\alpha t} dt + \sum_{i=1}^\infty k[\mathbf{a}(\theta_{i-1}), \mathbf{a}(\theta_i)] \mathbf{1}_{\theta_i < \tau} e^{-\alpha \theta_i} \right\}$$

where $\alpha > 0$ is the *discount factor*, $E_{x\mathbf{a}}\{\cdot\}$ is expectation over paths $y(t)$, $\mathbf{a}(t)$ starting in $x \in R^N$ and $\mathbf{a} \in A$, respectively, and $\theta_0 = 0$.

Problem statement. We wish to characterize the optimal cost

$$(2.10) \quad u_h(x, \mathbf{a}) = \inf \{J_{x\mathbf{a}}(\mathbf{v}): \mathbf{v} \text{ admissible}\}$$

as a function of the scheduling delay h and the relative costs $f(y, \mathbf{p} \circ \mathbf{a})/f(y, \mathbf{p} \circ \mathbf{b})$, $k(\mathbf{a}, \mathbf{b})/k(\mathbf{a}, \mathbf{c})$ for all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in A$.

The first question of interest is the existence of the optimal cost. Since the problem is possibly degenerate ($\det \sigma\sigma^T(x, \mathbf{a}) = 0$ for some x, \mathbf{a}) and irregular ($\Gamma_{\mathbf{a}}$ not closed), this is a potentially delicate issue. However, the results of [5] and [6] adapt to the present case with minor modifications. Since we are mainly interested in the qualitative features of the optimal scheduling problem, we shall present a minimal treatment of the existence question.

For each $\mathbf{a} \in A$ we associate the operators

$$(2.11) \quad \mathcal{L}_{\mathbf{a}} = -\frac{1}{2} \text{tr} [\sigma\sigma^T \partial_{xx}] - g \cdot \partial_x, \quad \tilde{\mathcal{L}}_{\mathbf{a}} = \mathcal{L}_{\mathbf{a}} + \alpha,$$

with the diffusion

$$(2.12) \quad dy^0(t) = g(y^0(t), \mathbf{a}) dt + \sigma(y^0(t), \mathbf{a}) dw(t).$$

(Here \mathbf{a} plays the role of a parameter.) Following [5] and [6] we use the integral formulation of $\hat{\mathcal{L}}_{\mathbf{a}}$; that is,

$$\hat{\mathcal{L}}_{\mathbf{a}}u(x, \mathbf{a}) \leq \hat{f}(x, \mathbf{a}) \quad \text{in } \bar{\mathcal{O}} - \Gamma_{\mathbf{a}}$$

if the process

$$(2.13) \quad x_t = \int_0^{\theta \wedge \tau^0} \hat{f}(y_x^0(s), \mathbf{a}) e^{-\alpha s} ds + u(y_x^0(t \wedge \tau^0), \mathbf{a}) e^{-\alpha(t \wedge \tau^0)}$$

is a \mathcal{F}_t -submartingale for each $x \in \bar{\mathcal{O}} - \Gamma_{\mathbf{a}}$.

Here

$$(2.14) \quad \hat{f}(x, \mathbf{a}) = \min \{f(x, \mathbf{p} \circ \mathbf{a}), \mathbf{p} \in P\}.$$

We shall also say that $\hat{\mathcal{L}}_{\mathbf{a}}u \leq \hat{f}$ in the martingale sense when (2.13) holds.

Define the operator M as

$$(2.15) \quad \begin{aligned} Mu(x, \mathbf{a}) = \min_{\mathbf{b} \neq \mathbf{a}} E_{x\mathbf{b}} \left\{ \int_0^{h \wedge \delta} f(y(t, \mathbf{b}), \mathbf{b}) e^{-\alpha t} dt \right. \\ \left. + k(\mathbf{a}, \mathbf{b}) + e^{-\alpha(h \wedge \tau)} u(y(h \wedge \tau, \mathbf{b}), \mathbf{b}) \right\}. \end{aligned}$$

If we set $\|v\| = \sup \{|v(x, \mathbf{a})|, x \in \bar{\mathcal{O}}, \mathbf{a} \in A\}$ for $v(\cdot, \mathbf{a})$ continuous on $\bar{\mathcal{O}}$, then M maps $C(\bar{\mathcal{O}})$ into itself and

$$(2.16) \quad \|Mu - Mv\| \leq e^{-\alpha h} \|u - v\|,$$

if $u(x, \mathbf{a}) = 0 = v(x, \mathbf{a})$ for all $x \in \Gamma_{\mathbf{a}}, \mathbf{a} \in A$.

The problem (2.10) can be formulated as follows.

Find a real, bounded, measurable function $u(x, \mathbf{a})$ on $\bar{\mathcal{O}} \times A$ such that

$$(2.17) \quad \begin{aligned} u &= 0 \quad \text{on } \Gamma_{\mathbf{a}}, \quad \forall \mathbf{a} \in A, \\ u &\leq Mu \quad \text{in } \bar{\mathcal{O}} - \Gamma_{\mathbf{a}}, \quad \forall \mathbf{a} \in A, \\ \hat{\mathcal{L}}_{\mathbf{a}}u &\leq \hat{f} \quad \text{in the martingale sense on } \bar{\mathcal{O}} - \Gamma_{\mathbf{a}}, \quad \forall \mathbf{a} \in A. \end{aligned}$$

We can reformulate the problem (2.10) or (2.17) as a quasi-variational inequality along the lines in [6, p. 724], but this takes us somewhat away from our main line of inquiry, and so, we will omit it.

We associate with (2.10) a sequence of stopping time problems as follows. Let

$$(2.18) \quad \hat{u}^0(x, \mathbf{a}) = E_{x\mathbf{a}} \left\{ \int_0^{\tau} \hat{f}(y(t, \mathbf{a}), \mathbf{a}) e^{-\alpha t} dt \right\}.$$

Given $\hat{u}^{n-1}(x, \mathbf{a})$, define $\hat{u}^n(x, \mathbf{a})$ by

$$(2.19) \quad \hat{u}^n(x, \mathbf{a}) = \inf_{\theta \geq 0} E_{x\mathbf{a}} \left\{ \int_0^{\theta \wedge \tau} \hat{f}(y_x(t, \mathbf{a}), \mathbf{a}) e^{-\alpha t} dt + \mathbf{1}_{\theta < \tau} e^{-\alpha \theta} M \hat{u}^{n-1}(y_x(\theta, \mathbf{a}), \mathbf{a}) \right\}.$$

Note that $\mathbf{a} \in A$ is a parameter in (2.18), (2.19). In abstract terms (2.18), (2.19) takes the following equivalent form: let $\hat{u}^0(x, \mathbf{a})$ be the bounded, continuous, nonnegative real function on $\bar{\mathcal{O}} \times A$ such that

$$(2.20) \quad \begin{aligned} \hat{u}^0(x, \mathbf{a}) &= 0 \quad \forall x \in \Gamma_{\mathbf{a}}, \quad \forall \mathbf{a} \in A, \\ \hat{\mathcal{L}}_{\mathbf{a}}\hat{u}^0 &= \hat{f} \quad \text{in the martingale sense on } \bar{\mathcal{O}} - \Gamma_{\mathbf{a}}, \quad \forall \mathbf{a} \in A, \end{aligned}$$

and, given \hat{u}^{n-1} , let \hat{u}^n be the bounded, continuous, nonnegative real function on $\bar{\mathcal{O}}$ which is the maximum solution of

$$(2.21) \quad \begin{aligned} u^n(x, \mathbf{a}) &= 0 \quad \forall x \in \Gamma_{\mathbf{a}} \quad \forall \mathbf{a} \in A, \\ u^n &\leq M\hat{u}^{n-1} \quad \text{in } \bar{\mathcal{O}}\Gamma_{\mathbf{a}}, \\ \mathcal{L}_{\mathbf{a}}u^n &\leq \hat{f} \quad \text{in the martingale sense on } \bar{\mathcal{O}} - \Gamma_{\mathbf{a}} \quad \forall \mathbf{a} \in A. \end{aligned}$$

The sequence of variational inequalities corresponds, in effect, to the sequence of stopping time problems—make n optimal decisions, startup or shutdown, and then stop.

LEMMA 2.1. *Under the stated hypotheses on g , σ , f , and k the problem (2.17) admits a maximum solution \hat{u} which is upper semicontinuous and given as the optimal cost in (2.10). Moreover,*

$$(2.22) \quad 0 \leq \hat{u}^{n+1} \leq \hat{u}^n \leq \dots \leq \hat{u}^0 \leq \frac{1}{\alpha} \|\hat{f}\| \quad \forall n = 1, 2, \dots,$$

$$(2.23) \quad 0 \leq \hat{u} = \hat{u}^n \leq \left[\frac{\exp(-n\alpha h)}{1 - \exp(-\alpha h)} \right] \|\hat{u}^1 - \hat{u}^0\|,$$

and if the set of regular points $\Gamma_{\mathbf{a}}$ is closed, then

$$(2.24) \quad \hat{u}^n(\cdot, \mathbf{a}), \hat{u}(\cdot, \mathbf{a}) \in C.$$

Proof. The first two results (2.22) and (2.23) follow from simple modifications of the arguments in Robin [4, pp. 279–283]. The third result (2.24) follows from the arguments in [5]. QED

THEOREM 2.1. *Under the stated hypotheses on f , g , σ , and k and the assumption of regularity ($\Gamma_{\mathbf{a}}$ closed $\forall \mathbf{a}$) there exists an optimal, admissible control policy.*

Proof. First, note that $\hat{u}(x, \mathbf{a})$ constructed as the limit of the sequence (2.19) via (2.23) satisfies the problem

$$(2.25) \quad \hat{u}(x, \mathbf{a}) = \inf_{\tau \geq 0} E_{x, \mathbf{a}} \left\{ \int_0^{\theta \wedge \tau} \hat{f}(y(t, \mathbf{a}), \mathbf{a}) e^{-\alpha t} dt + \mathbf{1}_{\theta < \tau} e^{-\alpha \theta} M\hat{u}(y(\theta, \mathbf{a}), \mathbf{a}) \right\}.$$

Let $\hat{b}(x, \mathbf{a})$ be defined by

$$(2.26) \quad \hat{b}(x, \mathbf{a}) = \arg \min_{\mathbf{b} \neq \mathbf{a}} \left[k(\mathbf{a}, \mathbf{b}) + E_{x, \mathbf{b}} \left\{ \int_0^{h \wedge \tau} \hat{f}(y(t, \mathbf{b}), \mathbf{b}) e^{-\alpha t} dt + e^{-\alpha(h \wedge \tau)} u(y(h \wedge \tau, \mathbf{b}), \mathbf{b}) \right\} \right]$$

with \hat{b} Borel measurable in $\bar{\mathcal{O}} \times A$.

The optimal policy $\hat{\mathbf{a}}(\cdot)$ is defined by

$$(2.27) \quad \hat{\mathbf{a}}(\cdot) = \mathbf{a} \mathbf{1}_{[0, \hat{\theta}_1)} + \sum_{i=1}^{\infty} \hat{\mathbf{a}}^i \mathbf{1}_{[\hat{\theta}_i, \hat{\theta}_{i+1})}$$

with the values $\hat{\mathbf{a}}^i$ selected as follows. Let

$$(2.28) \quad \hat{\theta}^0 = 0,$$

$$(2.29) \quad d\hat{y}^0(t) = g(\hat{y}^0(t), \mathbf{a}) dt + \sigma(\hat{y}^0(t), \mathbf{a}) dw(t), \quad \hat{y}^0(0) = x, \quad t \geq 0,$$

and for $i = 0, 1, 2, \dots$

$$(2.30) \quad \hat{\tau}^i = \begin{cases} \inf \{t \geq 0: \hat{y}^i(t) \in \mathcal{O}\} \\ \infty \text{ if the set is empty,} \end{cases}$$

$$(2.31) \quad \hat{\theta}_{i+1} = \begin{cases} \inf \{t \in [\tilde{\theta}^i, \hat{\tau}^i): \hat{y}^i(t) \notin \{\hat{u}(\cdot, \hat{\mathbf{a}}^i) < M\hat{u}(\cdot, \hat{\mathbf{a}}^i)\}\} \\ \infty \text{ if the set is empty,} \end{cases}$$

$$(2.32) \quad \tilde{\theta}^i = (\hat{\theta}_i + h) \wedge \hat{\tau}^i, \quad i = 1, 2, \dots,$$

$$(2.33) \quad \hat{\mathbf{a}}^{i+1} = \begin{cases} \hat{b}(\hat{y}^i(\hat{\theta}_{i+1}), \hat{\mathbf{a}}^i), & \text{if } \hat{\theta}_{i+1} < \infty \\ \hat{\mathbf{a}}^i, & \text{otherwise, } i = 0, 1, 2, \dots, \end{cases}$$

and

$$(2.34) \quad \begin{aligned} d\hat{y}^i(t) &= g(\hat{y}^i(t), \hat{\mathbf{a}}^i) dt + \sigma(\hat{y}^i(t), \hat{\mathbf{a}}^i) dw(t), \quad t \geq \hat{\theta}_i, \\ \hat{y}^i(t) &= \hat{y}^{i-1}(t), \quad t \leq \hat{\theta}_i, \quad i = 1, 2, \dots \end{aligned}$$

Using the Markov property, we have

$$(2.35) \quad \begin{aligned} u(x, \mathbf{a}) &= E \left\{ \int_0^{\hat{\theta}_n \wedge \hat{\tau}^{n-1}} \hat{f}(\hat{y}(t), \hat{\mathbf{a}}(t)) e^{-\alpha t} dt + \sum_{i=1}^n k(\hat{\mathbf{a}}^{i-1}, \hat{\mathbf{a}}^i) e^{-\alpha \hat{\theta}_i} \mathbf{1}_{\hat{\theta}_i < \hat{\tau}} \right\} \\ &\quad + E \{ e^{-\alpha \hat{\theta}_n} \hat{u}(\hat{y}(\hat{\theta}_n), \hat{\mathbf{a}}^n) \}, \end{aligned}$$

where

$$(2.36) \quad \hat{y}(t) = y(t, \hat{\mathbf{a}}(\cdot)) = \hat{y}^n(t) \quad \forall t \in [0, \hat{\theta}_n].$$

Since \hat{u} is bounded and $\hat{\theta} \rightarrow \infty$ (a.s.) as $n \rightarrow \infty$, we obtain

$$(2.37) \quad u(x, \mathbf{a}) = E \left\{ \int_0^{\hat{\tau}} \hat{f}(\hat{y}(t), \hat{\mathbf{a}}(t)) e^{-\alpha t} dt + \sum_{i=1}^{\infty} k(\hat{\mathbf{a}}^{i-1}, \hat{\mathbf{a}}^i) e^{-\alpha \hat{\theta}_i} \mathbf{1}_{\hat{\theta}_i < \hat{\tau}} \right\}.$$

Finally, let $\hat{\mathbf{p}}(x, \mathbf{a})$ measurable be such that

$$(2.38) \quad f(x, \mathbf{a} \circ \hat{\mathbf{p}}(x, \mathbf{a})) = \hat{f}(x, \mathbf{a}) \quad \forall (x, \mathbf{a}) \in \bar{\mathcal{O}} \times A.$$

Defining

$$(2.39) \quad \hat{\mathbf{p}}(t) = \hat{\mathbf{p}}(y(t), \hat{a}), \quad \hat{\mathbf{v}}(t) = \hat{\mathbf{p}}(t) \circ \hat{\mathbf{a}}(t)$$

completes the proof. QED

Remark. The function $\hat{u}^n(x, \mathbf{a})$ is the optimal cost given that n switchings are permitted.

3. Dependence of the cost on the scheduling delay. In this section we shall show that the optimal scheduling cost depends continuously on the delay h as $h \rightarrow 0$, if the hypotheses of Theorem 2.1 hold. To emphasize the dependence on h , let $\hat{u}_h(x, \mathbf{a})$ be the optimal cost in the problem (2.10), and let \hat{u}_h^n be the costs in the sequence (2.19). Also, let

$$(3.1) \quad M_h v(x, \mathbf{a}) = \min_{\mathbf{b} \neq \mathbf{a}} \left[k(\mathbf{a}, \mathbf{b}) + E_{x\mathbf{b}} \left\{ \int_0^{h \wedge \tau} \hat{f}(y(t, \mathbf{b}), \mathbf{b}) e^{-\alpha t} dt + e^{-\alpha(h \wedge \tau)} v(y(h, \mathbf{b}), \mathbf{b}) \right\} \right]$$

and

$$(3.2) \quad Mv(x, \mathbf{a}) = \min_{\mathbf{b} \neq \mathbf{a}} [k(\mathbf{a}, \mathbf{b}) + v(x, \mathbf{a})].$$

LEMMA 3.1. *Under the stated hypotheses on f , g , σ and k and the assumption of regularity we have*

$$(3.3) \quad \|M_h v - M v\| \leq \left(\frac{1 - e^{-\alpha h}}{\alpha} \right) \|\hat{f}\| + (1 - e^{-\alpha h}) \|v\|$$

for all Borel measurable v such that

$$(3.4) \quad v(x, \mathbf{a}) = 0 \quad \forall x \in \Gamma_{\mathbf{a}}, \quad \forall \mathbf{a} \in A.$$

Proof. This follows immediately from (3.1) (3.2). QED

LEMMA 3.2. *Under the conditions of Lemma 3.1 we have*

$$(3.5) \quad \|\hat{u}_h^n - \hat{u}_h\| \leq \frac{c}{n - m}, \quad n = m + 1, \quad m + 2, \dots$$

for some constant c independent of h and where m is the number of machines.

Proof. Let

$$(3.6) \quad \begin{aligned} N^+[0, T] &= \text{number of machine starts in } [0, T], \\ N^-[0, T] &= \text{number of machine shutdowns in } [0, T] \end{aligned}$$

associated with a policy $\mathbf{a}(t)$, $t \in [0, T]$. For each $T > 0$

$$(3.7) \quad N^+[0, T] + N^-[0, T] \leq 2N^+[0, T] + m.$$

Given any policy $\mathbf{a}(\cdot)$, let $\mathbf{a}^n(\cdot)$ be the policy whose first n switchings coincide with those of $\mathbf{a}(\cdot)$ and which is constant (at the value of the n th switching) throughout the remainder of the interval. Using the notation \hat{u}_h^n introduced earlier, we have

$$(3.8) \quad 0 \leq \hat{u}_h^n - \hat{u}_h \leq \sup_{\mathbf{a}(\cdot)} E \left\{ \int_{\theta_n \wedge \tau}^T \hat{f}(y^n(t), \mathbf{a}^n(t)) e^{-\alpha t} dt \right\},$$

and so,

$$(3.9) \quad 0 \leq \hat{u}_h^n - \hat{u}_h \leq \|\hat{f}\| \sup_{\mathbf{a}(\cdot)} E \{ e^{-\alpha(\theta_n \wedge \tau)} \mathbf{1}_{\theta_n < \tau} \}.$$

To estimate the expectation, we use (3.7) and an observation about the startup cost.

Clearly, for any $T > 0$

$$(3.10) \quad k_0 N^+[0, T] e^{-\alpha T} \leq \sum_{0 \leq t \leq T} k[\mathbf{a}(t-), \mathbf{a}(t)] e^{-\alpha t}.$$

Now from (3.7) for a policy with n switching

$$(3.11) \quad n \leq 2N^+[0, \theta_n] + m.$$

Using this in (3.10) with $T = \theta_n \wedge \tau$, we have

$$(3.12) \quad \frac{1}{2} k_0 (n - m) E \{ e^{-\alpha(\theta_n \wedge \tau)} \mathbf{1}_{\theta_n < \tau} \} \leq \sup_{\mathbf{a}(\cdot)} E \left\{ \sum_{i=1}^{\infty} k[\mathbf{a}(\theta_{i-1}), \mathbf{a}(\theta_i)] e^{-\alpha(\theta_i \wedge \tau)} \right\}.$$

To estimate the term on the right, recall from (2.9), the form of the cost of an admissible policy. Consider a suboptimal policy which involves no switching. The cost of such a policy is bounded above by $\|\hat{f}\|/\alpha$. It follows from (2.9) that we can restrict attention to policies in which

$$(3.13) \quad E \left\{ \sum_{0 \leq t \leq \tau} k[\mathbf{a}(t-), \mathbf{a}(t)] e^{-\alpha t} \right\} \leq \|\hat{f}\|/\alpha.$$

Using this on the right in (3.12), and the result in (3.9), we have the desired inequality (3.5) with

$$(3.14) \quad c = 2\|\hat{f}\|^2/\alpha k_0. \quad \text{QED}$$

The bound (3.5) and an inequality of Robin give us the desired continuity result. Let $\hat{u}_0(x, \mathbf{a})$ be the optimal cost in the system (2.10) with no delay, i.e., $h = 0$.

THEOREM 3.1. *Under the stated hypotheses on f , g , σ , and k and the assumption of regularity we have*

$$(3.15) \quad \lim_{h \downarrow 0} \hat{u}_h = \hat{u}_0 \quad \text{uniformly in } \bar{\mathcal{O}} \times A.$$

Proof. Let \hat{u}_0^n be the optimal cost in the problem (2.10) with no delay over admissible policies having at most n switchings. Then

$$(3.16) \quad \|\hat{u}_h - \hat{u}_0\| \leq \|\hat{u}_h - \hat{u}_h^n\| + \|\hat{u}_h^n - \hat{u}_0^n\| + \|\hat{u}_0^n - \hat{u}_0\|.$$

The first and third terms on the right may be bounded using Lemma 3.1. A bound for the second term is given in Robin's thesis [4], p. 235,

$$(3.17) \quad \|\hat{u}_h^n - \hat{u}_0^n\| \leq 2nh\|\hat{f}\|.$$

It follows that for any $n \geq m + 1$

$$(3.18) \quad \|\hat{u}_h - \hat{u}_0\| \leq (4\|\hat{f}\|^2/\alpha k_0)/(n - m) + 2\|\hat{f}\|nh.$$

Thus, taking $h \downarrow 0$ and then $n \rightarrow \infty$ leads to the desired result. QED

4. Scheduling with some expensive machines. Now suppose that one machine, or more generally, a group of machines is much more expensive to operate and/or start than the remaining machines. One would expect that the expensive machine would be used only in extreme circumstances, or not at all when its cost is very high. We show that the problem (2.10) has these properties.

Let $\mathbf{a}^e \in A$ be an "expensive" schedule. Recall the notation $\hat{f}(x, \mathbf{a}) = \min \{f(x, \mathbf{p} \circ \mathbf{a}), \mathbf{p} \in P\}$.

LEMMA 4.1. *Under the stated hypotheses on f , g , σ and k and the assumption of regularity the inequality*

$$(4.1) \quad \hat{f}(x, \mathbf{a}^e) > \hat{f}(x, \mathbf{b}) + [k(\mathbf{a}, \mathbf{b}) - k(\mathbf{a}, \mathbf{a}^e)] \left(\frac{\alpha}{1 - e^{-\alpha h}} \right) \quad \forall x \in \bar{\mathcal{O}}, \mathbf{a} \in A, \mathbf{b} \in A - \{\mathbf{a}^e\}$$

implies that the optimal policy $\hat{\mathbf{a}}(t)$, $0 \leq t \leq \hat{\tau}$, defined in Theorem 2.1 for the problem (2.10) has the property

$$(4.2) \quad P\{\hat{\mathbf{a}}(t) = \mathbf{a}^e, 0 \leq t \leq \hat{\tau} - h\} = 0.$$

Remark. In other words, the optimal policy $\hat{\mathbf{a}}(\cdot)$ switches to the expensive schedule only near the boundary Γ , and in that case, it switches once and then stops.

Proof. Let $\mathbf{a}(\cdot)$ be the optimal policy. Over an interval $[\hat{\theta}_i, \hat{\theta}_{i+1})$ the optimal cost increases by the amount

$$(4.3) \quad \Delta J_i = \int_{\hat{\theta}_i}^{\hat{\theta}_{i+1}} \hat{f}(\hat{y}(t), \hat{\mathbf{a}}^i) e^{-\alpha t} dt + k(\hat{\mathbf{a}}^{i-1}, \hat{\mathbf{a}}^i) e^{-\alpha \hat{\theta}_i}.$$

Suppose that $\hat{\mathbf{a}}^{i-1} \neq \mathbf{a}^e$ and $\hat{\mathbf{a}}^i = \mathbf{a}^e$. Then by (4.1) for any $\mathbf{b} \in A - \{\mathbf{a}^e\}$ we have

$$(4.4) \quad \Delta J_i > \int_{\hat{\theta}_i}^{\hat{\theta}_{i+1}} \hat{f}(\hat{y}(t), \mathbf{b}) e^{-\alpha t} dt + \left(\frac{1}{1 - e^{-\alpha h}} \right) (e^{-\alpha \hat{\theta}_i} - e^{-\alpha \hat{\theta}_{i+1}}) [k(\hat{\mathbf{a}}^{i-1}, \mathbf{b}) - k(\hat{\mathbf{a}}^{i-1}, \mathbf{a}^e)] + k(\hat{\mathbf{a}}^{i-1}, \mathbf{a}^e) e^{-\alpha \hat{\theta}_i}.$$

Since $\hat{\theta}_{i+1} \geq h + \hat{\theta}_i$ if $0 \leq t \leq \hat{\tau} - h$, it follows that

$$(4.5) \quad \Delta J_i > \int_{\hat{\theta}_i}^{\hat{\theta}_{i+1}} \hat{f}(\hat{y}(t), \mathbf{b}) e^{-\alpha t} dt + k(\hat{\mathbf{a}}^{i-1}, \mathbf{b}) e^{-\alpha \hat{\theta}_i},$$

and this must have probability zero since $\hat{\mathbf{a}}(t)$ is optimal. QED

Note that either a large “operating cost rate” $\hat{f}(x, \mathbf{a}^e)$ or a large “startup cost” $k(\mathbf{a}, \mathbf{a}^e)$ will cause (4.1) to be satisfied. Now suppose that the operating cost $\hat{f}(x, \mathbf{a}^e)$ becomes arbitrarily large

$$(4.6) \quad \hat{f}(x, \mathbf{a}^e) \geq 1/\varepsilon, \quad x \in \bar{\mathcal{O}}, \quad \varepsilon > 0 \text{ small.}$$

One would expect that in the limit as $\varepsilon \downarrow 0$ that the expensive state \mathbf{a}^e will never be used. We shall treat the cases $h > 0$ and $h = 0$ separately. Let

$$(4.7) \quad \hat{u}_h^\varepsilon(x, \mathbf{a}) = \inf \{J_{x\mathbf{a}}(\mathbf{v}): \mathbf{v}(t) = \mathbf{p}(t) \circ \mathbf{a}(t), \mathbf{a}(t) \neq \mathbf{a}^e, 0 \leq t \leq \hat{\tau}, h > 0\},$$

$$(4.8) \quad \hat{u}_0^\varepsilon(x, \mathbf{a}) = \inf \{J_{x\mathbf{a}}(\mathbf{v}): \mathbf{v}(t) = \mathbf{p}(t) \circ \mathbf{a}(t), \mathbf{a}(t) \neq \mathbf{a}^e, 0 \leq t \leq \hat{\tau}, h = 0\}.$$

Let

$$(4.9) \quad \hat{u}_h^\varepsilon(x, \mathbf{a}) = \inf \{J_{x\mathbf{a}}(\mathbf{v}): \mathbf{v}(\cdot) \text{ admissible, (4.6) holds, } h > 0\},$$

$$(4.10) \quad \hat{u}_0^\varepsilon(x, \mathbf{a}) = \inf \{J_{x\mathbf{a}}(\mathbf{v}): \mathbf{v}(\cdot) \text{ admissible, (4.6) holds, } h = 0\}.$$

THEOREM 4.1. *Under the stated hypotheses on f , g , σ and k , the assumption of regularity then*

$$(4.11) \quad \lim_{\varepsilon \downarrow 0} \hat{u}_h^\varepsilon = \hat{u}_h^\varepsilon \quad \text{uniformly in } x \in \bar{\mathcal{O}}, \quad h > 0.$$

If, in addition,

$$(4.12) \quad g = g(x), \quad \sigma = \sigma(x),$$

independent of $\mathbf{a} \in A$ then

$$(4.13) \quad \lim_{\varepsilon \downarrow 0} \hat{u}_0^\varepsilon = \hat{u}_0^\varepsilon \quad \text{uniformly in } x \in \bar{\mathcal{O}}, \quad h = 0.$$

Proof of (4.11) $h > 0$.

Suppose $\varepsilon > 0$ is small enough so that

$$(4.14) \quad \frac{1}{\varepsilon} > \hat{f}(x, \mathbf{b}) + [k(\mathbf{a}, \mathbf{b}) - k(\mathbf{a}, \mathbf{a}^e)] \left(\frac{\alpha}{1 - e^{-\alpha h}} \right) \quad \forall x \in \bar{\mathcal{O}}, \quad \mathbf{a} \in A, \quad \mathbf{b} \in A - \{\mathbf{a}^e\}.$$

Using (2.32) from the proof of Theorem 2.1, we have

$$(4.15) \quad \begin{aligned} \hat{u}_h^\varepsilon(x, \mathbf{a}) = & E \left\{ \int_0^{\hat{\theta}_n \wedge \hat{\tau}^{n+1}} \hat{f}(\hat{y}(t, \hat{\mathbf{a}}(t)), \hat{\mathbf{a}}(t)) e^{-\alpha t} dt + \sum_{i=1}^n k(\hat{\mathbf{a}}^{i-1}, \hat{\mathbf{a}}^i) e^{-\alpha \hat{\theta}_i} \right\} \\ & + E \{ e^{-\alpha \hat{\theta}_n} \hat{u}_h^\varepsilon(\hat{y}(\hat{\theta}_n, \hat{\mathbf{a}}^n), \hat{\mathbf{a}}^n) \}. \end{aligned}$$

Now for any admissible $\nu(\cdot) = \mathbf{p}(\cdot) \circ \mathbf{a}(\cdot)$ with $\mathbf{a}(t) \neq \mathbf{a}^e$, $t \geq 0$, we have

$$(4.16) \quad 0 \leq \hat{u}_h^e - \hat{u}_h^\varepsilon \leq J_{x\mathbf{a}}(\nu) - \hat{u}_h^\varepsilon(x, \mathbf{a}).$$

By Lemma 4.1 we can consider policies $\hat{\mathbf{a}}(t)$ which switch to \mathbf{a}^e in $[\hat{\tau} - h, \hat{\tau}]$ and $\hat{\mathbf{a}}(t) \neq \mathbf{a}^e$, $0 \leq t < \hat{\tau} - h$. It follows that

$$(4.17) \quad 0 \leq \hat{u}_h^e - \hat{u}_h^\varepsilon \leq \sup_{\mathbf{b} \neq \mathbf{a}^e} E \left\{ \int_{\hat{\tau}-h}^{\hat{\tau}} \hat{f}(\hat{\mathbf{y}}(\mathbf{b}), \mathbf{b}) e^{-\alpha t} dt \right\}.$$

Hence,

$$(4.18) \quad 0 \leq \hat{u}_h^e - \hat{u}_h^\varepsilon \leq (\sup_{\mathbf{b} \neq \mathbf{a}^e} \|\hat{f}(\cdot, \mathbf{b})\|)(e^{\alpha h} - 1)/\alpha.$$

But this can be improved.

Let

$$(4.19) \quad n_h = \inf \{n \geq 1: \hat{\theta}_n \geq \hat{\tau} - h\}, \quad \hat{\theta}_h = \hat{\theta}_{n_h} \quad \text{for } n = n_h.$$

Then we can replace $\hat{\tau} - h$ by $\hat{\theta}_h$ in (4.17). Since a switching to \mathbf{a}^e is assumed to occur in $[\hat{\tau} - h, \hat{\tau}]$ and since (4.6) and (4.14) hold, we have

$$(4.20) \quad 0 \leq \sup_{\mathbf{b} \neq \mathbf{a}^e} \left(E \left\{ \int_{\hat{\theta}_h}^{\hat{\tau}} \hat{f}(\hat{\mathbf{y}}(\mathbf{b}), \mathbf{b}) e^{-\alpha t} dt - \frac{1}{\varepsilon} \int_{\hat{\theta}_h}^{\hat{\tau}} e^{-\alpha t} dt - k(\mathbf{b}, \mathbf{a}^e) e^{-\alpha \hat{\theta}_h} \right\} \right).$$

This implies

$$(4.21) \quad \frac{1}{\varepsilon} E \int_{\hat{\theta}_h}^{\hat{\tau}} e^{-\alpha t} dt \leq \frac{1}{\alpha} \sup_{\mathbf{b} \neq \mathbf{a}^e} \|\hat{f}(\cdot, \mathbf{b})\|.$$

Using this in (4.17) (4.18) with $\hat{\theta}_h$ replacing $\hat{\tau} - h$ leads to

$$(4.22) \quad 0 \leq \hat{u}_h^e - \hat{u}_h^\varepsilon \leq \frac{1}{\alpha} (\sup_{\mathbf{b} \neq \mathbf{a}^e} \|\hat{f}(\cdot, \mathbf{b})\|)^2 \varepsilon.$$

Note that (4.22) and (4.18) hold only if ε and h satisfy (4.14); that is, they are not uniform. In any event (4.22) implies (4.11). QED

Proof of (4.13). $h = 0$. Let $\mathbf{a}(t)$ be the optimal policy associated with \hat{u}_0^ε . It exists by virtue of property (3.7). Write it as

$$(4.23) \quad \mathbf{a}(t) = \mathbf{a} \mathbf{1}_{[0, \theta_1)} + \sum_{i=1}^{\infty} \mathbf{a}^i \mathbf{1}_{[\hat{\theta}_i, \hat{\theta}_{i+1})}$$

where $\hat{\theta}_i$ were defined in the proof of Theorem 2.1 (with h set to zero). We define by induction

$$(4.24) \quad \tilde{\mathbf{a}}^i = \begin{cases} \mathbf{a}^i & \text{if } \mathbf{a}^i \neq \mathbf{a}^e, \\ \mathbf{a}^{i-1} & \text{if } \mathbf{a}^i = \mathbf{a}^e, \end{cases}$$

and if $\mathbf{a} \neq \mathbf{a}^e$

$$(4.25) \quad \tilde{\mathbf{a}}(t) = \mathbf{a} \mathbf{1}_{[0, \hat{\theta}_1)} + \sum_{i=1}^{\infty} \tilde{\mathbf{a}}^i \mathbf{1}_{[\hat{\theta}_i, \hat{\theta}_{i+1})}.$$

If $\mathbf{a} \neq \mathbf{a}^e$, then

$$(4.26) \quad J_{\mathbf{x}\mathbf{a}}(\hat{\mathbf{a}}(\cdot)) = E_x \left\{ \int_0^{\hat{\theta}_1 \wedge \hat{\tau}} \hat{f}(\hat{y}(t), \mathbf{a}) e^{-\alpha t} dt + \sum_{i=1}^{\infty} \int_{\hat{\theta}_i \wedge \hat{\tau}}^{\hat{\theta}_{i+1} \wedge \hat{\tau}} \hat{f}(\hat{y}(t), \hat{\mathbf{a}}^i) e^{-\alpha t} dt \right. \\ \left. + k(\mathbf{a}, \hat{\mathbf{a}}^1) e^{-\alpha(\hat{\theta}_1 \wedge \hat{\tau})} + \sum_{i=1}^{\infty} k(\hat{\mathbf{a}}^i, \hat{\mathbf{a}}^{i+1}) e^{-\alpha(\hat{\theta}_{i+1} \wedge \hat{\tau})} \right\}.$$

Since $k(\mathbf{b}, \mathbf{b}) = 0$ and $k(\mathbf{a}, \mathbf{b}) + k(\mathbf{b}, \mathbf{c}) \geq k(\mathbf{a}, \mathbf{c})$ for each \mathbf{a}, \mathbf{b} , and \mathbf{c} in A , we have

$$(4.27) \quad J_{\mathbf{x}\mathbf{a}}(\hat{\mathbf{a}}(\cdot)) \geq E_x \left\{ \int_0^{\hat{\theta}_1 \wedge \hat{\tau}} \hat{f}(\hat{y}(t), \mathbf{a}) e^{-\alpha t} dt + \sum_{i=1}^{\infty} \int_{\hat{\theta}_i \wedge \hat{\tau}}^{\hat{\theta}_{i+1} \wedge \hat{\tau}} \hat{f}(\hat{y}(t), \hat{\mathbf{a}}^i) e^{-\alpha t} dt \right. \\ \left. + k(\mathbf{a}, \hat{\mathbf{a}}^1) e^{-\alpha(\hat{\theta}_1 \wedge \hat{\tau})} + \sum_{i=1}^{\infty} k(\hat{\mathbf{a}}^i, \hat{\mathbf{a}}^{i+1}) e^{-\alpha(\hat{\theta}_{i+1} \wedge \hat{\tau})} \right\}.$$

Using this and the fact that $\hat{\mathbf{a}}(\cdot)$ is optimal leads to

$$(4.28) \quad 0 \leq J_{\mathbf{x}\mathbf{a}}(\tilde{\mathbf{a}}(\cdot)) - J_{\mathbf{x}\mathbf{a}}(\hat{\mathbf{a}}(\cdot)) \leq \sum_{\mathbf{a}^i = \mathbf{a}^e} E \left\{ \int_{\hat{\theta}_i \wedge \hat{\tau}}^{\hat{\theta}_{i+1} \wedge \hat{\tau}} [\hat{f}(\hat{y}(t), \hat{\mathbf{a}}^{i+1}) - \hat{f}(\hat{y}(t), \mathbf{a}^e)] e^{-\alpha t} dt \right\}.$$

Since $f \geq 0$, it follows that for $\mathbf{a} \neq \mathbf{a}^e$,

$$(4.29) \quad 0 \leq \hat{u}_0^e(x, \mathbf{a}) - \hat{u}_0^e(x, \mathbf{a}) \leq \left(\sup_{\mathbf{b} \neq \mathbf{a}^e} \|\hat{f}(\cdot, \mathbf{b})\| \right) \left(\sum_{\mathbf{a}^i = \mathbf{a}^e} E \left\{ \int_{\hat{\theta}_i \wedge \hat{\tau}}^{\hat{\theta}_{i+1} \wedge \hat{\tau}} e^{-\alpha t} dt \right\} \right).$$

But using (4.6) gives

$$(4.30) \quad \frac{1}{\varepsilon} \sum_{\mathbf{a}^i = \mathbf{a}^e} E \left\{ \int_{\hat{\theta}_i \wedge \hat{\tau}}^{\hat{\theta}_{i+1} \wedge \hat{\tau}} e^{-\alpha t} dt \right\} \leq E \left\{ \int_0^{\tau} \hat{f}(\hat{y}(t), \hat{\mathbf{a}}(t)) e^{-\alpha t} dt \right\} \leq \hat{u}_0^e(x, \mathbf{a}).$$

And

$$(4.31) \quad 0 \leq \hat{u}_0^e(x, \mathbf{a}) \leq \min_{\mathbf{b} \in A} \left\{ \frac{1}{\alpha} \|\hat{f}(\cdot, \mathbf{b})\| + k(\mathbf{a}, \mathbf{b}) \right\}.$$

It follows that if $\mathbf{a} \neq \mathbf{a}^e$, then

$$(4.32) \quad 0 \leq \hat{u}_0^e(x, \mathbf{a}) - \hat{u}_0^e(x, \mathbf{a}) \leq c\varepsilon$$

with

$$(4.33) \quad c = \frac{1}{\alpha} \left(\sup_{\mathbf{b} \neq \mathbf{a}^e} \|f(\cdot, \mathbf{b})\| \right) \|\hat{f}(\cdot, \mathbf{a})\|.$$

This implies (4.13) in case $\mathbf{a} \neq \mathbf{a}^e$.

If $\mathbf{a} = \mathbf{a}^e$, the argument is much the same. Define

$$(4.34) \quad \tilde{\mathbf{a}}(t) = \mathbf{a}^1 \quad \forall t \in [0, \hat{\theta}_1].$$

Since (4.30) and (4.31) still hold, we can deduce (4.32) by adding the term

$$(4.35) \quad E \{ k(\mathbf{a}^e, \mathbf{a}^1) (1 - e^{-\alpha(\hat{\theta}_1 \wedge \hat{\tau})}) \} = \alpha k(\mathbf{a}^e, \mathbf{a}^1) E \left\{ \int_0^{\hat{\theta}_1 \wedge \hat{\tau}} e^{-\alpha t} dt \right\}$$

to (4.29). In this case the constant in (4.32) is

$$(4.36) \quad c = \alpha \left[\sup_{\mathbf{b} \neq \mathbf{a}^e} \|\hat{f}(\cdot, \mathbf{b})\| + \alpha k(\mathbf{a}^e, \mathbf{b}) \right] \left[\min_{\mathbf{b} \in A} \frac{1}{\alpha} \|\hat{f}(\cdot, \mathbf{b})\| + k(\mathbf{a}^e, \mathbf{b}) \right].$$

This completes the argument. QED

Remark. Using similar techniques, we can consider systems with locally bounded coefficients in an unbounded domain \mathcal{O} . All the results can be extended to the associated time-dependent problem.

Acknowledgment. The first named author would like to thank J. S. Baras for useful discussions related to this work.

REFERENCES

- [1] A. BENSOUSSAN AND J. L. LIONS, *Nouvelle formulation de problèmes de contrôle impulsionnel et applications*, C. R. Acad. Sci. Paris, A-276 (1973), pp. 1189–1192.
- [2] ———, *Contrôle impulsionnel et systèmes d'inéquations quasi-variationnelles*, C. R. Acad. Sci. Paris, A-278 (1974), pp. 747–751.
- [3] J. GRUHL, F. SCHWEPPE AND M. RUANE, *Unit commitment scheduling of electric power systems*, Systems Engineering for Power: Status and Prospects, L. H. Fink and K. Carlsen, eds., US DOE Report NTIS-CONF-750867, 1975, pp. 116–129.
- [4] M. ROBIN, *Contrôle impulsionnel des processus de Markov*, Thèse de Etat, INRIA, Le Chesnay, France, 1977.
- [5] J.-L. MENALDI, *On the optimal stopping time problem for degenerate diffusions*, this Journal, 18 (1980), pp. 697–721.
- [6] ———, *On the optimal impulse control problem for degenerate diffusions*, this Journal, 18 (1980), pp. 722–739.
- [7] A. BENSOUSSAN AND J.-L. LIONS, *Inéquations quasi-variationnelles dépendant d'un paramètre*, Ann. Scuola Normale di Pisa, ser. IV, (1977), pp. 231–255.
- [8] J.-L. MENALDI, *A singular perturbation result for variational and quasi-variational inequalities*, Non-linear Anal., 5 (1981), pp. 381–400.
- [9] W. E. HOPKINS, JR. AND G. L. BLANKENSHIP, *Perturbation analysis of a system of quasi-variational inequalities for optimal stochastic scheduling*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 1054–1070.
- [10] J.-L. MENALDI, J.-P. QUADRAT AND E. ROFMAN, *On the role of the impulse fixed cost in stochastic optimal control: An application to the management of energy production*, Proc. 10th IFIP Conference on System Modelling Optimization, New York, September 1981, Springer-Verlag, to appear.
- [11] A. FRIEDMAN AND P.-L. LIONS, *The optimal strategy in the control problem associated with the Hamilton-Jacobi-Bellman equation*, this Journal, 18 (1980), pp. 191–198; *Corrigendum*, 20 (1982), pp. 153–154.
- [12] D. W. STROOCK AND S. R. S. VARADHAN, *On degenerate elliptic-parabolic operators of second order and their associated diffusions*, Comm. Pure Appl. Math., 25 (1972), pp. 651–713.

NEW RESULTS ON THE INTERPOLATION PROBLEM FOR CONTINUOUS-TIME STATIONARY INCREMENTS PROCESSES*

MICHELE PAVON†

Abstract. Explicit solutions to the interpolation problem for continuous-time stationary increments processes with a rational spectral density are derived. To do so we take a new approach to the problem relying on stochastic realization theory. In particular we show that the optimal interpolator is completely characterized by two steady-state Kalman–Bucy estimates.

Key words. interpolation problem, stationary process, stochastic realization theory

1. Introduction. Linear interpolation of continuous-time stationary processes, first studied by Karhunen in 1952 [1], has been described as a difficult subject by several specialists in the prediction theory of stationary processes, cf. Rozanov [2, p. 131], Masani [3, p. 1466], Dym and McKean [4, pp. 8–9] and Salehi [5, p. 841]. Indeed, the results as yet available in the literature on the problem of characterizing the optimal interpolator and the corresponding error intensity [6]–[8], [2, pp. 129–135], [4, §§ 4.13, 6.13, 6.14] are nonexplicit, hard to apply and of difficult statistical interpretation.

The purpose of this paper is to present the first explicit solution to the above problem. Our approach is new and makes essential use of some basic results and techniques from the recently developed *stochastic realization theory* [9]–[20] (see [16] and [20] for other references). We study the case of a multivariate process with stationary increments and rational spectral density, where the observation of the increments is not possible on a certain finite time interval. This problem appears to have potential applications to many diverse areas of the physical and engineering sciences. Indeed it models the rather common situation when a blackout has occurred in the stationary flow of information about a certain physical system, and the missing data have to be estimated from the known increments of the process. The rational spectral density case is of central importance for engineering applications, cf. e.g. [15], in particular because in this case the process admits a *finite dimensional Markovian representation* (see, for example, [9]). The latter fact allows us to derive a compact expression for the optimal least-squares interpolator in terms of two Kalman–Bucy estimates (see Theorem 4.4). These are generated by a forward and a backward steady-state filter, respectively. This representation holds under the assumption only that the process is purely nondeterministic.

The derivation, relying on a variant of a geometrical argument of stochastic realization theory (see [9] and [12]), is simple and illuminating. The key step consists in replacing a projection onto an infinite-dimensional space by a projection onto a finite-dimensional space which admits a nice basis (induced by the components of two Markov processes) (see Lemmas 4.1 and 4.2). Although the latter Hilbert space results can be easily established in the case of a general spectral density, further study is required in order to obtain satisfactory characterizations of the interpolation estimate in this case. Clearly our method also works for stationary processes, as we briefly indicate in § 6. Actually, it can also be applied to some nonstationary situations [33].

* Received by the editors April 29, 1982, and in revised form December 15, 1982. This research was conducted at the Department of Statistics, Florida State University, Tallahassee, Florida 32306, with support provided by a CNR fellowship.

† LADSEB-CNR, Corso Stati Uniti 4, 35100 Padova, Italy.

The process is assumed to have Gaussian increments, but the results hold for weakly stationary increments processes as well. No attention is given to the algorithmic aspect of the problem. However, we feel that our results are of computational interest since they only involve quantities which can be efficiently calculated from the spectral density via the deterministic and stochastic realization algorithms, cf. viz [21] and [15]. We refer the interested reader to [22] for an application of this method to a simple discrete-time interpolation problem. Other references on the continuous-time problem are [23] and [24].

The contents of the paper are as follows. Section 2 is devoted to introducing the relevant mathematical notation and to formulating the problem. In § 3 we record some basic results from stochastic realization theory. Here [10] is the main reference. These results are then applied to the interpolation problem in § 4 as means to derive the key representation for the optimal interpolation. Several other expressions for it are readily obtained in the following section, where we also make contact with some smoothing results derived in [25]. In § 6 we comment briefly on the case when the process is actually assumed to be stationary.

2. Mathematical notation and problem formulation. We shall be concerned with centered stationary (stationary increments) Gaussian processes defined on a fixed probability space (Ω, \mathcal{F}, P) . Let $\{\xi(t); t \in \mathbb{R}\}$ be such a process taking values in \mathbb{R}^p . Then $H_t(\xi)$ indicates the linear space induced by the random variables $\{\xi_1(t), \dots, \xi_p(t)\}$. We define the spaces $H_t^-(d\xi)$, $H_t^+(d\xi)$ and $H(d\xi)$ to be the *Gaussian spaces* [26, p. 53] induced by the increments $\{\xi(t) - \xi(s); t, s \in I\}$ where I is the set $(-\infty, t]$, $[t, +\infty)$ and \mathbb{R} , respectively. If $H \subset L^2(\Omega, \mathcal{F}, P)$ is a Gaussian space, $E\{\cdot|H\}$ denotes the orthogonal projection onto H . We write $E\{\cdot|\xi(t)\}$ instead of $E\{\cdot|H_t(\xi)\}$ and $E\{\nu|H\}$ for the vector with components $E\{\nu_i|H\}$. Let K be another Gaussian space. We indicate by $E\{K|H\}$ the linear hull of the projections of elements in K onto H . When $K \subset H$, we write $H \ominus K$ for the orthogonal complement of K in H . The identity matrix is denoted by I . If R is a symmetric positive (nonnegative) definite matrix, we write $R > 0$ ($R \geq 0$) and indicate by $R^{1/2}$ its positive (nonnegative) square root. Transposition is denoted by a prime. Vectors without a prime are column vectors. We write $\text{var}[\nu]$ for the variance matrix $E\{\nu\nu'\}$ of the vector ν .

Let $\{y(t); t \in \mathbb{R}\}$ be a purely nondeterministic, mean-square continuous stochastic process defined on the probability space (Ω, \mathcal{F}, P) and taking values in \mathbb{R}^m . We assume that y is centered and has Gaussian stationary increments. Its increments can then be represented as

$$(2.1) \quad y(t) - y(s) = \int_{-\infty}^{\infty} \frac{e^{i\omega t} - e^{i\omega s}}{i\omega} d\mu(\omega)$$

[27, p. 205], where μ is a vector orthogonal stochastic measure satisfying

$$E\{d\mu(\omega) d\mu(\omega)^\dagger\} = \frac{\Phi(i\omega)}{2\pi} d\omega,$$

the symbol \dagger denoting conjugation and transposition. We suppose that the spectral density Φ is a matrix of real rational functions such that $\Phi(i\omega) > 0$ for almost all ω . It follows [2] that $R = \Phi(\infty)$ is positive definite.

Let $0 < t < T$ and $\hat{y}(t) = E\{y(t)|H_0^-(dy) \vee H_T^+(dy)\}$, where $H_0^-(dy) \vee H_T^+(dy)$ denotes the smallest Gaussian space containing $H_0^-(dy)$ and $H_T^+(dy)$. We study the following *interpolation problem*: given $\{y(\tau) - y(\sigma); \tau, \sigma \in (-\infty, 0] \text{ or } \tau, \sigma \in [T, \infty)\}$ determine the increments of \hat{y} . Clearly, because of stationarity, it is no restriction to consider only intervals of the form $(0, T)$.

3. Preliminaries. Using partial fractions, we can write Φ in the form of

$$\Phi(z) = S(z) + S(-z)',$$

where S is a *positive real* function with McMillan degree equal to one-half the McMillan degree of Φ [28], [15]. We can then compute a *minimal realization* $[F, G, H, \frac{1}{2}R]$ of S employing one of the algorithms available in the systems theory literature (see e.g. [21]). Hence, in the sequel, we shall regard such a quadruplet as part of the data. Also notice that under the present assumptions $\text{Re} \{\lambda(F)\} < 0$, namely the eigenvalues of F lie in the open left half-plane [9], [15].

In [10] it was shown that there exist two Markovian representations (*stochastic realizations*) of the increments of y of the following form. The first,

$$(3.1a) \quad dx_* = Fx_* dt + B_* du_*,$$

$$(3.1b) \quad dy = Hx_* dt + R^{1/2} du_*,$$

is a *steady-state Kalman-Bucy filter*. The *innovations* process u_* is a standard m -dimensional Brownian motion defined on (Ω, \mathcal{F}, P) with $H_t^-(du_*) = H_t^-(dy)$, for all $t \in \mathbb{R}$. Inverting (3.1) we also get

$$(3.2) \quad dx_* = \Gamma_* x_* dt + B_* R^{-1/2} dy,$$

where the *feedback matrix* $\Gamma_* = F - B_* R^{-1/2} H$ satisfies $\text{Re} \{\lambda(\Gamma_*)\} \leq 0$ [15]. The second realization,

$$(3.3a) \quad d\bar{x}_* = -F' \bar{x}_* dt + \bar{B}_* d\bar{u}_*,$$

$$(3.3b) \quad dy = G' \bar{x}_* dt + R^{1/2} d\bar{u}_*,$$

is a *backward steady-state Kalman-Bucy filter*. The *backward innovation* \bar{u}_* is a process of the same type as u_* and satisfies $H_t^+(d\bar{u}_*) = H_t^+(dy)$, for all $t \in \mathbb{R}$. We also have

$$(3.4) \quad d\bar{x}_* = -\bar{\Gamma}_* \bar{x}_* dt + \bar{B}_* R^{-1/2} dy,$$

with $\bar{\Gamma}_* = F' + \bar{B}_* R^{-1/2} G'$ and $\text{Re} \{\lambda(\bar{\Gamma}_*)\} \leq 0$. There exist two other representations of particular interest. The first one,

$$(3.5a) \quad dx^* = Fx^* dt + B^* du^*,$$

$$(3.5b) \quad dy = Hx^* dt + R^{1/2} du^*,$$

corresponds to (3.3) in the sense of [10, Thm. 3.4]. The second,

$$(3.6a) \quad d\bar{x}^* = -F' \bar{x}^* dt + \bar{B}^* d\bar{u}^*,$$

$$(3.6b) \quad dy = G' \bar{x}^* dt + R^{1/2} d\bar{u}^*,$$

is a *backward model* (i.e. $H_t^-(d\bar{u}^*)$ is orthogonal to $H_t^+(\bar{x}^*)$) and corresponds to (3.1). The inputs u^* and \bar{u}^* are Brownian motions corresponding to \bar{u}_* and u_* , respectively, according to [10, (3.35a)]. Let $P_* = \text{var}[x_*(t)]$ and $P^* = \text{var}[x^*(t)]$ be the state variances in the forward models (3.1) and (3.5), respectively. Then

$$(3.7) \quad P^* - P_* \geq 0,$$

$$(3.8) \quad \bar{x}_*(t) = (P^*)^{-1} x^*(t),$$

$$(3.9) \quad \bar{x}^*(t) = P_*^{-1} x_*(t),$$

cf. [10], [15]. We refer the reader to [10] for further information on the correspondence

between forward and backward realizations. The filter property of (3.1) and (3.3) in particular gives

$$(3.10) \quad E\{x^*(t)|H_t^-(dy)\} = x_*(t),$$

$$(3.11) \quad E\{\bar{x}^*(t)|H_t^+(dy)\} = \bar{x}_*(t),$$

for all $t \in \mathbb{R}$.

4. The main representation. Let $0 < s < t < T$. We are interested in computing $\hat{y}(t) - \hat{y}(s)$. We shall show that the two vectors $x_*(0)$ and $\bar{x}_*(T)$ contain all the information in $H_0^-(dy) \setminus H_T^+(dy)$ which is useful in estimating $y(t) - y(s)$. To do so we need a preliminary result.

LEMMA 4.1. *The space $H_0^-(dy) \setminus H_T^+(dy)$ admits the following orthogonal decomposition,*

$$(4.1) \quad H_0^-(dy) \setminus H_T^+(dy) = N^- \oplus [H(x_*(0)) \setminus H(\bar{x}_*(T))] \oplus N^+,$$

where $N^- = H_0^-(dy) \ominus H(x_*(0))$ and $N^+ = H_T^+(dy) \ominus H(\bar{x}_*(T))$.

Proof. It is well known that $E\{H_0^+(dy)|H_0^-(dy)\} = H(x_*(0))$, cf. for example [16]. Thus, N^- is orthogonal to $H_0^+(dy)$. However, $H_T^+(dy) \subset H_0^+(dy)$ and therefore $N^- \perp H_T^+(dy)$. Similarly, one can show that $N^+ \perp H_0^-(dy)$ and the result is proved. \square

LEMMA 4.2. $H(\hat{y}(t) - \hat{y}(s)) \subset H(x_*(0)) \setminus H(\bar{x}_*(T))$.

Proof. Notice that $E\{\hat{y}(t) - \hat{y}(s)|H_0^-(dy)\} = E\{y(t) - y(s)|H_0^-(dy)\} = (\int_s^t H e^{F\tau} \times d\tau)x_*(0)$. In fact the first equality follows from the law of iterated conditioning (projecting), and the second from the integrated form of (3.1b),

$$(4.2) \quad y(t) - y(s) = \int_s^t Hx_*(\tau) d\tau + R^{1/2}[u_*(t) - u_*(s)],$$

and the expression

$$(4.3) \quad x_*(\tau) = e^{F\tau}x_*(0) + \int_0^\tau e^{F(\tau-\sigma)}B_* du_*(\sigma)$$

for the solution of (3.1a). This shows that the components of $\hat{y}(t) - \hat{y}(s) - (\int_s^t H e^{F\tau} d\tau)x_*(0)$ are orthogonal to $H_0^-(dy)$. Since $N^- \subset H_0^-(dy)$, we deduce that $H(\hat{y}(t) - \hat{y}(s)) \perp N^-$. On the other hand a similar argument, using the properties of the backward filter (3.3), yields $H(\hat{y}(t) - \hat{y}(s)) \perp N^+$. The conclusion now follows from (4.1). \square

LEMMA 4.3. *The components of $x^*(T) - e^{FT}x_*(0)$ are orthogonal to $H(x_*(0))$. Moreover the variance*

$$(4.4) \quad \Pi = P^* - e^{FT}P_* e^{F'T}$$

of $x^*(T) - e^{FT}x_*(0)$ is positive definite.

Proof. Using iterated conditioning, formula (4.3) for x^* , and (3.10) we get $E\{x^*(T) - e^{FT}x_*(0)|H(x_*(0))\} = E\{E\{x^*(T) - e^{FT}x_*(0)|H_0^-(dy)\}|H(x_*(0))\} = E\{e^{FT}x_*(0) - e^{FT}x_*(0)|H(x_*(0))\} = 0$ which proves the first assertion. It follows that $\Pi = P^* - e^{FT}P_* e^{F'T}$. Next observe that (4.3) gives

$$(4.5) \quad -e^{FT}P_* e^{F'T} = -P_* + \int_0^T e^{F(T-\sigma)}B_*B_*' e^{F'(T-\sigma)} d\sigma.$$

The controllability Gramian appearing in the right-hand side of (4.5) is positive definite since the pair (F, B_*) is controllable [21]. Inserting (4.5) into (4.4) and taking (3.7) into account, we conclude that $\Pi > 0$. \square

THEOREM 4.4. *The interpolation estimate \hat{y} is differentiable on $(0, T)$. Its derivative admits the following orthogonal decomposition:*

$$(4.6) \quad \frac{d\hat{y}}{dt} = H(t)x_*(0) + K(t)\Pi^{-1}[x^*(T) - e^{FT}x_*(0)],$$

where $H(t) = He^{Ft}$, $K(t) = G'e^{F'(T-t)} - He^{Ft}P_*e^{F'T}$ and Π is defined by (4.4). Let $\tilde{y}(t) = y(t) - \hat{y}(t)$ denote the estimation error. Then

$$(4.7) \quad \begin{aligned} \text{var} [\tilde{y}(t) - \tilde{y}(s)] &= \int_s^t \int_s^t \int_0^{\tau \wedge \tau'} [H(\tau - \sigma)B_*B'_*H(\tau' - \sigma)'] d\sigma d\tau d\tau' \\ &+ \int_s^t \int_s^\tau [H(\tau - \sigma)B_*R^{1/2} + R^{1/2}B'_*H(\tau - \sigma)'] d\sigma d\tau + R[t - s] \\ &- M(t - s)\Pi^{-1}M(t - s)', \end{aligned}$$

where $\tau \wedge \tau' = \min(\tau, \tau')$ and $M(t - s) = \int_s^t K(\tau) d\tau$.

Proof. By Lemma 4.2 and the law of iterated conditioning we have $\hat{y}(t) - \hat{y}(s) = E\{y(t) - y(s) | H(x_*(0)) \vee H(\bar{x}_*(T))\}$. Notice that, in view of (3.8), the components of $x_*(0)$ and $x^*(T) - e^{FT}x_*(0)$ span the space $H(x_*(0)) \vee H(\bar{x}_*(T))$.

Then Lemma 4.3 gives

$$\hat{y}(t) - \hat{y}(s) = E\{y(t) - y(s) | x_*(0)\} + E\{y(t) - y(s) | x^*(T) - e^{FT}x_*(0)\}.$$

As argued in the proof of Lemma 4.2, $E\{y(t) - y(s) | x_*(0)\} = (\int_s^t H(\tau) d\tau)x_*(0)$ with $H(\tau) = He^{F\tau}$. Since $\Pi > 0$ (Lemma 4.3), we can apply a standard projection formula and get

$$\begin{aligned} \hat{y}(t) - \hat{y}(s) &= \left(\int_s^t H(\tau) d\tau \right) x_*(0) \\ &+ E\{[y(t) - y(s)][x^*(T) - e^{FT}x_*(0)]\} \Pi^{-1} [x^*(T) - e^{FT}x_*(0)]. \end{aligned}$$

Employing (4.2) we readily obtain $E\{[y(t) - y(s)]x_*(0)' e^{F'T}\} = (\int_s^t H(\tau) d\tau)P_*e^{F'T}$, whereas the representation

$$(4.8) \quad y(t) - y(s) = \int_s^t G'\bar{x}_*(\tau) d\tau + R^{1/2}[\bar{u}_*(t) - \bar{u}_*(s)]$$

and (3.8) yield $E\{[y(t) - y(s)]x^*(T)\} = \int_s^t G'e^{F'(T-\tau)} d\tau$. We get

$$(4.9) \quad \begin{aligned} \hat{y}(t) - \hat{y}(s) &= \left(\int_s^t H(\tau) d\tau \right) x_*(0) \\ &+ \left(\int_s^t [G'e^{F'(T-\tau)} - H(\tau)P_*e^{F'T}] d\tau \right) \Pi^{-1} [x^*(T) - e^{FT}x_*(0)]. \end{aligned}$$

Dividing both sides of (4.9) by $t - s$ and taking the limit as s tends to t , we finally obtain (4.6). In order to prove (4.7), first note that (4.2) and (4.3) gives

$$(4.10) \quad \begin{aligned} y(t) - y(s) &= \left(\int_s^t H(\tau) d\tau \right) x_*(0) \\ &+ \left(\int_s^t \int_0^\tau He^{F(\tau-\sigma)} B_* du_*(\sigma) + R^{1/2}[u_*(t) - u_*(s)] \right). \end{aligned}$$

The result is now a consequence of the orthogonality between the error increments and the estimate increments, and the orthogonality between the two terms in the right-hand sides of (4.10) and (4.9) respectively. \square

Remark 4.5. The vector $(\int_s^t H(\tau) d\tau)x_*(0)$ in (4.9) is just the optimal predictor of $y(t) - y(s)$ given $H_0^-(dy)$. Therefore the last term in (4.9) represents the modification of the prediction estimate due to the new data $H_T^+(dy)$. Its positive definite variance describes how much our information on $y(t) - y(s)$ has increased.

Remark 4.6. We shall now outline an alternative derivation of (4.6), which, although not as simple and straightforward as the one given above, appears to be of some interest. The idea is to use, besides stochastic realization, a corollary to J. von Neumann's alternating projections theorem due to Aronszajn [29, p. 375]. This result has already been applied to the interpolation problem in [8], [30]. Let $T_1 = E\{\cdot | H_0^-(dy)\}$, $T_2 = E\{\cdot | H_T^+(dy)\}$ and $T = E\{\cdot | H_0^-(dy) \vee H_T^+(dy)\}$. Then Aronszajn's theorem asserts that the sequence $S_1 = T_1$, $S_2 = T_1 + T_2 - T_2T_1$, $S_3 = T_1 + T_2 - T_2T_1 - T_1T_2 + T_1T_2T_1, \dots$ converges strongly to T . It follows that $S_n[y(t) - y(s)]$ converges to $\hat{y}(t) - \hat{y}(s)$ in the L^2 norm topology. Exploiting (4.2), (4.8), and (3.10) and (3.11) repeatedly, we quickly get

$$\begin{aligned} S_{2n}[y(t) - y(s)] &= \left(\int_s^t H(\tau) d\tau \right) x_*(0) \\ &\quad + \left(\int_s^t K(\tau) d\tau \right) \left[\sum_{i=0}^{n-2} ((P^*)^{-1} e^{FT} P_* e^{F'T})^i (\bar{x}_*(T) - (P^*)^{-1} e^{FT} x_*(0)) \right. \\ &\quad \left. + ((P^*)^{-1} e^{FT} P_* e^{F'T})^{n-1} \bar{x}_*(T) \right], \\ S_{2n+1}[y(t) - y(s)] &= \left(\int_s^t H(\tau) d\tau \right) x_*(0) \\ &\quad + \left(\int_s^t K(\tau) d\tau \right) \left[\sum_{i=0}^{n-1} ((P^*)^{-1} e^{FT} P_* e^{F'T})^i (\bar{x}_*(T) - (P^*)^{-1} e^{FT} x_*(0)) \right] \end{aligned}$$

for $n \geq 1$. Now observe that the eigenvalues of $(P^*)^{-1} e^{FT} P_* e^{F'T}$ lie in the open unit disc. Indeed, they are in the open interval $(0, 1)$ since $(P^*)^{-1} e^{FT} P_* e^{F'T}$ and $I - (P^*)^{-1} F P_* F' = (P^*)^{-1} \Pi$ have positive eigenvalues being the product of two positive definite matrices [31, p. 92]. A standard formula for geometric series, cf. for example [31, p. 113], and (3.8) now yield (4.6).

5. Other representations. If we use the components of $\bar{x}_*(T)$ and of $\bar{x}^*(0) - e^{F'T} \bar{x}_*(T)$ as a basis for $H(x_*(0)) \vee H(\bar{x}_*(T))$, we obtain an expression similar to (4.6) but involving backward quantities:

$$(5.1) \quad \frac{d\hat{y}}{dt} = G(t)' \bar{x}_*(T) + \bar{K}(t) \bar{\Pi}^{-1} [\bar{x}^*(0) - e^{F'T} \bar{x}_*(T)],$$

where $G(t) = e^{Ft} G$, $\bar{K}(t) = H e^{Ft} - G' e^{F'(T-t)} (P^*)^{-1} e^{FT}$ and $\bar{\Pi} = P_*^{-1} - e^{F'T} (P^*)^{-1} e^{FT}$. From this a relation corresponding to (4.7) is also easily derived. However, the following symmetric representation is more useful.

LEMMA 5.1. *In the notation of (4.6) and (5.1) we have*

$$(5.2) \quad \frac{d\hat{y}}{dt} = \bar{K}(t) \bar{\Pi}^{-1} \bar{x}^*(0) + K(t) \Pi^{-1} x^*(T).$$

Proof. A simple calculation using (4.6) and (3.11) yields the result. \square

The following theorem provides several alternative expressions for $d\hat{y}/dt$, including one in terms of the data of the problem, namely the increments $\{y(\tau) - y(\sigma); \tau, \sigma \in (-\infty, 0] \text{ or } \tau, \sigma \in [T, \infty)\}$. These expressions are all direct consequence of (4.6) or (5.2).

THEOREM 5.2. *Under the present assumptions, we have*

$$(i) \quad \frac{d\hat{y}}{dt} = \int_{-\infty}^0 \bar{K}(t) \bar{\Pi}^{-1} P_*^{-1} e^{-F\tau} B_* du_*(\tau) + \int_0^T K(t) \Pi^{-1} P^* e^{F'(T-t)} \bar{B}_* d\bar{u}_*(\tau).$$

If we also assume that $\Phi(i\omega) > 0$ for all $\omega \in \mathbb{R}$, then

$$(ii) \quad \begin{aligned} \frac{d\hat{y}}{dt} = & \int_{-\infty}^0 \bar{K}(t) \bar{\Pi}^{-1} P_*^{-1} e^{-\Gamma_* \tau} B_* R^{-1/2} dy(\tau) \\ & + \int_0^T K(t) \Pi^{-1} P^* e^{\bar{\Gamma}_*(\tau-T)} \bar{B}_* R^{-1/2} dy(\tau); \end{aligned}$$

$$(iii) \quad \begin{aligned} \frac{d\hat{y}}{dt} = & \int_{-\infty}^0 H(t) e^{-F\tau} B_* du_*(\tau) + \int_0^T K(t) \Pi^{-1} e^{F(T-t)} B_* du_*(\tau) \\ & + \int_T^\infty K(t) \Pi^{-1} \Sigma e^{\Gamma'_*(\tau-T)} H' R^{-1/2} du_*(\tau), \end{aligned}$$

where $\Sigma = P^* - P_*$;

$$(iv) \quad \begin{aligned} \frac{d\hat{y}}{dt} = & \int_{-\infty}^\infty [\bar{K}(t) \bar{\Pi}^{-1} P_*^{-1} (i\omega I - F)^{-1} B_* W_*(i\omega)^{-1} \\ & + K(t) \Pi^{-1} P^* e^{i\omega T} (i\omega I + F')^{-1} \bar{B}_* \bar{W}_*(i\omega)^{-1}] d\mu(\omega), \end{aligned}$$

where $W_*(i\omega) = H(i\omega I - F)^{-1} B_* + R^{1/2}$ and $\bar{W}(i\omega) = G'(i\omega I + F')^{-1} \bar{B}_* + R^{1/2}$;

$$(v) \quad \begin{aligned} \frac{d\hat{y}}{dt} = & \int_{-\infty}^\infty [\bar{K}(t) \bar{\Pi}^{-1} P_*^{-1} (i\omega I - \Gamma_*)^{-1} B_* R^{-1/2} \\ & + K(t) \Pi^{-1} P^* e^{i\omega T} (i\omega I + \bar{\Gamma}_*)^{-1} \bar{B}_* R^{-1/2}] d\mu(\omega); \end{aligned}$$

$$(vi) \quad \begin{aligned} \frac{d\hat{y}}{dt} = & \int_{-\infty}^\infty [(H(t) + K(t) \Pi^{-1} (e^{i\omega T} - e^{FT})) (i\omega I - F)^{-1} B_* \\ & - K(t) \Pi^{-1} \Sigma e^{i\omega T} (i\omega I + \Gamma'_*)^{-1} H' R^{-1/2}] W_*(i\omega)^{-1} d\mu(\omega). \end{aligned}$$

Proof. Formula (i) follows from (5.2), (3.10)–(3.11) and (3.1a)–(3.3a) in view of $\text{Re} \{\lambda(F)\} < 0$. It is well known, cf. viz. [15, p. 96], that $\Phi(i\omega) > 0$ for all real ω implies that $\text{Re} \{\lambda(\Gamma_*)\} < 0$, $\text{Re} \{\lambda(\bar{\Gamma}_*)\} < 0$ and $\Sigma > 0$. Expression (ii) is now a consequence of (5.2), (3.10)–(3.11) and (3.2)–(3.4). To prove (iii) we rely on (4.6), (4.3) and the equation

$$d\bar{z} = -\Gamma'_* \bar{z} dt - H' R^{-1/2} du_*$$

for $\bar{z}(t) = \Sigma^{-1}[x^*(t) - x_*(t)]$ which was derived in [12]. The spectral results (iv)–(vi) follow from (i)–(iii) respectively, using (2.1) and the relations

$$\begin{aligned} u_*(t) - u_*(s) &= \int_{-\infty}^\infty \frac{e^{i\omega t} - e^{i\omega s}}{i\omega} W_*(i\omega)^{-1} d\mu(\omega), \\ \bar{u}_*(t) - \bar{u}_*(s) &= \int_{-\infty}^\infty \frac{e^{i\omega t} - e^{i\omega s}}{i\omega} \bar{W}_*(i\omega)^{-1} d\mu(\omega) \end{aligned}$$

for the increments of the forward and backward innovations. \square

Remark 5.3. Formula (v) agrees with a classical result of Yaglom in that the poles of the two terms in the square bracket of (v) coincide with the poles of the corresponding terms in [7, formula (4.9)]. In fact the poles in (v) are the eigenvalues of Γ_* and $-\bar{\Gamma}_*$, which are the zeros of the determinant of the *outer factor* $W_*(z)$ and of the *conjugate outer factor* $\bar{W}_*(z)$, cf. [10], [15], [32]. These, in turn, coincide with the zeros of the determinant of $\Phi \det \Phi$ in the left and right half-plane respectively. This follows from the scalar factorization $\det \Phi(z) = \det W_*(z) \det W_*(-z)'$ and the fact that the scalar conjugate outer factor $\det \bar{W}_*(z)$ has the same zeros as $\det W_*(-z)'$.

The increments of u_* from time zero on appearing in (iii) of the previous theorem are not obtainable from the data. However, such a representation is of some theoretical interest as we shall see below. Consider a Markovian representation of y of the form

$$\begin{aligned} dx &= Fx \, dt + B \, dv, \\ dy &= Hx \, dt + R^{1/2} \, du, \end{aligned}$$

with the increments of v and u uncorrelated. This condition corresponds to the standard assumption in the filtering literature of uncorrelated state and observation noises. Let $\hat{\xi}(t) = E\{Hx(t)|H(dy)\}$ be the smoothing estimate of the observation signal $Hx(t)$.

PROPOSITION 5.4. *The interpolation estimate $d\hat{y}/dt$ converges a.s. to $\hat{\xi}(0)$ as T (and consequently t) tends to zero.*

Proof. Recall the formula $W_*(i\omega)^{-1} = R^{-1/2} - R^{-1/2}H(i\omega I - \Gamma_*)^{-1}B_*R^{-1/2}$ and $\Phi(i\omega) = W_*(i\omega)W_*(-i\omega)'$ (see e.g. [15]). Then taking the limit in (vi), Theorem 5.2, we get

$$(5.3) \quad \lim_{T \rightarrow 0} \lim_{t \rightarrow 0} \frac{d\hat{y}}{dt} = \int_{-\infty}^{\infty} [I - R\Phi(i\omega)^{-1}] \, d\mu(\omega),$$

which is equal to $\hat{\xi}(0)$ because of [25, (4.7)]. Alternatively we can take the limit in (4.6) as t goes to zero and then as T goes to zero and get

$$\lim_{T \rightarrow 0} \lim_{t \rightarrow 0} \frac{d\hat{y}}{dt} = Hx_*(0) + R^{1/2}B_*'\bar{z}(0),$$

which is $\hat{\xi}(0)$ in view of [25, (4.2)]. \square

Formula (5.3) resembles a famous discrete-time formula (cf. for example [2, p. 102]), which was rederived in [22] using the discrete-time counterpart of (vi) in Theorem 5.2. For further details on smoothing, and in particular on the significance of the assumption of independent state and observation noises, from the stochastic realization viewpoint we refer the reader to [25, § 4].

6. The stationary case. Let y be a stationary process with rational spectral density satisfying the assumptions of § 2, and let $H_t^-(y)$ and $H_t^+(y)$ be the Gaussian spaces induced by the components of $y(s)$ at times $s \leq t$ and $s \geq t$ respectively. Then [13], [17], we can compute from the spectral density matrices $[F, G, H]$ such that there exist two Markovian representations of y

$$\begin{aligned} dx_* &= Fx_* \, dt + B_* \, du_*, & d\bar{x}_* &= -F'\bar{x}_* \, dt + \bar{B}_* \, d\bar{u}_*, \\ y(t) &= Hx_*(t), & y(t) &= G'\bar{x}_*(t), \end{aligned}$$

with $x_*(t)$ a basis in $E\{H_t^+(y)|H_t^-(y)\}$ and $\bar{x}_*(t)$ a basis in $E\{H_t^-(y)|H_t^+(y)\}$, for all real t . Let $P_* = \text{var}[x_*(t)]$, $P^* = (\text{var}[\bar{x}_*(t)])^{-1}$, and let $x^*(t) = P^*\bar{x}_*(t)$ be the state of the forward model corresponding to the backward filter. Then the geometric

argument of § 4 gives

$$(6.1) \quad E\{y(t)|H_0^-(y)\setminus H_T^+(y)\} = H(t)x_*(t) + K(t)\Pi^{-1}[x^*(T) - e^{FT}x_*(0)],$$

with $H(\cdot)$, $K(\cdot)$ and Π defined as in § 4. The variance of the interpolation error $\tilde{y}(t)$ is given by

$$(6.2) \quad \text{var} [\tilde{y}(t)] = \int_0^t H(t-\tau)B_*B_*'H(t-\tau)' d\tau - K(t)\Pi^{-1}K(t)'.$$

From formula (6.1), it is then straightforward to obtain expressions for the optimal interpolator corresponding to those of § 5.

7. Final remarks. A new approach to the interpolation problem has been presented. We feel that stochastic realization theory provides a natural framework for studying this problem. The geometric argument used in the derivation of the results appears to have applications to a variety of interpolation problems, including interpolation of nonstationary processes.

REFERENCES

- [1] K. KARHUNEN, *Zur Interpolation von stationären zufälligen Funktionen*, Ann. Acad. Sci. Fenn., 142 (1952), pp. 3–8.
- [2] YU. A. ROZANOV, *Stationary Random Processes*, Holden-Day, San Francisco, 1967.
- [3] P. MASANI, *Review of the book Stationary Random Processes by Yu. A. Rozanov*, Ann. Math. Statist., 42 (1971), pp. 1463–1467.
- [4] H. DYM AND H. P. MCKEAN, *Gaussian processes, Function Theory, and the Inverse Spectral Problem*, Academic Press, New York, 1976.
- [5] H. SALEHI, *Algorithms for linear interpolator and interpolation error for minimal stationary stochastic processes*, Ann. Probab., 5 (1979), pp. 840–846.
- [6] A. M. YAGLOM, *Extrapolation, interpolation and filtration of stationary random processes with rational spectral density*, Amer. Math. Soc. Sel. Transl. Math. Statist., 4 (1963), pp. 345–387.
- [7] ———, *Effective solutions of linear approximation problems or multivariate stationary processes with a rational spectrum*, Theory Prob. Appl., 5 (1960), pp. 239–264.
- [8] V. M. ADAMYAN AND D. Z. AROV, *A general solution of a problem in linear prediction of stationary processes*, Theory Prob. Appl., 13 (1968), pp. 394–407.
- [9] G. RUCKEBUSCH, *Représentations Markoviennes de processus Gaussiens stationnaires*, Thèse 3^e cycle, Lab. Calcul des Probabilités, Université Paris VI, 1975.
- [10] A. LINDQUIST AND G. PICCI, *On the stochastic realization problem*, this Journal, 17 (1979), pp. 365–389.
- [11] M. PAVON, *Stochastic realization and invariant directions of the matrix Riccati equation*, this Journal, 18 (1980), pp. 155–180.
- [12] F. BADAWI, A. LINDQUIST AND M. PAVON, *A stochastic realization approach to the smoothing problem*, IEEE Trans. Aut. Contr. AC-24, pp. 878–888.
- [13] A. LINDQUIST AND G. PICCI, *Realization theory for multivariate Gaussian processes*, II, Proc. 2nd International Conference on Information Sciences and Systems, Patras, Greece, 1979.
- [14] A. LINDQUIST, G. PICCI AND G. RUCKEBUSCH, *On splitting subspaces and Markovian representations*, Math. Syst. Theory, 12 (1979), pp. 271–279.
- [15] P. FAURRE, M. CLERGET AND F. GERMAIN, *Opérateurs rationnels positifs*, Dunod, Paris, 1979.
- [16] G. RUCKEBUSCH, *Théorie géométrique de la représentation markovienne*, Ann. Inst. H. Poincaré, 3 (1980), pp. 225–297.
- [17] A. LINDQUIST AND G. PICCI, *State space models for Gaussian processes*, Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., Reidel, Dordrecht, 1981.
- [18] A. LINDQUIST AND M. PAVON, *Markovian representations of discrete-time stationary stochastic vector processes*, Proc. IEEE Decision and Control Conference, San Diego, 1981.
- [19] A. LINDQUIST AND G. PICCI, *On a condition for minimality of Markovian splitting subspaces*, Syst. Control Lett., 4 (1982), pp. 264–269.

- [20] A. LINDQUIST, M. PAVON AND G. PICCI, *Recent trends in stochastic realization theory*, Harmonic Analysis and Prediction Theory (vol. dedicated to P. Masani), V. Mandrekar and H. Salehi, eds., North-Holland, Amsterdam, 1983.
- [21] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
- [22] M. PAVON, *A new algorithm for optimal interpolation of discrete-time stationary processes*, Proc. 5th International Conference on Analysis and Optimization of Systems, Versailles, France, Dec. 1982.
- [23] H. SALEHI, *On the Hellinger integrals and interpolation of q -variate stationary stochastic processes*, Ark. Mat., 9 (1969), pp. 1–6.
- [24] YU. A. ROZANOV, *Some problems in the linear theory of random functions*, Theory Prob. Appl., 26 (1981), pp. 689–702.
- [25] M. PAVON, *The conjugate process in stochastic realization theory*, Math. Progr. Study, 18 (1982), pp. 12–26.
- [26] J. NEVEU, *Discrete-Parameter Martingales*, North-Holland, Amsterdam, 1975.
- [27] I. I. GIKHMAN AND A. V. SKOROKHOD, *Introduction to the Theory of Random Processes*, Saunders, Philadelphia, 1969.
- [28] B. D. O. ANDERSON, *The inverse problem of stationary covariance generation*, J. Statist. Phys., 1 (1969), pp. 133–147.
- [29] N. ARONSZAJN, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.
- [30] H. SALEHI, *On bilateral linear prediction for minimal stationary stochastic processes*, SIAM J. Appl. Math., 26 (1974), pp. 502–507.
- [31] D. K. FADDEEV AND V. N. FADDEEVA, *Computational Methods of Linear Algebra*, W. H. Freeman, San Francisco, 1963.
- [32] H. HELSON, *Lecture Notes on Invariant Subspaces*, Academic Press, New York, 1964.
- [33] M. PAVON, *Optimal interpolation for linear stochastic systems*, this Journal, to appear.

OPTIMAL SWITCHING FOR ORDINARY DIFFERENTIAL EQUATIONS*

I. CAPUZZO DOLCETTA† AND L. C. EVANS‡

Abstract. We consider the problem of controlling an ordinary differential equation, subject to positive switching costs, and show in particular that the value functions form the “viscosity solution” (cf. [6], [7]) of the dynamic programming quasi-variational inequalities. This interpretation allows for a rigorous application of various dynamic programming techniques.

Key words. optimal control, dynamic programming, quasi-variational inequalities, viscosity solutions

1. Introduction. Consider a system whose state is modelled by the solution of an ordinary differential equation, determined at each moment by one of m different control settings. Suppose further we repeatedly change these control settings as the system evolves so as to minimize a (discounted) running cost, but incur thereby a positive switching cost each time we modify the setting. What is an optimal way to control the system?

This problem, which we state precisely in § 2, is *formally* amenable to dynamic programming techniques. We first define each value function $u^d(x)$ ($x \in \mathbb{R}^n$, $d = 1, \dots, m$) to be the infimum of the costs taken over all controls with initial setting d , given that the ODE starts at the point x . Then fairly standard procedures (§ 3) indicate the construction of an optimal control in terms of the u^d ($d = 1, \dots, m$). Finally we observe (§ 4) that formally the value functions solve a coupled system of quasivariational inequalities (QVI) and that, conversely, any regular solution of (QVI) must in fact be the value function and hence lead to the synthesis of optimal controls. This much is all more-or-less routine.

The trouble in practice is that the value functions are generally not continuously differentiable and hence are not classical solutions of (QVI): the dynamic programming derivation of (QVI) is not justified. There is a further difficulty in that, on the other hand, (QVI) does not generally have a C^1 solution and hence standard verification techniques cannot be employed to recover the value functions and hence the optimal controls from a study of (QVI) by PDE methods. Such objections are, of course, typical in various applications of dynamic programming methods in deterministic control theory.

Our new contribution in this paper is to show that nevertheless *the full range of formal dynamic programming techniques can in fact be made rigorous for the problem at hand*: the key is an observation that the value functions, although not C^1 , do however solve (QVI) in an appropriately weak sense. For this we modify some recent work of Crandall–Lions [6] (cf. Crandall–Evans–Lions [7]) on scalar nonlinear first order PDE, and, in particular, adapt to our problem the new notion of a *viscosity solution* of such PDE. We will demonstrate in §§ 4–6 that the value functions comprise a viscosity solution of (QVI), that such solutions are unique, and that therefore PDE techniques are applicable in constructing the value functions and so the optimal

* Received by the editors July 20, 1982.

† Istituto Matematico, G. Castelnuovo, Università di Roma, 00100 Roma, Italy. The research of this author was supported in part by a NATO-CNR grant during a visit to the Department of Mathematics, University of Maryland.

‡ Department of Mathematics, University of Maryland, College Park, Maryland 20742. The research of this author was supported in part by the Alfred P. Sloan Foundation and the National Science Foundation under grant MCS-81-02846. The author was a part-time member of the Institute for Physical Science and Technology while this work was begun.

controls. (P. L. Lions in [12] first observed the connection between viscosity solutions and control theory).

To our knowledge the problem we study here is the only general example in the control of ODE for which the full extent of the dynamic programming formalism applies rigorously. Of course dynamic programming does work for other problems with special structure, for example the linear-quadratic regulator, as well as for discrete time problems. (See [4], [5].) Furthermore, dynamic programming is completely effective in stopping time problems for ODE, but this is a special case of our results: see the remark at the end of § 2.

Next we note that a control problem with zero switching costs also leads to a dynamic programming equation which admits a (unique) viscosity solution (see § 7 and also Lions [12]). Here we can also prove [12] that the value function is this viscosity solution, but are unable to synthesize optimal controls as these in general do not exist. Nevertheless, we will prove in § 7 that the value functions with positive switching costs do converge to the proper limit as the switching costs tend to zero. This is an analogue of a principal assertion in Evans–Friedman [9], where we proved similar statements for stochastic control theory.

Let us finally remark on some essential technical differences between applications of dynamic programming in deterministic and in stochastic control. The dynamic programming equations for the control of (nondegenerate) stochastic differential equations are second order, uniformly elliptic (or parabolic) PDE with convex nonlinearities. Such PDE are difficult to study owing to their fully nonlinear structure, at least when the controls affect the “noise” disturbing the system. It happens nevertheless that such strong estimates for the solution and its first and second derivatives have now been obtained that it is fairly easy to apply more-or-less routine analytic methods to study the existence and uniqueness of solutions, their properties, etc. A key tool in all this, and especially in the derivation of estimates, is the classical maximum principle for elliptic and parabolic PDE. The reader should refer to Evans–Friedman [9], Lions [13], Evans–Lions [10], Evans [8], etc., for elaboration of these comments.

A different situation prevails, as we have seen, in applications of the dynamic programming method to problems of optimal control for ordinary differential equations. There the resulting PDE again usually have a convex, badly nonlinear structure, but are now of first order. In consequence there are usually no such strong estimates available for the solution as in the stochastic case, and hence there are deep analytic problems as to the existence and uniqueness of solutions and their control theoretic interpretation. *Viscosity solutions* do not typically solve the dynamic programming PDE in any classical sense, but are nevertheless regular enough to ensure uniqueness of solutions and to justify rigorously many formal calculations. Our paper is thus in part a companion piece to Evans–Friedman [9], where we investigated some analogous questions for the control of stochastic differential equations. The principal differences in technique and point of view between this paper and [9] are best understood in light of the remarks above: in [9] we could derive good estimates for the solution of the appropriate system of quasi-variational inequalities involving uniformly elliptic operators, whereas here no such estimates are possible and we must rely on the viscosity solution concept. Notice, in particular, the quite different technical applications of the maximum principle. Nevertheless, there are strong heuristic and procedural similarities underlying this paper and [9], and we invite the interested reader to make note of this.

This work is motivated in addition by the earlier papers of Capuzzo Dolcetta–Matzeu [4], Capuzzo Dolcetta–Matzeu–Menaldi [5], Menaldi [14], and by our desire

to understand them in light of the new theory of viscosity solutions. Note in particular in [5] that the rigorous application of dynamic programming methods to the discrete approximations of the ODE allows for an interpretation of the value functions as the (unique) maximal subsolution of (QVI). Some other papers on control with positive switching costs are Belbas [3] and Belbas–Lenhart [3]. After this work was completed we learned of similar results to appear in the thesis of Barles [16].

In closing, we note here that although the methods alluded to above for the second order PDE depend strongly on the convex structure of the nonlinearity, the viscosity solution techniques for first order PDE do not. Thus there are applications of the latter ideas to the Isaacs equations in deterministic differential game theory (cf. Friedman [11]); see the forthcoming papers [1] and [15] for this.

2. Control of ODE with switching costs: statement of the problem. We will consider in this section the problem of the optimal control of an ordinary differential equation, whose dynamics can be modified—at the price of a positive switching cost—into any one of m different settings. What is the best way to adjust continuously the dynamics so as to minimize the associated cost?

More precisely let us define an *admissible control* α to be a sequence of *switching times* θ_i and *control settings* (or *switching decisions*) d_i :

$$\alpha = \{\theta_i, d_i\}_{i=0}^{\infty},$$

where

$$\begin{aligned} \theta_i &\in [0, +\infty], \quad 0 = \theta_0 \leq \theta_1 \leq \dots \leq \theta_i \leq \theta_{i+1} \leq \dots, \quad \theta_i \rightarrow \infty, \\ d_i &\in \{1, \dots, m\}, \quad d_i \neq d_{i-1} \quad \text{if } \theta_i < \infty \quad (i = 0, 1, \dots). \end{aligned}$$

For each $d = 1, \dots, m$ we define also \mathcal{A}^d , the set of all *admissible controls with initial setting* d :

$$\mathcal{A}^d \equiv \{\alpha \mid \alpha \text{ is an admissible control, } d_0 = d\}.$$

Consider now a given mapping $g: \mathbb{R}^n \times \{1, \dots, m\} \rightarrow \mathbb{R}^n$ satisfying

$$(2.1) \quad |g(x, d)| \leq L, \quad |g(x, d) - g(\hat{x}, d)| \leq L|x - \hat{x}|,$$

for some constant L and all $x, \hat{x} \in \mathbb{R}^n$, $d = 1, \dots, m$. For given $x \in \mathbb{R}^n$, $d \in \{1, \dots, m\}$, and $\alpha \in \mathcal{A}^d$, the *response of the system to the control* α is the unique continuous solution $y_x(\cdot) = y_{x,\alpha}(\cdot)$ of the differential equation

$$(ODE) \quad \begin{aligned} \frac{d^+ y_x(t)}{dt} &= g(y_x(t), d_i), \quad \theta_i \leq t \leq \theta_{i+1}, \quad i = 0, 1, \dots \\ y_x(0) &= x. \end{aligned}$$

Corresponding to each such control and associated response we consider the *cost index*

$$(2.2) \quad J_x^d(\alpha) \equiv \sum_{i=1}^{\infty} \left(\int_{\theta_{i-1}}^{\theta_i} f(y_x(s), d_{i-1}) e^{-\lambda s} ds + k(d_{i-1}, d_i) e^{-\lambda \theta_i} \right),$$

where λ is a given positive constant called the *discount factor*, $f: \mathbb{R}^n \times \{1, \dots, m\} \rightarrow \mathbb{R}$ is the *running cost*, and the constant $k(d, \hat{d})$ is the *cost of switching* from setting d to \hat{d} ($d, \hat{d} = 1, \dots, m$). We will assume

$$(2.3) \quad |f(x, d)| \leq D, \quad |f(x, d) - f(\hat{x}, d)| \leq D|x - \hat{x}|,$$

for some constant D and all $x, \hat{x} \in \mathbb{R}^n$, $d = 1, \dots, m$, and also

$$(2.4) \quad \begin{aligned} k(d, \hat{d}) &> 0, & k(d, d) &= 0, \\ k(d, \hat{d}) &< k(d, \tilde{d}) + k(\tilde{d}, \hat{d}), & d, \hat{d}, \tilde{d} &\in \{1, \dots, m\}, \quad d \neq \tilde{d} \neq \hat{d}. \end{aligned}$$

We interpret this assumption to mean that it is always cheaper to switch directly from setting d to setting \hat{d} , than to switch through an intermediate setting \tilde{d} .

Finally for each $d = 1, \dots, m$ and $x \in \mathbb{R}^n$ we define the *value function*

$$(2.5) \quad u^d(x) \equiv \inf_{\alpha \in \mathcal{A}^d} J_x^d(\alpha);$$

this is the minimum cost, provided we start at x with the initial control setting d .

Our goal is to design for each $x \in \mathbb{R}^n$, $d = 1, \dots, m$, an *optimal control* $\alpha^* = \alpha_x^* \in \mathcal{A}^d$ such that

$$(2.6) \quad u^d(x) = J_x^d(\alpha^*) = \min_{\alpha \in \mathcal{A}^d} J_x^d(\alpha).$$

Remark. Notice that we can also allow stopping the (ODE) and a resultant cost as a special case: if we incur a cost $\psi(y_x(\theta))$ should we stop (ODE) at time θ , we merely augment the problem above by setting

$$\begin{aligned} g(x, m+1) &\equiv 0, & x &\in \mathbb{R}^n, \\ f(x, m+1) &\equiv \frac{\psi(x)}{\lambda}, & x &\in \mathbb{R}^n. \end{aligned}$$

Remark. The restriction that $k(d, \tilde{d}) > 0$ is not truly essential and can be removed to yield results along the lines of Belbas–Lenhart [3].

3. Dynamic programming; optimal controls. Our plan is to show that the functions $u^d(\cdot)$ ($d = 1, \dots, m$) satisfy certain inequalities in \mathbb{R}^n and can thus be regarded as a kind of weak solution of the appropriate QVI of dynamic programming. This in turn will lead to the design of an optimal control.

First we prove the value functions are uniformly bounded and Hölder continuous.

LEMMA 3.1. (a) *There exists a constant C such that*

$$(3.1) \quad |u^d(x)| \leq C, \quad x \in \mathbb{R}^n, \quad d = 1, \dots, m.$$

(b) *There exists for each*

$$(3.2) \quad 0 < \gamma < \min\left(\frac{\lambda}{L}, 1\right)$$

a constant C_γ such that

$$(3.3) \quad |u^d(x) - u^d(\hat{x})| \leq C_\gamma |x - \hat{x}|^\gamma, \quad x, \hat{x} \in \mathbb{R}^n, \quad d = 1, \dots, m.$$

Proof. (a) The control,

$$\alpha = \{\theta_i, d_i\}_{i=0}^\infty$$

belongs to \mathcal{A}^d provided $\theta_0 = 0$, $\theta_i = +\infty$ ($i = 1, \dots$), $d_0 = d$. Thus,

$$u^d(x) \leq J_x^d(\alpha) = \int_0^\infty f(y_x(s), d) e^{-\lambda s} ds \leq \frac{D}{\lambda}.$$

On the other hand, for any control, $\alpha = \{\theta_i, d_i\}_{i=0}^\infty \in \mathcal{A}^d$, we have

$$J_x^d(\alpha) \geq \sum_{i=1}^\infty \int_{\theta_{i-1}}^{\theta_i} f(y_x(s), d_{i-1}) e^{-\lambda s} ds \geq -\frac{D}{\lambda}.$$

(b) Fix $d \in \{1, \dots, m\}$, $x, \hat{x} \in \mathbb{R}^n$. Choose $\varepsilon > 0$ and then $\alpha \in \mathcal{A}^d$, $\alpha = \{\theta_i, d_i\}_{i=0}^\infty$, such that

$$J_{\hat{x}}^d(\alpha) \leq u^d(\hat{x}) + \varepsilon.$$

We may assume $u^d(x) - u^d(\hat{x}) > 0$. Therefore,

$$\begin{aligned} |u^d(x) - u^d(\hat{x})| &\leq u^d(x) - J_{\hat{x}}^d(\alpha) + \varepsilon \\ &\leq J_x^d(\alpha) - J_{\hat{x}}^d(\alpha) + \varepsilon \\ &= \sum_{i=1}^\infty \left[\int_{\theta_{i-1}}^{\theta_i} (f(y_x(s), d_{i-1}) - f(y_{\hat{x}}(s), d_{i-1})) e^{-\lambda s} ds \right] + \varepsilon. \end{aligned}$$

Now Gronwall's lemma, (2.1) and (ODE) imply

$$|y_x(s) - y_{\hat{x}}(s)| \leq |x - \hat{x}| e^{Ls} \quad (s \geq 0),$$

and thus (2.3) implies

$$|f(y_x(s), d_{i-1}) - f(y_{\hat{x}}(s), d_{i-1})| \leq 2D|x - \hat{x}|^\gamma e^{L\gamma s}.$$

We employ this estimate above and note

$$\int_0^\infty e^{(L\gamma - \lambda)s} ds < \infty. \quad \square$$

Remark. This proof is adapted from [4].

Next we calculate the dynamic programming *optimality conditions* in the next proposition.

PROPOSITION 3.2. For each $d = 1, \dots, m$ and $x \in \mathbb{R}^n$,

$$(a) \quad u^d(x) \leq \min_{\tilde{d} \neq d} (u^{\tilde{d}}(x) + k(d, \tilde{d})), \quad x \in \mathbb{R}^n.$$

$$(b) \quad u^d(x) \leq \int_0^t f(y_x(s), d) e^{-\lambda s} ds + u^d(y_x(t)) e^{-\lambda t}$$

for all $t \geq 0$, where $dy_x(s)/ds = g(y_x(s), d)$ ($0 \leq s \leq t$).

(c) If, furthermore, for some point $x_0 \in \mathbb{R}^n$ a strict inequality holds in (a), then there exists $t_0 = t_{x_0} > 0$ such that

$$(3.4) \quad u^d(x_0) = \int_0^t f(y_{x_0}(s), d) e^{-\lambda s} ds + u^d(y_{x_0}(t)) e^{-\lambda t}$$

for all $0 \leq t \leq t_0$.

Proof. (a) Fix any $\tilde{d} \in \{1, \dots, m\}$, $\tilde{d} \neq d$, and choose any control $\alpha \in \mathcal{A}^{\tilde{d}}$,

$$\alpha = \{\theta_i, d_i\}_{i=0}^\infty.$$

Define $\tilde{\alpha} \in \mathcal{A}^d$ by

$$\tilde{\alpha} = \{\tilde{\theta}_i, \tilde{d}_i\}_{i=0}^{\infty},$$

where

$$\tilde{\theta}_i = \theta_{i-1}, \quad \tilde{d}_i = d_{i-1} (i = 1, \dots), \quad \tilde{\theta}_0 = 0, \quad \tilde{d}_0 = d.$$

Then

$$u^d(x) \leq J_x^d(\tilde{\alpha}) = J_x^d(\alpha) + k(d, \tilde{d}).$$

This holds for all $\alpha \in \mathcal{A}^d$ and so

$$u^d(x) \leq u^{\tilde{d}}(x) + k(d, \tilde{d}), \quad \tilde{d} \neq d.$$

(b) Next choose any $t \geq 0$. Pick any

$$\alpha = \{\theta_i, d_i\}_{i=0}^{\infty} \in \mathcal{A}^d.$$

Define

$$\hat{\alpha} = \{\hat{\theta}_i, \hat{d}_i\}_{i=0}^{\infty} \in \mathcal{A}^d$$

by

$$\hat{\theta}_i = \theta_i + t, \quad \hat{d}_i = d_i (i = 1, \dots), \quad \hat{\theta}_0 = 0, \quad \hat{d}_0 = d.$$

Then

$$u^d(x) \leq J_x^d(\hat{\alpha}) = \int_0^t f(y_x(s), d) e^{-\lambda s} ds + J_{y_x(t)}^d(\alpha) e^{-\lambda t}.$$

(c) Assume now

$$(3.5) \quad \min_{\tilde{d} \neq d} (u^{\tilde{d}}(x_0) + k(d, \tilde{d})) - u^d(x_0) = \sigma > 0.$$

Choose $\alpha_\varepsilon = \{\theta_i^\varepsilon, d_i^\varepsilon\}_{i=0}^{\infty} \in \mathcal{A}^d$ such that

$$(3.6) \quad J_{x_0}^d(\alpha_\varepsilon) \leq u^d(x_0) + \varepsilon.$$

We claim that if ε is small enough, then

$$(3.7) \quad \theta_1^\varepsilon \geq t_0 > 0$$

for some suitably chosen t_0 , independent of ε . To see this, suppose to the contrary

$$0 \leq \theta_1^\varepsilon < t_0 \quad (\text{for } t_0 \text{ as selected below}).$$

Then since $d_1^\varepsilon \neq d$

$$J_{x_0}^d(\alpha_\varepsilon) = e^{-\lambda \theta_1^\varepsilon} k(d, d_1^\varepsilon) + e^{-\lambda \theta_1^\varepsilon} J_{y_{x_0}(\theta_1^\varepsilon)}^{d_1^\varepsilon}(\tilde{\alpha}_\varepsilon),$$

where $\tilde{\alpha}_\varepsilon = \{\tilde{\theta}_i^\varepsilon, \tilde{d}_i^\varepsilon\} \in \mathcal{A}^{d_1^\varepsilon}$, $\tilde{\theta}_i^\varepsilon = \theta_{i+1}^\varepsilon - \theta_1^\varepsilon$, $\tilde{d}_i^\varepsilon = d_{i+1}^\varepsilon$ ($i = 0, 1, \dots$). Thus

$$e^{\lambda \theta_1^\varepsilon} J_{x_0}^d(\alpha_\varepsilon) \geq k(d, d_1^\varepsilon) + u^{d_1^\varepsilon}(y_{x_0}(\theta_1^\varepsilon)).$$

Hence, (3.6) gives

$$(3.8) \quad k(d, d_1^\varepsilon) + u^{d_1^\varepsilon}(y_{x_0}(\theta_1^\varepsilon)) \leq e^{\lambda t_0} (u^d(x_0) + \varepsilon).$$

But Lemma 3.1 implies

$$|u^{d_1^\varepsilon}(y_{x_0}(\theta_1^\varepsilon)) - u^{d_1^\varepsilon}(x_0)| \leq C t_0^\gamma;$$

so that (3.8) yields

$$k(d, d_1^\varepsilon) + u^{d_1^\varepsilon}(x_0) \leq e^{\lambda t_0}(u^d(x_0) + \varepsilon) + Ct_0^\gamma,$$

with $d_1^\varepsilon \neq d$. This contradicts (3.5) for all small $\varepsilon > 0$, provided we first choose t_0 small enough. Estimate (3.7) is thus proved.

Hence for $0 \leq t \leq t_0$ and α_ε as above

$$J_{x_0}^d(\alpha_\varepsilon) = \int_0^t f(y_{x_0}(s), d) e^{-\lambda s} ds + e^{-\lambda t} J_{y_{x_0}(t)}^d(\hat{\alpha}_\varepsilon),$$

for $\hat{\alpha}_\varepsilon = \{\hat{\theta}_i^\varepsilon, \hat{d}_i^\varepsilon\} \in \mathcal{A}^d$, $\hat{\theta}_i^\varepsilon = \theta_i^\varepsilon - t$, $\hat{d}_i^\varepsilon = d_i^\varepsilon$ ($i = 1, \dots$), $\hat{\theta}_0^\varepsilon = 0$, $\hat{d}_0^\varepsilon = d$. Therefore,

$$u^d(x_0) + \varepsilon \geq J_{x_0}^d(\alpha_\varepsilon) \geq \int_0^t f(y_{x_0}(s), d) e^{-\lambda s} ds + e^{-\lambda t} u^d(y_{x_0}(t)).$$

Send $\varepsilon \searrow 0$ and recall (b) to complete the proof of (c). \square

Notation. In view of this proposition it will be convenient to define

$$M^d[u](x) \equiv \min_{\tilde{d} \neq d} (u^{\tilde{d}}(x) + k(d, \tilde{d})), \quad x \in \mathbb{R}^n, \quad d \in \{1, \dots, m\}.$$

Remark. By standard continuation arguments we have in fact

$$(3.9) \quad u^d(x_0) = \int_0^t f(y_{x_0}(s), d) e^{-\lambda s} ds + u^d(y_{x_0}(t)) e^{-\lambda t}$$

for all

$$(3.10) \quad 0 \leq t \leq t_0^* = \inf \{t > 0 | u^d(y_{x_0}(t)) = M^d[u](y_{x_0}(t))\}$$

(and $t_0^* = +\infty$ if the set in (3.10) is empty). \square

Next, we exploit the optimality conditions just proved to show the existence of optimal controls. For this fix $x \in \mathbb{R}^n$, $d \in \{1, \dots, m\}$. Let us define $\alpha^* = \{\theta_i, d_i\}_{i=0}^\infty \in \mathcal{A}^d$ this way:

$$(3.11) \quad \theta_0 = 0, \quad d_0 = d,$$

$$(3.12) \quad \theta_i = \begin{cases} \inf \{t > \theta_{i-1} | u^{d_{i-1}}(y_x(t)) = M^{d_{i-1}}[u](y_x(t))\}, \\ +\infty & \text{if the set above is empty,} \end{cases}$$

$$(3.13) \quad d_i = \begin{cases} \min \{d = 1, \dots, m, d \neq d_{i-1} | M^{d_{i-1}}[u](y_x(\theta_i)) \\ \quad = u^d(y_x(\theta_i)) + k(d_{i-1}, d)\} & \text{if } \theta_i < +\infty, \\ d_{i-1} & \text{if } \theta_i = +\infty. \end{cases}$$

(Notice that the special choice of d_i in (3.13) (in case $\theta_i < +\infty$) is made for the sake of definiteness. Actually any d_i such that

$$M^{d_{i-1}}[u](y_x(\theta_i)) = u^{d_i}(y_x(\theta_i)) + k(d_{i-1}, d_i),$$

will do.)

That $\lim_{i \rightarrow \infty} \theta_i = +\infty$ and hence α^* defined above is admissible is a consequence of the following estimate.

LEMMA 3.3. *There exists a constant $\sigma > 0$ such that*

$$(3.14) \quad \theta_i \geq \theta_{i-1} + \sigma \quad \text{for } i = 2, \dots$$

Proof. If $\theta_{i-1} = +\infty$, there is nothing to prove. Otherwise, assume

$$\theta_{i-1} \leq \theta_i \leq \theta_{i-1} + \sigma \quad (\sigma \text{ as selected below}).$$

Then there exists

$$(3.15) \quad \theta_i \leq t < \theta_{i-1} + \sigma$$

such that

$$u^{d_{i-1}}(y_x(t)) = M^{d_{i-1}}[u](y_x(t)) = u^{d_i}(y_x(t)) + k(d_{i-1}, d_i),$$

for some $d_i \neq d_{i-1}$. But also

$$\begin{aligned} u^{d_{i-2}}(y_x(\theta_{i-1})) &= M^{d_{i-2}}[u](y_x(\theta_{i-1})) \\ &= u^{d_{i-1}}(y_x(\theta_{i-1})) + k(d_{i-2}, d_{i-1}) \\ &\leq u^{d_i}(y_x(\theta_{i-1})) + k(d_{i-2}, d_i). \end{aligned}$$

We combine these estimates to obtain

$$\begin{aligned} k(d_{i-2}, d_{i-1}) + k(d_{i-1}, d_i) &\leq k(d_{i-2}, d_i) + u^{d_i}(y_x(\theta_{i-1})) - u^{d_i}(y_x(t)) \\ &\quad + u^{d_{i-1}}(y_x(t)) - u^{d_{i-1}}(y_x(\theta_{i-1})) \\ &\leq k(d_{i-2}, d_i) + 2C_\gamma L \sigma^\gamma, \end{aligned}$$

a contradiction to (2.4) if σ is small enough. \square

PROPOSITION 3.4. *The control α^* defined by (3.11)–(3.13) is optimal:*

$$(3.16) \quad u^d(x) = J_x^d(\alpha^*) = \min_{\alpha \in \mathcal{A}^d} J_x^d(\alpha).$$

Proof. If $\theta_{i-1} < \infty$, then by the Remark after Proposition 3.2 we have for each $0 < \varepsilon < \sigma$,

$$\begin{aligned} u^{d_{i-1}}(y_x(\theta_{i-1} + \varepsilon)) &= \int_0^{\theta_i - \theta_{i-1} - \varepsilon} f(y_x(\theta_{i-1} + \varepsilon + s), d_{i-1}) e^{-\lambda s} ds \\ &\quad + u^{d_{i-1}}(y_x(\theta_i)) e^{-\lambda(\theta_i - \theta_{i-1} - \varepsilon)}; \end{aligned}$$

and so

$$(3.17) \quad \begin{aligned} e^{-\lambda\theta_{i-1}} u^{d_{i-1}}(y_x(\theta_{i-1})) &= \int_{\theta_{i-1}}^{\theta_i} f(y_x(s), d_{i-1}) e^{-\lambda s} ds \\ &\quad + e^{-\lambda\theta_i} u^{d_{i-1}}(y_x(\theta_i)). \end{aligned}$$

On the other hand if $\theta_i < \infty$,

$$u^{d_{i-1}}(y_x(\theta_i)) = u^{d_i}(y_x(\theta_i)) + k(d_{i-1}, d_i);$$

so that

$$\begin{aligned} e^{-\lambda\theta_{i-1}} u^{d_{i-1}}(y_x(\theta_{i-1})) &= \int_{\theta_{i-1}}^{\theta_i} f(y_x(s), d_{i-1}) e^{-\lambda s} ds \\ &\quad + k(d_{i-1}, d_i) e^{-\lambda\theta_i} + e^{-\lambda\theta_i} u^{d_i}(y_x(\theta_i)). \end{aligned}$$

Sum this for $i = 1$ to infinity if $\theta_i < \infty$ for all i , and to N if $\theta_N < \infty$, $\theta_{N+1} = +\infty$; in the latter case add also (3.17) for $i = N + 1$. In either situation we obtain

$$\begin{aligned} u^d(x) &= \sum_{i=1}^{\infty} \int_{\theta_{i-1}}^{\theta_i} f(y_x(s), d_{i-1}) e^{-\lambda s} ds + k(d_{i-1}, d_i) e^{-\lambda \theta_i} \\ &= J_x^d(\alpha^*). \end{aligned} \quad \square$$

4. Viscosity solutions of the dynamic programming QVI. Proposition 3.4 asserts that the control prescribed by (3.11)–(3.13) is optimal (for any $x \in \mathbb{R}^n$, $d = 1, \dots, m$). In practice, however, this is not yet useful as the value functions u^d are themselves unknown. We therefore describe in this section and the next a (theoretical) construction of $u = (u^1, \dots, u^m)$ by PDE techniques.

The formal motivation for this is simple. If $x \in \mathbb{R}^n$ is fixed and if the value functions u^1, \dots, u^m were C^1 on \mathbb{R}^n , then Proposition 3.2(b) implies

$$(4.1) \quad \frac{u^d(x) - u^d(y_x(t))}{t} \leq \frac{1}{t} \int_0^t f(y_x(s), d) e^{-\lambda s} ds + u^d(y_x(t)) \left(\frac{e^{-\lambda t} - 1}{t} \right)$$

for all $t > 0$. Thus our sending $t \searrow 0$ yields $\lambda u^d(x) - g(x, d) \cdot Du^d(x) \leq f(x, d)$ for all $x \in \mathbb{R}^n$, $d = 1, \dots, m$. Since Proposition 3.2(c) implies an equality in (4.1) for $0 \leq t \leq t_0$ provided $u^d(x) < M^d[u](x)$, we obtain also

$$\lambda u^d(x) - g(x, d) \cdot Du^d(x) = f(x, d)$$

in that case.

We may summarize these conclusions by noting

$$(QVI) \quad \max \{ \lambda u^d - g^d \cdot Du^d - f^d, u^d - M^d[u] \} = 0 \quad \text{in } \mathbb{R}^n, \quad d = 1, \dots, m,$$

for $g^d = g(\cdot, d)$, $f^d = f(\cdot, d)$. This is a system of m first order differential inequalities called the *dynamic programming system of quasi-variational inequalities* (QVI).

This derivation of (QVI) is, however, not justified as we do not know that the value functions u^1, \dots, u^m are C^1 ; indeed this is generally false, as simple examples show. Now, conversely, it is not particularly difficult to show that any C^1 solution of (QVI) must in fact equal the value functions; but here again is a problem since (QVI) like other nonlinear first order PDE does not usually admit a C^1 solution.

As noted in § 1, we resolve this difficulty by identifying a new notion of weak solutions of (QVI); this is inspired by Crandall–Lions [6], Lions [12] and Crandall–Evans–Lions [7]. It will turn out that the value functions u^1, \dots, u^m are such a weak or “viscosity” solution and that these viscosity solutions are unique. Additionally, the PDE techniques to be presented in § 5 will yield a viscosity solution of (QVI), which—owing to the uniqueness assertion—must equal the value functions.

Denote by $BUC(\mathbb{R}^n)^m$, the space of bounded, uniformly continuous \mathbb{R}^m -valued functions on \mathbb{R}^n .

Motivated by [7], we make the following definition.

DEFINITION. A function $u = (u^1, \dots, u^m) \in BUC(\mathbb{R}^n)^m$ is called a *viscosity solution* of

$$(QVI) \quad \max \{ \lambda u^d - g^d \cdot Du^d - f^d, u^d - M^d[u] \} = 0 \quad \text{in } \mathbb{R}^n, \quad d = 1, \dots, m,$$

provided for each $d \in \{1, \dots, m\}$ and each $\phi \in C^1(\mathbb{R}^n)$,

(i) if $u^d - \phi$ attains a local maximum at $x_0 \in \mathbb{R}^n$, then

$$(4.2) \quad \max \{ \lambda u^d - g^d D\phi - f^d, u^d - M^d[u] \} \leq 0 \quad \text{at } x_0;$$

and

(ii) if $u^d - \phi$ attains a local minimum at $x_1 \in \mathbb{R}^n$, then

$$(4.3) \quad \max \{ \lambda u^d - g^d D\phi - f^d, u^d - M^d[u] \} \geq 0 \quad \text{at } x_1.$$

Note that the u^d need not have derivatives in any sense.

We observe next that the optimality conditions of dynamic programming imply that the value functions form a viscosity solution.

THEOREM 4.1. *Under the hypotheses in § 2, the value function $u = (u^1, \dots, u^m)$ (u^d defined by (2.5)) is a viscosity solution of (QVI).*

Proof. According to Lemma 3.1, $u \in \text{BUC}(\mathbb{R}^n)^m$. We must verify (i) and (ii) above. Thus let $\phi \in C^1(\mathbb{R}^n)$ and assume for some $d \in \{1, \dots, m\}$ that $u^d - \phi$ attains a local maximum at $x_0 \in \mathbb{R}^n$. Then

$$(4.4) \quad u^d(x_0) - \phi(x_0) \geq u^d(y_{x_0}(t)) - \phi(y_{x_0}(t)),$$

for all sufficiently small $t > 0$. Now Proposition 3.2(b) implies

$$\frac{u^d(x_0) - u^d(y_{x_0}(t))}{t} \leq \frac{1}{t} \int_0^t f(y_{x_0}(s), d) e^{-\lambda s} ds + u^d(y_{x_0}(t)) \left(\frac{e^{-\lambda t} - 1}{t} \right).$$

We employ (4.4) and then send $t \searrow 0$ to obtain from (ODE)

$$\lambda u^d(x_0) - g^d(x_0) \cdot D\phi(x_0) \leq f^d(x_0).$$

Since $u^d \leq M^d[u]$, $x \in \mathbb{R}^n$, according to Proposition 3.2(a), we have seen that (4.2) is valid.

If, on the other hand, $u^d - \phi$ attains a local minimum at $x_1 \in \mathbb{R}^n$, we have

$$(4.5) \quad u^d(x_1) - \phi(x_1) \leq u^d(y_{x_1}(t)) - \phi(y_{x_1}(t)),$$

should $t > 0$ be small enough. Now if

$$(4.6) \quad u^d(x_1) = M^d[u](x_1),$$

then (4.3) clearly holds and no further analysis is needed. Should instead a strict inequality obtain in (4.6) then according to Proposition 3.2(c) we have

$$\frac{u^d(x_1) - u^d(y_{x_1}(t))}{t} = \frac{1}{t} \int_0^t f(y_{x_1}(s), d) e^{-\lambda s} dt + u^d(y_{x_1}(t)) \left(\frac{e^{-\lambda t} - 1}{t} \right)$$

for all sufficiently small $t > 0$. Recall (4.5) and then send $t \searrow 0$ to obtain

$$\lambda u^d(x_1) - g^d(x_1) \cdot D\phi(x_1) \geq f^d(x_1).$$

This implies (4.3). \square

Remark. If $\lambda > L$, then the choice $\gamma = 1$ is feasible in Lemma 3.1. Hence u^d is Lipschitz continuous and therefore $u = (u^1, \dots, u^m)$ solves (QVI) almost everywhere.

Remark. P. L. Lions in [12] has already observed the relation of dynamic programming and viscosity solutions in a different context.

5. Uniqueness of viscosity solutions. For the following we assume for the moment $u = (u^1, \dots, u^m)$ is any viscosity solution of (QVI), and not necessarily the value functions (1.5).

LEMMA 5.1. *We have*

$$(5.1) \quad u^d(x) \leq M^d[u](x)$$

for each $x \in \mathbb{R}^n$, $d = 1, \dots, m$.

Proof. Suppose (5.1) fails for some point $x_0 \in \mathbb{R}^n$ and some $d \in \{1, \dots, m\}$. Then there exists $\tilde{d} \neq d$ such that

$$(5.2) \quad u^{\tilde{d}}(x) + k(d, \tilde{d}) < u^d(x)$$

for all x contained in some small ball B centered at x_0 . Now it is not difficult to show there exists a smooth function ϕ such that $u^d - \phi$ has a local maximum at some point $x_1 \in B$; hence, according to the definition of viscosity solution,

$$\max (\lambda u^d - g^d \cdot D\phi - f^d, u^d - M^d[u]) \leq 0 \quad \text{at } x_1.$$

Thus

$$u^d(x_1) \leq M^d[u](x_1) \leq u^{\tilde{d}}(x_1) + k(d, \tilde{d}),$$

a contradiction to (5.2). \square

The principal result of this section is the following uniqueness theorem. Our proof exploits some methods from [6], [7].

THEOREM 5.2. *Suppose $u = (u^1, \dots, u^m)$ and $v = (v^1, \dots, v^m)$ are viscosity solutions of (QVI). Then $u \equiv v$.*

Proof. Choose a smooth function $\gamma: \mathbb{R}^n \rightarrow \mathbb{R}$ satisfying

$$(5.3) \quad \begin{aligned} \gamma(0) &= 5N, & 0 \leq \gamma \leq 5N, & \quad |D\gamma| \leq 10N, \\ \gamma(x) &< 5N & \text{if } x \neq 0, & \quad \gamma(x) = 0 \quad \text{if } |x| \geq 1, \end{aligned}$$

and set

$$(5.4) \quad \gamma_\varepsilon(x) = \gamma\left(\frac{x}{\varepsilon}\right) \quad \text{for } \varepsilon > 0, \quad x \in \mathbb{R}^n.$$

Here

$$N = \max \{\|u\|, \|v\|\}.$$

Consider now the auxiliary functions $\Phi^d: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$,

$$\Phi^d(x, y) \equiv u^d(x) - v^d(y) + \gamma_\varepsilon(x - y) \quad (d = 1, \dots, m).$$

Next choose $(x_1, y_1) \in \mathbb{R}^{2n}$ such that

$$(5.5) \quad \max_{1 \leq d \leq m} \Phi^d(x_1, y_1) \geq \sup_{x, y} \max_{1 \leq d \leq m} \Phi^d(x, y) - \varepsilon.$$

Now select $\zeta: \mathbb{R}^{2n} \rightarrow \mathbb{R}$ such that

$$\begin{aligned} \zeta(x_1, y_1) &= 1, & 0 \leq \zeta \leq 1, & \quad |D\zeta| \leq 2, \\ \zeta(x, y) &< 1 & \text{if } (x, y) \neq (x_1, y_1), \\ \zeta(x, y) &= 0 & \text{if } |x - x_1|^2 + |y - y_1|^2 \geq 1. \end{aligned}$$

Finally, define

$$\begin{aligned} \Psi^d(x, y) &\equiv \Phi^d(x, y) + 2\varepsilon\zeta(x, y) \\ &= u^d(x) - v^d(y) + \gamma_\varepsilon(x - y) + 2\varepsilon\zeta(x, y) \quad (d = 1, \dots, m). \end{aligned}$$

Since

$$\Psi^d(x, y) = \Phi^d(x, y) \leq \sup_{x, y} \max_{1 \leq d \leq m} \Phi^d(x, y)$$

for each $d = 1, \dots, m$, $|x - x_1|^2 + |y - y_1|^2 \geq 1$, whereas,

$$\begin{aligned} \max_{1 \leq d \leq m} \Psi^d(x_1, y_1) &= \max_{1 \leq d \leq m} \Phi^d(x_1, y_1) + 2\varepsilon \\ &\geq \sup_{x, y} \max_{1 \leq d \leq m} \Phi^d(x, y) + \varepsilon \quad \text{by (5.5),} \end{aligned}$$

there exists a finite point (x_0, y_0) (with $|x_0 - x_1|^2 + |y_0 - y_1|^2 < 1$) and some $\tilde{d} \in \{1, \dots, m\}$, say $\tilde{d} = 1$, such that

$$(5.6) \quad \Psi^1(x_0, y_0) = \max_{x, y} \max_{1 \leq d \leq m} \Psi^d(x, y).$$

The mapping $x \mapsto \Psi^1(x, y_0) = u^1(x) - \phi(x)$, for $\phi(x) \equiv v^1(y_0) - \gamma_\varepsilon(x - y_0) - 2\varepsilon\zeta(x, y_0)$, thus attains its maximum at x_0 . Therefore

$$(5.7) \quad \begin{aligned} \max \{ \lambda u^1(x_0) + g^1(x_0) \cdot (D\gamma_\varepsilon(x_0 - y_0) + 2\varepsilon D\zeta(x_0, y_0)) - f^1(x_0), \\ u^1(x_0) - M^1[u](x_0) \} \leq 0. \end{aligned}$$

Analogously

$$y \mapsto -\Psi^1(x_0, y) = v^1(y) - \psi(y)$$

attains its minimum at y_0 for

$$\psi(y) \equiv u^1(x_0) + \gamma_\varepsilon(x_0 - y) + 2\varepsilon\zeta(x_0, y),$$

so that

$$(5.8) \quad \begin{aligned} \max \{ \lambda v^1(y_0) + g^1(y_0) \cdot (D\gamma_\varepsilon(x_0 - y_0) - 2\varepsilon D\zeta(x_0, y_0)) - f^1(y_0), \\ v^1(y_0) - M^1[v](y_0) \} \geq 0. \end{aligned}$$

We will return to inequalities (5.7) and (5.8) after we pause to prove

$$(5.9) \quad |x_0 - y_0| = o(\varepsilon) \quad \text{as } \varepsilon \searrow 0.$$

To see this, note first that if $|x_0 - y_0| > \varepsilon$, then

$$\Psi^1(x_0, y_0) \leq 2N + 2\varepsilon < 3N \quad \text{if } 2\varepsilon < N,$$

whereas for any x

$$\Psi^1(x, x) \geq 3N.$$

Thus $|x_0 - y_0| \leq \varepsilon$. We refine this estimate by observing

$$\begin{aligned} \Psi^1(x_0, y_0) &= u^1(x_0) - v^1(y_0) + \gamma_\varepsilon(x_0 - y_0) + 2\varepsilon\zeta(x_0, y_0) \\ &\geq \Psi^1(x_0, x_0) \\ &= u^1(x_0) - v^1(x_0) + \gamma_\varepsilon(0) + 2\varepsilon\zeta(x_0, x_0), \end{aligned}$$

whence

$$\gamma_\varepsilon(x_0 - y_0) \geq \gamma_\varepsilon(0) + v^1(y_0) - v^1(x_0) - 2\varepsilon = \gamma(0) - \omega(|x_0 - y_0|) - 2\varepsilon,$$

$\omega(\cdot)$ denoting the modulus of continuity of v^1 on \mathbb{R}^n . Since $|x_0 - y_0| \leq \varepsilon$,

$$\lim_{\varepsilon \rightarrow 0} \gamma\left(\frac{x_0 - y_0}{\varepsilon}\right) = \lim_{\varepsilon \rightarrow 0} \gamma_\varepsilon(x_0 - y_0) = 5N.$$

Hence (5.3) implies (5.9).

Now return to estimates (5.7) and (5.8) and consider two possibilities.

Case 1. Suppose $v^1(y_0) < M^1[v](y_0)$ in (5.8). Then (5.7) and (5.8) together imply

$$\lambda(u^1(x_0) - v^1(y_0)) \leq f^1(x_0) - f^1(y_0) + 8L\varepsilon + \frac{10}{\varepsilon}LN|x_0 - y_0| = o(1) \quad \text{as } \varepsilon \searrow 0.$$

For any x and any $d = 1, \dots, m$

$$\begin{aligned} u^d(x) - v^d(x) + \gamma_\varepsilon(0) + 2\varepsilon\zeta(x, x) &= \psi^d(x, x) \leq \psi^1(x_0, y_0) \\ &= u^1(x_0) - v^1(y_0) + \gamma_\varepsilon(x_0 - y_0) + 2\varepsilon\zeta(x_0, y_0), \end{aligned}$$

and so

$$u^d(x) - v^d(x) \leq o(1) \quad \text{as } \varepsilon \searrow 0.$$

This proves $u^d \leq v^d$ for all $d = 1, \dots, m$ and the opposite inequality follows by symmetry. Theorem 5.2 is proved should Case 1 obtain.

Case 2. Suppose $v^1(y_0) = M^1[v](y_0)$ in (5.8). Then there exists $\tilde{d} \in \{2, \dots, m\}$, say $\tilde{d} = 2$, such that

$$(5.10) \quad v^1(y_0) = v^2(y_0) + k(1, 2).$$

But since

$$\Psi^2(x_0, y_0) \leq \Psi^1(x_0, y_0),$$

we have

$$u^1(x_0) - u^2(x_0) \geq v^1(y_0) - v^2(y_0) = k(1, 2).$$

However,

$$u^1(x_0) \leq u^2(x_0) + k(1, 2)$$

according to Lemma 5.1; thus

$$u^1(x_0) = u^2(x_0) + k(1, 2).$$

Consequently this and (5.10) give

$$\Psi^2(x_0, y_0) = \Psi^1(x_0, y_0).$$

Now repeat the considerations above with the index 2 replacing 1. Should Case 1 hold we are done. Otherwise there exists $\tilde{d} \in \{1, \dots, m\}$, $\tilde{d} \neq 2$, such that

$$v^2(y_0) = v^{\tilde{d}}(y_0) + k(2, \tilde{d}).$$

Since $k(1, 2), k(2, \tilde{d}) > 0$, the possibility $\tilde{d} = 1$ is precluded by (5.10). Hence we may assume $\tilde{d} = 3$ and prove as above

$$\Psi^3(x_0, y_0) = \Psi^2(x_0, y_0) = \Psi^1(x_0, y_0).$$

Repeat the preceding calculations with the index 3 replacing 2, etc. After finitely many steps we reach an index $\tilde{d} \leq m$ for which Case 1 holds. \square

Remark. The conclusion of the proof is reminiscent of [3, Prop. 1.1]. \square

6. Existence of viscosity solutions. In this section we construct by PDE techniques the viscosity solution of (QVI). In light of the uniqueness assertion in § 5 this procedure must yield the value functions (2.5).

We will obtain our solution as the limit of the solution $u_\varepsilon = (u_\varepsilon^1, \dots, u_\varepsilon^m)$ of the *penalized system* (cf. [9]):

$$(6.1) \quad -\varepsilon \Delta u_\varepsilon^d + \lambda u_\varepsilon^d - g^d \cdot Du_\varepsilon^d + \sum_{\substack{\tilde{d}=1 \\ \tilde{d} \neq d}}^m \beta_\varepsilon(u_\varepsilon^d - u_\varepsilon^{\tilde{d}} - k(d, \tilde{d})) = f^d \quad (d = 1, \dots, m),$$

where

$$\beta_\varepsilon(x) = \beta\left(\frac{x}{\varepsilon}\right) \quad (\varepsilon > 0, x \in \mathbb{R}^1)$$

for some smooth function $\beta: \mathbb{R} \rightarrow \mathbb{R}$ satisfying

$$\begin{aligned} \beta(x) &= 0 \quad \text{if } x \leq 0, & \beta(x) &> 0 \quad \text{if } x > 0, \\ \beta'' &\geq 0, & 0 &\leq \beta' \leq 1. \end{aligned}$$

Standard PDE methods imply that $(6.1)_\varepsilon$ has a unique, smooth solution $u_\varepsilon = (u_\varepsilon^1, \dots, u_\varepsilon^m)$ (cf. [9]).

First we prove an estimate similar to that in Lemma 3.1.

LEMMA 6.1. (a) *There exists a constant C such that*

$$(6.2) \quad |u_\varepsilon^d(x)| \leq C \quad (x \in \mathbb{R}^n, \varepsilon > 0, d = 1, \dots, m).$$

(b) *There exists for each*

$$(6.3) \quad 0 < \gamma < \min\left(\frac{\lambda}{L}, 1\right)$$

a constant C_γ such that

$$(6.4) \quad |u_\varepsilon^d(x) - u_\varepsilon^d(\hat{x})| \leq C_\gamma |x - \hat{x}|^\gamma \quad (x \in \mathbb{R}^n, \varepsilon > 0, d = 1, \dots, m).$$

Proof. To simplify notation we delete the subscripts ε .

(a) If there exists a finite point $x_0 \in \mathbb{R}^n$ at which

$$(6.5) \quad u^{\hat{d}}(x_0) = \min_d \min_x u^d(x),$$

then

$$Du^{\hat{d}}(x_0) = 0, \quad -\Delta u^{\hat{d}}(x_0) \leq 0,$$

and

$$u^{\hat{d}}(x_0) - u^d(x_0) - k(\hat{d}, d) < 0 \quad \text{if } d \neq \hat{d}.$$

Thus $(6.1)_\varepsilon$ implies

$$\min_d \min_x u^d(x) = u^{\hat{d}}(x_0) \geq \frac{1}{\lambda} f^{\hat{d}}(x_0) \geq -\frac{D}{\lambda}.$$

If (6.5) does not occur at any point x_0 , we argue as in § 5 by considering

$$w^d = u^d - 2\delta\zeta \quad (d \in \{1, \dots, m\}),$$

where $\delta > 0$ and ζ is chosen to force the minimum to occur at a finite point. We apply the reasoning above to the w^d and later send $\delta \searrow 0$.

A proof that $u^d(x) \leq D/\lambda$ ($x \in \mathbb{R}^n$, $d = 1, \dots, m$) is similar.

(b) Define

$$\Phi^d(x, y) \equiv \frac{|u^d(x+y) - u^d(x)|}{|y|^\gamma}$$

and assume there exists a finite point (x_0, y_0) at which

$$(6.6) \quad \Phi^{\hat{d}}(x_0, y_0) = \max_d \max_{x,y} \Phi^d(x, y)$$

for some index \hat{d} . We may assume $u^{\hat{d}}(x_0+y_0) - u^{\hat{d}}(x_0) > 0$, $y_0 \neq 0$. Then

$$0 = \frac{\partial}{\partial x_i} \Phi^{\hat{d}}(x_0, y_0) = \frac{u_{x_i}^{\hat{d}}(x_0+y_0) - u_{x_i}^{\hat{d}}(x_0)}{|y_0|^\gamma} \quad (1 \leq i \leq n)$$

and

$$0 = \frac{\partial}{\partial y_i} \Phi^{\hat{d}}(x_0, y_0) = \frac{u_{x_i}^{\hat{d}}(x_0+y_0)}{|y_0|^\gamma} - \frac{\gamma(u^{\hat{d}}(x_0+y_0) - u^{\hat{d}}(x_0))y_0^i}{|y_0|^{\gamma+2}} \quad (1 \leq i \leq n).$$

Hence

$$(6.7) \quad \begin{aligned} Du^{\hat{d}}(x_0+y_0) &= Du^{\hat{d}}(x_0), \\ Du^{\hat{d}}(x_0+y_0) &= \frac{\gamma}{|y_0|^2} (u^{\hat{d}}(x_0+y_0) - u^{\hat{d}}(x_0))y_0. \end{aligned}$$

Furthermore

$$(6.8) \quad 0 \leq -\Delta_x \Phi^{\hat{d}}(x_0, y_0) = \frac{-(\Delta u^{\hat{d}}(x_0+y_0) - \Delta u^{\hat{d}}(x_0))}{|y_0|^\gamma}.$$

In addition, for each $d \neq \hat{d}$,

$$(6.9) \quad u^{\hat{d}}(x_0+y_0) - u^d(x_0+y_0) \geq u^{\hat{d}}(x_0) - u^d(x_0).$$

Evaluate (6.1)_ε for $d = \hat{d}$ at x_0 and x_0+y_0 , subtract the resulting expressions, and then simplify using (6.7) and (6.9) to obtain

$$\begin{aligned} \lambda |u^{\hat{d}}(x_0+y_0) - u^{\hat{d}}(x_0)| &\leq |g(x_0+y_0, \hat{d}) - g(x_0, \hat{d})| \frac{\gamma}{|y_0|} |u^{\hat{d}}(x_0+y_0) - u^{\hat{d}}(x_0)| \\ &\quad + |f(x_0+y_0, \hat{d}) - f(x_0, \hat{d})| \\ &\leq L\gamma |u^{\hat{d}}(x_0+y_0) - u^{\hat{d}}(x_0)| + C|y_0|^\gamma. \end{aligned}$$

As $L\gamma < \lambda$, we obtain

$$\max_{x,y} \max_{d=1,\dots,m} \Phi^d(x, y) = \Phi^{\hat{d}}(x_0, y_0) \leq C(\lambda - L\gamma)^{-1}.$$

If no point satisfying (6.6) exists, we consider

$$\Psi^d(x, y) = \Phi^d(x, y) + 2\delta\zeta(x, y) \quad (d = 1, \dots, m),$$

where $\delta > 0$ and ζ is selected to force the maximum to occur at a finite point. \square

THEOREM 6.2. As $\varepsilon \searrow 0$

$$(6.10) \quad u_\varepsilon^d \rightarrow u^d \text{ locally uniformly on } \mathbb{R}^n, d = 1, \dots, m,$$

where $u = (u^1, \dots, u^m)$ is the (unique) viscosity solution of (QVI).

Remark. In particular

$$u_\varepsilon^d(x) \rightarrow \inf_\alpha J_x^d(\alpha) \quad \text{as } \varepsilon \searrow 0.$$

Proof. In view of Lemma 6.1, there exists a subsequence $\varepsilon_i \searrow 0$ and bounded, Hölder continuous functions u^d such that

$$u_{\varepsilon_i}^d \rightarrow u^d \text{ locally uniformly, } d = 1, \dots, m.$$

We will prove $u = (u^1, \dots, u^m)$ is a viscosity solution of (QVI); by uniqueness then in fact

$$u_\varepsilon^d \rightarrow u^d \quad \text{as } \varepsilon \searrow 0.$$

First, we claim

$$(6.11) \quad u^d \leq M^d[u] \quad x \in \mathbb{R}^n, \quad d = 1, \dots, m.$$

If this were not so, there would exist $\hat{d} \neq d$ in $\{1, \dots, m\}$, $\delta > 0$, and some nonempty open ball B such that

$$u^d(x) \geq u^{\hat{d}}(x) + k(d, \hat{d}) + 2\delta \quad x \in \bar{B}.$$

As $u_{\varepsilon_i}^d \rightarrow u^d$, $u_{\varepsilon_i}^{\hat{d}} \rightarrow u^{\hat{d}}$ uniformly on \bar{B} , we have

$$(6.12) \quad u_{\varepsilon_i}^d(x) \geq u_{\varepsilon_i}^{\hat{d}}(x) + k(d, \hat{d}) + \delta, \quad x \in \bar{B},$$

for all sufficiently small ε_i .

Next, there exists a C^2 function ϕ such that $u^d - \phi$ attains a local maximum in B , say at x_0 . We may in fact assume $u^d - \phi$ to have a strict local maximum at x_0 (cf. [7]). Since $u_{\varepsilon_i}^d \rightarrow u^d$ uniformly on B , $u_{\varepsilon_i}^d - \phi$ also attains a local maximum at $x_{\varepsilon_i} \in B$, for all ε_i small enough. Now,

$$Du_{\varepsilon_i}^d(x_{\varepsilon_i}) = D\phi(x_{\varepsilon_i})$$

and

$$-\Delta(u_{\varepsilon_i}^d - \phi)(x_{\varepsilon_i}) \geq 0,$$

so that (6.1) _{ε} implies

$$\lambda u_{\varepsilon_i}^d - g^d \cdot D\phi + \sum_{\substack{d=1 \\ d \neq d}}^m \beta_{\varepsilon_i}(u_{\varepsilon_i}^d - u_{\varepsilon_i}^{\hat{d}} - k(d, \hat{d})) \leq f^d + \varepsilon_i \Delta\phi \quad \text{at } x_{\varepsilon_i}.$$

Hence (6.11) implies

$$\beta_{\varepsilon_i}(\delta) \leq \beta_{\varepsilon_i}(u_{\varepsilon_i}^d - u_{\varepsilon_i}^{\hat{d}} - k(d, \hat{d})) \leq C \quad \text{at } x_{\varepsilon_i},$$

for some constant C independent of ε_i . This is a contradiction, as $\beta_\varepsilon(x) \rightarrow +\infty$ as $\varepsilon \searrow 0$ if $x > 0$. This proves (6.7).

Now we verify (4.2) and (4.3). Fix $d \in \{1, \dots, m\}$ and suppose $u^d - \phi$ attains a strict local maximum at some point x_0 , where $\phi \in C^2(\mathbb{R}^n)$.

Then $u_{\varepsilon_i}^d - \phi$ attains a strict local maximum at some nearby point x_{ε_i} , whence—as above—

$$\lambda u_{\varepsilon_i}^d - g^d \cdot D\phi \leq f^d + \varepsilon_i \Delta\phi \quad \text{at } x_{\varepsilon_i}.$$

Let $\varepsilon \searrow 0$, $x_{\varepsilon_i} \rightarrow x_0$ to find

$$\lambda u^d - g^d \cdot D\phi - f^d \leq 0 \quad \text{at } x_0.$$

This and (6.7) prove (4.2) at x_0 . In the general case that $\phi \in C^1(\mathbb{R}^n)$ and x_0 is not a strict local maximum, we approximate as in [7] and apply the argument above.

Suppose now $\phi \in C^2$ and $u^d - \phi$ attains a strict local minimum at x_1 . If $u^d(x_1) = M^d[u](x_1)$, then (4.3) is valid. Otherwise

$$u_{\varepsilon_i}^d < u_{\varepsilon_i}^{\tilde{d}} + k(d, \tilde{d}) \quad \text{near } x_1$$

for all $\tilde{d} = 1, \dots, m$, $\tilde{d} \neq d$, and all sufficiently small ε_i . Thus (6.1) _{ε} implies

$$(6.12) \quad -\varepsilon_i \Delta u_{\varepsilon_i}^d + \lambda u_{\varepsilon_i}^d - g^d \cdot Du_{\varepsilon_i}^d = f^d \quad \text{near } x_1$$

for all small enough ε_i . Furthermore

$$u_{\varepsilon_i}^d - \phi$$

attains a local minimum at some point x_{ε_i} in the region where (6.12) holds, and $x_{\varepsilon_i} \rightarrow x_1$. Reasoning as above we have

$$\lambda u_{\varepsilon_i}^d - g^d \cdot D\phi \geq f^d + \varepsilon_i \Delta \phi \quad \text{at } x_{\varepsilon_i}.$$

Thus

$$\lambda u^d - g^d \cdot D\phi \geq f^d \quad \text{at } x_1,$$

and so (4.3) obtains in this case as well.

As before, we may approximate if $\phi \in C^1(\mathbb{R}^n)$ and x_1 is only a local minimum. \square

7. Convergence as switching costs tend to zero. If we return to our control theory problem in § 2 and now instead of (2.4) assume $k(d, \tilde{d}) = 0$, ($d, \tilde{d} = 1, \dots, m$), then it is not hard to check that $u^1 = u^2 = \dots = u^m \equiv u$; that is, the minimum cost does not depend on the initial setting d of the control. (This is because we could always immediately switch to the best setting at no cost.) Furthermore the formal calculations of dynamic programming imply that if u were C^1 , then it would solve the *Hamilton–Jacobi–Bellman* type equation

$$(HJB) \quad \max_{d=1, \dots, m} \{\lambda u - g^d \cdot Du - f^d\} = 0 \quad \text{in } \mathbb{R}^n$$

(cf. [9]). In general of course u is not C^1 , but it is the (unique) viscosity solution of (HJB): see P. L. Lions [12]. This means that, for each $\phi \in C^1(\mathbb{R}^n)$,

(i) if $u - \phi$ attains a local maximum at x_0 , then

$$(7.1) \quad \max_{d=1, \dots, m} \{\lambda u - g^d \cdot D\phi - f^d\} \leq 0 \quad \text{at } x_0,$$

and

(ii) if $u - \phi$ attains a local minimum at x_1 , then

$$(7.2) \quad \max_{d=1, \dots, m} \{\lambda u - g^d \cdot D\phi - f^d\} \geq 0 \quad \text{at } x_1.$$

We prove now that as the switching costs tend to zero, the value functions (2.5) each converge to u . This is a deterministic analogue to the principal result of Evans–Friedman [9].

THEOREM 7.1. *Suppose for each $\varepsilon > 0$, $u_\varepsilon = (u_\varepsilon^1, \dots, u_\varepsilon^m)$ is the viscosity solution of (QVI), with switching costs $\{k^\varepsilon(d, \hat{d})\}$ satisfying (2.4). If*

$$k^\varepsilon(d, \hat{d}) \rightarrow 0 \quad \text{as } \varepsilon \rightarrow 0, \quad d, \hat{d} = 1, \dots, m.$$

then

$$u_{\varepsilon}^d \rightarrow u \quad \text{as } \varepsilon \rightarrow 0, \quad d = 1, \dots, m,$$

where u is the viscosity solution of (HJB).

Proof. As in § 6 there exists a subsequence ε_i such that

$$u_{\varepsilon_i}^d \rightarrow u^d \text{ locally uniformly on } \mathbb{R}^n, \quad d = 1, \dots, m.$$

Since

$$u_{\varepsilon_i}^d \leq M^d[u_{\varepsilon_i}] \leq u_{\varepsilon_i}^{\tilde{d}} + k^{\varepsilon_i}(d, \tilde{d}) \quad (\tilde{d} \neq d),$$

we have

$$u^d \leq u^{\tilde{d}}$$

for all $\tilde{d} \in \{1, \dots, m\}$, $\tilde{d} \neq d$, $x \in \mathbb{R}^n$. Thus

$$u^1 = u^2 = \dots = u^m \equiv u.$$

We will prove that u is the viscosity solution of (HJB). Suppose $\phi \in C^1$ and $u - \phi$ attains a strict local maximum at x_0 . Then, for each $d \in \{1, \dots, m\}$ and each sufficiently small ε_i , $u_{\varepsilon_i}^d - \phi$ attains a local maximum at $x_{\varepsilon_i}^d$ near x_0 . Since u_{ε_i} is the viscosity solution of (QVI) with $\{k^{\varepsilon}(d, \tilde{d})\}$, (4.2) implies

$$\lambda u_{\varepsilon_i}^d - g^d \cdot D\phi - f^d \leq 0 \quad \text{at } x_{\varepsilon_i}^d.$$

Let $\varepsilon_i \searrow 0$:

$$\lambda u - g^d \cdot D\phi - f^d \leq 0 \quad \text{at } x_0,$$

for each $d \in \{1, \dots, m\}$. Hence,

$$\max_{d=1, \dots, m} \{\lambda u - g^d \cdot D\phi - f^d\} \leq 0.$$

Conversely, suppose $u - \phi$ has a strict local minimum at some point x_1 . Then, for each $d \in \{1, \dots, m\}$ and each ε_i small enough, $u_{\varepsilon_i}^d - \phi$ attains a strict local minimum near x_1 . Choose $d_i \in \{1, \dots, m\}$ and x_{ε_i} such that

$$(u_{\varepsilon_i}^{d_i} - \phi)(x_{\varepsilon_i}) = \min_x \min_{\tilde{d}=1, \dots, m} (u_{\varepsilon_i}^{\tilde{d}} - \phi)(x),$$

the first minimum taken over all x in some neighborhood of x_1 . Thus (4.3) implies

$$\max \{\lambda u_{\varepsilon_i}^{d_i} - g^{d_i} \cdot D\phi - f^{d_i}, u_{\varepsilon_i}^{d_i} - M^{d_i}[u_{\varepsilon_i}]\} \geq 0 \quad \text{at } x_{\varepsilon_i}.$$

But

$$(u_{\varepsilon_i}^{d_i} - \phi)(x_{\varepsilon_i}) \leq (u_{\varepsilon_i}^{\tilde{d}} - \phi)(x_{\varepsilon_i}), \quad \tilde{d} = 1, \dots, m,$$

and so

$$u_{\varepsilon_i}^{d_i} < m^{d_i}[u_{\varepsilon_i}] \quad \text{at } x_{\varepsilon_i}.$$

Hence

$$\lambda u_{\varepsilon_i}^{d_i} - g^{d_i} \cdot D\phi - f^{d_i} \geq 0 \quad \text{at } x_{\varepsilon_i}.$$

Passing if necessary to a further subsequence, we may assume

$$d_i \rightarrow d_0.$$

Thus, if we send $\varepsilon_i \rightarrow 0$, we deduce

$$\lambda u - g^{d_0} \cdot D\phi - f^{d_0} \geq 0 \quad \text{at } x_1.$$

Consequently

$$\max_{d=1, \dots, m} \{\lambda u - g^d \cdot D\phi - f^d\} \geq 0 \quad \text{at } x_1,$$

as required. \square

REFERENCES

- [1] N. E. BARRON, L. C. EVANS AND R. JENSEN, *Viscosity solution of Isaacs' equations and differential games with Lipschitz controls*, to appear.
- [2] S. A. BELBAS, *Optimal switching control of large scale energy generation systems: the method of quasi-variational inequalities*, to appear.
- [3] S. A. BELBAS AND S. LENHART, *A system of nonlinear partial differential equations arising in the optimal control of stochastic systems with switching costs*, to appear.
- [4] I. CAPUZZO DOLCETTA AND M. MATZEU, *On the dynamic programming inequalities associated with the optimal stopping problem in discrete and continuous time*, Num. Funct. Anal. and Optim., (1981), pp. 425–450.
- [5] I. CAPUZZO DOLCETTA, M. MATZEU AND J. L. MENALDI, *On a system of first order quasi-variational inequalities connected with the optimal switching problem*, to appear.
- [6] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., to appear.
- [7] M. G. CRANDALL, L. C. EVANS AND P. L. LIONS, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., to appear.
- [8] L. C. EVANS, *Classical solutions of the Hamilton-Jacobi-Bellman equations for uniformly elliptic operators*, to appear in Trans. Amer. Math. Soc.
- [9] L. C. EVANS AND A. FRIEDMAN, *Optimal stochastic switching and the Dirichlet problem for the Bellman equation*, Trans. Amer. Math. Soc., 253 (1979), pp. 365–389.
- [10] L. C. EVANS AND P. L. LIONS, *Deux résultats de régularité pour le problème de Bellman-Dirichlet*, C. R. Acad. Sci. Paris, 286 (1978), pp. 587–589.
- [11] A. FRIEDMAN, *Differential Games*, John Wiley, New York, 1971.
- [12] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, London, 1982.
- [13] ———, *Resolution analytique des problèmes de Bellman-Dirichlet*, Acta. Math., 146 (1981), pp. 151–166.
- [14] J. L. MENALDI, *Le problème de temps d'arrêt optimal déterministe et l'inéquation variationnelle du premier ordre associée*, Appl. Math. and Opt., 8 (1982), pp. 131–158.
- [15] P. E. SOUGANIDIS, to appear.
- [16] G. BARLES, *Thèse de 3^e cycle*, Univ. Paris IX—Dauphine, Paris, 1982–83.

LOCAL DUALITY OF NONLINEAR PROGRAMS*

O. FUJIWARA†, S.-P. HAN‡ AND O. L. MANGASARIAN§

Abstract. It is shown that the second order sufficient (necessary) optimality condition for the dual of a nonlinear program is equivalent to the inverse of the Hessian of the Lagrangian being positive definite (semidefinite) on the normal cone to the local primal constraint surface. This compares with the Hessian itself being positive definite (semidefinite) on the tangent cone on the local primal constraint surface for the corresponding second order condition for the primal problem. We also show that primal second order sufficiency (necessity) and dual second order necessity (sufficiency) is essentially equivalent to the Hessian of the Lagrangian being positive definite. This follows from the following interesting linear algebra result: a necessary and sufficient condition for a nonsingular symmetric $n \times n$ matrix to be positive definite is that for some subspace of R^n , the matrix must be positive definite on the subspace and its inverse be positive semidefinite on the orthogonal complement of the subspace.

Key words: nonlinear programming, second order optimality, duality

AMS (MOS) subject classifications. 90C30, 15A03

1. Introduction. We consider the following nonlinear program

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{(P)} & \\ \text{subject to} & g(x) \leq 0, \\ & h(x) = 0, \end{array}$$

and its Wolfe dual [16], [9]

$$\begin{array}{ll} \text{maximize} & L(x, u, v) \\ \text{(D)} & \\ \text{subject to} & \nabla_x L(x, u, v) = 0, \\ & u \geq 0, \end{array}$$

where $f: R^n \rightarrow R$, $g: R^n \rightarrow R^m$ and $h: R^n \rightarrow R^q$ are differentiable functions on R^n , $L(x, u, v) := f(x) + u^T g(x) + v^T h(x)$ is the standard Lagrangian and $\nabla_x L$ is the gradient with respect to x . The relationships between the above two problems have been extensively studied for the convex case [16], [9]. Our principal concern here is local duality results which in the absence of convexity assumptions require the use of second order optimality conditions.

In § 2 we give a geometrically meaningful second order sufficient optimality condition in Definition 2.1 for the dual problem (D) and prove in Theorem 2.2 that it is equivalent to the standard second order sufficient optimality condition [8], [2] applied to the dual problem (D). Thus, as is well known, the second order sufficient optimality condition for the primal problem (P) is the positive definiteness of the Hessian of the Lagrangian on the tangent cone to the local constraint surface. Our second order sufficient (necessary) optimality condition Theorem 2.2 (Theorem 2.5) for the dual problem (D) is that the *inverse* of the Hessian of the Lagrangian is positive

* Received by the editors January 20, 1982, and in revised form January 31, 1983. This research was sponsored by the U.S. Army under contract DAAG29-80-C-0041. This material is based on work supported by the National Science Foundation under grants ENG-7903881 and MCS-7901066.

† Asian Institute of Technology, P.O. Box 2754, Bangkok, Thailand.

‡ Department of Mathematics, University of Illinois, Urbana, Illinois 61801.

§ Computer Sciences Department, University of Wisconsin-Madison, Madison, Wisconsin 53706.

definite (semidefinite) on the *normal* cone to the primal local constraint surface. It is worthwhile to note that while positive definiteness of the Hessian of the Lagrangian ensures the satisfaction of the second order sufficiency condition for the primal problem, this is not the case for the dual problem where a constraint qualification is needed (Theorem 2.3) in order to ensure that dual second order sufficiency holds under positive definiteness of the Hessian of the Lagrangian.

In § 3 we characterize Karush–Kuhn–Tucker points of the primal problem that locally solve both the primal and dual problems simultaneously. We show (Theorems 3.2 and 3.3) that these points are essentially points where the Hessian of the Lagrangian with respect to the primal variables is positive definite. In order to establish these results we prove an interesting result of linear algebra (Theorem 3.1) which states that a necessary and sufficient condition for a nonsingular symmetric $n \times n$ matrix to be positive definite is that for some subspace S of R^n , A must be positive definite on S and A^{-1} must be positive semidefinite on the orthogonal complement S^\perp of S .

We briefly describe our notation now. All vectors will be column vectors unless transposed to a row vector by the superscript T . For x in the n -dimensional real Euclidean space R^n , $x_i, i = 1, \dots, n$, will denote its components. For an $m \times n$ real matrix we shall say that $A \in R^{m \times n}$, A_i will denote the i th row of A , and if $I \subset \{1, \dots, m\}$ then A_I will denote the submatrix with rows $A_i, i \in I$. For a differentiable function $g: R^n \rightarrow R^m$, $\nabla g(x)$ will denote the transpose of the $m \times n$ Jacobian matrix of g at x . For a twice differentiable function $L: R^{n+m} \rightarrow R$, $\nabla_x L(x, u)$ will denote the $n \times 1$ gradient with respect to x , $\nabla_u L(x, u)$ will denote the $m \times 1$ gradient with respect to u , $\nabla^2 L(x, u)$ will denote the $(n+m) \times (n+m)$ Hessian with respect to both x and u whose submatrix components are denoted as follows:

$$\nabla^2 L(x, u) = \begin{bmatrix} \nabla_{xx} L(x, u) & \nabla_{xu} L(x, u) \\ \nabla_{ux} L(x, u) & \nabla_{uu} L(x, u) \end{bmatrix}.$$

2. Geometrically meaningful second order optimality condition for the dual problem. In order to establish local duality results without any convexity assumptions, second order necessary and sufficient optimality conditions become essential. The second order sufficient condition for the primal program (P) was given by McCormick and Fiacco [8], [2] and has been extensively studied. Research on this topic continues (see, for example, [4], [15]). In this section we formulate geometrically meaningful second order necessary and sufficient optimality conditions for the dual problem (D) and study the relationship to the corresponding conditions for the primal.

Recall that an $(n+m+q)$ -vector $(\bar{x}, \bar{u}, \bar{v})$ is said to be a *Karush–Kuhn–Tucker triple* of the primal program (P) if the following conditions hold:

$$\begin{aligned} (a) \quad & \nabla_x L(\bar{x}, \bar{u}, \bar{v}) = 0, \\ (b) \quad & g(\bar{x}) \leq 0, \\ (2.1) \quad (c) \quad & h(\bar{x}) = 0, \\ (d) \quad & \bar{u} \geq 0, \\ (e) \quad & \bar{u}_i g_i(\bar{x}) = 0, \quad i = 1, \dots, m. \end{aligned}$$

Such a triple is said to satisfy *the primal second order sufficient optimality condition*

if f , g and h are twice continuously differentiable at \bar{x} and

$$(2.2) \quad \left. \begin{aligned} \nabla g_J(\bar{x})^T d &= 0 \\ \nabla g_K(\bar{x})^T d &\leq 0 \\ \nabla h(\bar{x})^T d &= 0 \\ d &\neq 0 \end{aligned} \right\} \Rightarrow d^T \nabla_{xx} L(\bar{x}, \bar{u}, \bar{v}) d > 0,$$

where

$$J := \{i | g_i(\bar{x}) = 0 \text{ and } \bar{u}_i > 0\}, \quad K := \{i | g_i(\bar{x}) = 0 \text{ and } \bar{u}_i = 0\}.$$

We shall refer to (2.2) as McCormick's second order sufficient optimality condition.

We now give a second order sufficient optimality condition for the dual program (D) which we shall justify by Theorem 2.2 below.

DEFINITION 2.1. A Karush–Kuhn–Tucker triple $(\bar{x}, \bar{u}, \bar{v})$ of the primal problem (P) is said to satisfy the *dual second order sufficient optimality condition* if f , g and h are twice continuously differentiable at \bar{x} , if the Hessian $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})$ is nonsingular and

$$(2.3) \quad \left. \begin{aligned} w &= \nabla g(\bar{x})y + \nabla h(\bar{x})z, \\ y_i &= 0, \quad i \in I, \\ y_i &\geq 0, \quad i \in K, \\ (y, z) &\neq 0 \end{aligned} \right\} \Rightarrow w^T \nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})^{-1} w > 0,$$

where $I := \{i | g_i(\bar{x}) < 0\}$ and $K := \{i | g_i(\bar{x}) = 0, \bar{u}_i = 0\}$.

The geometric relationship between the primal and the dual second order sufficient conditions is an interesting one. Let T be a tangent cone of the local primal constraint surface at the point \bar{x} induced by the second order optimality condition (2.2); that is,

$$(2.4) \quad T := \{d | \nabla g_J(\bar{x})^T d = 0, \nabla g_K(\bar{x})^T d \leq 0, \nabla h(\bar{x})^T d = 0\}.$$

Then the polar cone of T , denoted by N and called the normal cone at \bar{x} , is given by

$$(2.5) \quad N := \{w | w^T d \leq 0, \forall d \in T\} = \{w | w = \nabla g(\bar{x})y + \nabla h(\bar{x})z, y_I = 0, y_K \geq 0\}.$$

Therefore, the primal second order sufficient condition merely says that the Hessian $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})$ is positive definite on the tangent cone T , while the dual second order sufficient condition says that the inverse Hessian $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})^{-1}$ is positive definite on the normal cone N . It will be shown in § 3 that for both conditions to hold it is not only sufficient but also necessary that the Hessian $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})$ be positive definite on the whole space R^n . We also note here that it was proved in [4] that the tangent cone can also be expressed as

$$T = \{d | \nabla f(\bar{x})^T d = 0, \nabla g_A(\bar{x})^T d \leq 0, \nabla h(\bar{x})^T d = 0\},$$

where $A := \{i | g_i(\bar{x}) = 0\} = J \cup K$; that is, A is the index set of *all* active inequality constraints at \bar{x} . Consequently, the normal cone N can also be written as

$$N = \{w | w = \mu \nabla f(\bar{x}) + \nabla g(\bar{x})y + \nabla h(\bar{x})z, y_A \geq 0, y_I = 0\}.$$

These expressions contain the gradient $\nabla f(\bar{x})$ of the objective function and treat the index sets J and K on an equal footing.

We now justify Definition 2.1 by showing that the dual second order sufficient condition given in this definition is equivalent to the one derived by applying McCormick's second order sufficient optimality condition directly to the dual program (D).

THEOREM 2.2 (equivalence of dual second order sufficient optimality condition to McCormick's condition). *If $(\bar{x}, \bar{u}, \bar{v})$ is a Karush–Kuhn–Tucker triple of the primal program (P) and if f, g and h are twice continuously differentiable at \bar{x} then $(\bar{x}, \bar{u}, \bar{v})$ is a Karush–Kuhn–Tucker point of the dual program (D) with the $(n+m)$ -vector $(0, -g(\bar{x}))$ as its Lagrange multiplier. Furthermore, the vector $(\bar{x}, \bar{u}, \bar{v}, 0, -g(\bar{x}))$ satisfies McCormick's second order sufficient optimality condition for (D) if and only if $\nabla_{xx}L(\bar{x}, \bar{u}, \bar{v})$ is nonsingular and condition (2.3) holds.*

Proof. The first statement of the theorem follows immediately by direct verification. Notice that the Lagrangian for the dual program (D) is given by

$$M(x, u, v, s, t) := L(x, u, v) + s^T \nabla_x L(x, u, v) + t^T u.$$

Let ∇M and $\nabla^2 M$ denote respectively the gradient and the Hessian of M with respect to (x, u, v) only. Thus we have

$$\nabla M(x, u, v, s, t) = \begin{bmatrix} \nabla_x L(x, u, v) + \nabla_{xx} L(x, u, v)s \\ g(x) + \nabla g(x)^T s + t \\ h(x) + \nabla h(x)^T s \end{bmatrix}.$$

Let $\bar{s} = 0$ and $\bar{t} = -g(\bar{x})$. Then it follows that

$$(2.6) \quad \nabla^2 M(\bar{x}, \bar{u}, \bar{v}, \bar{s}, \bar{t}) = \begin{bmatrix} \nabla_{xx} L(\bar{x}, \bar{u}, \bar{v}) & \nabla g(\bar{x}) & \nabla h(\bar{x}) \\ \nabla g(\bar{x})^T & 0 & 0 \\ \nabla h(\bar{x})^T & 0 & 0 \end{bmatrix}.$$

Therefore, McCormick's second order sufficient optimality condition [8], [2] for problem (D) is that

$$(2.7) \quad \left. \begin{array}{l} \nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})x + \nabla g(\bar{x})y + \nabla h(\bar{x})z = 0, \\ y_I = 0, \\ y_K \geq 0, \\ (x, y, z) \neq 0 \end{array} \right\} \Rightarrow (x^T, y^T, z^T) \nabla^2 M(\bar{x}, \bar{u}, \bar{v}, \bar{s}, \bar{t}) \begin{bmatrix} x \\ y \\ z \end{bmatrix} < 0,$$

where as before the Hessian is with respect to (x, y, v) only,

$$I := \{i | g_i(\bar{x}) < 0\} \quad \text{and} \quad K := \{i | g_i(\bar{x}) = 0, \bar{u}_i = 0\}.$$

By (2.6) and the equality on the left-hand-side of (2.7), the inequality on the right-hand side of (2.7) is equivalent to $x^T \nabla^2 L(\bar{x}, \bar{u}, \bar{v})x > 0$. Hence condition (2.7) can be expressed as follows:

$$(2.8) \quad \left. \begin{array}{l} \nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})x + \nabla g(\bar{x})y + \nabla h(\bar{x})z = 0, \\ y_I = 0, \\ y_K \geq 0, \\ (x, y, z) \neq 0 \end{array} \right\} \Rightarrow x^T \nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})x > 0.$$

We claim now that condition (2.8) implies the nonsingularity of the Hessian $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})$. Suppose it is not true. Then there exists a nonzero $\hat{x} \in \mathbb{R}^n$ such that $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})\hat{x} = 0$. Let $(\hat{y}, \hat{z}) := (0, 0)$. Then $(\hat{x}, \hat{y}, \hat{z})$ satisfies the conditions in the

left-hand side of implication (2.8), and hence, by implication (2.8), we have $\hat{x}^T \nabla_{xx} L(\bar{x}, \bar{u}, \bar{v}) \hat{x} > 0$. This, however, contradicts $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v}) \hat{x} = 0$.

It now follows from the nonsingularity of $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})$ that the condition $(x, y, z) \neq 0$ in (2.8) can be replaced by $(y, z) \neq 0$, because $(y, z) = 0$ implies $x = 0$. Therefore, by defining $w := -\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})x$, condition (2.8) again can be rewritten as

$$(2.9) \quad \left. \begin{aligned} w &= \nabla g(\bar{x})y + \nabla h(\bar{x})z, \\ y_I &= 0, \\ y_K &\geq 0, \\ (y, z) &\neq 0 \end{aligned} \right\} \Rightarrow w^T \nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})^{-1} w > 0,$$

which is condition (2.3) of our Definition 2.1.

Conversely, it is obvious that condition (2.9) and the nonsingularity of $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})$ imply (2.8) which is equivalent to McCormick's condition (2.7). \square

It should be remarked here that because we used $\bar{s} = 0$ as a multiplier for the dual problem (D) we were able to get away without assuming that f , g and h are thrice differentiable at \bar{x} but merely twice differentiable.

It is important to note that unlike the situation for the primal problem, where the second order sufficiency implication (2.2) holds automatically when $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})$ is positive definite, the second order sufficiency implication (2.3) for the dual problem need not hold when $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})$ is positive definite because $w = 0$ may satisfy the conditions of the left-hand side of implication (2.3). However, under a slightly more stringent version of the standard constraint qualification of nonlinear programming [11], [9] we can show that positive definiteness of $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})$ does indeed ensure the satisfaction of the dual second order sufficiency implication (2.3) as follows.

THEOREM 2.3 (dual second order sufficiency under positive definiteness of the Hessian of the Lagrangian). *Let $(\bar{x}, \bar{u}, \bar{v})$ be a Karush–Kuhn–Tucker point of the primal problem (P), let f , g and h be twice continuously differentiable at \bar{x} , let $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})$ be positive definite and let the following primal constraint qualification hold at \bar{x} ;*

$$(2.10) \quad \nabla h_i(\bar{x}), i = 1, \dots, q, \nabla g_{i \in J}(\bar{x}), \text{ are linearly independent and there exists a } p \in \mathbb{R}^n \text{ such that } \nabla g_K(\bar{x})^T p < 0, \nabla g_J(\bar{x})^T p = 0, \nabla h(\bar{x})^T p = 0.$$

Then the second order sufficiency implication (2.3) holds.

Proof. Let (w, y, z) satisfy the conditions of the left-hand side of implication (2.3). If $w \neq 0$ then implication (2.3) holds because $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})^{-1}$ is positive definite. We now show that if $w = 0$ we contradict the constraint qualification (2.10). If $y_K = 0$ then $(y_J, z) \neq 0$ and we contradict the linear independence of $\nabla h_i(\bar{x}), i = 1, \dots, q, \nabla g_{i \in J}(\bar{x})$. If $y_K \neq 0$ then we have the contradiction

$$0 = y_K^T \nabla g_K(\bar{x})^T p + y_J^T \nabla g_J(\bar{x})^T p + z^T \nabla h(\bar{x})^T p < 0. \quad \square$$

Remark 2.4. It can be shown that the constraint qualification (2.10) implies the standard constraint qualification of nonlinear programming [9 (Def. 11.3.5)], [11] and (2.10) itself is implied by the often used [14], [10] linear independence assumption of all the active constraint gradients: $\nabla h_i(\bar{x}), i = 1, \dots, q, \nabla g_{i \in A}(\bar{x})$.

We now derive a second order necessary optimality condition for the dual problem which, besides having a geometrically meaningful interpretation, will be useful in characterizing simultaneous local solutions of the primal and dual problems. Recall that McCormick's second order *necessary* condition for the primal problem (P) is that

the Hessian $\nabla_{xx}L(\bar{x}, \bar{u}, \bar{v})$ be *positive semidefinite* on the cone $\{d | \nabla g_A(\bar{x})^T d = 0, \nabla h(\bar{x})^T d = 0\}$, where $A := \{i | g_i(\bar{x}) = 0\}$. As expected, the second order necessary condition for the dual problem (D) is that the inverse Hessian $\nabla_{xx}L(\bar{x}, \bar{u}, \bar{v})^{-1}$ be positive semidefinite on the normal cone N defined in (2.5). We give this result below. Note that, under our assumption of nonsingularity of the Hessian $\nabla_{xx}L(\bar{x}, \bar{u}, \bar{v})$, no constraint qualification is required as is the case in McCormick's second order necessary condition.

THEOREM 2.5 (dual second order necessity). *If $(\bar{x}, \bar{u}, \bar{v})$ is a local maximum point of the dual program (D), if f, g and h are twice continuously differentiable at \bar{x} , and if the Hessian $\nabla_{xx}L(\bar{x}, \bar{u}, \bar{v})$ is nonsingular, then*

$$(2.11) \quad \left. \begin{array}{l} w = \nabla g(\bar{x})y + \nabla h(\bar{x})z, \\ y_i = 0, i \in I, \\ y_i \geq 0, i \in K \end{array} \right\} \Rightarrow w^T \nabla_{xx}L(x, u, v)^{-1} w \geq 0,$$

where $I := \{i | g_i(\bar{x}) < 0\}$ and $K := \{i | g_i(\bar{x}) = 0 \text{ and } \bar{u}_i = 0\}$.

Proof. Let vectors w, y and z be fixed vectors satisfying the conditions of the left-hand side of (2.11). We consider the function $F: \mathbb{R}^{n+m+q+1} \rightarrow \mathbb{R}^{n+m+q}$ defined by

$$F(x, u, v, t) := \begin{bmatrix} \nabla_x L(x, u, v) \\ u - \bar{u} - ty \\ v - \bar{v} - tz \end{bmatrix}.$$

Clearly, we have $F(\bar{x}, \bar{u}, \bar{v}, 0) = 0$. Furthermore, it follows from the nonsingularity of $\nabla_{xx}L(\bar{x}, \bar{u}, \bar{v})$ that the Jacobian $\nabla_{x,u,v}F(\bar{x}, \bar{u}, \bar{v}, 0)$ is also nonsingular. Hence, by the implicit function theorem, there exist a positive number ε and continuously differentiable functions $x(t), u(t)$ and $v(t)$ defined on $(-\varepsilon, \varepsilon)$ such that $x(0) = \bar{x}, u(0) = \bar{u}, v(0) = \bar{v}$ and

$$(2.12) \quad \begin{array}{ll} (a) & \nabla_x L(x(t), u(t), v(t)) = 0, \\ (b) & u(t) = \bar{u} + ty, \\ (c) & v(t) = \bar{v} + tz. \end{array}$$

Differentiating (2.12a) at $t = 0$, we get that

$$\nabla_{xx}L(\bar{x}, \bar{u}, \bar{v})x'(0) + \nabla g(\bar{x})y + \nabla h(\bar{x})z = 0,$$

which implies, for $w = \nabla g(\bar{x})y + \nabla h(\bar{x})z$, that

$$w = -\nabla_{xx}L(\bar{x}, \bar{u}, \bar{v})x'(0).$$

Let $\theta(t) := L(x(t), u(t), v(t))$. Notice that for sufficiently small $t \in [0, \varepsilon)$ the vector $(x(t), u(t), v(t))$ is feasible to the dual program (D) and hence $\theta(0) \geq \theta(t)$ for all sufficiently small nonnegative t . On the other hand, we have that

$$\theta'(0) = \nabla_x L(\bar{x}, \bar{u}, \bar{v})^T x'(0) + g(\bar{x})^T y + h(\bar{x})^T z = 0.$$

Therefore it follows that $\theta''(0) \leq 0$. By direct verification, we have that

$$\theta''(0) = -w^T \nabla_{xx}L(\bar{x}, \bar{u}, \bar{v})^{-1} w.$$

Hence, the proof is complete. \square

3. Characterization of simultaneous local solutions of the primal and dual problems. In this section we characterize Karush–Kuhn–Tucker points of the primal

problem that locally solve both the primal and dual programs simultaneously. We will show that these points are essentially those Karush–Kuhn–Tucker points at which the Hessian of the Lagrangian with respect to the primal variables is positive definite. To establish this we need a preliminary fundamental result, Theorem 3.1, which appears to be an interesting linear algebra result in its own right. Related results have appeared in [12].

THEOREM 3.1 (geometric characterization of positive definiteness of a nonsingular matrix). *The nonsingular symmetric matrix A in $R^{n \times n}$ is positive definite if and only if it is positive definite on some subspace of R^n*

$$S = \{x | Bx = 0\}, \quad \text{where } B \in R^{m \times n},$$

and A^{-1} is positive semidefinite on the orthogonal complement S^\perp of S :

$$S^\perp = \{y | y = B^T u\}.$$

Proof. The “only if” part is obvious. The “if” part follows from the facts that $A + \alpha B^T B$ is positive definite for all positive α sufficiently large because A is positive definite on S [3], [1], $BA^{-1}B^T$ is positive semidefinite because A^{-1} is positive semidefinite on S^\perp , and the Sherman–Morrison–Woodbury identity [13] is

$$A^{-1} = (A + \alpha B^T B)^{-1} + \alpha A^{-1} B^T (I + \alpha B A^{-1} B^T)^{-1} B A^{-1}. \quad \square$$

Other proofs of this interesting theorem are also possible. For example it can be established by using an inertia theorem for partitioned matrices [6, p. 75]. In [5] a projection induced by the inner product $x^T A y$ is used to generalize the theorem by replacing S by a closed convex cone in R^n . A referee has also given a proof of Theorem 3.1 by projecting a general point in R^n on AS using the inner product $x^T A^{-1} y$.

We remark here that the matrix A may not even be positive semidefinite when A is positive definite on both the space S and its orthogonal complement S^\perp . This can be seen from the example:

$$A := \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix}, \quad S := \{x \in R^2 | x_2 = 0\}.$$

We are now ready to present our characterization results.

THEOREM 3.2 (positive definiteness of Hessian of Lagrangian under primal second order sufficiency (necessity) and dual second order necessity (sufficiency)). *Let $(\bar{x}, \bar{u}, \bar{v})$ be a Karush–Kuhn–Tucker triple of the primal program (P) that satisfies the strict complementarity condition: $\bar{u}_1 > 0$ whenever $g_i(\bar{x}) = 0$, and let the Hessian $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})$ be nonsingular. If $(\bar{x}, \bar{u}, \bar{v})$ satisfies the primal second order sufficient (necessary) optimality condition and the dual second order necessary (sufficient) optimality condition, then $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})$ is positive definite.*

Proof. Notice that when the strict complementarity condition holds, K is empty and the tangent cone T defined in § 2 is a subspace and the normal cone N is its orthogonal complement. When the primal second order sufficient (necessary) condition is satisfied, the Hessian $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})$ is positive definite (semidefinite) on T . While, when the dual second order necessary (sufficient) condition holds, the inverse Hessian $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})^{-1}$ is positive semidefinite (definite) on N . Therefore, it follows from Theorem 3.1 that the Hessian $\nabla_{xx} L(\bar{x}, \bar{u}, \bar{v})$ must be positive definite. \square

In light of the above theorem, it is natural to expect that the Hessian would be positive semidefinite when both the primal and the dual second order necessary conditions hold. Curiously this turns out not to be true as can be seen from the

following example:

$$\begin{array}{ll} \min & -x_1x_2 \\ \text{s.t.} & x_1 = 0. \end{array}$$

The vector $\bar{x}^T = (0, 1)$ together with $\bar{v} = 1$ constitute a Karush–Kuhn–Tucker point. Both the Hessian of the Lagrangian and its inverse are $\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$. We also have the tangent space $T = \{x \in \mathbb{R}^2 | x_1 = 0\}$ and the normal space $N = \{x \in \mathbb{R}^2 | x_2 = 0\}$. Therefore, both the primal and the dual second order necessary conditions hold. But the matrix $\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$ is not positive semidefinite.

The positive definiteness of the Hessian $\nabla_{xx}L(\bar{x}, \bar{u}, \bar{v})$ clearly implies the primal second order sufficient condition. By Theorem 2.3 under the constraint qualification (2.10), the positive definiteness of $\nabla_{xx}L(\bar{x}, \bar{u}, \bar{v})$ also implies the dual second order sufficient optimality condition. Therefore, we have a converse to Theorem 3.2 which extends and sharpens Luenberger's local duality result [7].

THEOREM 3.3 (primal and dual second order sufficiency under positive definiteness of Hessian of Lagrangian and constraint qualification). *Let \bar{x} be a local minimum point of (P) satisfying the constraint qualification (2.10), let f, g and h be twice continuously differentiable at \bar{x} and let (\bar{u}, \bar{v}) be a Lagrange multiplier associated with \bar{x} . If $\nabla_{xx}L(\bar{x}, \bar{u}, \bar{v})$ is positive definite then $(\bar{x}, \bar{u}, \bar{v})$ satisfies both the primal and the dual second order sufficient optimality conditions.*

REFERENCES

- [1] G. DEBREU, *Definite and semidefinite quadratic forms*, *Econometrica*, 20 (1952), pp. 295–300.
- [2] A. V. Fiacco AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.
- [3] P. FINSLER, *Über das Vorkommen definiter und semidefiniter Formen in Scharen quadratischer Formen*, *Commentarii Mathematici Helvetici*, 9 (1937), pp. 188–192.
- [4] S.-P. HAN AND O. L. MANGASARIAN, *Exact penalty functions in nonlinear programming*, *Math. Programming*, 17 (1979), pp. 251–269.
- [5] ———, *Conjugate cone characterization of positive definite and semidefinite matrices*, Tech. Rep. 471, Computer Sciences Dep., Univ. of Wisconsin–Madison, March 1982, to appear in *Linear Algebra and Appl.*
- [6] E. V. HAYNSWORTH, *Determination of the inertia of a partitioned Hermitian matrix*, *Linear Algebra and Appl.*, 1 (1968), pp. 73–81.
- [7] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.
- [8] G. P. MCCORMICK, *Second order conditions for constrained minima*, *SIAM J. Appl. Math.*, 15 (1967), pp. 641–652.
- [9] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.
- [10] ———, *Unconstrained Lagrangians in nonlinear programming*, this Journal, 13 (1975), pp. 772–791.
- [11] O. L. MANGASARIAN AND S. F. FROMOVITZ, *The Fritz–John necessary optimality conditions in the presence of equality constraints*, *J. Math. Anal. Appl.*, 17 (1967), pp. 34–47.
- [12] D. H. MARTIN AND D. H. JACOBSON, *Copositive matrices and definiteness of quadratic forms subject to homogeneous linear inequality constraints*, *Linear Algebra and Appl.*, 35 (1981), pp. 227–258.
- [13] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
- [14] T. PIETRZYKOWSKI, *An exact potential method for constrained maxima*, *SIAM J. Numer. Anal.*, 6 (1969), pp. 299–304.
- [15] S. M. ROBINSON, *Generalized equations and their solutions; Part II: Applications to nonlinear programming*, Tech. Summary Report 2048, Math. Research Center, Univ. of Wisconsin–Madison, 1980, *Math. Programming Studies*, to appear.
- [16] P. WOLFE, *A duality theorem for nonlinear programming*, *Quart. Appl. Math.*, 19 (1961), pp. 239–244.

REACHABILITY, OBSERVABILITY, AND REALIZABILITY OF CONTINUOUS-TIME POSITIVE SYSTEMS*

YOSHITO OHTA,[†] HAJIME MAEDA[†] AND SHINZO KODAMA[†]

Abstract. This paper discusses reachability, observability, and realizability of single-input, single-output linear time-invariant systems, in which state variables and/or input (output) functions are restricted to be nonnegative to reflect physical constraints frequently encountered in real systems. We define a set reachable from the origin with nonnegative inputs, and also a set observable with nonnegative outputs. We investigate geometrical structures of the sets through convex analysis, and a duality relation between them is established. Next we consider positive realization of a given transfer function. Using the reachable set and the observable set, we give a necessary and sufficient condition for positive realizability. An example is given to demonstrate that a positive realizable transfer function does not in general have a jointly controllable and observable positive realization.

Key words. positive systems, controllability, observability, positive realization, convex analysis

1. Introduction. This paper discusses the reachability, observability, and realizability of a class of linear time-invariant systems, in which state variables and/or the input (output) are restricted to be nonnegative to reflect physical constraints frequently encountered in real systems in engineering, medicine and economics. For example, in tracer kinetics in medicine, state variables may represent concentrations of certain tracer substances and therefore take nonnegative values; the input (output) corresponds to the injected (measured) tracer substance and takes also nonnegative values. In control problems of such systems, controllability and reachability with nonnegative input is a natural as well as a fundamental question. With regard to controllability of linear time-invariant systems with nonnegative inputs, several authors have studied the complete-controllability problem. It is known that a necessary and sufficient condition for complete controllability is that the system is complete-controllable in the usual sense, and moreover the system eigenvalues have nonzero imaginary parts [12] [3]. Complete observability with nonnegative outputs has been considered by Brammer and Jacobson [4] [10] in which a duality relationship between complete controllability and complete observability is established.

In this paper we deal with single-input, single-output linear time-invariant continuous-time systems. We first examine the reachable set from the origin with nonnegative inputs. Specifically we investigate the geometrical structure through convex analysis, and derive conditions which ensure that the reachable set is pointed or n -dimensional. We define the observable set with nonnegative output as the set of initial states which cause nonnegative zero-input responses. From this definition, it is shown that there is a duality relation between the reachable set and the observable set.

In the above problems, state variables are not restricted to be nonnegative. In connection with the nonnegativity restraints on state variables, a fundamental problem is the positive realization problem. The problem is to find, from a given transfer function, a state equation in which state variables and the output take nonnegative values whenever initial states and inputs are nonnegative. If there is no nonnegativity restriction, the problem is trivial; it is well known that every proper rational function has a realization, and the minimal dimension of realizations coincides with the McMillan degree of the transfer function. There is no guarantee that these properties hold

* Received by the editors June 2, 1982, and in revised form January 20, 1983. This work was supported in part by the Grant in Aid for Scientific Research of the Ministry of Education, Science, and Culture of Japan under Grant (C) 00555162 (1981).

[†] Department of Electronic Engineering, Osaka University, Suita, Osaka 565, Japan.

for the positive realization problem. Related problems have been investigated in [6] for continuous systems, and in [7], [8], [9] for discrete time systems. In this paper, we treat the positive realization problem with the aid of the reachable and the observable set, and give a necessary and sufficient condition for positive realizability in terms of these sets.

The format of the paper is as follows. In § 2, we define the reachable set and the observable set, and establish a duality between them. With the aid of this duality, the geometric structure of the sets are examined in detail. In § 3, we deal with the positive realization problem, and give a necessary and sufficient condition for positive realizability. An example is given to show that the minimal dimension is not necessarily equal to the McMillan degree.

Throughout the paper, R_+ denotes the set of nonnegative real numbers, and R_+^n denotes the nonnegative orthant, the set of all nonnegative vectors in n -dimensional Euclidean space R^n . Cone \mathbf{X} denotes the smallest convex cone containing \mathbf{X} , i.e., the set consisting of all finite nonnegative linear combinations of elements of \mathbf{X} . The dual of \mathbf{X} , \mathbf{X}^* , is defined by $\mathbf{X}^* = \{y \mid x^T y \geq 0, \forall x \in \mathbf{X}\}$. It is known that \mathbf{X}^* is a closed set for any \mathbf{X} , and $\mathbf{X}^{**} = \mathbf{X}$ holds if and only if \mathbf{X} is a closed convex cone [2]. The closure of \mathbf{X} is denoted by $\text{cl } \mathbf{X}$. The dimension of a convex cone \mathbf{X} is that of the smallest linear subspace containing \mathbf{X} , i.e., that of $\mathbf{X} + (-\mathbf{X})$. An n -dimensional convex cone in R^n is often called to be *solid*. A solid cone, by its definition, has a topological interior point. A convex cone is said to be *pointed* if and only if $x \in \mathbf{X}$ and $-x \in \mathbf{X}$ together imply $x = 0$. A pointed and solid cone is called *proper*. It is known that the notions of solidness and pointedness are dual in the sense that a closed convex cone $\mathbf{X} \subset R^n$ is solid if and only if \mathbf{X}^* is pointed [2].

2. Reachable set and observable set. Consider a single-input, single-output linear time-invariant system

$$(2.1) \quad \dot{x} = Ax + bu, \quad y = cx,$$

where $A \in R^{n \times n}$, $b \in R^{n \times 1}$, $c \in R^{1 \times n}$.

Let $\mathbf{R}_\infty(A, b)$ be the set of all points to which the states are steered within finite time from the origin by nonnegative inputs, i.e.,

$$(2.2) \quad \mathbf{R}_\infty = \mathbf{R}_\infty(A, b) = \left\{ x \mid x = \int_0^t e^{A(t-\tau)} bu(\tau) d\tau, t \geq 0, u: R_+ \rightarrow R_+, \text{integrable} \right\}.$$

It is well known that \mathbf{R}_∞ is a convex cone (not necessarily closed). We refer to the *reachable set* \mathbf{R} as

$$(2.3) \quad \mathbf{R} = \mathbf{R}(A, b) = \text{cl } \mathbf{R}_\infty(A, b).$$

Let \mathbf{S} be the set of initial states which cause the output to be nonnegative for all $t \geq 0$ if $u(t) \equiv 0$, i.e.,

$$(2.4) \quad \mathbf{S} = \mathbf{S}(c, A) = \{x \mid c e^{At} x \geq 0, \forall t \geq 0\}.$$

Note that \mathbf{S} is a closed convex cone, and is called the *observable set*, since as will be shown there is a duality between \mathbf{R} and \mathbf{S} .

To derive the duality and to obtain specific geometric features of the sets \mathbf{R} and \mathbf{S} , we shall first characterize \mathbf{R} in the following form.

LEMMA 1.

$$(2.5) \quad \mathbf{R} = \text{cl } \mathbf{U},$$

where

$$(2.6) \quad \mathbf{U} = \text{Cone} \{x \mid x = e^{A^t}b, t \geq 0\}.$$

Proof. A proof is easily established by applying the standard continuity argument and by considering the piecewise constant inputs which converge to Dirac delta functions, and details are omitted here. \square

From this, we can establish the following *duality* theorem.

THEOREM 1.

$$(2.7) \quad \mathbf{R}(A, b)^* = \mathbf{S}(b^T, A^T),$$

$$(2.8) \quad \mathbf{S}(c, A)^* = \mathbf{R}(A^T, c^T).$$

Proof. Since $\mathbf{R} = \mathbf{R}^{**}$, (2.8) follows from (2.7). Therefore we only verify (2.7).

$$\begin{aligned} \mathbf{R}(A, b)^* &= [\text{cl Cone} \{x \mid x = e^{A^t}b, t \geq 0\}]^* \\ &= \{x \mid x = e^{A^t}b, t \geq 0\}^* \\ &= \{y \mid y^T e^{A^t}b \geq 0, \forall t \geq 0\} \\ &= \{y \mid b^T e^{A^T t} y \geq 0, \forall t \geq 0\} \\ &= \mathbf{S}(b^T, A^T). \end{aligned} \quad \square$$

In what follows, we shall investigate geometric structures of the reachable set \mathbf{R} and the observable set \mathbf{S} in detail. The duality of Theorem 1 is fully used for establishing the following theorems.

THEOREM 2. (i) $\mathbf{R}(A, b)$ is solid if and only if the pair (A, b) is controllable. (ii) $\mathbf{S}(c, A)$ is pointed if and only if the pair (c, A) is observable.

Proof. By the duality of pointedness and solidness, and by Theorem 1, it suffices to verify (ii). Suppose \mathbf{S} is not pointed. Then there exists a nonzero x such that $x \in \mathbf{S}$ and $-x \in \mathbf{S}$, which implies $c e^{A^t}x \geq 0$ and $c e^{A^t}(-x) \geq 0, \forall t \geq 0$, and hence $c e^{A^t}x = 0, \forall t \geq 0$. This shows that the pair (c, A) is not observable. Conversely, suppose the pair (c, A) is not observable. Then there exists a nonzero x such that $c e^{A^t}x = 0, \forall t \geq 0$, which implies $x \in \mathbf{S}$ and $-x \in \mathbf{S}$, and hence \mathbf{S} is not pointed. \square

By this result, we can get a geometric meaning of one of the conditions, i.e., the rank condition, for complete controllability with nonnegative inputs derived by [12] [3]. To get a geometric meaning of the other condition for complete controllability with nonnegative inputs, i.e., the oscillatory condition, we further investigate the geometric structure of $\mathbf{R}(\mathbf{S})$. We say that the maximal real eigenvalue γ of A (set $\gamma = -\infty$ if there is no real eigenvalue) is *dominant* if A has no eigenvalues in $\text{Re } s > \gamma$ and if $\deg \lambda_i \leq \deg \gamma$ whenever $\text{Re } \lambda_i = \gamma$, where $\deg \lambda_i$ is the multiplicity of eigenvalue λ_i in the minimal polynomial of A .

THEOREM 3. (i) Let the pair (A, b) be controllable. Then \mathbf{R} is pointed if and only if A has at least one real eigenvalue and the maximal real eigenvalue is dominant. (ii) Let the pair (c, A) be observable. Then \mathbf{S} is solid if and only if A has at least one real eigenvalue and the maximal real eigenvalue is dominant.

Proof. Necessity. Note that, by definition, \mathbf{R} and \mathbf{S} are invariant under e^{A^t} , i.e.,

$$(2.9) \quad e^{A^t}\mathbf{R} \subset \mathbf{R}, e^{A^t}\mathbf{S} \subset \mathbf{S} \quad \text{for all } t \geq 0,$$

and by hypothesis, \mathbf{R} and \mathbf{S} are proper. From [2, p. 6], it follows that e^{A^t} ($t \geq 0$) has the maximal real eigenvalue which is equal to $\rho(e^{A^t})$, the spectral radius of e^{A^t} , and

the degree of an eigenvalue with modulus $\rho(e^{A_t})$ is equal to or less than $\deg \rho(e^{A_t})$. Since the multiplicity of $e^{\lambda_i t}$ in the minimal polynomial of e^{A_t} ($t > 0$) is equal to that of λ_i in the minimal polynomial of A , the necessity part follows.

Sufficiency. To complete the sufficiency part of Theorem 3, we only show the sufficiency of (i) because of the duality theorem (Theorem 1). Suppose \mathbf{R} is not pointed, i.e., suppose there is a nonzero x such that $x \in \mathbf{R}$ and $-x \in \mathbf{R}$. Then by Lemma 1, there are sequences $\{x_k^+\}$ and $\{x_k^-\}$ converging to x and $-x$ respectively, such that

$$(2.10) \quad x_k^\pm = \sum_{r=1}^n \alpha_{kr}^\pm \exp(At_{kr}^\pm)b, \quad \alpha_{kr}^\pm \geq 0, \quad t_{kr}^\pm \geq 0.$$

We use the following modal decomposition of e^{A_t} [13, p. 307]:

$$(2.11) \quad e^{A_t} = \sum_{k=1}^{\sigma} \sum_{i=0}^{m_k-1} Y_{ki} \frac{t^i}{i!} e^{\lambda_k t} = \sum_{i=1}^{n'} f_i(t) M_i,$$

where $m_i = \deg \lambda_i$, $n' = \sum_{i=1}^{\sigma} m_i$ is the degree of the minimal polynomial of A , M_i are real constant matrices, and $f_i(t)$ are functions of the following forms: $t^j \exp(\lambda_i t)$, $t^j \exp(\operatorname{Re} \lambda_i t) \cos(\operatorname{Im} \lambda_i t)$, $t^j \exp(\operatorname{Re} \lambda_i t) \sin(\operatorname{Im} \lambda_i t)$; in particular $f_i(t) \geq 0$ ($t \geq 0$) if λ_i is real. Noting that $n = n'$ by controllability, and substituting (2.11) into (2.10), we have

$$(2.12) \quad x_k^\pm = \sum_{i=1}^n \left(\sum_{r=1}^n \alpha_{kr}^\pm f_i(t_{kr}^\pm) \right) M_i b = \sum_{i=1}^n \xi_{ki}^\pm M_i b.$$

Since $M_i b$ ($i = 1, \dots, n$) are linearly independent by controllability, x has a representation

$$(2.13) \quad x = \sum_{i=1}^n \xi_i M_i b,$$

and it can be shown $\xi_{ki}^\pm \rightarrow \pm \xi_i$ as $k \rightarrow \infty$. If λ_i is real then $\xi_{ki}^\pm \geq 0$ for all k , and hence $\pm \xi_i \geq 0$, i.e.,

$$(2.14) \quad \xi_i = 0 \quad \text{if } \lambda_i \text{ is real.}$$

If λ_i is a complex (not a real) eigenvalue, then $|f_i(t)| \leq t^j \exp(\operatorname{Re} \lambda_i t)$ holds for all $t \geq 0$, and either $\operatorname{Re} \lambda_i < \gamma$ or $\operatorname{Re} \lambda_i = \gamma$ with $\deg \lambda_i \leq \deg \gamma$ holds. From these facts, we can show that, for each complex λ_i , $|\xi_i|$ is dominated by at least one of the ξ_j 's corresponding to the real eigenvalue λ_j , and hence

$$(2.15) \quad \xi_i = 0 \quad \text{if } \lambda_i \text{ is complex.}$$

From (2.14) and (2.15), it follows that $x = 0$, which is a contradiction. \square

From this theorem, we see that the proper cones \mathbf{R} and \mathbf{S} are invariant under e^{A_t} if e^{A_t} has a dominant (in modulus) eigenvalue. A related result was obtained in [2, p. 8], where an invariant proper cone is constructed for A (not necessarily of the form e^{A_t}) by a different procedure.

By this theorem, we see that \mathbf{R} is not pointed if either A has no real eigenvalue or if the maximal real eigenvalue is not dominant. In such cases, the state of (2.1) can be freely controlled along a linear subspace by nonnegative inputs. The following theorem shows that the dimension of such a linear subspace in \mathbf{R} is determined solely by μ , the index of A , defined by

$$(2.16) \quad \mu = \sum_{i \in I_1} n_i + \sum_{i \in I_2} (n_i - m_i),$$

(summation over the empty set is considered to be zero), where n_i is the multiplicity of λ_i in the minimal polynomial of A , and in particular m is that of γ , and

$$I_1 = \{i | \operatorname{Re} \lambda_i > \gamma\}, \quad I_2 = \{i | \operatorname{Re} \lambda_i = \gamma \text{ and } n_i > m\}.$$

THEOREM 4. (i) *If the pair (A, b) is controllable, then the dimension of the maximal linear subspace contained in \mathbf{R} is μ .* (ii) *If the pair (c, A) is observable, then the dimension of \mathbf{S} is $n - \mu$ (μ is the index of A).*

Proof. For the proof of Theorem 4, we need the following.

LEMMA 2. *Let \mathbf{X} be a closed convex cone in R^n . Then the following two conditions are equivalent.* (i) *The dimension of the maximal linear subspace contained in \mathbf{X} is μ .* (ii) *The dimension of \mathbf{X}^* is $n - \mu$.*

Proof of Lemma 2. Note that $\mathbf{X} \cap (-\mathbf{X})$ is the maximal linear subspace contained in \mathbf{X} . Since $(\mathbf{X} \cap (-\mathbf{X}))^* = \mathbf{X}^* + (-\mathbf{X})^* = \mathbf{X}^* + (-\mathbf{X}^*)$ [11, p. 146], and since the dual operation is equivalent to the orthogonal operation when applied to a subspace, the lemma follows. \square

We now return to the proof of Theorem 4. By Lemma 2 and by the duality theorem (Theorem 1), it suffices to show that (a) the dimension of \mathbf{S} is not greater than $n - \mu$ and (b) \mathbf{R} does not contain a linear subspace of dimension greater than μ .

(a) In the modal decomposition of $c e^{At}$, (2.11), n -vectors cM_i 's are linearly independent because the pair (c, A) is observable. There are exactly μ cM_i 's such that for the corresponding $f_i(t)$ either $\operatorname{Re} \lambda_i > \gamma$ holds or $\operatorname{Re} \lambda_i = \gamma$ and the power of t is larger than $m - 1$. Hence if the dimension of \mathbf{S} is greater than $n - \mu$, we can choose $x \in \mathbf{S}$ such that $cM_{ix} \neq 0$ where cM_i is one of the cM_i 's mentioned above. With this x , $c e^{At}x = f(t)\{g(t) + h(t)\}$, where $f(t) > 0$ for $t > 0$, $g(t)$ is an almost periodic function with zero mean and is not identically zero, and $h(t) \rightarrow 0$, as $t \rightarrow \infty$. As was shown in [12], there is an instant t such that $c e^{At}x < 0$, a contradiction.

(b) If $x \in \mathbf{R}$ and $-x \in \mathbf{R}$, then by a similar argument in the proof of Theorem 3, we have $\xi_i = 0$ for the expression $x = \sum_{i=1}^n \xi_i M_i b$ if λ_i is real, or if $\operatorname{Re} \lambda_i < \gamma$, or if $\operatorname{Re} \lambda_i = \gamma$ and the power of t in $f_i(t)$ is less than m . Thus x must be in the subspace spanned by $M_i b$'s, $i \in I_1 \cup I_2$. Since the maximal subspace in \mathbf{R} must be contained in this μ -dimensional subspace, (b) follows. \square

From Theorem 2 and Theorem 4, we see that the system is completely reachable with nonnegative inputs, i.e. $\mathbf{R} = R^n$, if and only if the pair (A, b) is controllable and A is oscillatory (has no real eigenvalues), which was first obtained by [12]. Also we may interpret the condition for complete observability with nonnegative outputs proposed in [4] from the viewpoint of geometric features of \mathbf{S} . For a single-output system to be completely observable in the sense of [4], it is necessary and sufficient that $-\mathbf{S}$ is $\{0\}$, since $-\mathbf{S}$ is the unobservable set in the sense of [4]. In fact, from Theorem 2 and Theorem 4, $-\mathbf{S} = \{0\}$ holds if and only if the pair (c, A) is observable and $n - \mu = 0$, i.e., A has no real eigenvalues.

3. Positive realizability condition. In this section, we deal with the positive realization problem through convex cone analysis; specifically we shall examine the realizability condition via the reachable set and the observable set. A strictly proper rational function $\hat{H}(s)$ is said to be *positive realizable* if there exist a matrix A with nonnegative off-diagonal elements and nonnegative vectors b, c such that

$$(3.1) \quad \hat{H}(s) = c(sI - A)^{-1}b.$$

Such a realization $\{A, b, c\}$ is called the *positive realization*, since it yields nonnegative state (and output) responses whenever initial states and inputs are nonnegative. More

specifically, one can show that $\exp(At) \geq 0$ for all $t \geq 0$ if and only if every off-diagonal element is nonnegative [1, p. 172]. The realization problem of compartmental systems is obviously in this category [6].

In usual realization problems without sign restriction, it is well known that any proper rational function $\hat{H}(s)$ has a realization $\{F, g, h\}$, i.e.,

$$(3.2) \quad \dot{z} = Fz + gu, \quad y = hz,$$

such that the pair (F, g) is controllable, the pair (h, F) is observable and the dimension of z is equal to the McMillan degree of $\hat{H}(s)$.

In what follows, we shall investigate the positive realization problem by using triplet $\{F, g, h\}$ rather than $\hat{H}(s)$.

Note that the impulse response function $H(t)$ of a positive realization is nonnegative for all $t \geq 0$, hence the condition

$$(3.3) \quad H(t) \geq 0, \quad t \geq 0,$$

is necessary for $\hat{H}(s)$ to be positive realizable. One can show, from Lemma 1 and a continuity argument, that (3.3) is equivalent to

$$(3.4) \quad \mathbf{R} \subset \mathbf{S},$$

where \mathbf{R} and \mathbf{S} are the reachable set and the observable set of (3.2) respectively. This is a necessary condition but not in general sufficient one for positive realizability. A necessary and sufficient condition will be given in a modified form of (3.4). To this end, we first give further properties of \mathbf{R} and \mathbf{S} .

LEMMA 3. $\mathbf{R}(A, b) = \mathbf{R}(A + \lambda I, b)$, $\mathbf{S}(c, A) = \mathbf{S}(c, A + \lambda I)$ for all real λ .

Proof. This may be easily shown by definition of \mathbf{R} and Theorem 1. \square

THEOREM 5. Let $\hat{H}(s)$ be a strictly proper rational function of degree n , and let $\{F, g, h\}$ be a minimal realization in the usual sense, i.e., $\hat{H}(s) = h(sI - F)^{-1}g$, where $\{F, g, h\}$ is jointly completely controllable and completely observable. Then $\hat{H}(s)$ is

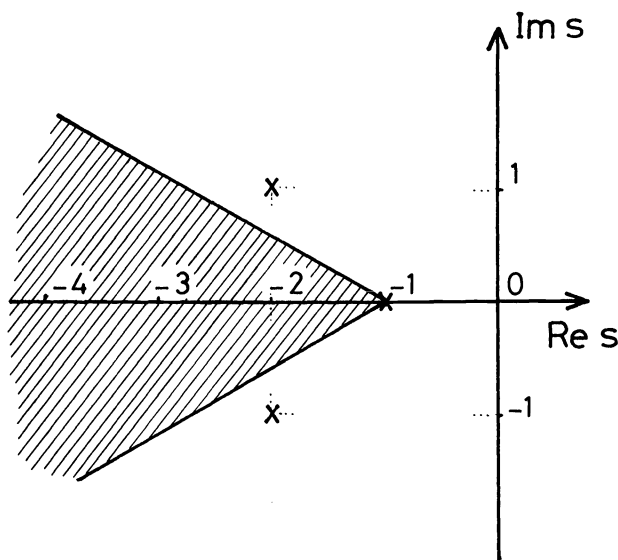


FIG. 1. Restricted area of eigenvalues of 3×3 -matrices with nonnegative off-diagonal elements if the maximal real eigenvalue is -1 .

positive realizable if and only if there exists a polyhedral convex cone \mathbf{P} such that

- (i) $(F + \lambda I)\mathbf{P} \subset \mathbf{P}$ for some $\lambda \geq 0$,
- (ii) $\mathbf{R} \subset \mathbf{P} \subset \mathbf{S}$.

Proof. Sufficiency. A polyhedral convex cone \mathbf{P} can be expressed as $\mathbf{P} = PR_+^N$ or $\mathbf{P} = \text{Cone}\{P\}$ where P is an $n \times N$ matrix and $\{P\}$ denotes the set consisting of the column vectors of P . Since \mathbf{P} is $(F + \lambda I)$ -invariant, there exists a nonnegative matrix \bar{A} satisfying $(F + \lambda I)P = P\bar{A}$. Furthermore, from $\mathbf{R} \subset \mathbf{P}$ it follows that $g \in \mathbf{P}$, and hence there exists a nonnegative vector $b \in R_+^N$ such that $Pb = g$. Finally, from $\mathbf{P} \subset \mathbf{S}$ it follows that $h^T \in \mathbf{R}(F^T, h^T) = \mathbf{S}(h, F)^* \subset \mathbf{P}^*$, and hence $c = hP \in R^{1 \times N}$ is a nonnegative vector. Let $A = \bar{A} - \lambda I$; then A is an off-diagonally nonnegative matrix and $FP = PA$ holds. Thus

$$he^{Ft}g = h\left(\sum_{i=0}^{\infty} \frac{F^i t^i}{i!}\right)Pb = h\left(\sum_{i=0}^{\infty} PA^i \frac{t^i}{i!}\right)b = hP\left(\sum_{i=0}^{\infty} A^i \frac{t^i}{i!}\right)b = ce^{At}b,$$

and hence $\{A, b, c\}$ is a positive realization of $\hat{H}(s)$.

Necessity. Suppose $\{A, b, c\}$ is a positive realization of $\hat{H}(s)$ of dimension N . Let $\lambda = \max\{-a_{ii} (i = 1, \dots, N), 0\}$ where a_{ii} 's are the diagonal elements of A , and let $\bar{A} = A + \lambda I$ and $\bar{F} = F + \lambda I$. Note that \bar{A} is a nonnegative matrix and both $\{\bar{A}, b, c\}$ and $\{\bar{F}, g, h\}$ are the realizations of $\hat{H}(s - \lambda)$. Furthermore $\{\bar{F}, g, h\}$ is a minimal realization. By Lemma 3, $\mathbf{R}(F, g) = \mathbf{R}(\bar{F}, g)$ and $\mathbf{S}(h, F) = \mathbf{S}(h, \bar{F})$. Thus to prove the necessity, we have only to show that there is a polyhedral convex cone $\mathbf{P} \subset R^n$ satisfying $\bar{F}\mathbf{P} \subset \mathbf{P}$ and $\mathbf{R}(\bar{F}, g) \subset \mathbf{P} \subset \mathbf{S}(h, \bar{F})$. By the well-known canonical decomposition theorem, $\{\bar{A}, b, c\}$ is transformed into controllable part and observable part:

$$(3.5) \quad \begin{aligned} T\bar{A}T^{-1} &= \begin{pmatrix} A_{aa} & 0 \\ A_{ba} & A_{bb} \end{pmatrix}, \quad Tb = \begin{pmatrix} b_a \\ b_b \end{pmatrix}, \quad cT^{-1} = (c_a \quad 0), \\ A_{aa} &= \begin{pmatrix} \bar{F}^T & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad b_a = \begin{pmatrix} h^T \\ 0 \end{pmatrix}, \quad c_a = (g^T \quad c_2), \\ T &= \begin{pmatrix} T_a \\ T_b \end{pmatrix}. \end{aligned}$$

Let $\mathbf{K} = (\text{Cone}\{T_a\})^*$, or $\mathbf{K}^* = \text{Cone}\{T_a\}$. Since \bar{A} is a nonnegative matrix, $A_{aa}\mathbf{K}^* = \text{Cone}\{A_{aa}T_a\} = \text{Cone}\{T_a\bar{A}\} \subset \text{Cone}\{T_a\} = \mathbf{K}^*$, i.e., \mathbf{K}^* is A_{aa} -invariant, and hence \mathbf{K} is A_{aa}^T -invariant. Since $b_a = T_a b$ and b is a nonnegative vector, we have $b_a \in \mathbf{K}^*$. Similarly from $c_a T_a = c$, we have $c_a \in \mathbf{K}^{**} = \mathbf{K}$. Since \mathbf{K} is a polyhedral convex cone, \mathbf{K} can be expressed as $\mathbf{K} = \text{Cone}\{K\}$ for some $K \in R^{n \times N''}$. Let

$$(3.6) \quad K = \begin{pmatrix} P \\ * \end{pmatrix}, \quad P \in R^{n \times N''}, \quad \mathbf{P} = \text{Cone}\{P\}.$$

Then $A_{aa}\mathbf{K} \subset \mathbf{K}$, $c_a^T \in \mathbf{K}$, and $b_a \in \mathbf{K}^*$ imply $\bar{F}\mathbf{P} \subset \mathbf{P}$, $g \in \mathbf{P}$, and $h^T \in \mathbf{P}^*$ respectively. From $\bar{F}\mathbf{P} \subset \mathbf{P}$ and $g \in \mathbf{P}$, we have $\bar{F}^i g \in \mathbf{P} (i = 0, 1, 2, \dots)$. Thus $e^{\bar{F}t}g = \sum_{i=0}^{\infty} (t^i/i!) \bar{F}^i g \in \mathbf{P}$, for all $t \geq 0$. Hence $\mathbf{R}(\bar{F}, g) \subset \mathbf{P}$. From $\bar{F}\mathbf{P} \subset \mathbf{P}$, $\bar{F}^T \mathbf{P}^* \subset \mathbf{P}^*$. This, together with $h^T \in \mathbf{P}^*$, implies $(\bar{F}^T)^i h^T \in \mathbf{P}^* (i = 0, 1, 2, \dots)$. Thus by the same argument, $\mathbf{R}(\bar{F}^T, h^T) \subset \mathbf{P}^*$. Hence $\mathbf{S}(h, \bar{F}) = \mathbf{R}(\bar{F}^T, h^T)^* \supset \mathbf{P}^{**} = \mathbf{P}$. \square

Remark 1. Whether or not $\hat{H}(s)$ is positive realizable does not depend on the initial choice of minimal realization $\{F, g, h\}$ of $\hat{H}(s)$ since the minimal realization is unique within a nonsingular transformation.

Remark 2. The polyhedral cone satisfying (ii) is proper since \mathbf{R} is solid and \mathbf{S} is pointed, and hence the edge vectors of \mathbf{P} are unique within positive factors and the renumbering. The number of edges gives the dimension of realization.

It seems difficult at the present stage to obtain a necessary and sufficient condition for positive realizability in terms of the input-output function, say $\hat{H}(s)$ or the impulse response $H(t)$. However, it is possible to give a partial answer. If \mathbf{R} (or \mathbf{S}) was a polyhedral and $\mathbf{R}(\mathbf{S})$ was $(F + \lambda I)$ -invariant then the positive realizability condition (i) and (ii) would be reduced to a simple condition (3.4). Unfortunately, although \mathbf{R} (\mathbf{S}) is e^{Ft} -invariant, they are not necessarily $(F + \lambda I)$ -invariant. The following lemma gives a condition for \mathbf{R} (\mathbf{S}) to be $(F + \lambda I)$ -invariant.

LEMMA 4. Let \mathbf{P} be a polyhedral cone in R^n and $F \in R^{n \times n}$. Then $e^{Ft}\mathbf{P} \subset \mathbf{P}$ for any $t \geq 0$ if and only if $(F + \lambda I)\mathbf{P} \subset \mathbf{P}$ for some $\lambda \geq 0$.

Proof. We first show the sufficiency. We need only prove $e^{Ft}\mathbf{P} \subset \mathbf{P}$ for $t > 0$. Note that if $(F + \lambda I)\mathbf{P} \subset \mathbf{P}$ for some $\lambda \geq 0$, then $(F + \eta I)\mathbf{P} \subset \mathbf{P}$ for all $\eta \geq \lambda$. Now let $x \in \mathbf{P}$ and define $x_k = (I + Ft/k)^k x$. Then for sufficiently large k satisfying $k/t \geq \lambda$, $x_k = (t/k)^k \{F + (k/t)I\}^k x \in \mathbf{P}$. Hence $\lim_{k \rightarrow \infty} x_k = e^{Ft}x \in \mathbf{P}$, which shows that $e^{Ft}\mathbf{P} \subset \mathbf{P}$ for any $t > 0$. To prove the necessity, denote $\mathbf{P} = \text{Cone}\{P\}$, $P \in R^{n \times N}$, $P = (p_1, \dots, p_N)$, $\mathbf{P}^* = \text{Cone}\{Q\}$, $Q \in R^{n \times N'}$, $Q = (q_1, \dots, q_{N'})$. Note that by the definition of dual cone $q_j^T p_i \geq 0$ ($i = 1, \dots, N, j = 1, \dots, N'$), and note also that, by assumption, $a_{ij}(t) = q_j^T e^{Ft} p_i \geq 0$ ($i = 1, \dots, N, j = 1, \dots, N'$), hold for any $t \geq 0$. By Taylor expansion, we have $a_{ij}(t) = q_j^T p_i + q_j^T F p_i t + O(t^2)$; thus $q_j^T p_i = 0$ implies $q_j^T F p_i \geq 0$. Define λ_{ij} so that $\lambda_{ij} = 0$, if $q_j^T F p_i \geq 0$, and $\lambda_{ij} = -q_j^T F p_i / q_j^T p_i$, if $q_j^T F p_i < 0$ and let $\lambda = \max_{i,j} \lambda_{ij}$. Then $q_j^T (\lambda I + F) p_i \geq 0$, ($i = 1, \dots, N, j = 1, \dots, N'$), hold, hence $(F + \lambda I)\mathbf{P} \subset \mathbf{P}^{**} = \mathbf{P}$. \square

By Theorem 5 and by Lemma 4, we can obtain a positive realizability condition when the degree of $\hat{H}(s)$ is two.

COROLLARY. Let $\hat{H}(s)$ be a strictly proper rational function of degree 2. $\hat{H}(s)$ is positive realizable if and only if the impulse response function $H(t)$ is nonnegative for all $t \geq 0$.

Proof. Let $\{F, g, h\}$ be a jointly controllable and observable realization of $\hat{H}(s)$ and \mathbf{R}, \mathbf{S} be its reachable and observable set, respectively. The condition that $H(t) \geq 0$, $t \geq 0$, is equivalent to the condition $\mathbf{R} \subset \mathbf{S}$. \mathbf{R} is a closed convex cone in R^2 , and hence it is a polyhedral cone, being $(F + \lambda I)$ -invariant for some $\lambda \geq 0$ by Lemma 4. Regarding \mathbf{R} as \mathbf{P} in Theorem 5, it follows that $\hat{H}(s)$ is positive realizable. \square

When the degree of a transfer function is less than or equal to two, we know that a positive realizable transfer function always has a jointly controllable and observable positive realization. But if the degree is larger than two, i.e., $\deg \hat{H}(s) \geq 3$, then the minimal dimension of realizations is not necessarily equal to the degree of $\hat{H}(s)$, as in the case of compartmental systems [6]. To illustrate this and to see how Theorem 5 works out, we give an example.

Example. Consider

$$\hat{H}(s) = \frac{2s^2 + 7s + 8}{(s+1)(s^2 + 4s + 5)},$$

whose poles are $-1, -2 \pm j$. It is remarked that $\hat{H}(s)$ does not have three-dimensional positive realizations, since the eigenvalues of 3×3 -matrices with nonnegative off-diagonal elements are restricted in the area $D = \{s \mid \text{Im } s \geq [\tan(\pi/6)](\text{Re } s + 1), \text{Im } s \leq -[\tan(\pi/6)](\text{Re } s + 1)\}$ [5].

Let $\{F, g, h\}$ be the controllable companion form of $\hat{H}(s)$, i.e.,

$$F = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -5 & -9 & -5 \end{pmatrix}, \quad g = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad h = (8 \quad 7 \quad 2).$$

Let $\lambda = 2$ and $\mathbf{P} = \text{Cone} \{p_1, p_2, p_3, p_4\}$, where $p_1^T = (-1, 1, 2)$, $p_2^T = (-1, 4, 10)$, $p_3^T = (2, -2, -1)$ and $p_4^T = (2, -5, 11)$. Then $(F + \lambda I)p_i \in \mathbf{P}$ ($i = 1, \dots, 4$), hence \mathbf{P} is $(F + \lambda I)$ -invariant. Moreover one can show $e^{Ft}g \in \mathbf{P}$, $t \geq 0$, and $h e^{Ft}p_i \geq 0$, $t \geq 0$, $i = 1, \dots, 4$, from which $\mathbf{R} \subset \mathbf{P} \subset \mathbf{S}$. Thus by Theorem 5, $\hat{H}(s)$ has a positive realization. In fact,

$$A = \begin{pmatrix} -2 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & -2 & 1 \\ 1 & 0 & 0 & -2 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad c = (1 \quad 1 \quad 0 \quad 1)$$

is a positive realization of $\hat{H}(s)$.

Remark 3. It should be emphasized that a positive realizable transfer function does not in general have a jointly controllable and observable positive realization as shown by this example; it means that the dimension greater than the McMillan degree may be required for realization to satisfy the nonnegativity restraints.

Remark 4. At this stage, a systematic way to find a polyhedral cone satisfying the condition of Theorem 5 is not known except in the two-dimensional case.

4. Conclusion. We have studied problems concerning reachability, observability, and realizability for a class of single-input, single-output linear time-invariant systems in which input (output) and/or state variables are restricted to be nonnegative. We first introduced a reachable and an observable set and showed that there is a duality relation between them in the sense that those two sets are dual cones each other. Using this duality, we derived necessary and sufficient conditions for the reachable set to be pointed or solid. We then derived a necessary and sufficient condition for positive realizability in terms of these sets. Although it is preferable to derive realizability conditions directly from an input-output relation, the problem remains an open question.

5. Acknowledgment. The authors are grateful to the reviewer for valuable comments.

REFERENCES

- [1] R. BELLMAN, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.
- [2] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.
- [3] R. F. BRAMMER, *Controllability in linear autonomous systems with positive controllers*, this Journal, 10 (1972), pp. 339–353.
- [4] ———, *Geometrically constrained observability*, this Journal, 12 (1974), pp. 449–459.
- [5] F. I. KARPELEVICH, *On the characteristic roots of matrices with nonnegative elements*, Izv. Akad. Nauk SSSR Ser. Mat., 15 (1951), pp. 361–383. (In Russian.)
- [6] H. MAEDA, S. KODAMA AND F. KAJIYA, *Compartmental system analysis: Realization of a class of linear systems with physical constraints*, IEEE Trans. Circuits Syst., CAS-24 (1977), pp. 8–14.
- [7] H. MAEDA AND S. KODAMA, *Reachability, observability and realizability of linear systems with positive constraints*, Trans. IECE, 63-A (1980), pp. 688–694. (In Japanese.)

- [8] ———, *Positive realization of difference equations*, IEEE Trans. Circuits Syst., CAS-28 (1981), pp. 39–47.
- [9] J. W. NIEUWENHUIS, *About nonnegative realizations*, Systems & Control Letters, 1 (1982), pp. 283–287.
- [10] M. PACHTER AND D. H. JACOBSON, *Observability with a conic observation set*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 632–633.
- [11] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.
- [12] S. H. SAPERSTONE AND J. A. YORKE, *Controllability of linear oscillatory systems using positive controls*, this Journal, 9 (1971), pp. 253–262.
- [13] L. A. ZADEH AND C. A. DESOER, *Linear System Theory: The State Space Approach*, McGraw-Hill, New York, 1963.

LOCAL STABILITY AND OPTIMALITY IN CONTINUOUS GAMES*

D. J. GATES† AND M. WESTCOTT†

Abstract. We study games with continuous payoff functions $J_i(\sigma_1, \dots, \sigma_N)$, $i = 1, \dots, N$, where a strategy for player i is to choose a real number σ_i and is pure. Small adjustments by coalitions and responses by other coalitions are analyzed on the basis that one coalition may discipline another if the latter makes an adjustment which is favorable to itself. This concept provides a natural definition for defensive coalitions and for optimality based on a pair of coalitions which are both defensible. More general optimal states for the whole game, under limited information conditions, contain all the well-known optima and are closely related to the limiting states of an explicit time-dependent adjustment process. The optima are supported by the data of Fouraker and Siegel and the sequences of adjustments in their experiments are clarified.

Key words. game theory, optimality, stability, coalition, adjustment process, Perron–Frobenius theorem, competition, economic theory

1. Introduction. The extended optima for continuous games introduced by us recently (Gates and Westcott (1981a, b, c), henceforth referred to as GW.I, GW.II and GW.III respectively) were based on a new bargaining concept which was interpreted intuitively as *discipline*. Roughly speaking, this concept accounts for the ultimate reduction in payoff that a player, who makes a payoff-increasing adjustment, may suffer due to a subsequent joint adjustment by some coalition which excludes that player, the joint adjustment restoring the payoff of that coalition. Four optima were defined in terms of this concept and represent various information conditions. These optima and an associated dynamical process¹ were designed to include the behavior of players who are less sophisticated and less well informed than those usually considered in game theory. They were aimed at modelling existing competition in markets, rather than providing improved strategies for firms or modelling ideal competition.

The new optima comprise sets of strategies which contain, and are relatively large compared to, those of previous game-theory solutions. Consequently they include specialized bargaining behavior, but are much easier to achieve in practice.

The new optima agree with the experimental results of Fouraker and Siegel (1961) more satisfactorily than existing game theory solutions, and in particular provide a simple and reasonable description of the adjustment sequence in the experiments, as shown in GW.III.

The present paper extends the concept of discipline to include the possibility that a coalition, rather than a player, might be disciplined by another coalition. This leads to the concept of two balanced coalitions, which provides a new solution for a duopoly game between those two coalitions within the n -person game. More general optima are also defined and evaluated. The most significant set of optima defined in GW.I, called O_1 , has a natural extension here, called O_1^* , which is shown to essentially coincide with O_1 for an important class of games. Thus O_1^* inherits the properties of O_1 established in GW.I and GW.II: principally, (i) it is given by a set of simple constraints on the signs of the Jacobian determinant and principal minor determinants of the payoff functions; (ii) it coincides with the set of solutions of some matrix

* Received by the editors October 1, 1982, and in revised form March 8, 1983.

† Commonwealth Scientific and Industrial Research Organization, Division of Mathematics and Statistics, Canberra, A.C.T., Australia.

¹ Gates, Rickard and Wilson (1977). See also Gates, Rickard and Wilson (1978) and Gates, Rickard and Westcott (1981).

equations for the payoff functions which in turn (iii) contain the equilibrium solutions of the dynamical adjustment process of Gates et al. (1977), and (iv) it agrees with the data of Fouraker and Siegel.

The results (ii) and (iii) are mathematically surprising, as pointed out after Theorem 6, because the adjustment process involves no strategic concepts.

The results are stronger and deeper for O_1^* , and require rather different methods of proof. In particular, the Perron–Frobenius theorem for positive matrices now, surprisingly, plays a fundamental role in the determination of the set O_1^* . Normally this deep and powerful theorem is relevant only to dynamic models or asymptotic results, so that our application of the theorem offers new insights into optimality. We emphasize that these results are not minor generalizations of Gates and Westcott (1981a, b, c). By analogy, the results of Lucas (1966) and Owen (1968) on $(N-1)$ -person coalitions do not generalize in a minor way to general coalitions.

To further motivate these developments, we consider experiment 10 of Fouraker and Siegel which involves a quantity variation duopoly with complete information. The payoffs were

$$(1.1) \quad J_i(\sigma_1, \sigma_2) = 0.04\sigma_i(60 - \sigma_1 - \sigma_2), \quad i = 1, 2,$$

which were known, in effect, to both players from tables of the functions. There were 16 pairs of players and 25 transactions between every pair. The quantity choices (σ_1, σ_2) for all 400 transactions are plotted on Fig. 1: here the size of dot indicates

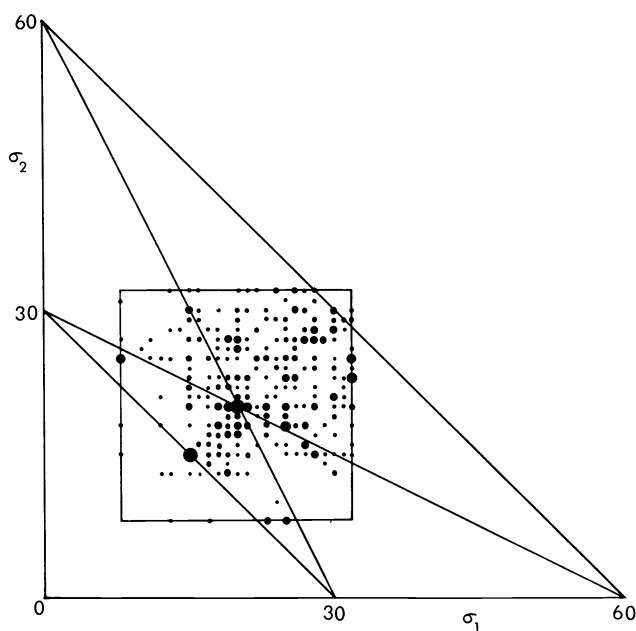


FIG. 1

the number of times the point occurs. We have superimposed all the data because this reveals properties of the “average behavior” of players which are not so clear from data on individual transactions or individual players. Outputs were confined by law to integer values 8, 9, \dots , 32 and hence to the square in Fig. 1.

The well-known game theory solutions—or “presolutions”—the Pareto optima, Edgeworth contract curve, imputations, core, stable set, bargaining set, nucleolus, kernel, cooperative solutions of von Neumann and Morgenstern with or without side

payments—lie on the line

$$(1.2) \quad \sigma_1 + \sigma_2 = 30.$$

The efficient point, and the threat solutions of Nash with and without side payments, lie on the line

$$(1.3) \quad \sigma_1 + \sigma_2 = 60.$$

The Cournot solution is the point

$$(1.4) \quad \sigma_1 = \sigma_2 = 20.$$

The inadequacy of all these solutions in explaining the spread of points in Fig. 1 is evident, and a little analysis (GW.III) supports this observation.

Although this is a complete information game, rather few players seem to have used the information, so that our limited information optima are appropriate. Perhaps the players were confused by the sheer volume of information, as concluded by Fouraker and Siegel. Our extended optima comprise two triangles,

$$(1.5) \quad \begin{aligned} &\{\sigma_1 + \sigma_2 \geq 30, \sigma_1 + 2\sigma_2 \leq 60, 2\sigma_1 + \sigma_2 \leq 60\} \quad \text{and} \\ &\{\sigma_1 + \sigma_2 \leq 60, \sigma_1 + 2\sigma_2 \geq 60, 2\sigma_1 + \sigma_2 \geq 60\}. \end{aligned}$$

The agreement with the data in Fig. 1 is fairly clear and a statistical comparison (GW.III) confirms this. GW.III also shows that several alternative theories are less satisfactory in explaining the data. Similar conclusions can be drawn for the other quantity variation experiments of Fouraker and Siegel which include triopoly games. The latter admit the possibility that a coalition of 2 players be disciplined by the third, so that a more complete analysis of such games involves the developments in this paper.

Our work makes little contact with popular developments in N -person game theory, which largely concentrate on discrete games in characteristic function form inspired by the early, celebrated work of von Neumann and Morgenstern. Thus the analytical richness of pure, continuous games has been largely overlooked. Solution concepts such as the bargaining set (Aumann and Maschler (1964)) degenerate into less interesting sets, such as (1.2), in the continuous case and comprise much smaller sets than our extended optima.

Popular solution concepts are more restrictive than Pareto optima, and arose from various philosophical considerations. To quote Aubin (1979, p. 293), "Since our aim is to devise procedures yielding as small a subset of strategies as possible, the problem of *selecting* Pareto strategies arises." In our case, the data demand an extension or *weakening* of the Pareto concept. In fact our aims are rather different from much of mainstream game theory epitomized by Jones (1980, p. 16): "The question asked by a game theorist is very different. It is: what would each player do, if all the players were doing as well for themselves as they possibly could." By contrast, we are more interested in asking, "What do real competitors do, and why?". A similar point of view has been taken by Case (1979) and others.

There are numerous dynamical models in the literature, most of them artificially contrived to achieve a particular equilibrium solution, such as the models of Stearns (1968), Billera (1972) and Kalai, Maschler and Owen (1975) whose equilibria are bargaining sets. These models have little resemblance to ours. In our case the dynamics (Gates et al. (1977)) came first and were based on a realistic adjustment process (Weinberg (1961)). The extended optima arose from attempts to characterize the resulting equilibrium.

Section 2 gives definitions and basic results for the discipline concept. Section 3 discusses the resulting extended optima for general payoffs. Section 4 relates these optima to the time-dependent adjustment process mentioned above. Section 5 gives a more complete description of the optima for an important class of payoffs; § 6 discusses alternative definitions of discipline, and the remaining sections contain proofs of theorems.

Our sharpest results and consequently those most relevant in applications, such as Fouraker and Siegel (1960), are contained in § 5.

2. Discipline and balanced coalitions. We consider general games in strategic form among N players, where player i 's strategy is to choose a real number σ_i , receiving a payoff $J_i(\sigma)$, where $\sigma = (\sigma_1, \dots, \sigma_N)$. The J_i are assumed continuously differentiable with respect to each σ_i and not constant in any neighborhood of any point σ . The set of all players $\{1, \dots, N\}$ will be denoted by Λ .

For nonnegative σ_i , this game is equivalent to quantity variation competition among N firms where σ_i and J_i are the output and profit respectively of firm i . We normally think of outputs and profits as applying to one business period (a month perhaps) while adjustments—small changes in σ_i 's—are made at the beginning of each business period.

We consider only pure strategies σ since we are aiming at modelling existing competition among firms. Although there may be random elements in the behavior of firms, we know of no examples where firms deliberately randomize their activities for strategic advantages of the game-theory type. Besides this, we have no need, in our case, for the mathematical tractability endowed on discrete games by the introduction of mixed strategies. Mixed strategies are clearly artificial, as pointed out by their inventor von Neumann (1928).

We assume that payoffs are not transferrable among players so that side payments in the economics context are excluded. This is the case in the experiments of Fouraker and Siegel.

For any set $S \subseteq \Lambda$, let \mathbf{J}_S be the vector with components $J_i (i \in S)$, and \mathbf{e}_S be the N -vector with components $\varepsilon_i (i \in S)$ and 0 elsewhere. Throughout this paper, inequalities between vectors and matrices are taken to hold componentwise.

DEFINITION 2.1. The game is said to be in *state* σ when the strategies are σ .

DEFINITION 2.2. A *coalition* is a subset of Λ .

DEFINITION 2.3. In state σ , coalition B is said to be *disciplinable* by coalition C , where $C \cap B = \emptyset$, if

$$(2.1) \quad J_i(\sigma + \mathbf{e}_B) < J_i(\sigma) \quad \text{for at least one } i \in B$$

for every sufficiently small $\mathbf{e}_B \neq \mathbf{0}$; or if, for every sufficiently small $\mathbf{e}_B \neq \mathbf{0}$, such that

$$(2.2) \quad \mathbf{J}_B(\sigma + \mathbf{e}_B) \geq \mathbf{J}_B(\sigma),$$

with strict inequality in at least one component, there exists $\mathbf{e}_C \neq \mathbf{0}$ such that

$$(2.3) \quad \mathbf{J}_B(\sigma + \mathbf{e}_B + \mathbf{e}_C) \leq \mathbf{J}_B(\sigma),$$

with strict inequality in at least one component,

$$(2.4) \quad \mathbf{J}_C(\sigma + \mathbf{e}_B + \mathbf{e}_C) = \mathbf{J}_C(\sigma),$$

and $\mathbf{e}_C \rightarrow \mathbf{0}$ as $\max_{i \in B} \varepsilon_i \rightarrow 0$.

In other words, a small adjustment by players in B either results directly in a reduction in payoff for at least one of them or can be countered by coalition C from the other players whose members can restore their original payoffs and leave no player in B with a net gain in payoff and at least one player in B with a net reduction in payoff. We denote the set of such states σ by $D_{B,C}$. It is a direct extension of the case $B = \{i\}$ in GW.I.

Note that (2.1) is just the statement that σ is Pareto optimal for the coalition B . Thus (2.2) to (2.4) are a natural extension of this familiar type of optimality. Actually (2.1) is mathematically redundant but we include it for clarity.

One might argue that it would be more natural to allow C to “at least restore its payoff”, i.e. to admit the possibility of gains in payoffs for C . Such a definition is in fact less useful, as shown in § 6.

Our notion of players forming a *defensive* coalition seems to be new in game theory. Previously, coalitions which give a benefit to the partners have been widely studied following the basic work of J. von Neumann and O. Morgenstern. The work of Vickrey (1959) involves some defensive or “policing” notions, but there is no technical similarity.

If C is also disciplinable by B there is a deadlock in which neither coalition can safely make payoff-increasing adjustments. Thus adjustments would tend not to be made, resulting in a kind of stable state or optimum, with states comprising a set

$$(2.5) \quad O_{B,C} \equiv O_{C,B} = D_{B,C} \cap D_{C,B}.$$

In such a state both coalitions are protected by their disciplining power, that is, they are both *defendable*. We refer to them as *balanced coalitions*. For unsophisticated players (as in the Fouraker and Siegel experiments) or players with inadequate information, each coalition can be held in $O_{B,C}$ over successive transactions through being accidentally disciplined by the others for unfavorable adjustments, just as players are attracted to $O_{1,2}$ in Fig. 1.

We now look for conditions for disciplinability which are independent of the ε 's and δ 's. Let

$$J_{ij} = \frac{\partial J_i}{\partial \sigma_j}, \quad i, j \in \Lambda.$$

For $S, T \subseteq \Lambda$, let J_{ST} denote the matrix of the elements J_{ij} , $i \in S, j \in T$, while $\det(J_{SS}) = \Delta_S$. Thus $\Delta_i = J_{ii}$ while Δ_Λ is the determinant of the whole J_{ij} matrix. Let $|S|$ denote the cardinality of S .

Define

$$Q = I_{BB} - J_{BC} J_{CC}^{-1} J_{CB} J_{BB}^{-1},$$

provided the inverses exist, where I_{BB} is the $|B| \times |B|$ unit matrix. Finally, $\mathbf{0}$ represents a zero vector or matrix throughout, with dimensions clear from context.

THEOREM 1. (a) If $\sigma \in D_{B,C}$ then $\Delta_B \Delta_C = 0$ or

$$(2.6) \quad \Delta_B \Delta_C \neq 0 \quad \text{and} \quad Q < \mathbf{0}.$$

(b) If (2.6) holds then $\sigma \in D_{B,C}$.

(A proof is given in § 7.)

COROLLARY 1.1. If $B = \{i\}$,

$$(2.6) \equiv \Delta_i \Delta_C \Delta_{C \cup i} < 0.$$

The corollary follows from the general result

$$\det Q = \frac{\Delta_{B \cup C}}{\Delta_B \Delta_C}$$

(Rao (1973, p. 32, Ex. 2.4)). It is proved, by a rather different method, in GW.I.

We note here that, in general, (2.6) implies nothing about $\text{sgn}(\Delta_{B \cup C})$; in particular it can be zero or nonzero. When $|B| = 1$, Corollary 1.1 shows that more can be said, and further examples may be found in Theorems 3 and 7.

Ideally one would like a one-to-one characterization of $D_{B,C}$ in terms of the J 's, but such a characterization would be extremely cumbersome in the ensuing applications (one would need to continually keep account of the vanishing of higher-order derivatives of the J_i 's). The set where $\Delta_B \Delta_C = 0$ which is not wholly identified with $D_{B,C}$ has, in any case, dimension $< N$.

With this limitation we therefore have explicit algebraic constraints on the J_{ij} 's and hence on σ , defining $D_{B,C}$. Theorem 7 below gives more explicit conditions for an important class of J_i 's.

To sidestep this limitation and emphasize the central role of (2.6), we follow GW.I in introducing weak and strong versions of $D_{B,C}$.

DEFINITION 2.4. Coalition B is said to be *weakly disciplinable* by C if $\Delta_B \Delta_C = 0$ or $\Delta_B \Delta_C \neq 0$ and $Q < 0$.

DEFINITION 2.5. Coalition B is said to be *strongly disciplinable* by C if $\Delta_B \Delta_C \neq 0$ and $Q < 0$.

Because of Corollary 1.1, these definitions agree with those of GW.I when $B = \{i\}$. The corresponding sets of states σ are denoted by $D_{B,C}^W$ and $D_{B,C}^S$, so that

$$(2.7) \quad D_{B,C}^S \subseteq D_{B,C} \subseteq D_{B,C}^W.$$

We also define

$$(2.8) \quad O_{B,C}^W \equiv D_{B,C}^W \cap D_{C,B}^W.$$

Typically, $D_{B,C}^W$ consists of $D_{B,C}$ and its boundary points, which gives added appeal to the notion of weak discipline. Similarly, $D_{B,C}^S$ usually comprises $D_{B,C}$ without its boundary points. Unfortunately, this correspondence is not always valid, as shown in GW.I. We can, however, equate $D_{B,C}^S$ to a set of linear constraints.

THEOREM 2. $\sigma \in D_{B,C}^S$ if and only if, given any $|B|$ -vector $\mu \geq 0$, $\mu \neq 0$, with which there is associated a unique $|B|$ -vector α_B defined by

$$(2.9) \quad J_{BB} \alpha_B = \mu,$$

there exists a unique $|C|$ -vector α_C such that

$$(2.10) \quad J_{BB} \alpha_B + J_{BC} \alpha_C < 0,$$

$$(2.11) \quad J_{BC} \alpha_B + J_{CC} \alpha_C = 0.$$

(A proof is given in § 8.)

Despite this pleasing equivalence, it turns out that weak discipline is the more useful concept. This will become especially apparent when we meet the generalized optima in the next section; for the present, we give one property of $D_{B,C}^W$.

THEOREM 3. For any $C \subset \Lambda$ and $i \in \Lambda - C$,

$$D_{C,i}^W \subseteq D_{i,C}^W.$$

(The theorem is proved in § 9.)

COROLLARY 3.1.

$$O_{C,i}^W = D_{i,C}^W.$$

Theorem 3 is an extension of the property

$$D_{1,2} = D_{2,1}$$

for $N = 2$, proved by Gates et al. (1977). It describes the *limited power of an individual*; a single player who has power to discipline a coalition is automatically subject to the discipline of that coalition.

Corollary 3.1 follows directly from Theorem 3 and (2.7) and identifies $O_{i,C}^W$ with the set $\{\Delta_i \Delta_C \Delta_{C \cup i} \leq 0\}$.

3. The generalized optima. In GW.I we introduced four new optima based on the concept of discipline. They all have natural analogues in the present setting. However, since almost all our results concern the least restrictive optimum (Type I), we state formal definitions and properties only for this type. The other types are briefly defined as they occur.

DEFINITION 3.1. A state σ is a *generalized type I optimum* if at least one coalition can be disciplined by another.

One can see that idealized players in a bargaining situation, knowing only that σ is type I, would be reluctant to make adjustments. Each coalition (including the single players) settles for this optimum in order to avoid the risk that it might be the susceptible coalition.

Such states are “optimal” only in the weak sense that every coalition is getting the best it can expect (locally) without risking what it has. The concept is perhaps somewhere between conventional “stability” (in a kinetic sense) and traditional “optimality”.

Unsophisticated or inadequately informed players can be attracted to generalized type I optima through accidental discipline, as described for $O_{B,C}$.

The other types of optima are more appropriate when players have more information (see the discussion in GW.I).

We denote the set of states σ corresponding to a generalized type I optimum by O_I^* , so that

$$(3.1) \quad O_I^* = \bigcup_{C \subset \Lambda} \bigcup_{B \subseteq \Lambda - C} D_{B,C}.$$

Similarly there is a set of *weak generalized type I optima* O_I^{*W} , defined by

$$(3.2) \quad O_I^{*W} = \bigcup_{C \subset \Lambda} \bigcup_{B \subseteq \Lambda - C} D_{B,C}^W;$$

we shall not be concerned with the strong version in this paper. It follows from (2.7) that

$$(3.3) \quad O_I^* \subseteq O_I^{*W}.$$

Further, the type I and weak type I optima of GW.I, defined as in (3.1), (3.2) only for $|B| = 1$, clearly satisfy

$$(3.4) \quad O_I \subseteq O_I^*, \quad O_I^W \subseteq O_I^{*W}.$$

An important property of the weak and weak generalized type I optima is their connection with the adjustment process of Gates et al. (1977). This is covered in §§ 4 and 5 of the paper. We now give a few other properties of the new optima. Let PO' denote the set of σ 's which are Pareto optimal and for which $\text{rank}(J_{\Lambda\Lambda}) = N - 1$. It is clear that

$$O_I^* = O_I, \quad O_I^{*W} = O_I^W \quad \text{for } N = 2.$$

THEOREM 4. (a) For $N = 3$,

$$O_I^{*W} = O_I^W.$$

(b) The Cournot and PO' optima belong to O_I^* .

(c) The Cournot and Pareto optima belong to O_I^{*W} .

Proof. The proof of part (a) follows from the observation that, for $N = 3$,

$$O_I^* = O_I, \quad O_I^{*W} = O_I^W \quad \text{for } N = 2.$$

together with Theorem 3. To prove (b), note that GW.I (Theorem 4) proved that the Cournot and PO' optima belong to O_I and use (3.4). Since the Cournot and Pareto optima belong to O_I^W (GW.I, Theorem 3), (3.4) also establishes (c).

Note that there are a number of explicit results about O_I^W for $N = 2$ and 3 in GW.II, which transfer directly to the generalized optima by Theorem 4(a) and the remark preceding it. In particular, (a) shows that, in a 3-player game, weak generalized optimality is determined entirely by the susceptibility of individual players to discipline. This equivalence of O_I^W and O_I^{*W} does not extend to $N > 3$.

Parts (b) and (c) relate the new generalized optima to the traditional optima of game theory, (b) being the deeper result. Note that the points σ which are Pareto optimal but not in PO' have relatively zero measure, since they belong to the sets where $\Delta_{\Lambda-i} = 0$ for at least one i , which have an $(N - 2)$ -dimensional intersection with the set of Pareto optimal states, where $\Delta_{\Lambda} = 0$. For example, they comprise only points in the 2-player case.

4. Equilibrium states of an adjustment process. In Gates et al. (1977) (see also GW.II) an explicit dynamic adjustment process of the form

$$(4.1) \quad \sigma(t+1) = F(\sigma(1), \sigma(2), \dots, \sigma(t))$$

was analyzed. Here $\sigma(t)$ denotes the strategies chosen by the N players at time t , or transaction t , where $t = 1, 2, \dots$. In this process each player attempted to maximize a least squares estimate of his payoff function which he constructed entirely from his past σ_i 's and payoffs. Thus each player had no direct knowledge (or made no use) of the payoff functions $J_i(\sigma)$ (even his own) nor of the σ_j 's chosen nor payoffs received by his competitors. The process models an accounting procedure actually used by small firms (see for example Weinberg (1961)).

The main result of Gates et al. (1977) was to prove for a substantial class of functions $J_i(\sigma)$ that any equilibrium solution (i.e. limiting solution as $t \rightarrow \infty$) of (3.1) satisfies the condition:

There exists a nonzero, positive semidefinite, symmetric matrix A_{ij} such that

$$(4.2) \quad \sum_{j \in \Lambda} A_{ij} J_{ij} = 0 \quad \text{for all } i \in \Lambda.$$

Thus it was proved that the set E of *equilibrium states* σ of (4.1) belongs to the set M of states σ defined by (4.2):

$$(4.3) \quad E \subseteq M.$$

Some exact solutions of (4.1) were obtained. A numerical study of (4.1) also strongly indicated that $E = M$ for $N = 2$. Further numerical study in GW.II likewise indicated that $E = M$ for $N = 3$. Convergence properties of a special case of (4.1) were derived in Gates et al. (1978), (1981), but the general case of (4.1) is unresolved.

In GW.I (Thm. 5) we provided a link between M and the weak optima, namely

$$(4.4) \quad O_I^W \subseteq M,$$

so that the numerical studies mentioned above suggest the relation $O_I^W \subseteq E$. We can now establish the following much deeper result.

THEOREM 5.

$$(4.5) \quad O_I^{*W} \subseteq M.$$

The proof, given in § 10, requires the Perron–Frobenius theorem applied to the positive matrix $(-Q)$. Theorem 5, of course, implies (4.4) via (3.4), and it also suggests that $O_I^{*W} \subseteq E$, i.e. all weak generalized optimum states are equilibrium states. Such relationships are discussed further after Theorem 6 in the next section.

While (4.2) defines M precisely it is very inexplicit, and Theorem 5 does not help in providing an exact delineation. It is known (Gates et al. (1977, Thm. 2)) that, for $N = 2$, M is equivalent to the condition

$$(4.6) \quad \Delta_1 \Delta_2 \Delta_{12} \leq 0$$

involving only the J_{ij} 's; that is, $M = O_I^W$ and hence, by Theorem 4(a), $M = O_I^{*W}$. For general payoff functions no similar result is available if $N > 2$, though there are some useful but incomplete constraints on M given in Theorems 7 and 8 of GW.I. However, for an important class of payoff functions (see (5.1) following), the identity $M = O_I^{*W}$ continues to hold, and this is the content of the next section.

5. A class of payoff functions. Following GW.II we shall consider payoff functions of the form

$$(5.1) \quad J_i(\sigma) = f_i\{\sigma_i, \phi(\sigma)\}, \quad i = 1, \dots, N.$$

They generalize widely studied models of competing firms in quantity-variation markets without product differentiation, which assume

$$J_i(\sigma) = \sigma_i \phi(\sigma) - C_i(\sigma_i), \quad i = 1, \dots, N.$$

In these formulae, ϕ is the common price per unit of output σ_i and C_i is the cost to firm i of producing σ_i . Results much stronger than the preceding can be obtained for payoffs of the form (5.1).

THEOREM 6. *For J_i of the form (5.1), and for all N ,*

$$(5.2) \quad O_I^{*W} = O_I^W = M.$$

Proof. The theorem follows from Theorem 5 and (3.4), which give

$$(5.3) \quad O_I^W \subseteq O_I^{*W} \subseteq M,$$

together with Theorem 1 of GW.II which states that, for these J_i ,

$$(5.4) \quad O_I^W = M.$$

Since $E \subseteq M$ we have $E \subseteq O_I^{*W}$, although this is weaker than the known result $E \subseteq O_I^W$, which states that every equilibrium state of the adjustment process outlined in § 4 is a generalized optimum. This is remarkable because identification by players of a state in O_I^{*W} requires at least a knowledge of all the payoff functions, while players can arrive at a state in E without such knowledge. There is presumably a hidden learning process implied by the adjustment process, but we are unable to clarify this. It does, however, support the kind of belief, which is common in much of economic theory, that firms acting independently with no knowledge of the profit functions can arrive at an optimal solution of the underlying market game (Day (1975)).

Note that if the conjecture $E = M$ mentioned in § 4 is valid, then for payoff functions of the form (5.1) we have $O_I^{*W} = O_I^W = E$, that is, complete identification of the equilibria with weak optima.

COROLLARY 6.1. *For J_i of the form (5.1) and $N = 3$,*

$$(5.5) \quad O_I^{*W} = O_{II}^W \cup O_{III}^W,$$

where

$$(5.6) \quad O_{II}^W = \bigcup_{i \in \Lambda} D_{i, \Lambda-i}^W$$

and

$$(5.7) \quad O_{III}^W = \bigcap_{i \in \Lambda} \bigcup_{C \subseteq \Lambda-i} D_{i,C}^W.$$

Proof. The corollary is an immediate consequence of Theorem 6 and Theorem 2 of GW.II which states that $O_I^W = O_{II}^W \cup O_{III}^W$ for $N = 3$. The corollary shows that although O_I^{*W} includes all weak disciplining sets for $N = 3$, it nevertheless contains no more than those states where either (O_{II}^W) at least one player can be weakly disciplined by all the rest or (O_{III}^W) every player can be weakly disciplined.

Another bonus of Theorem 6 is that it completely identifies M , for this class of J_i , since Theorem 4 of GW.II provides an explicit description of O_I^W in terms of constraints on the y_i 's defined in (5.8) below. This description is also valid for O_I^{*W} , of course. In GW.II we derived this description from another explicit formula for $D_{i,C}^W$ in terms of the y_i , and it is of interest to provide an analogous result for $D_{B,C}^W$.

We define the new variables

$$(5.8) \quad y_i = -\frac{h_i \phi_i}{g_i}, \quad i = 1, \dots, N,$$

for $g_i \neq 0$, where

$$(5.9) \quad \begin{aligned} g_i(\sigma) &= G_i\{\sigma_i, \phi(\sigma)\}, & h_i(\sigma) &= H_i\{\sigma_i, \phi(\sigma)\}, \\ \phi_i(\sigma) &= \frac{\partial \phi(\sigma)}{\partial \sigma_i}, & G_i(u, v) &= \frac{\partial f_i(u, v)}{\partial u}, & H_i(u, v) &= \frac{\partial f_i(u, v)}{\partial v}, \end{aligned}$$

and f_i and ϕ are the functions defining J_i in (5.1). In (GW.II, Thm. 3) we showed that $D_{i,C}^W$ is essentially equivalent to the condition

$$(5.10) \quad (1 - y_i)(1 - y_C)(1 - y_{C \cup i}) \leq 0,$$

where

$$(5.11) \quad y_C = \sum_{j \in S} y_j.$$

This simple form makes the optimum O_1^W relatively simple to specify in the y -space rather than the σ -space (GW.II, Thm. 4). To deal with $D_{B,C}^W$ we define

$$\Gamma_S = \{\sigma: g_j \neq 0 \forall j \in S\}$$

and

$$Z_{i,S} = \{\sigma: g_i = 0, h_i \phi_i \neq 0 (i \in S), g_j \neq 0 \forall j \in S - i\}$$

with the remaining set of σ 's conditioning g_j and $h_j \phi_j$ for $j \in S$ denoted by $E_S (S \subseteq \Lambda)$. We write $E = E_B \cup E_C$ and \bar{E} for its complement.

THEOREM 7. For J_i of the form (5.1), $D_{B,C}^W$ is the union of E and the following disjoint sets:

(i) those $\sigma \in (\Gamma_B \cup \Gamma_C) \cap \bar{E}$ for which either $(1 - y_B)(1 - y_C) = 0$ or

$$(5.12) \quad \text{and} \quad \frac{y_i y_C}{(1 - y_B)(1 - y_C)} > 1 \quad \forall i \in B,$$

$$g_i g_j \phi_i \phi_j > 0 \quad \forall i, j \in B, \quad i \neq j;$$

(ii) those $\sigma \in (\Gamma_B \cap \bigcup_{k \in C} Z_{k,C}) \cap \bar{E}$ for which either $(1 - y_B) = 0$ or

$$(5.13) \quad \text{and} \quad \frac{y_i}{1 - y_B} < 1 \quad \forall i \in B,$$

$$g_i g_j \phi_i \phi_j > 0 \quad \forall i, j \in B, \quad i \neq j;$$

(iii) those $\sigma \in (\Gamma_C \cap \bigcup_{k \in B} Z_{k,B}) \cap \bar{E}$ for which either $(1 - y_C) = 0$ or

$$(1 - y_C) < 0 \quad \text{and} \quad |B| = 1.$$

(A proof is given in § 11.) Although this description may appear formidable, it is effectively (i) plus constraints which apply only to a set of zero N -dimensional measure in the σ -space.

The theorem provides a relatively simple explicit specification of the set $O_{B,C}^W$ of (2.8) where B and C are balanced. It also in principle specifies O_1^{*W} via (3.2), though of course we already know this specification through O_1^W .

Some observations on Theorem 7 follow.

(a) If $|B| = 1$, say $B = \{i\}$, then (i) $\equiv \Delta_i \Delta_C \Delta_{C \cup i} \leq 0$, (ii) $\equiv 1 - y_i \leq 0$, and (iii) $\equiv 1 - y_C \leq 0$ and $E \equiv E_{C \cup i}$, which is GW.II [Thm. 3].

(b) In many applications ϕ_i and ϕ_j will have the same sign (cf. (1.1)), in which cases $g_i g_j \phi_i \phi_j > 0$ reduces to $g_i g_j > 0$. This in turn is always satisfied by payoffs like (1.1).

(c) The first condition in (5.12) implies the constraint

$$\Delta_B \Delta_C \Delta_{B \cup C} < 1 - |B| \leq 0;$$

to prove this, sum the inequalities over $i \in B$ as in the proof of Theorem 3. Thus there is a definite connection between the $Q < 0$ part of the definition of $D_{B,C}^W$ and the "naive" generalization of $D_{i,C}^W$ to $\Delta_B \Delta_C \Delta_{B \cup C} \leq 0$. Unfortunately, this connection appears to have few applications. Note that the proof of Theorem 3 shows that this implication is true for arbitrary J_i when $|C| = 1$.

(d) It is easy to see that $D_{B,C}^S$ is specified precisely by the "or" parts of (i)–(iii) of the theorem.

6. Alternative definitions of $D_{B,C}$. As mentioned in § 2, an alternative to $D_{B,C}$ which might appear more natural is the set $\bar{D}_{B,C}$ defined like $D_{B,C}$ except that (2.4) is replaced by

$$(6.1) \quad \mathbf{J}_C(\boldsymbol{\sigma} + \boldsymbol{\varepsilon}_B + \boldsymbol{\varepsilon}_C) \geq \mathbf{J}_C(\boldsymbol{\sigma}).$$

For such states, C can discipline B and *at least* restore its own payoffs. However, we shall show below that, perhaps paradoxically, $\bar{D}_{B,C}$ is in fact less conceptually attractive, and less mathematically and practically relevant, than $D_{B,C}$.

First, $\bar{D}_{B,C}$ is less relevant in practice because the optima derived from it do not share the properties of O_I^{*W} which are central to this work. In particular, they do not agree so well with the data of Fouraker and Siegel nor coincide with the set M defining the equilibrium states of our dynamical model. To see this, note that the set of type I optimum points based on $\bar{D}_{B,C}$ must at least contain O_I^{*W} , and hence, for J_i of the form (5.1), must contain M as well. That the inclusion is strict follows from Theorem 10 below, and it can be easily examined directly in simple cases. For example, for J_i of the form (1.1) ($i = 1, 2$) the set O_I^{*W} is given by (1.5), while the set of type I optima based on $\bar{D}_{B,C}$ includes the region $\{\sigma_1 + \sigma_2 \leq 30\}$ as well.

Second, there is a trio of mathematical results which supports the theoretical and conceptual attractiveness of $D_{B,C}$. In all three theorems, we assume $\Delta_B \Delta_C \neq 0$ for ease of presentation; the conclusions still apply almost everywhere in N -space. The proofs are in § 12.

THEOREM 8. *If $\boldsymbol{\sigma} \in D_{B,C}$ and $\Delta_B \Delta_C \neq 0$, then (2.1), (2.2), (2.3) and*

$$(6.2) \quad \mathbf{J}_C(\boldsymbol{\sigma} + \boldsymbol{\varepsilon}_B + \boldsymbol{\varepsilon}_C) > \mathbf{J}_C(\boldsymbol{\sigma})$$

apply for some $\boldsymbol{\varepsilon}_C$ differing from that in the $D_{B,C}$ definition.

Thus if C can discipline B in our original sense, it can also benefit while disciplining B . This shows the apparent extra flexibility in $\bar{D}_{B,C}$ is unnecessary.

THEOREM 9. *Suppose $\boldsymbol{\sigma} \in D_{B,C}$, $\Delta_B \Delta_C \neq 0$ and there exist $\boldsymbol{\varepsilon}_B$ and $\boldsymbol{\varepsilon}_C$ such that*

$$(6.3) \quad \mathbf{J}_B(\boldsymbol{\sigma} + \boldsymbol{\varepsilon}_B) \geq \mathbf{J}_B(\boldsymbol{\sigma}),$$

with strict inequality in at least one component, and

$$(6.4) \quad \mathbf{J}_C(\boldsymbol{\sigma} + \boldsymbol{\varepsilon}_B + \boldsymbol{\varepsilon}_C) = \mathbf{J}_C(\boldsymbol{\sigma})$$

for sufficiently small $\boldsymbol{\varepsilon}_B, \boldsymbol{\varepsilon}_C$. Then necessarily

$$(6.5) \quad \mathbf{J}_B(\boldsymbol{\sigma} + \boldsymbol{\varepsilon}_B + \boldsymbol{\varepsilon}_C) < \mathbf{J}_B(\boldsymbol{\sigma}).$$

Thus C need only restore its payoffs, in response to a payoff-increasing adjustment by B , in order to be certain that it will strictly discipline B . In a game with limited information, this gives extra security to C and consequently stability to O_I^* . Again, this shows that our original definition is conceptually and mathematically adequate.

THEOREM 10. *If $\boldsymbol{\sigma} \in \bar{D}_{B,C} - D_{B,C}$ and $\Delta_B \Delta_C \neq 0$ then for any $\boldsymbol{\varepsilon}_B$ such that*

$$\mathbf{J}_B(\boldsymbol{\sigma} + \boldsymbol{\varepsilon}_B) \geq \mathbf{J}_B(\boldsymbol{\sigma}),$$

with strict inequality in at least one component, there exists $\boldsymbol{\varepsilon}_C$ such that

$$J_i(\boldsymbol{\sigma} + \boldsymbol{\varepsilon}_B + \boldsymbol{\varepsilon}_C) > J_i(\boldsymbol{\sigma}) \quad \text{for at least one } i \in B$$

and

$$\mathbf{J}_C(\boldsymbol{\sigma} + \boldsymbol{\varepsilon}_B + \boldsymbol{\varepsilon}_C) = \mathbf{J}_C(\boldsymbol{\sigma}).$$

Thus states σ in $\bar{D}_{B,C}$ but not in $D_{B,C}$ have the undesirable property that restoration of payoff by C does *not* guarantee that it has disciplined B . This can easily be checked directly for the example of § 1, when the region $\bar{D}_{B,C} - D_{B,C}$ is $\{\sigma_1 + \sigma_2 \leq 30\}$.

Less radical alternative definitions can also be derived by modifying the other inequalities in the definition of $D_{B,C}$. For example, (2.3) could either have a strict $<$ in all components or permit equality in all components. These would lead to new D 's and O_I 's which bracket $D_{B,C}$ and O_I^* . Further, various intermediate steps between $D_{B,C}$ and $D_{B,C}^W$ in the proof of Theorem 1, involving linear equations like (2.9)–(2.11), could be considered as definitions, though these would not be as intuitively appealing. However, it is worth noting that, on the set $\Delta_B \Delta_C \neq 0$, *all these definitions will coincide*, by Theorem 1, since then $D_{B,C}^S \doteq D_{B,C}^W$. That is, all these definitions agree almost everywhere in N -space, so we can reasonably take whichever we find most natural. From all points of view, we feel this is $D_{B,C}$.

7. Proof of Theorem 1. From the definition of $D_{B,C}$ there are two possibilities. One is that (2.1) holds, so that $\Delta_B = 0$ (see for example GW.I [Thm. 3]), in agreement with Theorem 1. The other is that (2.2), (2.3) and (2.4) hold. Let α_S denote a real-valued $|S|$ -vector. Then for any $\alpha_B \neq 0$ such that

$$(7.1) \quad J_{BB}\alpha_B \geq 0$$

there exists $\alpha_C \neq 0$ such that

$$(7.2) \quad J_{BB}\alpha_B + J_{BC}\alpha_C \leq 0,$$

$$(7.3) \quad J_{CB}\alpha_B + J_{CC}\alpha_C = 0.$$

If either of Δ_B or Δ_C is zero then Theorem 1 is satisfied. If both are nonzero then inequalities (7.1) and (7.2) must be strict and J_{BB}^{-1}, J_{CC}^{-1} must exist. From (7.3) we have

$$(7.4) \quad \alpha_C = J_{CC}^{-1} J_{CB}\alpha_B$$

which reduces (7.2) to

$$(7.5) \quad (J_{BB} - J_{BC}J_{CC}^{-1}J_{CB})\alpha_B < 0.$$

Putting

$$(7.6) \quad \mu = J_{BB}\alpha_B > 0$$

reduces this to

$$(7.7) \quad (I_{BB} - J_{BC}J_{CC}^{-1}J_{CB}J_{BB}^{-1})\mu < 0$$

for any $\mu > 0$, since α_B is arbitrary subject to the positivity of μ . It is clear from (7.7) that Q can have no nonnegative elements, which completes the proof of (a).

Conversely, suppose that (2.6) holds. Given any $\mu > 0$ define α_B by (7.6) and α_C by (7.4). Then (7.1) with strict inequality and (7.3) are recovered, while $Q < 0$ implies (7.7), which in turn implies (7.2) with strict inequality. The statements (2.2) to (2.4) defining $D_{B,C}$ are then deduced from the implicit function theorem following the argument in GW.I [Thm. 1]. This completes the proof of (b).

8. Proof of Theorem 2. The implication from $\sigma \in D_{B,C}^S$ to the linear equations is effectively proved in the course of proving Theorem 1(b); the uniqueness of α_B, α_C comes directly from Δ_B and Δ_C nonzero. Conversely, if (2.9)–(2.11) hold, the uniqueness assumptions imply Δ_B and Δ_C are nonzero and then the argument in Theorem 1(a) proves that (2.6) holds, that is, $\sigma \in D_{B,C}^S$.

9. Proof of Theorem 3. Take $C \subseteq \Lambda - i$. Suppose $\sigma \in D_{C,i}^W$ and $\Delta_C \Delta_i = 0$; then, automatically $\sigma \in D_{i,C}^W$. Otherwise, Q is well defined and

$$(9.1) \quad Q = \frac{1}{\Delta_C \Delta_i} \{ \Delta_C \Delta_i I_{CC} - [J_{ji} J_{ik}] [\text{cof } J_{kj}^{(ii)}] \},$$

where $J_{ab}^{(cd)}$ is the (a, b) th element of $J_{C \cup i, C \cup i}$ omitting row c and column d , cof denotes cofactor and j and k run over C in the $[\]$ matrices.

Now $[J_{ji} J_{ik}] [\text{cof } J_{kj}^{(ii)}]$ has (j, k) th element

$$(9.2) \quad J_{ji} \sum_{l \in C} J_{il} \text{cof } J_{kl}^{(ii)}.$$

But

$$\text{cof } J_{kl}^{(ii)} = (-1)^{i+k-1} \text{cof } J_{il}^{(ki)},$$

so (9.2) becomes

$$(9.3) \quad J_{ji} \sum_{l \in C} J_{il} \text{cof } J_{il}^{(ki)} (-1)^{i+k-1} = J_{ji} (-1)^{i+k-1} (-1)^{i+k} \text{cof } J_{ki} = -J_{ji} \text{cof } J_{ki}.$$

So from (9.1) and (9.3)

$$Q \equiv \frac{X}{\Delta_C \Delta_i},$$

where $X = [x_{jk}]$ and, for $j, k \in C$,

$$x_{jk} = \begin{cases} J_{ii} \text{cof } J_{ii} + J_{ji} \text{cof } J_{ji} & (j = k), \\ J_{ji} \text{cof } J_{ki} & (j \neq k). \end{cases}$$

Since $\sigma \in D_{C,i}^W$ and $\Delta_C \Delta_i \neq 0$, we must have $Q < 0$. Summing the diagonal elements of Q gives

$$0 > \frac{1}{\Delta_C \Delta_i} \sum_{j \in C} x_{jj} = (|C| - 1) + \frac{\Delta_{C \cup i}}{\Delta_C \Delta_i},$$

which implies

$$\frac{\Delta_{C \cup i}}{\Delta_C \Delta_i} < -(|C| - 1) \leq 0;$$

that is, $\sigma \in D_{i,C}^W$, and the theorem is proved.

10. Proof of Theorem 5. If $\sigma \in O_1^{*W}$, then $\sigma \in D_{B,C}^W$ for at least one pair of coalitions B, C . If $\Delta_S = 0$ for S one of B and C , chooses

$$A_{jk} = \begin{cases} x_j x_k, & j, k \in S, \\ 0, & \text{otherwise,} \end{cases}$$

where the x_j are a nonzero solution of the equations

$$\sum_{k \in S} x_k J_{jk} = 0 \quad \text{for all } j \in S.$$

Clearly $A = [A_{jk}]$ is positive semidefinite and satisfies (4.2). If $\Delta_B = 0$ and $\Delta_C = 0$ hold simultaneously, one has a choice of the corresponding A_{jk} .

It remains to consider $\Delta_B \Delta_C \neq 0$, when $Q < 0$ by (2.6). Write $Q_+ = (-Q)$, so Q_+ is a positive matrix. Hence, by the Perron–Frobenius theorem (Gantmacher (1974, p. 53)) there is a unique, real, positive eigenvalue η of Q_+ with associated eigenvector $\mu > 0$, i.e.,

$$(10.1) \quad Q_+ \mu = \eta \mu.$$

Defining α_B by (7.6) and α_C by (7.4) leads as before to (7.3), while

$$J_{BB} \alpha_B + J_{BC} \alpha_C = -Q_+ \mu,$$

which, with (10.1) and (7.6), becomes

$$(10.2) \quad J_{BB}(1 + \eta) \alpha_B + J_{BC} \alpha_C = 0.$$

Multiplying the i th indexed equation in the set of equations (10.2) and (7.3) by α_i reduces them to the form (4.2) with

$$A_{ij} = \begin{cases} (1 + \eta) \alpha_i \alpha_j & \text{if } i, j \in B, \\ \alpha_i \alpha_j & \text{if } i, j \in B \cup C \text{ but } i, j \text{ not both in } B, \\ 0 & \text{otherwise.} \end{cases}$$

Now note that, for real $u_i (i \in \Lambda)$,

$$\sum_{i,j} u_i A_{ij} u_j = \left(\sum_{i \in B \cup C} \alpha_i u_i \right)^2 + \eta \left(\sum_{i \in B} \alpha_i u_i \right)^2,$$

which verifies that A_{ij} is positive-semidefinite. Thus $D_{B,C}^W \subseteq M$ for any B and C , which establishes Theorem 5.

11. Proof of Theorem 7. With the definitions (5.8) and (5.9) we have

$$(11.1) \quad J_{ij} = \begin{cases} g_i + h_i \phi_i & (i = j), \\ h_i \phi_j & (i \neq j), \end{cases} \quad i, j \in \Lambda.$$

Consider any set $S \subseteq \Lambda$. It can be shown, by routine calculations, that

$$(11.2) \quad \Delta_S = \prod_{i \in S} g_i + \sum_{i \in S} \left(\prod_{j \in S-i} g_j \right) h_i \phi_i$$

and, if $\Delta_S \neq 0$, that

$$(11.3) \quad J_{SS}^{-1} = \Delta_S^{-1} \Xi,$$

where

$$(11.4) \quad \Xi = [\xi_{ij}], \quad \xi_{ij} = \begin{cases} \prod_{k \in S-i} g_k + \sum_{k \in S-i} \left(\prod_{l \in S-i-k} g_l \right) h_k \phi_k & (i = j), \\ - \left(\prod_{k \in S-i-j} g_k \right) h_i \phi_j & (i \neq j). \end{cases}$$

Thus, if $\Delta_B \Delta_C \neq 0$, we find from (11.4) that

$$(11.5) \quad J_{BC} J_{CC}^{-1} = \frac{1}{\Delta_C} \left(h_i \phi_j \prod_{k \in C-j} g_k \right)_{i \in B, j \in C},$$

$$(11.6) \quad J_{CB} J_{BB}^{-1} = \frac{1}{\Delta_B} \left(h_i \phi_j \prod_{k \in B-j} g_k \right)_{i \in C, j \in B},$$

whence

$$(11.7) \quad Q = I_{BB} - \frac{1}{\Delta_B \Delta_C} \left\{ h_i \phi_j \sum_{p \in C} h_p \phi_p \left(\prod_{k \in C-p} g_k \right) \left(\prod_{k \in B-j} g_k \right) \right\}_{i,j \in B}.$$

Now for any $S \subseteq \Lambda$, by definition,

$$(11.8) \quad \{\sigma\} = \Gamma_S \cup \bigcup_{t \in S} Z_{t,S} \cup E_S.$$

We must consider all possible combinations of the sets in (11.8), for $S = B$ and C , in specifying $D_{B,C}^W$.

First, using (11.2), we have

$$(11.9) \quad \{\sigma: \Delta_S \neq 0\} \cap \Gamma_S = \{(1 - y_S) \neq 0\} \cap \Gamma_S,$$

$$(11.10) \quad \{\sigma: \Delta_S \neq 0\} \cap Z_{t,S} = Z_{t,S} \quad \forall t \in S,$$

$$(11.11) \quad \{\sigma: \Delta_S \neq 0\} \cap E_S = \emptyset.$$

Case 1. $\sigma \in E$. This implies $\Delta_B = \Delta_C = 0$, by (11.11), so that

$$(11.12) \quad E \cap D_{B,C}^W = E.$$

Case 2. $\sigma \in (\Gamma_B \cap \Gamma_C) \cap \bar{E}$. In this case, either $\Delta_B \Delta_C = 0$, which is equivalent to $(1 - y_B)(1 - y_C) = 0$ by (11.9), or, from (11.7), for $i, j \in B$,

$$(11.13) \quad Q_{ij} = \begin{cases} 1 - \frac{y_i y_C}{(1 - y_B)(1 - y_C)} & (i = j), \\ -\frac{y_i y_C}{(1 - y_B)(1 - y_C)} \frac{g_i \phi_j}{g_j \phi_i} & (i \neq j), \end{cases}$$

provided $\phi_i \neq 0$. From (11.13), the subset of these σ 's which is in $D_{B,C}^W$ is given precisely by $(1 - y_B)(1 - y_C)(1 - y_{B \cup C}) = 0$ or by (5.12); the constraint $\phi_i \neq 0$ disappears since it is implied by the requirement $Q_{ii} < 0$.

Case 3. $\sigma \in (\Gamma_B \cap Z_{t,C}) \cap \bar{E}$ for some $t \in C$. In this case, either $\Delta_B \Delta_C = 0$, which is equivalent to $(1 - y_B) = 0$ by (11.9), (11.10), or, from (11.7), for $i, j \in B$,

$$(11.14) \quad Q_{ij} = \begin{cases} 1 + \frac{y_i}{1 - y_B} & (i = j), \\ -\frac{y_i}{1 - y_B} \frac{g_i \phi_j}{g_j \phi_i} & (i \neq j), \end{cases}$$

provided $\phi_i \neq 0$.

From (11.14), the subset of these σ 's which is in $D_{B,C}^W$ is given precisely by $(1 - y_B) = 0$ or by (5.13). Again, $\phi_i \neq 0$ is necessarily satisfied when $Q_{ii} < 0$.

Case 4. $\sigma \in (Z_{t,B} \cap \Gamma_C) \cap \bar{E}$ for some $t \in B$. In this case, either $\Delta_B \Delta_C = 0$, which is equivalent to $1 - y_C = 0$ by (11.9), (11.10), or, from (11.7), for $i, j \in B$,

$$(11.15) \quad Q_{ij} = \begin{cases} 1 + \delta_{it} \frac{y_C}{1 - y_C} & (i = j), \\ \delta_{jt} \frac{y_C}{1 - y_C} \frac{h_i}{h_t} & (i \neq j). \end{cases}$$

Here δ_{ab} is the Kronecker delta, and division by h_t is allowed since, in $Z_{t,B}$, $h_t \neq 0$. Since $Q_{ii} = 1$ for $i \neq t$, it is clear that the intersection of this set of σ 's with $D_{B,C}^W$ is empty unless $|B| = 1$, when $Q_{ii} < 0 \equiv 1 - y_C < 0$.

Case 5. $\sigma \in (Z_{t,B} \cap Z_{u,C}) \cap \bar{E}$ for some $t \in B$, $u \in C$. In this case, $\Delta_B \Delta_C \neq 0$ by (11.10), while from (11.7), for $i, j \in B$,

$$(11.16) \quad Q_{ij} = \begin{cases} 1 - \delta_{it} & (i = j), \\ -\delta_{jt} \frac{h_i}{h_t} & (i \neq j). \end{cases}$$

Since $Q_{ii} = 1$ for $i \neq t$, this set of σ 's also has a null intersection with $D_{B,C}^W$.

Because the conclusions drawn in Cases 3–5 are true irrespective of the values of t and u , $\sigma \in E$ follows from Case 1, (i) from Case 2, (ii) from Case 3 and (iii) from Case 4. This proves the theorem.

12. Proofs of Theorems 8, 9, 10. We begin by proving Theorem 9. This follows directly from Theorem 1, since if $\Delta_B \Delta_C \neq 0$ at σ then $D_{B,C} \equiv \{Q < 0\}$ which automatically makes the inequality in (2.3) strict.

To prove Theorem 8, put

$$(12.1) \quad \omega = J_B(\sigma) - J_B(\sigma + \epsilon_B + \epsilon_C) > 0$$

subject to the definition of $D_{B,C}$. That $\omega > 0$ follows from Theorem 9. If $\Delta_C \neq 0$ at σ then, by the implicit function theorem, we can choose a $|C|$ -vector $\gamma > 0$, and an N -vector δ , with $\delta_i = 0$ unless $i \in C$, such that

$$(12.2) \quad J_C(\sigma + \epsilon_B + \epsilon_C + \delta) = J_C(\sigma) + \gamma,$$

where $\delta \rightarrow 0$ as $\gamma \rightarrow 0$. Choosing γ small enough therefore ensures that δ is small enough so that

$$(12.3) \quad J_B(\sigma + \epsilon_B + \epsilon_C + \delta) < J_B(\sigma + \epsilon_B + \epsilon_C) + \omega$$

by continuity. We deduce from (12.1) and (12.3) that

$$(12.4) \quad J_B(\sigma + \epsilon_B + \epsilon'_C) < J_B(\sigma)$$

where $\epsilon'_C = \epsilon_C + \delta$, while $J_C(\sigma + \epsilon_B + \epsilon'_C) > J_C(\sigma)$ from (12.2), which proves Theorem 8.

For Theorem 10, if $\sigma \in \bar{D}_{B,C} - D_{B,C}$ then for any ϵ_B such that (2.2) holds there exists an ϵ_C such that (2.3) and

$$(12.5) \quad J_C(\sigma + \epsilon_B + \epsilon_C) > J_C(\sigma)$$

hold, but no ϵ'_C such that (2.3) and

$$(12.6) \quad J_C(\sigma + \epsilon_B + \epsilon'_C) = J_C(\sigma)$$

hold. Thus varying ϵ_C to reduce (12.5) to an equality must result in violation of (2.3), thus proving Theorem 10.

REFERENCES

- [1] J. P. AUBIN (1979), *Mathematical Methods of Game and Economic Theory*, Studies in Mathematics and its Applications, 7, North-Holland, Amsterdam.
- [2] R. J. AUMANN AND M. MASCHLER (1964), *The bargaining set for cooperative games*, in *Advances in Game Theory*, M. Dresher, L. S. Shapley and A. W. Tucker, eds., Princeton Univ. Press, Princeton, NJ.
- [3] L. J. BILLERA (1972), *Global stability in n -person games*, Trans. Amer. Math. Soc., 172, pp. 45–56.
- [4] J. H. CASE (1979), *Economics and the competitive process*, New York Univ. Press, New York.
- [5] R. H. DAY (1975), *Adaptive processes and economic theory*, in *Adaptive Economic Models*, R. H. Day and T. Groves, eds., Academic Press, New York.
- [6] L. E. FOURAKER AND S. SIEGEL (1960), *Bargaining Behaviour*, McGraw-Hill, New York.
- [7] F. R. GANTMACHER (1974), *The Theory of Matrices*, Vol. II. Chelsea Publ. Co., New York.
- [8] D. J. GATES, J. A. RICKARD AND M. WESTCOTT (1982), *Exact cooperative solutions of a duopoly model without cooperation*, J. Math. Econom., 9, pp. 27–35.
- [9] ——— (1977), *A convergent adjustment process for firms in competition*, Econometrica, 45, pp. 1349–1364.
- [10] ——— (1978), *Convergence of a market related game strategy*, J. Math. Econom., 5, pp. 97–109.
- [11] D. J. GATES AND M. WESTCOTT (1981a), *Extended optima and equilibria for continuous games, I. General results*, J. Austral. Math. Soc. Ser. B, 22, pp. 291–307.
- [12] ——— (1981b), *Extended optima and equilibria for continuous games, II. A class of economic models*, J. Austral. Math. Soc. Ser. B, 23, pp. 187–209.
- [13] ——— (1981c), *Extended optima and equilibria for continuous games, III. Comparison with bargaining experiments*, J. Austral. Math. Soc. Ser. B, 23, pp. 210–227.
- [14] A. J. JONES (1980), *Game Theory: Mathematical Models of Conflict*, Ellis Horwood, London.
- [15] G. KALAI, M. MASCHLER AND G. OWEN (1975), *Asymptotic stability and other properties of trajectories and transfer sequences leading to bargaining sets*, Internat. J. Games Theory, 4, pp. 193–213.
- [16] W. F. LUCAS (1966), *n -person games with only 1, $n - 1$ and n -person coalitions*, Z. Wahrsch. Verw. Gebiete, 6, pp. 287–292.
- [17] J. VON NEUMANN (1928), *On the theory of games*, Ann. Math. Studies, 40 (1959), pp. 13–42; English transl. of 1928 paper.
- [18] G. OWEN (1968), *n -person games with only 1, $n - 1$ and n -person coalitions*, Proc. Amer. Math. Soc., 19, pp. 1258–1261.
- [19] C. R. RAO (1973), *Linear Statistical Inference and its Applications*, 2nd ed., John Wiley, New York.
- [20] R. E. STEARNS (1968), *Convergent transfer schemes for n -person games*, Trans. Amer. Math. Soc., 134, pp. 449–459.
- [21] W. VICKREY (1959), *Self policing properties of certain imputation sets*, Ann. Math. Studies, 40, pp. 213–246.
- [22] E. S. WEINBERG (1961), *The uses and limitations of mathematical models for market planning*, in *Mathematical Models and Methods in Marketing*, Irwin Series in Quantitative Analysis for Business, Homewood, Irwin, IL.

AFFINE INCENTIVE SCHEMES FOR STOCHASTIC SYSTEMS WITH DYNAMIC INFORMATION*

TAMER BAŞAR†

Abstract. In this paper we study the derivation of optimal incentive schemes in two-agent stochastic decision problems with a hierarchical decision structure, in a general Hilbert space setting. The agent at the top of the hierarchy is assumed to have access to the value of other agent's decision variable as well as to some common and private information, and the second agent's loss function is taken to be strictly convex. In this set-up, it is shown that there exists, under some fairly mild structural restrictions, an optimal incentive policy for the first agent, which is affine in the dynamic information and generally nonlinear in the static (common and private) information. Certain special cases are also discussed and a numerical example is solved.

Key words. stochastic systems, decision problems with multiple decision makers, incentive schemes, hierarchical information patterns, stochastic nonzero-sum games, Stackelberg solution

1. Introduction. Consider the general class of two-agent stochastic dynamic decision problems with a hierarchical decision structure, wherein one of the agents (called the leader) has access to both the decision value and observation of the other agent (called the follower), and the objective is verification of existence and derivation of optimal strategies for the leader under which the follower's optimal response (based on the minimization of his expected cost function) leads to a desired "optimal" performance for the leader. Such problems are known as Stackelberg problems [1]–[5] or incentive design problems [8], [20]–[23] and have recently attracted considerable attention in the literature, because of the nonstandard nature of the optimization problem faced by the leader, when he has access to dynamic information [6]–[18]; for a survey and unification of some of the available results in the literature on deterministic and stochastic dynamic Stackelberg problems we refer to [8], [12] and [19], and also to [26] for a general discussion.

A recent reference [15] has shown that in deterministic dynamic incentive problems with perfect or partial dynamic information, and when the follower's cost function is strictly convex (but not necessarily quadratic), there exists an optimal incentive strategy for the leader which is affine in the dynamic information. The object of this paper is to provide a nontrivial extension of this result to stochastic decision problems in which there is available some common information on the unknown state of Nature to both agents as well as some private information to the leader; the leader has also access to the value of the follower's decision variable. The problem is formulated in general Hilbert spaces with the follower's loss function taken to be strictly convex in both agents' decision variables. In this general framework, we establish existence of an optimal incentive strategy for the leader, which is affine in the dynamic information, and in general nonlinear in the static (common and private) information; we also obtain an analytic expression for the optimal solution and consider some special cases of the general problem.

* Received by the editors May 28, 1982, and in revised form January 20, 1983. This work was supported in part by the Joint Services Electronics Program under contract N00014-79-C-0424 and in part by the Electric Energy Systems Division, Department of Energy under contract DE-AC01-81RA-50658 with Dynamic Systems, Urbana, Illinois 61801. It was presented at the American Automatic Control Conference, Arlington, Virginia, June 14–16, 1982.

† Department of Electrical Engineering and Coordinated Science Laboratory, University of Illinois, Urbana, Illinois 61801.

Section 2 provides a precise problem formulation for the case when only common information is available to the leader, whose solution is obtained in § 3 (cf. Proposition 1). Section 4 extends the formulation and results of §§ 2 and 3 to the more general case when the leader has also access to some private information, and a characterization of the complete affine solution is provided in Proposition 2. Section 5 contains a numerical example that serves to illustrate some salient aspects of the solution, and the paper ends with the concluding remarks of § 6.

2. Problem formulation. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be an underlying probability space, on which are defined two random variables x and z taking values in \mathbb{R}^n and \mathbb{R}^m respectively. Let U and V be given Hilbert spaces, denoting the decision spaces of DM1 (the leader) and DM2 (the follower) respectively, and let Γ_1 and Γ_2 be the corresponding strategy spaces defined as

$$(1) \quad \Gamma_2 = \{\text{measurable } \gamma_2: \mathbb{R}^m \rightarrow V, \text{ such that } E_z \{ \langle \gamma_2(z), \gamma_2(z) \rangle \} < \infty\},$$

$$(2) \quad \Gamma_1 = \{\text{measurable } \gamma_1: V \times \mathbb{R}^m \rightarrow U, \text{ such that } E_z \{ \langle \gamma_1[\gamma_2(z), z], \gamma_1[\gamma_2(z), z] \rangle_u \} < \infty \forall \gamma_2 \in \Gamma_2\}.$$

Furthermore let $\Gamma_1^s \subset \Gamma_1$ denote the set of all static policies for the leader, i.e.

$$(3) \quad \Gamma_1^s = \{\text{measurable } \gamma_1: \mathbb{R}^m \rightarrow U, \text{ such that } E_z \{ \langle \gamma_1(z), \gamma_1(z) \rangle_u \} < \infty\}.$$

Here, $\langle \cdot, \cdot \rangle_u$ (respectively, $\langle \cdot, \cdot \rangle_v$) denotes the inner product associated with the Hilbert space U (respectively, V), and the measurable transformations are restricted by the further (implicit) condition that the expectations of the related expressions are well defined. With this construction, Γ_1 , Γ_1^s , and Γ_2 become Hilbert spaces under the natural inner products derived from those defined on U , V , and V , respectively. Note that, to each pair (γ_1, γ_2) in $\Gamma_1 \times \Gamma_2$, there corresponds an unique element $\beta_1 \in \Gamma_1^s$, defined by $\beta_1(z) = \gamma_1[\gamma_2(z), z]$.

We now introduce two functions, $L_1: \mathbb{R}^n \times U \times V \rightarrow \mathbb{R}$, $L_2: \mathbb{R}^n \times U \times V \rightarrow \mathbb{R}$, as the loss functions of DM1 and DM2 respectively, and further introduce

$$J_i: \mathbb{R}^m \times \Gamma_1 \times \Gamma_2 \rightarrow \mathbb{R}, \quad i = 1, 2,$$

as

$$(4) \quad J_i(z, \gamma_1, \gamma_2) = E_{x|z} \{ L_i(x, u, v) \mid u = \gamma_1(v, z), v = \gamma_2(z) \},$$

where $E_{x|z}$ denotes expectation over the statistics of x with conditioning on the observed value of z . Finally, we let $\bar{J}_i(\gamma_1, \gamma_2)$, $i = 1, 2$, defined by

$$(5) \quad \bar{J}_i(\gamma_1, \gamma_2) = E_z \{ J_i(z, \gamma_1, \gamma_2) \},$$

denote the expected cost of DM i , under the policy pair $\gamma_1 \in \Gamma_1$, $\gamma_2 \in \Gamma_2$. We assume at this point that the follower's loss functional $L_2(x, u, v)$ is strictly convex on $U \times V$, for every $x \in \mathbb{R}^n$.

The problem faced by the leader is to find a strategy (synonymously, an incentive scheme) which, by also taking into account rational (expected-cost minimizing)

responses of the follower, leads to a most favorable performance for the leader. This performance may be defined as the global minimum value of $\bar{J}_1(\gamma_1, \gamma_2)$ over $\Gamma_1 \times \Gamma_2$, or equivalently over $\Gamma_1^s \times \Gamma_2$ [assuming that it exists]:

$$(6) \quad \bar{J}_1 = \min_{(\gamma_1, \gamma_2) \in \Gamma_1^s \times \Gamma_2} \bar{J}_1(\gamma_1, \gamma_2) = \bar{J}_1(\gamma_1^t, \gamma_2^t)$$

which corresponds to some specific choices of $\gamma_1 \in \Gamma_1^s$ and $\gamma_2 \in \Gamma_2$ (in this case, $\gamma_1 = \gamma_1^t$ and $\gamma_2 = \gamma_2^t$); or, more generally, there may exist some pair in $\Gamma_1^s \times \Gamma_2$ (denoted again (γ_1^t, γ_2^t)) which is chosen according to some criterion and is considered to be most favorable to the leader. The question we address, in the next section, is the existence and derivation of an “optimal” incentive scheme $\gamma_1^0 \in \Gamma_1$ for the leader, under which the best (\bar{J}_2 -minimizing) policy for the follower is $\gamma_2^t \in \Gamma_2$, and a corresponding element in Γ_1^s for the leader is $\gamma_1^t = \gamma_1^0[\gamma_2^t(z), z]$. Note that this is a meaningful problem because, given a pair $(\gamma_1^t, \gamma_2^t) \in \Gamma_1^s \times \Gamma_2$, there exists a plethora of elements γ_1 in Γ_1 with the property $\gamma_1[\gamma_2^t(z), z] = \gamma_1^t(z)$. Furthermore, $\gamma_1^0 \in \Gamma_1$, in this case, is clearly a Stackelberg strategy for the leader [12].

3. Optimal affine incentive schemes for the leader. Let us first introduce, for each $z \in \mathbb{R}^m$, the set

$$(7) \quad \Omega_t(z) = \{(u, v) \in U \times V \mid \tilde{J}_2(z, u, v) \leq \tilde{J}_2(z, u_z^t, v_z^t)\}$$

where

$$(8) \quad \tilde{J}_2(z, u, v) \triangleq E_{x|z} \{L_2(x, u, v) \mid z\}$$

and (u, v) are taken as (deterministic) elements in $U \times V$. Since $L_2(x, \cdot, \cdot)$ was taken to be strictly convex on $U \times V$, $\Omega_t(z)$ is also strictly convex for each $z \in \mathbb{R}^m$, with $\{u_z^t = \gamma_1^t(z), v_z^t = \gamma_2^t(z)\}$ being a boundary point. This implies that, for each $z \in \mathbb{R}^m$, there exists a hyperplane passing through (u_z^t, v_z^t) , and if, further, $\tilde{J}_2(z, \cdot, \cdot)$ is Fréchet differentiable on $U \times V$, for each $z \in \mathbb{R}^m$, the equation of this supporting hyperplane can be written as

$$(9) \quad \langle \nabla_u \tilde{J}_2(z, u_z^t, v_z^t), u - u_z^t \rangle_u + \langle \nabla_v \tilde{J}_2(z, u_z^t, v_z^t), v - v_z^t \rangle_v = 0$$

where $\nabla_u \tilde{J}_2(z, u_z^t, v_z^t) \in U^*$ is the Fréchet derivative of \tilde{J}_2 with respect to u , evaluated at the point (u_z^t, v_z^t) , and U^* is the Hilbert space adjoint to U ; $\nabla_v \tilde{J}_2$ is analogously defined as an element of the adjoint space V^* . Now, assuming that, for every $z \in \mathbb{R}^m$, $\nabla_u \tilde{J}_2(z, u_z^t, v_z^t) \neq 0$, it follows by utilizing the Hahn–Banach theorem [25] (see also [15, Lemma 1]) that there exists a bounded linear operator $Q_z^*: U^* \rightarrow V^*$ satisfying

$$(10) \quad Q_z^* \nabla_u \tilde{J}_2(z, u_z^t, v_z^t) = \nabla_v \tilde{J}_2(z, u_z^t, v_z^t),$$

so that

$$(11) \quad u_z = u_z^t - Q_z(v - v_z^t)$$

lies, for each $z \in \mathbb{R}^m$, on the hyperplane described by (9) and passes through the point (u_z^t, v_z^t) . Here, $Q_z: V \rightarrow U$ is the bounded linear operator that is adjoint to Q_z^* , for each fixed $z \in \mathbb{R}^m$.

The next question is whether (11) is a well-defined strategy for the leader, i.e. whether it belongs to Γ_1 , which requires Q_z to be a measurable function of z . We now establish an even stronger regularity property for Q_z under some regularity conditions on \tilde{J}_2 , u_z^t and v_z^t :

LEMMA 1. Let $\nabla_u \tilde{J}_2(z, u_z^t, v_z^t)$ and $\nabla_v \tilde{J}_2(z, u_z^t, v_z^t)$ be weakly continuous¹ in z . Then, there exists a linear bounded operator $Q_z: V \rightarrow U$, weakly continuous in z , whose adjoint satisfies the linear equation (10).

Proof. For brevity in notation, let $\nabla_u \tilde{J}_2(z, u_z^t, v_z^t) \triangleq \tilde{u}_z^*$, and $\nabla_v \tilde{J}_2(z, u_z^t, v_z^t) \triangleq \tilde{v}_z^*$. For each fixed $z \in \mathbb{R}^m$, introduce a bounded linear operator $P_z: U^* \rightarrow V^*$ by

$$(12) \quad P_z u^* = \frac{\langle \tilde{u}_z^*, u^* \rangle_{u^*}}{\|\tilde{u}_z^*\|_{u^*}^2} \tilde{v}_z^*, \quad \text{with } u^* \in U^*,$$

which clearly satisfies (10), when substituted for Q_z^* . Now, for any $u^* \in U^*$ and $v \in V$,

$$\langle v, P_z u^* \rangle_v = \langle \tilde{u}_z^*, u^* \rangle_{u^*} \frac{\langle v, \tilde{v}_z^* \rangle_v}{\|\tilde{u}_z^*\|_{u^*}^2},$$

and the latter expression is a continuous functional of z by virtue of the weak continuity of \tilde{u}_z^* and \tilde{v}_z^* . This then implies that P_z is weakly continuous in z , and thereby P_z^* is also weakly continuous in z [24]. Now, taking $Q_z = P_z^*$, it readily follows that there exists a version of Q_z (satisfying (10)) that is weakly continuous in z . \square

The following proposition now summarizes the solution to the incentive problem formulated in § 2.

PROPOSITION 1. For the incentive problem of § 2, if (i) $J_2(z, u, v)$ is Fréchet differentiable on $U \times V$, (ii) for every $z \in \mathbb{R}^m$, $\nabla_u \tilde{J}_2(z, u_z^t, v_z^t) \neq 0$, and (iii) $\nabla_u \tilde{J}_2(z, u_z^t, v_z^t)$ and $\nabla_v \tilde{J}_2(z, u_z^t, v_z^t)$ are weakly continuous² in z , there exists an optimal incentive strategy $(\gamma_1^0(v, z))$ for the leader, in the form

$$(13) \quad u_z^0 = \gamma_1^0(v, z) = u_z^t - Q_z(v - v_z^t),$$

where the linear operator $Q_z: V \rightarrow U$ is chosen according to (10) and is weakly continuous in z .

Remark 1. If U and V are finite-dimensional spaces, $U^* = U$ and $V^* = V$, and consequently $\nabla_u \tilde{J}_2$ and $\nabla_v \tilde{J}_2$ are (column) vectors of appropriate dimensions for each $z \in \mathbb{R}^m$. Then Q_z becomes a matrix-valued function of z , and can be chosen as

$$(14) \quad Q_z = \nabla_u \tilde{J}_2(z, u_z^t, v_z^t) [\nabla_v \tilde{J}_2(z, u_z^t, v_z^t)]' / \|\nabla_u \tilde{J}_2(z, u_z^t, v_z^t)\|^2.$$

Note that, under the hypotheses of Proposition 1, every element of Q_z will be a continuous function of z .

As a special class of problems, let us consider now the case when $L_1(z, u, v)$ is quadratic on $\mathbb{R}^n \times U \times V$:

$$(15) \quad L_2(x, u, v) = \frac{1}{2} \langle u, A_{11} u \rangle_u + \langle u, A_{12} v \rangle_u + \frac{1}{2} \langle v, A_{22} v \rangle_v + \langle u, C_1 x \rangle_u + \langle v, C_2 x \rangle_v,$$

where A_{ij} and C_i are linear bounded operators, A_{22} is strongly positive, and $A_{11} - A_{12}(A_{22})^{-1}A_{21}^*$ is also strongly positive. Then,

$$\tilde{J}_2(z, u, v) = \frac{1}{2} \langle u, A_{11} u \rangle_u + \langle u, A_{12} v \rangle_u + \frac{1}{2} \langle v, A_{22} v \rangle_v + \langle u, C_1 \hat{x} \rangle_u + \langle v, C_2 \hat{x} \rangle_v$$

¹ See [24] for a definition of weak continuity.

² A set of sufficient conditions for this is that i) $\tilde{J}_2(z, u, v)$ be continuously Fréchet differentiable in u and v , and be continuous in z , and ii) u_z^t and v_z^t be weakly continuous in z .

where $\hat{x} = E[x|z]$. Given a point $(u_z^t, v_z^t) \in U \times V$, for each $z \in \mathbb{R}^m$, the Fréchet derivatives at this operating point can easily be determined to be $\langle \cdot, \tilde{u}_z \rangle_u$ and $\langle \cdot, \tilde{v}_z \rangle_v$, where $\tilde{u}_z \in U$ and $\tilde{v}_z \in V$ are

$$(16a) \quad \tilde{u}_z = A_{11}u_z^t + A_{12}v_z^t + C_1\hat{x},$$

$$(16b) \quad \tilde{v}_z = A_{12}^*u_z^t + A_{22}v_z^t + C_2\hat{x}.$$

This then leads to the following Corollary (to Proposition 1) in view of (12).

COROLLARY 1. *Let L_2 be given by (15), and \tilde{u}_z and \tilde{v}_z be defined by (16a) and (16b), respectively. If u_z^t, v_z^t are weakly continuous in z , $\hat{x} = E[x|z]$ is continuous in z , and, for every $z \in \mathbb{R}^m$, $\tilde{u}_z \neq 0$, there exists an optimal incentive strategy for the leader which is affine in v and weakly continuous in z , and is given by*

$$(17) \quad \gamma_1^0(v, z) = u_z^t - \frac{\tilde{u}_z}{\langle \tilde{u}_z, \tilde{u}_z \rangle_u} \langle \tilde{v}_z, v - v_z^t \rangle_v.$$

Proof. In view of the discussion preceding Corollary 1, the proof will be complete if we show that $Q_z: V \rightarrow U$ in (13) (and (10)) is given by $[\langle \tilde{v}_z, \cdot \rangle_v / \|\tilde{u}_z\|_u^2] \tilde{u}_z$. Towards this end, we first observe from (12) that a possible solution of (10) is given by

$$Q_z^* u^* = \frac{\langle \tilde{u}_z^*, u^* \rangle_{u^*}}{\|\tilde{u}_z^*\|_{u^*}^2} \tilde{v}_z^* \quad \text{with } u^* \in U^*,$$

where \tilde{u}_z^* and \tilde{v}_z^* are the Fréchet derivatives belonging to U^* and V^* , respectively. Since U^* and V^* are Hilbert spaces, corresponding to \tilde{u}_z^* and \tilde{v}_z^* there are unique elements $\tilde{u}_z \in U$ and $\tilde{v}_z \in V$, with the property $\langle \tilde{u}_z^*, u^* \rangle_{u^*} = \langle \tilde{u}_z, u^* \rangle_u$ and $\langle \tilde{v}_z^*, v^* \rangle_{v^*} = \langle \tilde{v}_z, v^* \rangle_v$ for all $u^* \in U^*$, $v^* \in V^*$ and every fixed $z \in \mathbb{R}^m$ (see [25]). These elements \tilde{u}_z and \tilde{v}_z can explicitly be determined in our case (because of the specific structure of L_2) and are given by (16a) and (16b), respectively. Hence, we have

$$\langle v, Q_z^* u^* \rangle_v = \frac{\langle \tilde{u}_z, u^* \rangle_u}{\|\tilde{u}_z\|_u^2} \langle v, \tilde{v}_z \rangle_v = \langle Q_z v, u^* \rangle_u$$

whereby

$$Q_z v = \frac{\langle v, \tilde{v}_z \rangle_v}{\|\tilde{u}_z\|_u^2} \tilde{u}_z,$$

which establishes the desired results. \square

4. A more general formulation: Leader acquires private information. We now extend the analysis and results of the previous section to a more general class of incentive problems wherein the leader observes, in addition to z , the output of a second random variable \tilde{y} (taking values in \mathbb{R}^p). This random variable will in general be correlated with x and z ; however, we assume (for technical reasons) existence of a measurable transformation $f: \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}^p$, so that the random variables $y = f(\tilde{y}, z)$ and z are statistically independent, and the sigma fields generated by (\tilde{y}, z) and (y, z) are the same. (For example, if y and z have a joint Gaussian distribution, $f(\tilde{y}, z) \triangleq \tilde{y} - E[\tilde{y}|z]$.) Therefore, we henceforth assume that $u = \gamma_1(v, z, y)$, $v = \gamma_2(z)$, and z and y are statistically independent.

For this problem, we now first modify the definitions of the strategy sets (2) and (3) to read

$$(18) \quad \Gamma_1 = \{\text{measurable } \gamma_1: V \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow U, \text{ such that} \\ E_{z,y} \{ \langle \gamma_1[\gamma_2(z), z, y], \gamma_1[\gamma_2(z), z, y] \rangle_u \} < \infty \forall \gamma_2 \in \Gamma_2\}$$

and

$$(19) \quad \Gamma_1^s = \{\text{measurable } \gamma_1: \mathbb{R}^m \times \mathbb{R}^p \rightarrow U, \text{ such that } E_{z,y} \{ \langle \gamma_1(z, y), \gamma_1(z, y) \rangle_u \} < \infty\}.$$

Furthermore, we redefine J_i and \bar{J}_i as

$$(20) \quad J_i(z, \gamma_1, \gamma_2) = E_{x,y|z} \{L_i(x, u, v) | u = \gamma_1(v, z, y), v = \gamma_2(z)\},$$

$$(21) \quad \bar{J}_i(\gamma_1, \gamma_2) = E_z \{J_i(z, \gamma_1, \gamma_2)\},$$

and let $\{u_{z,y}^t = \gamma_1^t(z, y), v_z^t = \gamma_2^t(z)\}$ denote a pair in $\Gamma_1^s \times \Gamma_2$ that (globally) minimizes \bar{J}_1 . For each fixed $z \in \mathbb{R}^m$, we let

$$(22) \quad B(z) = \{\text{measurable } \beta_z: \mathbb{R}^p \rightarrow U, \text{ so that } E_{y|z} \{ \langle \beta_z(y), \beta_z(y) \rangle_u \} < \infty\},$$

and note that to each $\gamma_1 \in \Gamma_1^s$ and for fixed $z \in \mathbb{R}^m$ there will correspond a unique $\beta_z \in B(z)$ such that $\gamma_1(z, \cdot) = \beta_z(\cdot)$.

Now utilizing the statistical independence of z and y , let us introduce as a counterpart (7), and for each $z \in \mathbb{R}^m$, the set

$$(23) \quad \Omega_t(z) = \{(\beta, v) \in B(z) \times V | \tilde{J}_2(z, \beta, v) \leq \tilde{J}_2(z, \beta_z^t, v_z^t)\},$$

where

$$(24) \quad \tilde{J}_2(z, \beta, v) = E_{x,y|z} \{L_2(x, \beta(y), v) | z\},$$

and β_z^t is the restriction of $\gamma_1^t \in \Gamma_1^s$ to $B(z)$. It is worth to note, at this point, that

$$(25) \quad E_z \{ \tilde{J}_2(z, \beta_z^t, v_z^t) \} = E_z \left\{ E_{x,y|z} \{L_2(x, \beta_z^t(y), \gamma_2^t(z)) | z\} \right\} \\ = E_{x,y,z} \{L_2(x, \gamma_1^t(x, \gamma_1^t(z, \gamma), \gamma_2^t(z)))\} \equiv \bar{J}_2(\gamma_1^t, \gamma_2^t).$$

We can now proceed with the derivation of an optimal incentive scheme by following the analysis of § 3, with U replaced by $B(z)$, where the latter can be made a Hilbert space under the inner product

$$\langle \beta_1, \beta_2 \rangle_\beta = E_y \{ \langle \beta_1(y), \beta_2(y) \rangle_u \}, \quad \beta_i \in B(z).$$

It is easy to see that strict convexity of $L_2(x, \cdot, \cdot)$ on $U \times V$ implies strict convexity of $\tilde{J}_2(z, \cdot, \cdot)$ on $B(z) \times V$, for each $z \in \mathbb{R}^m$, and hence assuming that the latter is Fréchet differentiable on $B(z) \times V$, the equation of the hyperplane supporting $\Omega_t(z)$ at (β_z^t, v_z^t) is

$$(26) \quad \langle \nabla_\beta \tilde{J}_2(z, \beta_z^t, v_z^t), \beta - \beta_z^t \rangle_\beta + \langle \nabla_v \tilde{J}_2(z, \beta_z^t, v_z^t), v - v_z^t \rangle_v = 0$$

where $\nabla_{\beta} \tilde{J}_2(z, \beta_z^t, v_z^t) \in B(z)^*$ is the Fréchet derivative of \tilde{J}_2 with respect to β , and evaluated at the point (β_z^t, v_z^t) . Since there is a natural counterpart of Lemma 1 in this framework, validity of the following counterpart of Proposition 1 can readily be established:

PROPOSITION 2. *For the incentive problem formulated in this section, if (i) $\tilde{J}_2(z, \beta, v)$ is Fréchet differentiable on $B(z) \times V$, (ii) for every $z \in \mathbb{R}^m$, $\nabla_{\beta} \tilde{J}_2(z, \beta_z^t, v_z^t) \neq 0$, and (iii) $\nabla_{\beta} \tilde{J}_2(z, \beta_z^t, v_z^t)$ and $\nabla_v \tilde{J}_2(z, \beta_z^t, v_z^t)$ are weakly continuous in z , there exists an optimal incentive strategy $\gamma_1^0(v, z, y)$ for the leader, given by*

$$(27) \quad u_{z,y}^0 = \gamma_1^0(v, z, y) = u_{z,y}^t - Q_z(v - v_z^t)(y),$$

where $Q_z: V \rightarrow B(z)$ is a linear bounded operator which is weakly continuous in z , and whose adjoint satisfies the linear equation

$$(28) \quad Q_z^* \nabla_{\beta} \tilde{J}_2(z, \beta_z^t, v_z^t) = \nabla_v \tilde{J}_2(z, \beta_z^t, v_z^t),$$

which is defined on V^* .

For the special case when L_2 is quadratic, as given by (15), $\tilde{J}_2(z, \beta, v)$ can be written as

$$(29) \quad \begin{aligned} \tilde{J}_2(z, \beta, v) &= E_{x,y|z} \{ \frac{1}{2} \langle \beta(y), A_{11} \beta(y) \rangle_u + \langle \beta(y), A_{12} v \rangle_u \\ &\quad + \frac{1}{2} \langle v, A_{22} v \rangle_v + \langle \beta(y), C_1 x \rangle_u + \langle v, C_2 x \rangle_v \} \\ &= \frac{1}{2} E_y \{ \langle \beta(y), A_{11} \beta(y) \rangle_u \} + \langle \hat{\beta}, A_{12} v \rangle_u \\ &\quad + \frac{1}{2} \langle v, A_{22} v \rangle_v + E_{x,y|z} \{ \langle \beta(y), C_1 x \rangle_u + \langle v, C_2 \hat{x} \rangle_v \} \end{aligned}$$

where

$$\hat{\beta}_z = E_{y|z} \{ \beta_z(y) | z \} = E_y \{ \beta_z(y) \},$$

$$\hat{x}(z) \triangleq E_{x,y|z} \{ x | z \} = E_{x|z} \{ x | z \}.$$

For fixed $z \in \mathbb{R}^m$, the Fréchet (or Gateaux) differential [25] of \tilde{J}_2 with respect to β , and at the point (β_z^t, v_z^t) is

$$\begin{aligned} \delta_{\beta} \tilde{J}_2(z, \beta_z^t, v_z^t; h_z) &= E_y \{ \langle \beta_z^t(y), A_{11} h_z(y) \rangle_u \} \\ &\quad + \langle A_{12} v_z^t, \hat{h}_z \rangle_u + E_{x,y|z} \{ \langle C_1 x, h_z(y) \rangle_u \}, \end{aligned}$$

where $h_z \in B(z)$ is an admissible variation and $\hat{h}_z \triangleq E_y \{ h_z(y) \}$. Since

$$E_{x,y|z} \{ \cdot \} = E_y \left\{ E_{x|y,z} \{ \cdot \} \right\},$$

this expression can be written as

$$\delta_{\beta} \tilde{J}_2(z, \beta_z^t, v_z^t; h_z) = E_y \{ \langle A_{11} \beta_z^t(y) + A_{12} v_z^t + C_1 \hat{x}, h_z(y) \rangle_u \}$$

where

$$\hat{x}(z, y) \triangleq E_{x|y,z} \{ x | y, z \},$$

and it readily follows from this expression that the Fréchet derivative of \tilde{J}_2 with respect to $\beta_z \in B(z)$ is $\langle \cdot, \tilde{\beta}_z \rangle_\beta$ where $\tilde{\beta}_z \in B(z)$ is given by

$$(30) \quad \tilde{\beta}_z = A_{11}\beta_z^t + A_{12}v_z^t + C_1\hat{x}(z, \cdot).$$

The Fréchet derivative with respect to $v \in V$, on the other hand, readily follows from (29) to be $\langle \cdot, \tilde{v}_z \rangle_v$ where $\tilde{v}_z \in V$ is given by

$$(31) \quad \tilde{v}_z = A_{12}^* E_y \{\beta_z^t(y)\} + A_{22}v_z^t + C_2\hat{x}(z).$$

In view of these relations, a possible solution for Q_z , whose adjoint satisfies (28), is

$$(32) \quad Q_z(\cdot) = \frac{\langle \cdot, \tilde{v}_z \rangle_v}{\|\tilde{\beta}_z\|_\beta^2} \tilde{\beta}_z,$$

which follows by following the arguments used in the proof of Corollary 1 in § 3. This then leads to the following corollary (to Proposition 2):

COROLLARY 2. *Let L_2 be given by (15), and $\tilde{\beta}_z$ and \tilde{v}_z be defined by (30) and (31), respectively. If $\gamma_1^t(z, y)$ is weakly continuous in z and y , $\gamma_2^t(z)$ is weakly continuous in z , $\hat{x}(z, y)$ is continuous in z and y , $\hat{x}(z)$ is continuous in z , and, for every $z \in \mathbb{R}^m$, $\tilde{\beta}_z \neq 0$, there exists an optimal incentive strategy for the leader which is affine in v , and weakly continuous in z and y , and is given by*

$$(33) \quad \gamma_1^0(v, z, y) = \gamma_1^t(z, y) - \frac{\tilde{\beta}_z(y)}{E_y \{\|\tilde{\beta}_z(y)\|_\beta^2\}} \langle \tilde{v}_z, v - \gamma_2^t(z) \rangle_v.$$

Remark 2. An important observation that can be made from (33) is that the dynamic part of the leader's optimal policy depends not only on the common information z (about x) but also on the leader's "private" information y .

5. A scalar example. To illustrate Corollary 2, and especially the structural dependence of γ_1^0 on the common and private information (z and y), we consider in this section a structurally simple numerical example. Let $n = m = p = 1$, and $U = V = \mathbb{R}$. Let x , w_1 and w_2 be independent zero-mean Gaussian random variables with variance 1. Define $z = x + w_1$ and $\tilde{y} = x + w_2$, in which case

$$y = \tilde{y} - E[\tilde{y}|z] = \tilde{y} - \frac{1}{2}z.$$

Assume that $\gamma_1^t(z, y)$ and $\gamma_2^t(y)$ are in the structural form (where $\alpha_1, \alpha_2, \alpha_3$ are known scalars)

$$\gamma_1^t(z, y) = \alpha_1 z + \alpha_2 y, \quad \gamma_2^t(z) = \alpha_3 z, \quad \alpha_1, \alpha_2, \alpha_3 \in \mathbb{R},$$

which would constitute a globally minimizing solution to a quadratic cost function for the leader.

Now let L_2 be given as

$$L_2(x, u, v) = \frac{1}{2}(u)^2 + uv + 2(v)^2 + ux + 2vx,$$

which is strictly convex in the pair (u, v) . Then $\tilde{\beta}_z(y)$ and \tilde{v}_z can be computed to be [from (30) and (31), respectively]

$$\tilde{\beta}_z(y) = \alpha_1 z + \alpha_2 y + \alpha_3 z + E[x|z, y] = (\alpha_1 + \alpha_3 + \frac{1}{2})z + (\alpha_2 + \frac{1}{3})y \triangleq \bar{\alpha}_1 z + \bar{\alpha}_2 y,$$

$$\tilde{v}_z = E_{y|z} [\alpha_1 z + \alpha_2 y] + 4\alpha_3 z + 2E[x|z] = (\alpha_1 + 4\alpha_3 + 1)z \triangleq \bar{\alpha}_3 z.$$

Under the parametric restriction $\alpha_2 \neq -\frac{1}{3}$, all the hypotheses of Corollary 2 are satisfied; and since

$$E_{y|z} \{[\tilde{\beta}_z(y)]^2\} = \bar{\alpha}_1^2 z^2 + \frac{3}{2} \bar{\alpha}_2^2,$$

an optimal incentive scheme for the leader is

$$(34) \quad \gamma_1^0(v, z, y) = \alpha_1 z + \alpha_2 y - \frac{(\bar{\alpha}_1 z + \bar{\alpha}_2 y) \bar{\alpha}_3 z}{\bar{\alpha}_1^2 z^2 + 3 \bar{\alpha}_2^2 / 2} [v - \alpha_3 z].$$

Note that this policy is nonlinear in the common information z , and the private information y also enters the gain term multiplying v . We show in Appendix 1, by direct verification, that (34) indeed constitutes an optimal incentive scheme for the leader, forcing the follower to the desired solution $\gamma_2^t(z) = \alpha_3 z$.

An important question that can be raised at this point is whether (34) constitutes a *unique* solution to the problem under consideration within the class of incentive schemes that are affine in v , or equivalently, whether the gain term in (34) solves (28) uniquely. We address this question in the sequel and show that the affine solution is *not* unique.

Towards this end, let us first assume that there is no private information for the leader, and $\gamma_1^t(z) = \alpha_1 z$. Then, an optimal solution can be obtained from Corollary 1 as

$$(35) \quad \gamma_1^0(v, z) = \alpha_1 z - \left(\frac{\bar{\alpha}_3}{\bar{\alpha}_1} \right) [v - \alpha_3 z]$$

which is in fact the unique one in the class of affine policies, and is linear also in the static information z , thus corroborating a result obtained in [16] for linear-quadratic problems with hierarchical decision structure. Now, if an additional information y comes in to the leader, which is statistically independent of the random variable z characterizing the common information, there seems, at the outset, no particular reason for the gain term in (35) to change, since v is measurable only with respect to the sigma field generated by z . Hence, intuitively, one expects the policy

$$(36) \quad \gamma_1^{00}(v, z, y) = \alpha_1 z + \alpha_2 y - \left(\frac{\bar{\alpha}_3}{\bar{\alpha}_1} \right) [v - \alpha_3 z]$$

to constitute an optimal incentive scheme when both z and y are acquired by the leader. This is indeed true, and the validity of this intuitive result has been established in Appendix 1 by showing that

$$\arg \min_v E_{x,y|z} \{L_2(x, \gamma_1^{00}(v, z, y), v) | z\} = \gamma_2^t(z).$$

Hence, the conclusion is that the scalar example of this section admits at least two affine optimal incentive schemes one of which is also linear in the static information (z, y) .

6. Concluding remarks. By adopting a functional analytic framework, we have obtained optimal incentive strategies (for the leader) in a general class of hierarchical two-agent stochastic Stackelberg problems in which the leader has access to the follower's decision, to some common information, and also to some private information. The main conclusion of this analysis is that, under some fairly mild structural restrictions, there exists an optimal incentive policy for the leader, which is affine in the dynamic information and generally nonlinear in the static (common and private) information.

Even though we have used a general Hilbert space setting for the control (decision) spaces, we have assumed the random variables to take values in finite-dimensional spaces. We have chosen this framework in order to display salient features of the derivation without being distracted by the additional technical restrictions that would be required otherwise. However, our results (embodied in Propositions 1 and 2) are valid in a more general framework which allows the random variables to be weak random variables (cf. [24]) defined on (infinite-dimensional) Hilbert spaces, which includes, for example, the case of stochastic processes.

Appendix. In this appendix we show, by direct verification, that both γ_1^0 and γ_1^{00} , given by (34) and (35), respectively, solve the stochastic incentive problem of § 5.

Starting with the functional form

$$u = \alpha_1 z + \alpha_2 y - Q(z, y)[v - \alpha_3 z],$$

which is clearly a dynamic representation of the static policy $\gamma_1^t(z, y)$ at the desired equilibrium (γ_1^t, γ_2^t) , we substitute this into $L_2(x, u, v)$ and take the expected value conditioned on z , with Q being an arbitrary function measurable in z and y . The result is the function

$$\begin{aligned} J(v, z) &= E_{x,y|z} \left\{ \frac{1}{2} [u^t - Q(v - v^t)]^2 + [u^t - Q(v - v^t)](v + x) + 2v^2 + 2vx \mid z \right\} \\ &= \frac{1}{2} E\{Q^2 \mid z\} (v - v^t)^2 - E[u^t Q \mid z] (v - v^t) \\ &\quad + \frac{1}{2} E\{u_t^2 \mid z\} + \alpha_1 z v + \frac{\alpha_1}{2} z^2 + E\{\alpha_2 y x \mid z\} \\ &\quad - E\{Q \mid z\} (v - v^t) v - E\{Qx \mid z\} (v - v^t) + 2v^2 + vx, \end{aligned}$$

where $u^t = \gamma_1^t(z, y)$, $v^t = \gamma_2^t(z)$.

Since $\frac{1}{2} E\{Q^2 \mid z\} - E\{Q \mid z\} + 2 > 0$ a.e. \mathcal{P}_z , $J(v, z)$ is strictly convex in v a.e. \mathcal{P}_z , and hence $v = v^t$ constitutes the unique minimizing solution to J if and only if $\partial J(v^t, z)/\partial v = 0$ a.e. \mathcal{P}_z . This leads to the following equation to be satisfied by $Q(y, z)$:

$$\begin{aligned} (A1) \quad & [\alpha_1 - E\{Q(y, z) \mid z\}(\alpha_3 + \alpha_1) + 4\alpha_3 + 1]z \\ & - \alpha_2 E\{yQ(y, z) \mid z\} - E\{xQ(y, z) \mid z\} = 0. \end{aligned}$$

Let us now consider the following two choices for Q :

- 1) $Q(y, z) = \bar{\alpha}_3 / \bar{\alpha}_1$;
- 2) $Q(y, z) = (\bar{\alpha}_1 z + \bar{\alpha}_2 y) \bar{\alpha}_3 z / [\bar{\alpha}_1^2 z^2 + 3\bar{\alpha}_2^2 / 2]$.

In the former case, (A1) reads

$$\left[\alpha_1 - \frac{\bar{\alpha}_3(\alpha_3 + \alpha_1)}{\bar{\alpha}_1} + 4\alpha_3 + 1 - \frac{\bar{\alpha}_3}{2\bar{\alpha}_1} \right] z = 0,$$

which can easily be shown to be an identity, by making use of the definitions of $\bar{\alpha}_1$ and $\bar{\alpha}_3$. Hence γ_1^{00} given by (35) is indeed an optimal incentive scheme.

In the latter case, (A-1) reads

$$\left[\bar{\alpha}_3 - (\bar{\alpha}_1 - \frac{1}{2}) \frac{\bar{\alpha}_1 \bar{\alpha}_3 z^2}{\bar{\alpha}_1^2 z^2 + 3\bar{\alpha}_2^2/2} \right] z - \frac{3\bar{\alpha}_2 \bar{\alpha}_2 \bar{\alpha}_3 z}{2\bar{\alpha}_1^2 z^2 + 3\bar{\alpha}_2^2} - E[xQ(y, z)|z] = 0$$

$$\Leftrightarrow \frac{\bar{\alpha}_3[\bar{\alpha}_2 z + \bar{\alpha}_1 z^3]}{2\bar{\alpha}_1^2 z^2 + 3\bar{\alpha}_2^2} - \frac{\bar{\alpha}_3 \bar{\alpha}_1 z^3}{2\bar{\alpha}_1^2 z^2 + 3\bar{\alpha}_2^2} - \frac{\bar{\alpha}_2 \bar{\alpha}_3 z}{2\bar{\alpha}_1^2 z^2 + 3\bar{\alpha}_2^2} = 0$$

since $E[x|z] = \frac{1}{2}z$ and $E[xy|z] = \frac{1}{2}$. The latter equation is an identity, thus corroborating the optimality of the incentive strategy γ_1^0 given by (34).

REFERENCES

- [1] H. VON STACKELBERG, *Markform und Gleichgewicht*, Springer, Vienna, 1934; or *The Theory of the Market Economy*, Oxford Univ. Press, Oxford, 1952.
- [2] C. I. CHEN AND J. B. CRUZ, JR., *Stackelberg solution for two-person games with biased information patterns*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 791–798.
- [3] M. SIMAAN AND J. B. CRUZ, JR., *On the Stackelberg strategy in nonzero-sum games*, J. Optim. Theory Appl., 11 (1973), pp. 533–555.
- [4] ———, *Additional aspects of the Stackelberg strategy in nonzero-sum games*, J. Optim. Theory Appl., 11 (1973), pp. 613–626.
- [5] J. B. CRUZ, JR., *Leader-follower strategies for multilevel systems*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 244–255.
- [6] T. BAŞAR AND H. SELBUZ, *Closed-loop Stackelberg strategies with applications in the optimal control of multilevel systems*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 166–179.
- [7] Y. C. HO, P. B. LUH AND R. MURALIDHARAN, *Information structure, Stackelberg games and incentive controllability*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 454–460.
- [8] Y. C. HO, P. B. LUH AND G. J. OLSDER, *A control theoretic view on incentives*, Automatica, 18 (1982), pp. 167–179.
- [9] G. P. PAPAVALLOPOULOS AND J. B. CRUZ, JR., *Nonclassical control problems and Stackelberg games*, IEEE Trans. Automat. Control, 24 (1979), pp. 155–166.
- [10] B. TOLWINSKI, *Closed-loop Stackelberg solution to multi-stage linear-quadratic game*, J. Optim. Theory Appl., 34 (1981), pp. 485–501.
- [11] G. P. PAPAVALLOPOULOS AND J. B. CRUZ, JR., *Sufficient conditions for Stackelberg and Nash strategies with memory*, J. Optim. Theory Appl., 31 (1980), pp. 233–260.
- [12] T. BAŞAR, *A general theory for Stackelberg games with partial state information*, Large Scale Systems, 3 (1982), pp. 47–56. An earlier version appeared in Proc. 4th International Conference on the Analysis and Optimization of Systems, Springer, 1980, New York.
- [13] ———, *Equilibrium strategies in dynamic games with multilevels of hierarchy*, Automatica, 17 (1981), pp. 749–754.
- [14] ———, *Performance bounds for hierarchical systems under partial dynamic information*, J. Optim. Theory Appl., 39 (1983), pp. 67–87.
- [15] Y. P. ZHENG AND T. BAŞAR, *Existence and derivation of optimal affine incentive schemes for Stackelberg games with partial information: A geometric approach*, Int. J. Control, 35 (1982), pp. 997–1011.
- [16] T. BAŞAR, *Hierarchical decision making under uncertainty*, in Dynamic Optimization and Mathematical Economics, P. T. Liu, ed., Plenum, New York, 1980, pp. 205–221.
- [17] ———, *Stochastic multicriteria decision problems with multi levels of hierarchy*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 549–553.
- [18] T. S. CHANG AND Y. C. HO, *Incentive problems: A class of Stackelberg closed-loop dynamic games*, Systems and Control Letters (1981), pp. 16–21.
- [19] T. BAŞAR AND J. B. CRUZ, JR., *Concepts and methods in multiperson coordination and control*, in Optimization and Control of Dynamic Operational Research Models, S. G. Tzafestas, ed., North-Holland, Amsterdam, 1982, Ch. 11, pp. 351–387.
- [20] T. GROVES AND M. LOEB, *Incentives in a divisionalized firm*, Management Sci., 25 (1979), pp. 221–230.
- [21] L. HURWICZ AND L. SHAPIRO, *Incentive structures maximizing residual gain under incomplete information*, Bell J. Economics, 9 (1978), pp. 180–191.

- [22] L. P. JENNERGREN, *On the design of incentives in business firms—A survey of some recent research*, Management Sci., 26 (1980), pp. 180–201.
- [23] M. SALMAN AND J. B. CRUZ, JR., *An incentive model of duopoly with government coordination*, Automatica, 17 (1981), pp. 821–830.
- [24] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer, New York, 1976.
- [25] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
- [26] T. BAŞAR AND G. J. OLSDER, *Dynamic Noncooperative Game Theory*, Academic Press, London, 1982.

A SYSTEMATIC APPROACH TO HIGHER-ORDER NECESSARY CONDITIONS IN OPTIMIZATION THEORY*

DENNIS S. BERNSTEIN†

Abstract. Necessary conditions for an abstract optimization problem are derived under weak assumptions. The presence of a generalized critical direction in these conditions is the basis for deriving necessary conditions of arbitrary order for various concrete problems. Two applications are considered in detail. The first concerns first- and second-order necessary conditions for a constrained optimization problem in an infinite-dimensional vector space where the cost, equality and inequality functions possess differentials of a finite-dimensional one-sided character. The second application concerns first-, second- and third-order necessary conditions for a constrained optimization problem in a Banach space with Fréchet differentiability hypotheses. In both applications normality conditions are not required. Several well-known results are generalized.

Key words. higher-order necessary conditions, normality condition, convex constraint set, mathematical programming

1. Introduction. We consider the following optimization problem (OP). Minimize $\phi_0(e)$ subject to

$$(1.1) \quad e \in E,$$

$$(1.2) \quad \tilde{\phi}(e) \leq 0,$$

$$(1.3) \quad \psi(e) = 0,$$

where: \mathcal{E} is a set, $E \subset \mathcal{E}$, $k \in \{1, 2, \dots\}$, $\psi: \mathcal{E} \rightarrow \mathbb{R}^k$, \mathcal{X}_0 and \mathcal{Z} are topological vector spaces, $\phi_0: \mathcal{E} \rightarrow \mathcal{X}_0$ and $\tilde{\phi}: \mathcal{E} \rightarrow \mathcal{Z}$. Let $Z_0 \subset \mathcal{X}_0$ and $\tilde{Z} \subset \mathcal{Z}$ be closed convex cones with nonempty interior such that $Z_0 \neq \mathcal{X}_0$ and $\tilde{Z} \neq \mathcal{Z}$. For $z, \hat{z} \in \mathcal{X}_0$, $z < \hat{z}$ means $z - \hat{z} \in \text{int } Z_0$ and $z \leq \hat{z}$ means $z - \hat{z} \in Z_0$. Identical notation applies if $z, \hat{z} \in \mathcal{Z}$. The vector spaces are defined over the real field. The element e is feasible if (1.1)–(1.3) are satisfied and a feasible element \bar{e} solves OP if there is no feasible element e such that $\phi_0(e) < \phi_0(\bar{e})$.

The purpose of this paper is to present a systematic approach for deriving higher-order necessary conditions for a solution of OP. By first deriving necessary conditions for OP under weak assumptions, we obtain results for more specialized problems by successively incorporating stronger hypotheses. New results are obtained and several well-known results are generalized.

This work was motivated by two factors. The first of these was the appearance in the literature of second-order necessary conditions with significantly different assumptions. In particular, [35, Thm. 2.3] involves an infinite-dimensional constraint space without a topology, a convex constraint set, directional differentials and a full-range normality assumption. On the other hand, [28, Thm. 6] involves a finite-dimensional Euclidean constraint space, a conical approximation to the constraint set, continuous differentiability and no normality assumption. The second motivating factor was the series of papers [34], [35] and [36] which are based entirely

* Received by the editors November 30, 1981, and in revised form December 20, 1982. This research was performed at the University of Michigan, Ann Arbor, Michigan, Program in Computer, Information and Control Engineering. Support was provided by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under grant AFOSR-77-3158.

† Control Systems Engineering Group, Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Massachusetts 02173.

on [35, Thm. 2.3]. Weak, strong and “hybrid” variations are systematically exploited in these papers to derive a trio of second-order necessary conditions for optimal control. Thus, the usefulness of [35, Thm. 2.3] for optimal control and the possibility of its being generalized were evident.

The approach of the present paper follows in the spirit of [29]. [29, Thm. 3.1] contains the essential features of a wide variety of optimization problems and yields several well-known first-order necessary conditions as special cases. In the present paper, the formulation of [29] is expanded in the Main Theorem (§ 2) to include a generalized critical direction. Although the Main Theorem itself involves no differentiability hypotheses, this extra feature is the key ingredient for obtaining necessary conditions of arbitrary order when differentiability assumptions are present. Specifically, the generalized critical direction Y and accompanying convex set K correspond respectively to the intermediate- and highest-order terms in a power series expansion.

To illustrate the role of Y and K in deriving higher-order necessary conditions, let \mathcal{E} be a subset of \mathbb{R}^n , $\mathcal{X}_0 = \mathbb{R}$, $\mathcal{X} = \mathbb{R}^m$, $\mathcal{Z}_0 = \mathbb{R}_-$ and $\mathcal{Z} = \mathbb{R}^m$, where \mathbb{R}_- denotes the nonpositive real numbers. Also, let $\bar{e} \in E$, $y \in E - \bar{e}$ and define $\Phi = (\phi_0, \phi, \psi)$. Then in the second-order expansion

$$\Phi(\bar{e} + \alpha y + \alpha^2 x) \approx \Phi(\bar{e}) + \alpha \Phi'(\bar{e})y + \frac{\alpha^2}{2} [2\Phi'(\bar{e})x + y^T \Phi''(\bar{e})y],$$

where $x \in E - \bar{e}$ and $\alpha > 0$, Y and K are given by

$$Y = (\phi_0, \tilde{\phi})'(\bar{e})y,$$

$$K = \{\Phi'(\bar{e})x + \frac{1}{2}y^T \Phi''(\bar{e})y : x \in E - \bar{e}\}.$$

Since we seek directions y which have inferior cost and which approximately satisfy (1.2), we require that $Y \in \bar{\mathbb{R}}^{m+1}$. Also, because of (1.3) the term $\psi'(\bar{e})y$ is required to be zero and thus does not appear in the definition of Y . As will be seen, Y leads to the generalized complementary slackness condition $l_\phi(Y) = 0$, where $l = (l_\phi, l_\psi)$ is the Lagrange multiplier, and K yields the Lagrangian condition $l(h) \geq 0$, $h \in K$. Note that first-order necessary conditions are obtained by setting $y = 0$ (and hence $Y = 0$). Third- and higher-order conditions can be obtained in a straightforward manner. For example, we can consider a pair of directions y and \hat{y} in $E - \bar{e}$ to obtain the third-order expansion

$$\begin{aligned} \Phi(\bar{e} + \alpha y + \alpha^2 \hat{y} + \alpha^3 x) \approx & \Phi(\bar{e}) + \alpha \Phi'(\bar{e})y + \frac{\alpha^2}{2} [2\Phi'(\bar{e})\hat{y} + y^T \Phi''(\bar{e})y] \\ & + \frac{\alpha^3}{6} [6\Phi'(\bar{e})x + 6y^T \Phi''(\bar{e})\hat{y} + \Phi'''(\bar{e})(y)^3]. \end{aligned}$$

Now Y and K are given by

$$Y = (\phi_0, \tilde{\phi})'(\bar{e})y + (\phi_0, \tilde{\phi})'(\bar{e})\hat{y} + \frac{1}{2}y^T (\phi_0, \tilde{\phi})''(\bar{e})y,$$

$$K = \{\Phi'(\bar{e})x + y^T \Phi''(\bar{e})\hat{y} + \frac{1}{6}\Phi'''(\bar{e})(y)^3 : x \in E - \bar{e}\}.$$

In the Main Theorem ϕ_0 , $\tilde{\phi}$ and ψ are assumed to possess weak approximation-like properties (see the Main Condition in § 2). As in [10, Thm. 13.1], these properties are stated solely in terms of the elements of the image spaces of ϕ_0 , $\tilde{\phi}$ and ψ , and hence the set E requires neither topological nor algebraic structure. Because of this fundamental setting, the proof of the Main Theorem given in § 3 is quite simple and succinct. Further simplification is obtained by utilizing a refined separation theorem from [19].

By imposing additional structure (but still no differentiability hypotheses), our next result, Theorem 4.1, follows from the Main Theorem. In the hypotheses for Theorem 4.1 (Condition 4.1), an auxiliary vector space is introduced and the cost and constraint functions are assumed to possess n th-order polynomial expansions. Because Condition 4.1 retains much of the generality of the Main Condition, results from the literature involving conical approximations (see, e.g., [7], [28]) can be obtained as corollaries. In our development, Theorem 4.1 serves as a convenient intermediate step to the results of §§ 5 and 6. For example, the polynomial expansions of Condition 4.1 are given concrete realizations in terms of directional differentials in § 5 and in terms of Fréchet derivatives in § 6.

Section 5 begins with a generalization of the Fréchet derivative (the \tilde{F} -derivative) due to Warga (Definition 5.1). This definition allows us to work with one-sided directional differentials when several directions appear simultaneously. Theorems 5.1 and 5.2 contain first- and second-order necessary conditions, respectively. The relationship between these results can be rather complex. Examples are given to illustrate the following points: 1) the first-order conditions of Theorem 5.1 may not follow from the second-order conditions of Theorem 5.2, and 2) Theorem 5.2 may yield necessary conditions that have the appearance of first-order necessary conditions but which are unobtainable from Theorem 5.1. The discussion generalizes and clarifies some remarks made in [28]. Further specialization leads to Theorem 5.4 which generalizes [35, Thm. 2.3]. It is shown that for each critical direction y there exists a multiplier l satisfying both the first-order necessary conditions and an additional second-order condition.

In § 6, E is a subset of Banach space and ϕ_0 , $\tilde{\phi}$ and ψ are assumed to be Fréchet differentiable. OP now closely resembles a nonlinear programming problem (NP). First-, second- and third-order necessary conditions for this problem are given in Theorem 6.1. The third-order conditions are derived directly from Theorem 4.1 to illustrate the usefulness of the Main Theorem in obtaining higher-order necessary conditions. The relationship between Theorem 6.1 and various higher-order necessary conditions in the literature is discussed.

The appendix contains notation and results pertaining to Fréchet and \tilde{F} -derivatives. We state a version of Taylor's theorem for \tilde{F} -derivatives and a result on converting a multivariable expansion into a one-parameter expansion with remainder satisfying a uniform convergence condition. These results allow n th-order generalizations of the results of §§ 5 and 6.

It is important to point out that neither normality assumptions nor constraint qualifications appear in the statements of the necessary conditions. The absence of these hypotheses leads to nonuniqueness of the multiplier l and hence dependence of l on the critical direction y (or critical directions y, \hat{y}, \dots for third- and higher-order conditions). This idea seems to have first appeared (without proof) in [27] and was apparently rediscovered (in a more general version) in [28]. Subsequent related results can be found in [3], [4], [5], [14], [16], [18], [22], [23], [24], [32].

Before continuing, it is convenient to collect here general notation and definitions and some results concerning vector spaces and topological vector spaces. The empty set is denoted by \emptyset . If S_1 and S_2 are sets, then $S_1/S_2 \triangleq \{s \in S_1: s \notin S_2\}$. If A and B are sets, $A_1 \subset A$, $B_1 \subset B$ and $f: A \rightarrow B$ then $f(A_1) \triangleq \{f(a): a \in A_1\}$ and $f^{-1}(B_1) \triangleq \{a \in A: f(a) \in B_1\}$. Let $\mathbb{N} \triangleq \{1, 2, \dots\}$, $\mathbb{R} \triangleq$ real field, $\mathbb{R}_+ \triangleq \{\alpha \in \mathbb{R}: \alpha > 0\}$, $\mathbb{R}_- \triangleq \{\alpha \in \mathbb{R}: \alpha < 0\}$, $\bar{\mathbb{R}}_+ \triangleq \mathbb{R}_+ \cup \{0\}$ and $\bar{\mathbb{R}}_- \triangleq \mathbb{R}_- \cup \{0\}$. The results obtained here do not depend on the choice of norm for \mathbb{R}^m ; for convenience, the norm of $\alpha \triangleq (\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m$ is taken to be $|\alpha| = \sum_{i=1}^m |\alpha_i|$. For $\beta > 0$, $\mathbb{B}^m(\beta) \triangleq \{\alpha \in \mathbb{R}^m: |\alpha| < \beta\}$ and $\mathbb{P}^m(\beta) \triangleq \{\alpha \in \bar{\mathbb{R}}_+^m: |\alpha| < \beta\}$. Define $\Delta^m \triangleq \{\mu \triangleq (\mu_1, \dots, \mu_{m+1}) \in \bar{\mathbb{R}}_+^{m+1}: |\mu| = 1\}$. When not implied by context, the origin of \mathbb{R}^m is denoted by 0_m .

Let \mathcal{V} be a vector space and $V, \hat{V} \subset \mathcal{V}$. Define $\text{co } V \triangleq$ convex hull of V , $\text{cone } V \triangleq \{\alpha v: \alpha > 0, v \in V\}$ and $\text{coco } V \triangleq \text{co cone } V$. V is a cone if $V = \text{cone } V$ and a convex cone if $V = \text{coco } V$. If $v \in V$ and $V = \text{cone } \{v\}$ then V is a ray. Let $\alpha V \triangleq \{\alpha v: v \in V\}$, where $\alpha \in \mathbb{R}$, $V + \hat{V} \triangleq \{v + \hat{v}: v \in V, \hat{v} \in \hat{V}\}$ and $V - \hat{V} \triangleq V + (-1)\hat{V}$. For $v \in \mathcal{V}$, $\text{cone } v \triangleq \text{cone } \{v\}$, $V + v \triangleq V + \{v\}$ and $V - v \triangleq V + (-v)$. If V is a cone then $\alpha V = V$, $\alpha > 0$; if V is a convex cone then $\alpha V + \beta V = V$, $\alpha > 0, \beta > 0$. If V is convex the dimension of V is $\dim V$.

Let \mathcal{V} and \mathcal{W} be vector spaces, $V \subset \mathcal{V}$ and $f: V \rightarrow \mathcal{W}$. The vectors $v_1, \dots, v_m \in \mathcal{V}$ are linearly independent if $\alpha \in \mathbb{R}^m$ and $\sum_{i=1}^m \alpha_i v_i = 0$ imply that $\alpha = 0$; they are affinely independent if $\alpha \in \mathbb{R}^m$, $\sum_{i=1}^m \alpha_i v_i = 0$ and $\sum_{i=1}^m \alpha_i = 0$ imply that $\alpha = 0$. V is an m -simplex if $m \in \{0, 1, 2, \dots\}$ and $V = \text{co } \{v_1, \dots, v_{m+1}\}$, where v_1, \dots, v_{m+1} are affinely independent. The points v_1, \dots, v_{m+1} are the (unique) vertices of V and, for $v \triangleq \sum_{i=1}^{m+1} \mu_i v_i \in V$, where $\mu \in \Delta^m$, the numbers μ_1, \dots, μ_{m+1} are the (unique) barycentric coordinates of v . f is positively homogeneous if V is a cone and $f(\alpha v) = \alpha f(v)$, $\alpha > 0$, $v \in V$; f is affine if V is convex and for every $m \in \mathbb{N}$, $\mu \in \Delta^{m-1}$ and $v_1, \dots, v_m \in V$ it follows that $f(\sum_{i=1}^m \mu_i v_i) = \sum_{i=1}^m \mu_i f(v_i)$. Let W be a cone and for $w_1, w_2 \in \mathcal{W}$ let $w_1 \leq w_2$ denote $w_1 - w_2 \in W$. f is W -convex if V is convex and for every $m \in \mathbb{N}$, $\mu \in \Delta^{m-1}$ and $v_1, \dots, v_m \in V$ it follows that $f(\sum_{i=1}^m \mu_i v_i) \leq \sum_{i=1}^m \mu_i f(v_i)$.

Suppose that $\mathcal{V}, \mathcal{V}_1, \dots, \mathcal{V}_n$ and \mathcal{W} are topological vector spaces and $V \subset \mathcal{V}$. The closure of V is $\text{cl } V$, the interior of V is $\text{int } V$ and the boundary of V is $\text{bd } V \triangleq (\text{cl } V) \setminus (\text{int } V)$. V is solid if $\text{int } V \neq \emptyset$. $\mathcal{B}(\mathcal{V}_1, \dots, \mathcal{V}_n; \mathcal{W})$ denotes the vector space of all continuous multilinear mappings from $\mathcal{V}_1 \times \dots \times \mathcal{V}_n$ into \mathcal{W} . If $\mathcal{V}_1 = \dots = \mathcal{V}_n$ then we write $\mathcal{B}_n(\mathcal{V}; \mathcal{W})$ for $\mathcal{B}(\mathcal{V}_1, \dots, \mathcal{V}_n; \mathcal{W})$. Recall (see, e.g., [11, p. 318]) that if $\mathcal{V}_1, \dots, \mathcal{V}_n$ and \mathcal{W} are Banach spaces with norms $|\cdot|_{\mathcal{V}_1}, \dots, |\cdot|_{\mathcal{V}_n}$ and $|\cdot|_{\mathcal{W}}$, respectively, then $\mathcal{B}(\mathcal{V}_1, \dots, \mathcal{V}_n; \mathcal{W})$ is a Banach space with norm $\|F\| \triangleq \sup \{|F(v_1, \dots, v_n)|_{\mathcal{W}}: v_i \in \mathcal{V}_i, |v_i|_{\mathcal{V}_i} = 1, i = 1, \dots, n\}$. Define the dual space $\mathcal{V}^* \triangleq \mathcal{B}(\mathcal{V}; \mathbb{R})$ and the conjugate cone $V^\ominus \triangleq \{l \in \mathcal{V}^*: l(v) \leq 0, v \in V\}$.

If \mathcal{V}_1 and \mathcal{V}_2 are topological vector spaces, then $\mathcal{V}_1 \times \mathcal{V}_2$ is assumed to be the topological vector space possessing the product topology. $(\mathcal{V}_1 \times \mathcal{V}_2)^*$ and $\mathcal{V}_1^* \times \mathcal{V}_2^*$ are in one-to-one correspondence in the sense that $l \in (\mathcal{V}_1 \times \mathcal{V}_2)^*$ if and only if there exist $l_1 \in \mathcal{V}_1^*$ and $l_2 \in \mathcal{V}_2^*$ such that $l(v) = l_1(v_1) + l_2(v_2)$, $v \triangleq (v_1, v_2) \in \mathcal{V}_1 \times \mathcal{V}_2$. Specifically, $l_1(v_1) \triangleq l(v_1, 0)$, $v_1 \in \mathcal{V}_1$, and $l_2(v_2) \triangleq l(0, v_2)$, $v_2 \in \mathcal{V}_2$. We denote this correspondence by $l = (l_1, l_2)$. If $\Omega \subset (\mathcal{V}_1 \times \mathcal{V}_2)^*$ and $\hat{\Omega} \subset \mathcal{V}_1^* \times \mathcal{V}_2^*$ then the relations $\Omega \subset \hat{\Omega}$, $\Omega = \hat{\Omega}$, etc., can be interpreted in this sense. Analogous remarks apply to $(\mathcal{V}_1 \times \dots \times \mathcal{V}_n)^*$ and $\mathcal{V}_1^* \times \dots \times \mathcal{V}_n^*$, where $\mathcal{V}_1, \dots, \mathcal{V}_n$ are topological vector spaces.

For the following results let \mathcal{X} and \mathcal{W} be vector spaces, $M \subset \mathcal{X}$ and $f: M \rightarrow \mathcal{W}$.

PROPOSITION 1.1. *Assume M is a convex cone, f is positively homogeneous and affine and $x_1, \dots, x_m \in M$. If $f(x_1), \dots, f(x_m)$ are linearly independent then x_1, \dots, x_m are linearly independent.*

PROPOSITION 1.2. *Assume M is convex, f is affine and $x_1, \dots, x_m \in M$. If $f(x_1), \dots, f(x_m)$ are affinely independent then x_1, \dots, x_m are affinely independent.*

Now let $\mathcal{V}, \mathcal{V}_1$ and \mathcal{V}_2 be topological vector spaces, $V, \hat{V} \subset \mathcal{V}$, $V_1 \subset \mathcal{V}_1$ and $V_2 \subset \mathcal{V}_2$.

PROPOSITION 1.3. *If V is an m -simplex with vertices v_1, \dots, v_{m+1} , then the map $\mu \rightarrow \sum_{i=1}^{m+1} \mu_i v_i: \Delta^m \rightarrow V$ is a homeomorphism.*

PROPOSITION 1.4. *If V is solid and convex, then $\text{cl int } V = \text{cl } V$ and $\text{int cl } V = \text{int } V$.*

PROPOSITION 1.5. $V^\ominus = (\text{cl } V)^\ominus = (\text{co } V)^\ominus = (\text{cone } V)^\ominus$.

PROPOSITION 1.6. *If V is solid and convex, then $V^\ominus = (\text{int } V)^\ominus$.*

PROPOSITION 1.7. *If V is a solid closed convex cone, then $V + \text{int } V = \text{int } V$.*

PROPOSITION 1.8. *If V is open, $l \in V^\ominus$ and $0 \in l(V)$, then $l = 0$.*

PROPOSITION 1.9. $V^\ominus \cap \hat{V}^\ominus = (V + \text{cone } \hat{V})^\ominus$.

PROPOSITION 1.10. $V_1^\ominus \times V_2^\ominus = (V_1 \times \text{cone } V_2)^\ominus$.

Proofs. Propositions 1.1 and 1.2 follow from the definitions of linear and affine independence. Propositions 1.3 and 1.4 can be found in [30, pp. 25, 27] and [17, p. 59], respectively. Proposition 1.5 follows from the linearity and continuity of the elements of V^\ominus . Using Propositions 1.4 and 1.5 we have $V^\ominus = (\text{cl } V)^\ominus = (\text{cl int } V)^\ominus = (\text{int } V)^\ominus$, which proves Proposition 1.6. Proposition 1.7 is a consequence of [30, p. 28], and Proposition 1.4. Proposition 1.8 follows from [30, p. 34], Proposition 1.5 and the fact that cone V is open. To prove Proposition 1.9, let $l \in (V + \text{cone } \hat{V})^\ominus$ and suppose $l \notin V^\ominus$, i.e., there exists $v_1 \in V$ such that $l(v_1) > 0$. For each $\hat{v} \in \hat{V}$ there exists $\alpha > 0$ (sufficiently small) such that $l(v_1 + \alpha \hat{v}) > 0$, which is a contradiction. Now suppose $l \notin \hat{V}^\ominus$, i.e., there exists $v_2 \in \hat{V}$ such that $l(v_2) > 0$. Then, for each $v \in V$ there exists $\alpha > 0$ (sufficiently large) such that $l(v + \alpha v_2) > 0$, which is also a contradiction. The reverse inclusion follows from Proposition 1.5 and the obvious fact $V^\ominus \cap \hat{V}^\ominus \subset (V + \hat{V})^\ominus$. Identical arguments can be used to prove Proposition 1.10. \square

We conclude this section with some comments about the orderings \leq and $<$ on \mathcal{X}_0 (similar remarks apply to \mathcal{Z}). Clearly, \leq is reflexive and both orderings are transitive. The ordering \leq is antisymmetric only when $Z_0 \cap (-Z_0) = \{0\}$, i.e., when Z_0 contains no lines. Because $\text{int } Z_0$ does not contain any lines, the relations $z < \hat{z}$ and $\hat{z} < z$ are never both satisfied.

Note that both orderings are compatible with the linear structure on \mathcal{X}_0 in the sense that if $z \leq \hat{z}$ then $\alpha z \leq \alpha \hat{z}$, $\alpha \geq 0$, and if $z \leq \hat{z}$ and $z' \leq \hat{z}'$ then $z + z' \leq \hat{z} + \hat{z}'$ (and similarly for $<$ with $\alpha = 0$ excluded). Finally, Proposition 1.7 leads to the fact that $z \leq 0$ and $z < 0$ imply $z + \hat{z} < 0$.

2. The Main Theorem. First, we introduce some notation and conventions for OP which simplify the statement of the necessary conditions in this and subsequent sections. Without loss of generality we assume in this section and in §§ 4 and 5 that the solution \bar{e} of OP satisfies $\phi_0(\bar{e}) = 0$. This convention, which simplifies the notation considerably, can be removed by replacing ϕ_0 by $\phi_0 - \phi_0(\bar{e})$ wherever it appears. In order to deal efficiently with various special cases of OP we adopt the convention that $\mathcal{Z} = \emptyset$ and $k = 0$ denote the absence of, respectively, (1.2) and (1.3). For the case $\mathcal{Z} \neq \emptyset$ (i.e., (1.2) is present) define $\mathcal{Z} \triangleq \mathcal{X}_0 \times \mathcal{Z}$, $Z \triangleq Z_0 \times \hat{Z}$ and $\phi \triangleq (\phi_0, \hat{\phi})$; when $\mathcal{Z} = \emptyset$ let $\mathcal{Z} \triangleq \mathcal{X}_0$, $Z \triangleq Z_0$ and $\phi \triangleq \phi_0$. For $k \neq 0$ (i.e., (1.3) is present) define $\mathcal{Z}' \triangleq \mathcal{Z} \times \mathbb{R}^k$ and for $k = 0$ let $\mathcal{Z}' \triangleq \mathcal{Z}$. For $h \in \mathcal{Z} \times \mathbb{R}^k$ let $\pi_\phi h \triangleq$ component of h in \mathcal{Z} and $\pi_\phi \triangleq$ component of \mathbb{R}^k . If $H \subset \mathcal{Z} \times \mathbb{R}^k$ then $\pi_\phi H$ and $\pi_\phi H$ are defined in the obvious way.

The assumptions required for the Main Theorem are contained in the Main Condition (MC) which follows. Roughly speaking, this condition involves: a rep-

resentation of E (through a set K and a map Θ) “near” the feasible point \bar{e} , a generalized “critical direction” Y , which is the basis for higher-order necessary conditions, and some differentiability-like conditions ((2.4) and (2.6)) on ϕ_0 , $\tilde{\phi}$ and ψ . It is shown in later sections that MC includes as special cases conditions commonly satisfied in optimization problems. The paper [29] contains closely related assumptions and considerable comment about them.

MAIN CONDITION (MC). There exist a feasible element \bar{e} , $Y \in Z$ and a nonempty convex set $K \subset \mathcal{X}'$ such that the following property is satisfied. (Let σ and η be positive real numbers, γ be a nonzero real number, S be a k -simplex in K , N be a neighborhood of the origin in \mathcal{Z} , $A: \mathcal{X} \rightarrow \mathcal{Z}$ and $\Theta: S \rightarrow E$.) For all N and η and for all S and σ satisfying

$$(2.1) \quad 0 \in \text{int } \pi_\psi S \text{ (omit if } k = 0),$$

$$(2.2) \quad \sigma(\phi(\bar{e}) + Y) + \pi_\phi S \subset \text{int } Z$$

there exist γ , A and Θ satisfying

$$(2.3) \quad A^{-1}(\text{int } Z) \subset \text{int } Z,$$

$$(2.4) \quad A \circ \phi \circ \Theta(h) \in \sigma(\phi(\bar{e}) + Y) + \pi_\phi S + Z + N, \quad h \in S,$$

$$(2.5) \quad \psi \circ \Theta: S \rightarrow \mathbb{R}^k \text{ is continuous (omit if } k = 0),$$

$$(2.6) \quad |\gamma \psi \circ \Theta(h) - \pi_\psi h| < \eta, \quad h \in S \text{ (omit if } k = 0).$$

MAIN THEOREM (MT). Suppose MC is satisfied. If \bar{e} solves OP then there exists $l \triangleq (l_\phi, l_\psi) \in \mathcal{X}^* \times \mathbb{R}^k$ when $k > 0$ and $l \triangleq l_\phi \in \mathcal{X}^*$ when $k = 0$ such that

$$(2.7) \quad l \neq 0,$$

$$(2.8) \quad l_\phi \in Z^\ominus,$$

$$(2.9) \quad l_\phi(\phi(\bar{e}) + Y) = 0,$$

$$(2.10) \quad l(h) \geq 0, \quad h \in K.$$

Remark 2.1. When $k = 0$ the simplex S in MC consists of a single element h . In this case the role of Θ and S can be handled by a single element $e \triangleq \Theta(h) \in E$ and (2.4) can be replaced by

$$(2.4)' \quad A \circ \phi(e) \in \sigma(\phi(\bar{e}) + Y) + h + Z + N.$$

The distinction between (2.4) and (2.4)' accounts for the pair of conditions [29, 3.1 and 3.2].

Remark 2.2. The focus in [29] is on first-order necessary conditions for a very general extremal problem which includes OP as a special case. MC and MT can be reformulated to apply to this problem although this is not pursued here. In the context of OP, MC both weakens and generalizes [29, conditions 3.1 and 3.2]. These conditions involve the introduction of an auxiliary vector space \mathcal{Y} , a convex set $M \subset \mathcal{Y}$ and a map $f: M \rightarrow \mathcal{X}'$. Necessary conditions are then stated in terms of the elements of M . MC is weaker since \mathcal{Y} does not appear. Instead MC involves elements of the set K which corresponds to the image of M under f . A similar idea appears in [10, Thm. 13.1], p. 46, which follows from MT with $\mathcal{X}_0 = \mathbb{R}$, $\tilde{\mathcal{X}} = \emptyset$, $Y = 0$ and K a cone. The term Y , which leads to higher-order necessary conditions, has no counterpart in either [10] or [29].

Remark 2.3. Because $\phi(\bar{e}) \in Z$, $Y \in Z$ and $l_\phi \in Z^\ominus$, it follows that (2.9) is equivalent to the pair of conditions

$$(2.11) \quad l_\phi(\phi(\bar{e})) = 0,$$

$$(2.12) \quad l_\phi(Y) = 0.$$

Note that (2.11) is a complementary slackness condition in that it yields additional information concerning l_ϕ . By virtue of the additional condition (2.12), (2.9) may be viewed as a generalized complementary slackness condition.

Sometimes MC is verified with a qualification of S which is weaker than (2.1) and (2.2). Although this results in a stronger version of MC it may be more suitable for applications. For example, either or both of the conditions (2.1) and (2.2) may be omitted. A common situation is the subject of the following easily proved result.

PROPOSITION 2.1. *Suppose MC is satisfied and*

$$(2.13) \quad \phi(\bar{e}) + Y \in \text{bd } Z.$$

Then (2.1) and (2.2) imply

$$(2.14) \quad \text{The vertices of } S \text{ are linearly independent.}$$

Remark 2.4. Because $\phi_0(\bar{e}) = 0$, (2.13) is satisfied when $Y = 0$. This is the situation in [10, Thm. 13.1]. There (2.1) and (2.2) are omitted and the (weaker) linear independence condition (2.14) appears.

Remark 2.5. The set \mathcal{E} plays no role in the Main Theorem or its proof. In §§ 5 and 6, \mathcal{E} is used to define differentiability properties which cannot be stated solely in terms of the elements of E .

3. Proof of the Main Theorem. The following notation and definitions are needed. Let \mathcal{V} be a vector space and $V \subset \mathcal{V}$. If $V + v$ is a subspace of \mathcal{V} for some $v \in \mathcal{V}$ then V is an affine subset of \mathcal{V} . The codimension of an affine subset V , $\text{codim } V$, is the dimension of a subspace $\tilde{V} \subset \mathcal{V}$ such that the direct sum of \tilde{V} and the subspace $V + v$ is \mathcal{V} . The affine hull of an arbitrary set V , $\text{aff } V$, is the smallest affine subset containing V .

Suppose now that \mathcal{V} is a topological vector space and $V, A, B \subset \mathcal{V}$. The interior of V relative to $\text{aff } V$ is denoted by $\text{ri } V$. $l \in \mathcal{V}^*$ separates A and B if $l \neq 0$ and there exists $\alpha \in \mathbb{R}$ such that $l(a) \leq \alpha \leq l(b)$, $a \in A$, $b \in B$. If either A or B is a cone then α can be chosen to be zero without loss of generality and thus $l \in A^\ominus \cap (-B)^\ominus$.

We will also require the following easily verified results. If \mathcal{V}_1 and \mathcal{V}_2 are vector spaces, $V_1 \subset \mathcal{V}_1$ and $V_2 \subset \mathcal{V}_2$ then $\text{aff}(V_1 \times V_2) = (\text{aff } V_1) \times (\text{aff } V_2)$. If in addition \mathcal{V}_1 and \mathcal{V}_2 are topological vector spaces then $\text{ri}(V_1 \times V_2) = (\text{ri } V_1) \times (\text{ri } V_2)$.

The proof of MT rests upon the following separation lemma. This result is a generalization of a well-known theorem which follows when $\text{int } A \neq \emptyset$ (see, e.g., [7, p. 63], or [31, p. 24]).

LEMMA 3.1. *Let A and B be convex subsets of a topological vector space \mathcal{V} such that $\text{aff } A$ is closed and has finite codimension, $\text{ri } A \neq \emptyset$ and $(\text{ri } A) \cap B = \emptyset$. Then there exists $l \in \mathcal{V}^*$ separating A and B .*

Lemma 3.1 is a corollary of an algebraic separation theorem stated without proof in the survey paper [19, p. 253]. In the algebraic setting \mathcal{V} is assumed to be a vector space and the “intrinsic core” of A plays the role of $\text{ri } A$. The additional assumption that $\text{aff } A$ is closed implies that l is continuous (see [19, pp. 240–1]). A proof of the algebraic separation theorem can be obtained by means of induction on $n = \text{codim aff } A$ using a method similar to that used in [31, proof of Thm. 2.9].

We consider the full proof of the Main Theorem only for the case $\tilde{\mathcal{Z}} \neq \emptyset$ and $k > 0$ since the proofs of the remaining cases involve similar arguments. A large portion of the proof is contained in the following two lemmas. Suppose MC is satisfied and define $\hat{Z} \triangleq (\text{int } Z) - \text{cone}(\phi(\bar{e}) + Y)$. Note that \hat{Z} is an open convex cone.

LEMMA 3.2. *If K and $\hat{Z} \times \{0\}$ are separated by a continuous linear functional then there exists $l = (l_\phi, l_\psi) \in \mathcal{X}^* \times \mathbb{R}^k$ satisfying (2.7)–(2.10).*

Proof. Since $\hat{Z} \times \{0\}$ is a cone there exists $l = (l_\phi, l_\psi) \in \mathcal{X}^* \times \mathbb{R}^k$ satisfying (2.7) and such that $l \in (-K)^\circ \cap (\hat{Z} \times \{0\})^\circ$. Thus l satisfies (2.10). By Proposition 1.10, $l_\phi \in \hat{Z}^\circ$ and, by Proposition 1.9, $l_\phi \in (\text{int } Z)^\circ$ and $l_\phi(\phi(\bar{e}) + Y) \geq 0$. By Proposition 1.6 $(\text{int } Z)^\circ = Z^\circ$ and thus (2.8) holds. Since $\phi(\bar{e}) + Y \in Z$, $l_\phi(\phi(\bar{e}) + Y) \leq 0$, which implies (2.9). \square

LEMMA 3.3. *If K and $\hat{Z} \times \{0\}$ are not separated by a continuous linear functional then there exist a k -simplex $S \subset K$ and a positive real number σ satisfying (2.1) and (2.2).*

Proof. We first show that $0 \in \text{int } \pi_\psi K$. If this is not true, then it follows (see, e.g., [30, Thm. I.5.19] that there exists $\xi \in \mathbb{R}^k$, $\xi \neq 0$, such that $\xi \cdot v \geq 0$, $v \in \pi_\psi K$. Then $l = (0, \xi) \in \mathcal{X}^* \times \mathbb{R}^k$ separates K and $\hat{Z} \times \{0\}$, which is a contradiction. Since $0 \in \text{int } \pi_\psi K$ there exists a k -simplex $U \triangleq \text{co}\{u_1, \dots, u_{k+1}\} \subset \pi_\psi K$ such that $0 \in \text{int } U$. For each $i \in \{1, \dots, k+1\}$ let $s_i \in K$ satisfy $\pi_\psi s_i = u_i$.

Since $\text{aff}(\hat{Z} \times \{0\}) = (\text{aff } \hat{Z}) \times (\text{aff } \{0\}) = \mathcal{X} \times \{0\}$ is closed and has finite codimension k and since $\text{ri}(\hat{Z} \times \{0\}) = (\text{ri } \hat{Z}) \times (\text{ri } \{0\}) = \hat{Z} \times \{0\} \neq \emptyset$, Lemma 3.1 implies that $(\hat{Z} \times \{0\}) \cap K \neq \emptyset$. Thus, there exists $s \triangleq (\hat{z} - \sigma'(\phi(\bar{e}) + Y), 0) \in K$, where $\hat{z} \in \text{int } Z$ and $\sigma' > 0$. Since $\hat{z} \in \text{int } Z$ we can choose $\lambda \in (0, 1)$ sufficiently close to 1 so that $\lambda \hat{z} + (1-\lambda)\pi_\phi s_i \in \text{int } Z$, $i \in \{1, \dots, k+1\}$. Define $h_i \triangleq \lambda s + (1-\lambda)s_i$, $i \in \{1, \dots, k+1\}$, and $S \triangleq \text{co}\{h_1, \dots, h_{k+1}\}$. Letting $\sigma \triangleq \lambda \sigma'$, and $\tilde{z} \triangleq \sigma(\phi(\bar{e}) + Y)$ it is easy to see that $h_i = (\lambda \hat{z} + (1-\lambda)\pi_\phi s_i - \tilde{z}, (1-\lambda)u_i) \in K$, $S \subset K$ and $\pi_\phi h_i \in (\text{int } Z) - \tilde{z}$. Since $\pi_\psi S = (1-\lambda)U$ and $0 \in \text{int } U$, (2.1) must hold. Since u_1, \dots, u_{k+1} are affinely independent and $(1-\lambda)^{-1}\pi_\psi h_i = u_i$ it follows from Proposition 1.2 that h_1, \dots, h_{k+1} are affinely independent and hence S is a k -simplex. Finally, since $\tilde{z} + \pi_\phi h_i \in \text{int } Z$, $i \in \{1, \dots, k+1\}$, (2.2) is satisfied. \square

We can now proceed with the proof of MT. Suppose that the theorem is false. By Lemmas 3.2 and 3.3 there exist a k -simplex $S = \text{co}\{h_1, \dots, h_{k+1}\} \subset K$ and a positive real number σ satisfying (2.1) and (2.2). Since $\sigma(\phi(\bar{e}) + Y) + \pi_\phi h_i \in \text{int } Z$, $i \in \{1, \dots, k+1\}$, the open set $N \triangleq \bigcap_{i=1}^{k+1} ((\text{int } Z) - \sigma(\phi(\bar{e}) + Y) - \pi_\phi h_i)$ is a neighborhood of the origin in \mathcal{X} . It follows easily that $\sigma(\phi(\bar{e}) + Y) + \pi_\phi S + N \subset \text{int } Z$. Because of (2.1) we can choose $\eta > 0$ so that $\mathbb{B}^k(\eta) \subset \pi_\psi S$. For S , σ , N and η thus defined there exist γ , A and Θ with the properties specified in MC.

From (2.3), (2.4), Proposition 1.7 and the choice of N we have

$$\begin{aligned} \phi \circ \Theta(S) &\subset A^{-1}(\sigma(\phi(\bar{e}) + Y) + \pi_\phi S + Z + N) \\ &\subset A^{-1}((\text{int } Z) + Z) = A^{-1}(\text{int } Z) \subset \text{int } Z. \end{aligned}$$

This implies that, for all $h \in S$, $\phi_0 \circ \Theta(h) < 0$ and $\tilde{\phi} \circ \Theta(h) < 0$.

Let $\tilde{\pi}_\psi: S \rightarrow \pi_\psi S$ be defined by $\tilde{\pi}_\psi(h) = \pi_\psi h$. Since S and $\pi_\psi S$ are k -simplexes, $\tilde{\pi}_\psi$ associates points with the same barycentric coordinates. Proposition 1.3 implies that $\tilde{\pi}_\psi^{-1}: \pi_\psi S \rightarrow S$ is continuous. Define $G: \pi_\psi S \rightarrow \mathbb{R}^k$ by $G(u) = -\gamma \psi \circ \Theta \circ \tilde{\pi}_\psi^{-1}(u) + u$. From (2.5) it follows that G is continuous and, from (2.6) and the choice of η , $G: \pi_\psi S \rightarrow \pi_\psi S$. Since $\pi_\psi S$ is compact and convex, the Brouwer fixed point theorem implies that there exists $u^* \in \pi_\psi S$ such that $G(u^*) = u^*$. Since $\gamma \neq 0$ it follows that $\psi(e^*) = 0$, where $e^* \triangleq \Theta(h^*) \in E$ and $h^* \triangleq \tilde{\pi}_\psi^{-1}(u^*) \in S$. Since e^* also satisfies $\phi_0(e^*) < 0$ and $\tilde{\phi}(e^*) < 0$, \bar{e} does not solve OP, which is a contradiction.

The proof for $\tilde{\mathcal{Z}} = \emptyset$ is almost identical to the above proof. It is only necessary to delete the details pertaining to $\tilde{\phi}$. In the case $k = 0$ the proof is considerably shorter. $\tilde{\mathcal{Z}}$ plays the role of $\tilde{\mathcal{Z}} \times \{0\}$ in Lemma 3.2 and, in Lemma 3.3, the k -simplex S is a singleton. The remainder of the proof proceeds along the same lines as the above proof except now the development relating to ψ is not needed.

4. A specialization of the Main Theorem. In this section we specialize the Main Theorem to obtain Theorem 4.1. For our purposes here, Theorem 4.1 may be regarded as a convenient intermediate step to the results of §§ 5 and 6. It may have independent interest, however, in other applications.

Theorem 4.1 is obtained by imposing additional structure on MT in three ways. First, we introduce an auxiliary vector space as discussed in Remark 2.2. This compensates for the lack of assumptions on \mathcal{E} . Second, we assume that the cost and constraint functions possess one-parameter polynomial expansions. These expansions may be regarded as a primitive form of the variational and power series required for the derivation of higher-order necessary conditions in §§ 5 and 6. Third, we impose additional structure on the inequality constraint (1.2). Specifically, we assume that there exists $j \in \mathbb{N}$ and, for each $i \in \{1, \dots, j\}$, there exist a topological vector space \mathcal{Z}_i , a mapping $\phi_i: \mathcal{E} \rightarrow \mathcal{Z}_i$ and a solid closed convex cone $Z_i \subset \mathcal{Z}_i$ not equal to \mathcal{Z}_i such that $\tilde{\mathcal{Z}} = \mathcal{Z}_1 \times \dots \times \mathcal{Z}_j$, $\tilde{\phi} = (\phi_1, \dots, \phi_j)$ and $\tilde{\mathcal{Z}} = Z_1 \times \dots \times Z_j$. For $i \in \{1, \dots, j\}$ and $z, \hat{z} \in \mathcal{Z}_i$ we define $z \leq \hat{z}$ and $z < \hat{z}$ in the obvious way. Now (1.2) becomes

$$(1.2)' \quad \phi_i(e) \leq 0, \quad i \in \{1, \dots, j\}.$$

As discussed below, writing (1.2) in the form (1.2)' extends the applicability of the Main Theorem.

Condition 4.1, which is used in Theorem 4.1, incorporates the above aspects. This rather complex condition may be motivated by the following comments. Roughly speaking, we assume that ϕ_0, \dots, ϕ_j have expansions of the form

$$(4.1) \quad \phi_i(e) = \phi_i(\bar{e}) + \sum_{r=1}^{m_i} \alpha^r Y_{ir} + o(\alpha^{m_i}),$$

where $m_i \in \mathbb{N} \cup \{0\}$, $Y_{ir} \in Z_i$ for all $r \in \{1, \dots, m_i\}$, $\alpha > 0$ and $\alpha^{-m_i} o(\alpha^{m_i}) \rightarrow 0$ as $\alpha \rightarrow 0^+$. In the proof of Theorem 4.1 it turns out that if for all $\alpha > 0$ sufficiently small

$$(4.2) \quad \phi_i(\bar{e}) + \sum_{r=1}^{m_i} \alpha^r Y_{ir} < 0,$$

then the term $o(\alpha^{m_i})$ plays no role in the higher-order necessary conditions. This is advantageous if either: (1) ϕ_i possesses an expansion of order m_i and not of order $m_i + 1$, or (2) the term Y_{i, m_i+1} is not an element of Z_i . The form of the constraint (1.2)' is used so that the order m_i of the expansion can depend on i .

By adding $\sum_{r=1}^{m_i} (1 - \alpha^r) Y_{ir}$ to the left side of (4.2) (assuming $\alpha < 1$) and using Proposition 1.7 it follows that (4.2) implies

$$(4.3) \quad \phi_i(\bar{e}) + \sum_{r=1}^{m_i} Y_{ir} < 0.$$

Similarly, it can be seen that the reverse implication is true. The equivalence of (4.2) and (4.3) accounts for the appearance of (4.3) in Condition 4.1 (via the set I' defined below). In the proof of Theorem 4.1 the i th component of Y in MT is given by $\sum_{r=1}^{m_i} Y_{ir}$.

Additional aspects of Condition 4.1 are: the “auxiliary” vector space \mathcal{X} , a convex set of “variations” $M \subset \mathcal{X}$, a mapping $V: M \rightarrow \mathcal{Z}'$ which provides a representation for K in MC (specifically, $K = \text{co } V(M)$) and, for each k -simplex $X \subset M$, a mapping $\zeta: X \rightarrow E$ which corresponds to Θ in MC.

In what follows we use the conventions $\{a_i\}_{i=1}^0 = \emptyset$, $\sum_{i=1}^0 a_i = 0$ and let $j = 0$ when the inequality constraints (1.2)' are absent. As in § 2 we assume for convenience that $\phi_0(\bar{e}) = 0$. For $\bar{e} \in E$ define the index sets $I' \triangleq \{i \in \{0, \dots, j\}: \phi_i(\bar{e}) + \sum_{r=1}^{m_i} Y_{ir} < 0\}$ and $I'' \triangleq \{0, \dots, j\}/I'$. Finally, let $l \in \mathcal{Z}'^*$ denote $l = (l_0, \dots, l_j, l_\psi) \in \mathcal{Z}_0^* \times \dots \times \mathcal{Z}_j^* \times \mathbb{R}^k$ when $k > 0$ and $l = (l_0, \dots, l_j) \in \mathcal{Z}_0^* \times \dots \times \mathcal{Z}_j^*$ when $k = 0$.

Condition 4.1. There exist a feasible element \bar{e} , a vector space \mathcal{X} , a nonempty convex set $M \subset \mathcal{X}$, $n \in \mathbb{N}$, $m_i \in \{0, \dots, n-1\}$ and $\{Y_{ir}\}_{r=1}^{m_i} \subset Z_i$ for all $i \in \{0, \dots, j\}$, and $V \triangleq (V_0, \dots, V_j, V_\psi): M \rightarrow \mathcal{Z}'$ (omit V_ψ when $k = 0$) such that $V_i: M \rightarrow \mathcal{Z}_i$ is Z_i -convex for all $i \in I''$, $V_\psi: M \rightarrow \mathbb{R}^k$ is affine and such that the following property is satisfied. (Let $X = \text{co } \{x_1, \dots, x_{k+1}\} \subset M$ be a k -simplex, N_i be a neighborhood of the origin in \mathcal{Z}_i for all $i \in \{0, \dots, j\}$, $\eta > 0$, $\tau > 0$, $\alpha \in (0, \tau)$ and $\zeta: X \rightarrow E$.) For all X, N_0, \dots, N_j , η and τ there exist α and ζ satisfying

$$(4.4) \quad \alpha^{-m_i} \left[\phi_i \circ \zeta(X) - \phi_i(\bar{e}) - \sum_{r=1}^{m_i} \alpha^r Y_{ir} \right] \subset Z_i + N_i, \quad i \in I',$$

$$(4.5) \quad \alpha^{-(m_i+1)} \left[\phi_i \circ \zeta(X) - \phi_i(\bar{e}) - \sum_{r=1}^{m_i} \alpha^r Y_{ir} \right] \subset (\text{co } V_i(X)) + Z_i + N_i, \quad i \in I'',$$

$$(4.6) \quad \mu \rightarrow \psi \circ \zeta \left(\sum_{i=1}^{k+1} \mu_i x_i \right): \Delta^k \rightarrow \mathbb{R}^k \text{ is continuous (omit if } k = 0),$$

$$(4.7) \quad |\alpha^{-n} \psi \circ \zeta(x) - V_\psi(x)| < \eta, \quad x \in X \text{ (omit if } k = 0).$$

THEOREM 4.1. Suppose Condition 4.1 is satisfied. If \bar{e} solves OP then there exists $l \in \mathcal{Z}'^*$ such that

$$(4.8) \quad l \neq 0,$$

$$(4.9) \quad l_i \in Z_i^\ominus, \quad i \in \{0, \dots, j\},$$

$$(4.10) \quad l_i = 0, \quad i \in I',$$

$$(4.11) \quad l_i \left(\phi_i(\bar{e}) + \sum_{r=1}^{m_i} Y_{ir} \right) = 0, \quad i \in I'',$$

$$(4.12) \quad l(V(x)) \geq 0, \quad x \in M.$$

Remark 4.1. The additional structure of (1.2)' accounts for the form of (4.10) and (4.11). Condition (4.10) generalizes the “complementary slackness” condition pertaining to the “inactive” inequality constraints ($\phi_i(\bar{e}) < 0$) in first-order necessary conditions. Note that $\phi_i(\bar{e}) + \sum_{r=1}^{m_i} Y_{ir} \neq 0$ for $i \in I''$ is possible when $\dim Z_i > 1$. In this way (4.11) may yield information regarding l_i .

Remark 4.2. By setting $n = 1$ in Condition 4.1 it is easy to see that Theorem 4.1 implies [29, Thm. 3.1] when this result is specialized to OP. Note that in Condition 4.1, X is a k -simplex whereas in [29, cond. 3.1], X is an i -simplex, where $i \in \{1, \dots, k\}$. Furthermore, because inactive inequality constraints require only a trivial expansion of the form (4.4) (since $m_i = 0$), (4.4) and (4.5) allow a more general treatment of the inequality constraints than is possible by [29, (3.2)].

Proof of Theorem 4.1. We give the details for the case $k > 0$ only since the proof for $k = 0$ is similar. Since the functions V_i , $i \in I'$, play no role in Condition 4.1 or the results of the theorem, we define $V_i(x) = 0$, $x \in M$, $i \in I'$. The main idea of the proof is to show that MC holds with

$$(4.13) \quad Y = \left(\sum_{r=1}^{m_0} Y_{0r}, \dots, \sum_{r=1}^{m_j} Y_{jr} \right)$$

and

$$(4.14) \quad K = \text{co } V(M).$$

To show that (2.7)–(2.10) imply (4.8)–(4.12), note that (4.9) is merely a rewriting of (2.8) and that (4.11) and (4.12) follow from (2.9) and (2.10), respectively. Also, (4.10) follows from (2.9), (4.9) and Propositions 1.6 and 1.8.

We now show that Condition 4.1 implies MC. Let $S = \text{co } \{h_1, \dots, h_{k+1}\}$, σ , N and η have the properties specified in MC. Since $h_i \in K$, $i \in \{1, \dots, k+1\}$, and $K = \text{co } V(M)$, we have $h_i = \sum_{r=1}^{\nu_i} \mu_{ir} V(x_{ir})$, where $\nu_i \in \mathbb{N}$, $(\mu_{i1}, \dots, \mu_{i,\nu_i}) \in \Delta^{\nu_i-1}$ and $x_{i1}, \dots, x_{i,\nu_i} \in M$. Define $x_i \triangleq \sum_{r=1}^{\nu_i} \mu_{ir} x_{ir}$, $i \in \{1, \dots, k+1\}$, and $X \triangleq \text{co } \{x_1, \dots, x_{k+1}\}$. Since V_ψ is affine,

$$(4.15) \quad V_\psi(x_i) = \pi_\psi h_i, \quad i \in \{1, \dots, k+1\}.$$

From (2.1) it follows that $\pi_\psi h_1, \dots, \pi_\psi h_{k+1}$ are affinely independent. Proposition 1.2 and (4.15) thus imply that X is a k -simplex. For each $i \in \{0, \dots, j\}$ choose N_i to be a neighborhood of the origin in \mathcal{Z}_i so that

$$(4.16) \quad \sigma(N_0 \times \dots \times N_j) \subset N.$$

Also, let $\tau \triangleq \sigma^{-1}$.

With X , N_0, \dots, N_j , ζ and τ now specified Condition 4.1 implies that there exist $\alpha \in (0, \tau)$ and $\zeta: X \rightarrow E$ satisfying (4.4)–(4.7). We now exhibit γ , A and Θ so that (2.3)–(2.6) are satisfied.

Let $\Theta: S \rightarrow E$ be defined by $\Theta = f \circ g$ where $g: S \rightarrow \Delta^k$, $f: \Delta^k \rightarrow E$, $g(\sum_{i=1}^{k+1} \mu_i h_i) \triangleq \mu$ and $f(\mu) \triangleq \zeta(\sum_{i=1}^{k+1} \mu_i x_i)$. From (4.6) it follows that $\psi \circ f: \Delta^k \rightarrow \mathbb{R}^k$ is continuous. Since, by Proposition 1.3, g is continuous, $\psi \circ \Theta = \psi \circ f \circ g: S \rightarrow \mathbb{R}^k$ is continuous. This proves (2.5). For use below note that

$$(4.17) \quad \Theta \left(\sum_{i=1}^{k+1} \mu_i h_i \right) = \zeta \left(\sum_{i=1}^{k+1} \mu_i x_i \right), \quad \mu \in \Delta^k,$$

$$(4.18) \quad \Theta(S) = \zeta(X).$$

For $h \in Z_0 \times \dots \times Z_j \times \mathbb{R}^k$, $H \subset Z_0 \times \dots \times Z_j \times \mathbb{R}^k$ and $i \in \{0, \dots, j\}$ let $\pi_i h \triangleq$ component of h in Z_i and $\pi_i H \triangleq \{\pi_i h: h \in H\}$.

Since V_i is Z_i -convex, $i \in I''$, we have

$$(4.19) \quad V_i(x_p) \leq \sum_{r=1}^{\nu_p} \mu_{pr} V_i(x_{pr}) = \pi_i h_p, \quad p \in \{1, \dots, k+1\},$$

which implies that, for all $\mu \in \Delta^k$,

$$(4.20) \quad V_i \left(\sum_{p=1}^{k+1} \mu_p x_p \right) \leq \sum_{p=1}^{k+1} \mu_p V_i(x_p) \leq \sum_{p=1}^{k+1} \mu_p \pi_i h_p \in \pi_i S.$$

Consequently,

$$(4.21) \quad \text{co } V_i(X) \subset \pi_i S + Z_i, \quad i \in I''.$$

To simplify what follows we assume without loss of generality that $\sigma > 1$. To see that this is possible note that if $\sigma \leq 1$ then we can add $\rho(\phi(\bar{e}) + Y)$, where $\rho > 1 - \sigma$, to each side of (2.2). Since $\rho(\phi(\bar{e}) + Y) \in Z$, Proposition 1.7 can be used to replace $(\text{int } Z) + \rho(\phi(\bar{e}) + Y)$ by $\text{int } Z$. Then $\sigma + \rho$ can be relabelled as σ to yield the desired result. Note that $\sigma > 1$ implies $\alpha < 1$.

Since Z_i is a convex cone it follows that, for all $i \in \{0, \dots, j\}$,

$$(4.22) \quad \alpha_0 \phi_i(\bar{e}) + \sum_{r=1}^{m_i} \alpha_i Y_{ir} \in Z_i, \quad \alpha_0, \dots, \alpha_{m_i} > 0.$$

From (4.13) it follows that

$$(4.23) \quad \pi_i Y = \sum_{r=1}^{m_i} Y_{ir}, \quad i \in \{0, \dots, j\},$$

and, since $V_i(x) = 0$, $i \in I'$, we have

$$(4.24) \quad \pi_i S = \{0\}, \quad i \in I'.$$

Since $\sigma \alpha^{-r} \geq \sigma$, $r \in \{0, 1, \dots\}$, (4.4), (4.18), (4.22), (4.23) and (4.24) imply

$$(4.25) \quad \sigma \alpha^{-m_i} \phi_i \circ \Theta(S) \subset \sigma(\phi_i(\bar{e}) + \pi_i Y) + \pi_i S + Z_i + \sigma N_i, \quad i \in I'.$$

Since $\alpha^{-r} > \sigma$, $r \in \mathbb{N}$, (4.5), (4.18), (4.21), (4.22) and (4.23) imply

$$(4.26) \quad \alpha^{-(m_i+1)} \phi_i \circ \Theta(S) \subset \sigma(\phi_i(\bar{e}) + \pi_i Y) + \pi_i S + Z_i + N_i, \quad i \in I''.$$

For $z \triangleq (z_0, \dots, z_j) \in Z$ define $A(z) \triangleq (a_0 z_0, \dots, a_j z_j)$, where $a_i \triangleq \sigma \alpha^{-m_i}$, $i \in I'$, and $a_i \triangleq \alpha^{-(m_i+1)}$, $i \in I''$, which satisfies (2.3). Condition (2.4) follows from (4.16), (4.25), (4.26) and the fact that $N_i \subset \sigma N_i$. Since V_ψ is affine, (4.15) implies

$$(4.27) \quad V_\psi\left(\sum_{i=1}^{k+1} \mu_i x_i\right) = \pi_\psi \sum_{i=1}^{k+1} \mu_i h_i, \quad \mu \in \Delta^k.$$

Letting $\gamma \triangleq \alpha^{-n}$, (4.7), (4.17) and (4.27) imply (2.6), which completes the proof. \square

Remark 4.3. If $k > 0$ and $\dim M < k$ then M does not contain a k -simplex and Condition 4.1 holds trivially. In this case the necessary conditions (4.8)–(4.12) can be satisfied by choosing $l = (0, \dots, 0, l_\psi)$ where $l_\psi \in (-V_\psi(M))^\ominus$ and $l_\psi \neq 0$. Such l_ψ exists since by Proposition 1.2 $\dim V_\psi(M) < k$.

Remark 4.4. Suppose Condition 4.1 is satisfied and, in addition, $I'' \neq \emptyset$, M is a cone and V is positively homogeneous and affine. Then Theorem 4.1 remains valid with a slightly weakened version of Condition 4.1 in which x_1, \dots, x_{k+1} are assumed to be linearly independent. To see this, note first that $I'' \neq \emptyset$ is equivalent to (2.13) with (4.13). By (2.14) h_1, \dots, h_{k+1} are linearly independent. Thus Proposition 1.2 and the fact that $V(x_i) = h_i$, $i \in \{1, \dots, k+1\}$, imply that x_1, \dots, x_{k+1} are linearly independent. This weakened version of Condition 4.1 is thus satisfied trivially when M does not contain $k+1$ linearly independent elements.

5. Applications involving directional differentials. We now assume that the mappings ϕ_0, \dots, ϕ_j and ψ satisfy certain first- and second-order one-sided differentiability conditions. These assumptions lead to the principal results of this section, Theorems 5.1 and 5.2, which contain the first- and second-order necessary conditions for OP. A rather extensive investigation of the relationship between these results leads to Theorem 5.4 which generalizes [35, Thm. 2.3]. The proof of Theorem 5.2 is then given along with remarks pointing out how the results of this section can be generalized.

In order to state the differentiability assumptions for the cost and constraint functions, it is necessary to introduce a generalization of the Fréchet derivative. This definition, which is due to Warga [33, p. 167] allows a function to have a derivative at a point which is not in the interior of its domain. Some consequences of this definition needed in the proofs of this section are discussed in the Appendix.

DEFINITION 5.1. Let \mathcal{X} be a Banach space with norm $|\cdot|$, \mathcal{Y} be a topological vector space, $\hat{A} \subset \mathcal{X}$ and $f: \hat{A} \rightarrow \mathcal{Y}$. f is \tilde{F} -differentiable at $\bar{x} \in \hat{A}$ and has the \tilde{F} -derivative $f^{(1)}(\bar{x}) \triangleq f'(\bar{x}) \in \mathcal{B}(\mathcal{X}; \mathcal{Y})$ if \bar{x} is contained in a solid convex subset of \hat{A} and

$$(5.1) \quad \lim_{\substack{x \rightarrow \bar{x} \\ x \in \hat{A}/\{\bar{x}\}}} |x - \bar{x}|^{-1} [f(x) - f(\bar{x}) - f'(\bar{x})(x - \bar{x})] = 0.$$

If $f'(x)$ exists for all $x \in \hat{A}$ then f is \tilde{F} -differentiable.

The second derivative requires a topology on $\mathcal{B}(\mathcal{X}; \mathcal{Y})$. This is handled by assuming that \mathcal{Y} is a Banach space and defining a norm on $\mathcal{B}(\mathcal{X}; \mathcal{Y})$ as in § 1.

DEFINITION 5.2. Let \mathcal{X} , \hat{A} , \mathcal{Y} and f be as in Definition 5.1 and assume further that \mathcal{Y} is a Banach space. f is twice \tilde{F} -differentiable at $\bar{x} \in \hat{A}$ and has the second \tilde{F} -derivative $f^{(2)}(\bar{x}) \triangleq f''(\bar{x}) \in \mathcal{B}(\mathcal{X}; \mathcal{B}(\mathcal{X}; \mathcal{Y}))$ if f is \tilde{F} -differentiable and the mapping $x \rightarrow f'(x): \hat{A} \rightarrow \mathcal{B}(\mathcal{X}; \mathcal{Y})$ is \tilde{F} -differentiable at \bar{x} .

First- and second-order one-sided directional differentials appear in the theorem statements. Their definition is a simple application of the preceding definitions.

DEFINITION 5.3. Let \mathcal{X} be a vector space, \mathcal{Y} be a topological vector space, $A \subset \mathcal{X}$, $F: A \rightarrow \mathcal{Y}$, $\bar{x} \in A$ and $h \in \mathcal{X}$. Suppose that there exists $\beta > 0$ such that $\bar{x} + \alpha h \in A$, $\alpha \in [0, \beta)$, and define $f: [0, \beta) \rightarrow \mathcal{Y}$ by $f(\alpha) = F(\bar{x} + \alpha h)$. If $f'(0)$ exists then $DF(\bar{x}; h) \triangleq f'(0)$ is the one-sided directional differential of f at \bar{x} in the direction h . If \mathcal{Y} is a Banach space and $f''(0)$ exists, then $D^2F(\bar{x}; h) \triangleq f''(0)$ is the second-order one-sided directional differential of F at \bar{x} in the direction h .

Note that \hat{A} in Definition 5.1 is given by $[0, \beta)$ in Definition 5.3. It can be seen that if $DF(\bar{x}; h)$ exists then $DF(\bar{x}; \alpha h)$ exists for all $\alpha > 0$. Thus, although $DF(\bar{x}; h)$ as defined is an element of $\mathcal{B}(\mathbb{R}; \mathcal{Y})$, we regard $DF(\bar{x}; \cdot)$ as a map from cone h into \mathcal{Y} .

The following notation concerning a feasible element \bar{e} simplifies the statement of what follows. If $j > 0$ let $I_N \triangleq \{i \in \{1, \dots, j\}; \phi_i(\bar{e}) < 0\}$, $I_A \triangleq \{1, \dots, j\}/I_N$ and $I_{A0} \triangleq I_A \cup \{0\}$; if $j = 0$ let $I_N \triangleq I_A \triangleq \emptyset$ and $I_{A0} \triangleq \{0\}$. Note that if $j > 0$ and $\dim Z_i > 1$ for some $i \in \{1, \dots, j\}$, then I_A does not necessarily coincide with the set $\{i \in \{1, \dots, j\}; \phi_i(\bar{e}) = 0\}$. We define

$$(5.2) \quad \Phi \triangleq \begin{cases} (\phi_0, \dots, \phi_j, \psi), & k > 0, \\ (\phi_0, \dots, \phi_j), & k = 0. \end{cases}$$

Finally, we recall from § 4 the meaning of the notation $l \in \mathcal{X}'^*$ and the convention $\phi_0(\bar{e}) = 0$.

Condition 5.1. E is a subset of a vector space \mathcal{X} and there exist a feasible element \bar{e} and a nonempty convex subset M of \mathcal{X} such that the following property is satisfied.

For each $X = \text{co}\{x_1, \dots, x_{k+1}\} \subset M$ there exists $\beta > 0$ such that

$$(5.3) \quad \bar{e} + \alpha X \subset E, \quad \alpha \in [0, \beta),$$

$$(5.4) \quad \mu \rightarrow \psi\left(\bar{e} + \alpha \sum_{i=1}^{k+1} \mu_i x_i\right): \Delta^k \rightarrow \mathbb{R}^k \text{ is continuous, } \alpha \in (0, \beta) \text{ (omit if } k = 0),$$

$$(5.5) \quad \alpha \rightarrow \Phi\left(\bar{e} + \sum_{i=1}^{k+1} \alpha_i x_i\right): \mathbb{P}^{k+1}(\beta) \rightarrow \mathcal{Z}' \text{ is } \tilde{F}\text{-differentiable at } \alpha = 0.$$

Furthermore, if $k = 0$ then

$$(5.6) \quad x \rightarrow D\phi_i(\bar{e}; x): M \rightarrow \mathcal{Z}_i \text{ is } Z_i\text{-convex, } i \in \{0, \dots, j\}.$$

THEOREM 5.1. *Suppose Condition 5.1 is satisfied. If \bar{e} solves OP then there exists $l \in \mathcal{Z}'^*$ such that*

$$(5.7) \quad l \neq 0,$$

$$(5.8) \quad l_i \in Z_i^\ominus, \quad i \in \{0, \dots, j\},$$

$$(5.9) \quad l_i = 0, \quad i \in I_N,$$

$$(5.10) \quad l_i(\phi_i(\bar{e})) = 0, \quad i \in I_A,$$

$$(5.11) \quad l(D\Phi(\bar{e}; x)) \geq 0, \quad x \in M.$$

Condition 5.2. $\mathcal{Z}_0, \dots, \mathcal{Z}_j$ are Banach spaces, \mathcal{E} is a subset of a vector space \mathcal{X} and there exist a feasible element \bar{e} , a vector $y \in \mathcal{X}$ and a nonempty convex set $M' \subset \mathcal{X}$ such that the following property is satisfied. For each $X = \text{co}\{x_1, \dots, x_{k+1}\} \subset M'$ there exists $\beta > 0$ such that

$$(5.12) \quad \bar{e} + \alpha_1 y + \alpha_2 X \subset \mathcal{E}, \quad (\alpha_1, \alpha_2) \in \mathbb{P}^2(\beta),$$

$$(5.13) \quad \bar{e} + \alpha y + \alpha^2 X \subset E, \quad (\alpha, \alpha^2) \in \mathbb{P}^2(\beta),$$

$$(5.14) \quad \alpha \rightarrow \Phi\left(\bar{e} + \alpha_1 y + \sum_{i=2}^{k+2} \alpha_i x_{i-1}\right): \mathbb{P}^{k+2}(\beta) \rightarrow \mathcal{Z}'$$

is twice \tilde{F} -differentiable at $\alpha = 0$.

Furthermore,

$$(5.15) \quad D\phi_i(\bar{e}; y) \leq 0, \quad i \in I_{A0},$$

$$(5.16) \quad D\psi(\bar{e}; y) = 0 \text{ (omit if } k = 0)$$

and, if $k = 0$, (5.6) is satisfied with M replaced by M' .

THEOREM 5.2. *Suppose Condition 5.2 is satisfied. If \bar{e} solves OP then there exists $l \in \mathcal{Z}'^*$ such that*

$$(5.17) \quad l \neq 0,$$

$$(5.18) \quad l_i \in Z_i^\ominus, \quad i \in \{0, \dots, j\},$$

$$(5.19) \quad l_i = 0, \quad i \in I'_0 \triangleq I_N \cup \{i \in I_{A0}: \phi_i(\bar{e}) + D\phi_i(\bar{e}; y) < 0\},$$

$$(5.20) \quad l_i(\phi_i(\bar{e}) + D\phi_i(\bar{e}; y)) = 0, \quad i \in I''_0 \triangleq \{0, \dots, j\} / I'_0,$$

$$(5.21) \quad l(D\Phi(\bar{e}; x) + \frac{1}{2}D^2\Phi(\bar{e}; y)) \geq 0, \quad x \in M'.$$

The proofs of Theorems 5.1 and 5.2, given at the end of this section, are based on Theorem 4.1. In brief, the reasons for Conditions 5.1 and 5.2 are as follows. Items (5.3) and (5.12) set up the differentiability domains for Φ . Theorem A.1 (see Appendix), a version of Taylor's theorem, is used with (5.5) and (5.14) to obtain expansions for Φ . These expansions in the variable α can be reduced to one-parameter expansions by means of Proposition A.1. The correct properties for the intermediate terms of these expansions follow from (5.15) and (5.16). In this way (4.4), (4.5), and (4.7) are obtained. The purpose of (5.4) is to guarantee (4.6). In Theorem 5.2, (5.14) implies (4.6). Condition (5.6) is required so that when $k=0$ the map V in Condition 4.1, which is given by $V(x) \triangleq D\Phi(\bar{e}; x)$, is Z -convex. No assumption analogous to (5.6) is needed when $k>0$ since (5.5) and Theorem A.2 imply that V in Condition 4.1 is affine. Similar remarks apply to Condition 5.2. In the proof of Theorem 5.1 $\zeta(x) = e + \alpha x$ which, by (5.3), is an element of E . Theorem 5.2 involves feasible elements of the form $\zeta(x) = \bar{e} + \alpha y + \alpha^2 x$ which, according to (5.13), are in E .

Remark 5.1. Because the origin of \mathbb{R}^n does not lie in the interior of $\mathbb{P}^n(\beta)$, (5.5) and (5.14) rely on Definitions 5.1 and 5.2 instead of the classical definition of the Fréchet derivatives. Since $\mathbb{P}^n(\beta) \subset \mathbb{B}^n(\beta)$, weaker versions of Theorems 5.1 and 5.2 are obtained by replacing “ \mathbb{P} ” by “ \mathbb{B} ” in (5.5) and (5.14). Proofs of these weaker results depend on the classical definition of the Fréchet derivative along with a classical Taylor theorem result. In this case, Theorem A.1 is not needed and a weakened version of Proposition A.1 suffices.

The necessary conditions of Theorems 5.1 and 5.2 are not necessarily satisfied by a common l . Thus, one should not jump to conclusions about the relationship between the first- and second-order results. Much of the development in the remainder of this section involves an examination of this issue. Of particular importance is the relationship between M and E in Theorem 5.1 and among y, M', E and \mathcal{E} in Theorem 5.2. The latter case is more complex because of the “quadratic” nature of (5.13). For example, y and M' may be chosen to characterize features of E such as a curved boundary.

The following two examples will be useful for illustrative purposes. It is easy to see that (5.12) and (5.13) are satisfied in both cases.

Example 5.1. $\mathcal{X} = \mathcal{E} = \mathbb{R}^2$, $k \in \mathbb{N}$, $\bar{e} = (0, 0)$, $y = (1, 0)$, $M' = \{0\} \times [1, 2]$, $E = \{(s_1, s_2) \in \mathbb{R}^2 : 0 \leq s_1 \leq 1, s_1^2 \leq s_2 \leq 2s_1^2\}$.

Example 5.2. $\mathcal{X}, \mathcal{E}, k, \bar{e}, y$ as in Example 5.1, $M' = \{(s_1, s_2) \in \mathbb{R}^2 : s_1 \leq 0, 1 - s_1 \leq s_2 \leq 2 - 2s_1\}$, $E = \{(s_1, s_2) \in \mathbb{R}^2 : s_1^2 \leq s_2, s_2^2 \leq s_1\}$.

Remark 5.2. Note that if Condition 5.1 holds then it remains valid if M is replaced by cone M . Thus, without loss of generality, M can be assumed to be a convex cone. Examples 5.1 and 5.2 show that it is not always possible to replace M' by cone M' in Condition 5.2 since (5.13) may not be satisfied.

Example 5.1 shows that Condition 5.2 does not imply that Condition 5.1 is meaningfully satisfied. For this example the only set M satisfying (5.3) is $M = \{(0, 0)\}$ which yields trivial necessary conditions. Even when Conditions 5.1 and 5.2 are both meaningfully satisfied, M and M' may be disjoint as in Example 5.2 where M must be a subset of $\mathbb{R}_+^2 \cup \{(0, 0)\}$. Thus, it is not surprising that there may be no l satisfying both (5.7)–(5.11) and (5.17)–(5.21). The following assumption about the structure of M' leads to the existence of a common l . Consider

$$(5.22) \quad M' = Q + R, \quad Q \text{ convex}, \quad R \text{ a convex cone.}$$

We use (5.22) to obtain the following result.

PROPOSITION 5.1. *Suppose the assumptions of Theorem 5.2 are valid with M' given by (5.22). If $k > 0$ then (5.21) is equivalent to*

$$(5.23) \quad l(D\Phi(\bar{e}; x)) \geq 0, \quad x \in R,$$

$$(5.24) \quad l(D\Phi(\bar{e}; x) + \frac{1}{2}D^2\Phi(\bar{e}; y)) \geq 0, \quad x \in Q.$$

If $k = 0$ then (5.21) implies (5.23) and (5.24). The converse in this case is valid if either $Q = \{0\}$ or the map $x \rightarrow D\Phi(\bar{e}; x): M' \rightarrow \mathcal{X}$ is affine.

Proof. Define $V_1 \triangleq -(D\Phi(\bar{e}; Q) + \frac{1}{2}D^2\Phi(\bar{e}; y))$, $V_2 \triangleq -(D\Phi(\bar{e}; R))$ and $V_3 \triangleq -(D\Phi(\bar{e}; M') + \frac{1}{2}D^2\Phi(\bar{e}; y))$ and note that (5.24), (5.23) and (5.21) are equivalent to $l \in V_1^\ominus$, $l \in V_2^\ominus$ and $l \in V_3^\ominus$, respectively. For the case $k > 0$ we have $V_1 + V_2 = V_3$ since $D\Phi(\bar{e}; \cdot)$ is positively homogeneous and affine (see Theorem A.2). Since V_2 is a cone, Proposition 1.9 implies $V_1^\ominus \cap V_2^\ominus = V_3^\ominus$, as desired. For the case $k = 0$ it can be shown that (5.6) with M replaced by M' implies $V_1 + V_2 \subset V_3 + Z$. Thus, $(V_3 + Z)^\ominus \subset (V_1 + V_2)^\ominus$ and, by Proposition 1.9, $V_3^\ominus \cap Z^\ominus \subset V_1^\ominus \cap V_2^\ominus$. Since (5.18) is satisfied, (5.21) implies (5.23) and (5.24). If either $Q = \{0\}$ or $D\Phi(\bar{e}; \cdot)$ is affine, $V_1 + V_2 = V_3$ and thus $V_1^\ominus \cap V_2^\ominus = V_3^\ominus$, which completes the proof. \square

Notice that (5.22) can be satisfied trivially with $Q = M'$ and $R = \{0\}$. However, this does not produce new results. Situations of general interest are: $Q = \{0\}$, $R = M'$ (i.e., when M' is a cone) and $Q \neq \{0\}$, $R \neq \{0\}$ (e.g., Example 5.2, where $Q = \{0\} \times [1, 2]$ and $R = \{(s_1, s_2) \in \mathbb{R}^2: -s_1 \leq s_2 \leq -2s_1\}$). The following result is a consequence of Theorem 5.2 and Proposition 5.1.

THEOREM 5.3. *Suppose Condition 5.2 is satisfied with M' given by (5.22) and let $M \subset R$. If \bar{e} solves OP then each l satisfying (5.17)–(5.21) also satisfies (5.7)–(5.11).*

Proof. By Proposition 5.1, (5.21) implies (5.23) and (5.24). Since $M \subset R$, (5.23) implies (5.11). Condition (5.9) follows immediately from (5.19); (5.10) follows from (5.19) and (5.20) using (5.15) and (5.18). \square

Theorem 5.3 shows that if the set M' in Condition 5.2 is given by (5.22) and M is an arbitrary convex subset of R , then the second-order necessary conditions (5.17)–(5.20), (5.23) and (5.24) imply the first-order necessary conditions (5.7)–(5.11). The strongest version of (5.7)–(5.11) is obtained when $M = R$. However, these “first-order necessary conditions” may be unobtainable from Theorem 5.1 because Condition 5.1 may not hold for certain choices of $M \subset R$. For instance, in Example 5.2 it is necessary to choose $M = \{(0, 0)\}$ in order to satisfy both (5.3) and the requirement $M \subset R$. The reason Theorem 5.2 gives stronger “first-order necessary conditions” than Theorem 5.1 is that (5.13) takes the curvature of E into account while (5.3) does not. Note that y must be nonzero since otherwise (5.13) and (5.3) are equivalent.

Remark 5.3. The observation that second-order necessary conditions may contain “first-order necessary conditions” that are stronger than actual first-order necessary conditions has been made previously in [28, Remark 2, p. 278]. In their treatment of OP $j = 0$, $\mathcal{X} = \mathbb{R}^n$, ϕ_0 and ψ are continuously differentiable and it is assumed from the outset that (5.22) holds with $Q = \{q\}$.

Remark 5.4. The relationship between the first- and second-order necessary conditions can be thought of in the following way. If Condition 5.1 is satisfied, then the first-order necessary conditions involve finding an element l in \mathcal{X}'^* satisfying (5.7)–(5.11). If, in addition, Condition 5.2 is satisfied with (5.22) and $M \subset R$, then the second-order necessary conditions are equivalent to the existence of an element l in \mathcal{X}'^* satisfying both (5.7)–(5.11) and the additional conditions (5.19), (5.20), (5.23) and (5.24). In this case, Theorem 5.2 supplements Theorem 5.1 and verification of the necessary conditions can be thought of as a two-stage process.

Remark 5.5. Suppose Conditions 5.1 and 5.2 are satisfied with (5.22) and $M \subset R$ and define

$$(5.25) \quad \Gamma \triangleq \{l \in \mathcal{L}^{**}: l \text{ satisfies (5.7)–(5.11)}\}.$$

Remark 5.4 shows that it is of particular interest to determine whether Γ is a ray. Specifically, if Γ is a ray then the search for l satisfying the second-order necessary conditions is simplified since l is determined uniquely to within a scalar multiple by the first-order necessary conditions alone.

We now consider a subset of M that is useful in verifying the first-order necessary conditions and understanding the relationship between Theorems 5.1 and 5.2. Define

$$(5.26) \quad \mathcal{D} \triangleq \{y \in M: y \text{ satisfies (5.15) and (5.16)}\},$$

$$(5.27) \quad \mathcal{J}(y) \triangleq \{i \in I_{A0}: \phi_i(\bar{e}) + D\phi_i(\bar{e}; y) < 0\},$$

where $y \in \mathcal{D}$, and

$$(5.28) \quad \mathcal{J} \triangleq \bigcup_{y \in \mathcal{D}} \mathcal{J}(y).$$

PROPOSITION 5.2. *If l satisfies (5.7)–(5.11) then l also satisfies*

$$(5.29) \quad l_i = 0, \quad i \in I_N \cup \mathcal{J},$$

and

$$(5.30) \quad l_i(\phi_i(\bar{e}) + D\phi_i(\bar{e}; y)) = 0, \quad i \in \{0, \dots, j\}/(I_N \cup \mathcal{J}), \quad y \in \mathcal{D}.$$

Proof. Since $\mathcal{D} \subset M$, set $x = y \in \mathcal{D}$ in (5.11). From (5.16) we obtain

$$(5.31) \quad \sum_{i \in I_{A0}} l_i(D\phi_i(\bar{e}; y)) \geq 0.$$

It follows from (5.31), (5.8) and (5.15) that $l_i(D\phi_i(\bar{e}; y)) = 0$, $i \in I_{A0}$. The result now follows from (5.10) and Proposition 1.8. \square

Remark 5.6. Proposition 5.2 sharpens Remark 5.4 in the following way. Suppose Conditions 5.1 and 5.2 are satisfied with (5.22) and $M \subset R$. If y in Condition 5.2 is an element of M (which implies $y \in \mathcal{D}$) then (5.19) and (5.20) are a consequence of (5.29) and (5.30) and thus do not strengthen (5.7)–(5.11).

Remark 5.7. The ideas used in the proof of Proposition 5.2 could also be used in other contexts in the optimization literature to show that additional multiplier components are zero. However, this approach does not appear to have been used before.

We now consider some consequences of strengthening Condition 5.1. Sometimes the set E is sufficiently “large” near \bar{e} so that

$$(5.32) \quad \bar{e} + \alpha_1 y + \alpha_2 X \subset E, \quad (\alpha_1, \alpha_2) \in \mathbb{P}^2(\beta).$$

This strengthening of Condition 5.2 is formalized in Condition 5.3.

Condition 5.3. Condition 5.2 is satisfied with (5.12) replaced by (5.32) and with (5.13) omitted.

The following results are easily verified.

PROPOSITION 5.3. *If Condition 5.3 is satisfied then it is also satisfied with M' replaced by cone M' .*

PROPOSITION 5.4. *If Condition 5.3 is satisfied, then Condition 5.1 is satisfied with $M = M'$.*

The importance of Condition 5.3 lies in Proposition 5.4 which guarantees that nontrivial first-order necessary conditions can be obtained independently of the second-order necessary conditions. Proposition 5.3 shows that (5.22) is satisfied with $R = \text{cone } M'$ and $Q = \{0\}$.

Sometimes Conditions 5.2 and 5.3 are equivalent. For example, this occurs if either $\mathcal{E} = E$ or $y = 0$. A more interesting case is contained in the following proposition which leads ultimately to Theorem 5.4, a simply stated specialization of Theorem 5.1 and 5.2.

PROPOSITION 5.5. *Suppose Condition 5.2 is satisfied with (5.13) deleted and with the additional conditions*

$$(5.33) \quad 0 \in M',$$

$$(5.34) \quad y \in M',$$

$$(5.35) \quad \mathcal{E} \cap (M' + \bar{e}) \subset E.$$

Then Condition 5.3 is satisfied.

Proof. It suffices to show that (5.33)–(5.35) and (5.12) imply (5.32). Let $\beta \in (0, 1)$ and $X = \text{co}\{x_1, \dots, x_{k+1}\} \subset M'$. From (5.33) and (5.34) it follows that $\{\bar{e}, \bar{e} + \beta y, \bar{e} + \beta x_1, \dots, \bar{e} + \beta x_{k+1}\} \subset M' + \bar{e}$. Thus $\text{co}\{\bar{e}, \bar{e} + \beta y, \bar{e} + \beta x_1, \dots, \bar{e} + \beta x_{k+1}\} \subset M' + \bar{e}$ which is equivalent to $\bar{e} + \alpha_1 y + \alpha_2 X \subset M' + \bar{e}$, $(\alpha_1, \alpha_2) \in \mathbb{P}^2(\beta)$. The desired result now follows from (5.12) and (5.35). \square

It is easy to verify that the following conditions imply Conditions 5.1 and 5.2, respectively.

Condition 5.4. E is a subset of a vector space \mathcal{X} and there exist a feasible element \bar{e} and a nonempty convex subset M of \mathcal{X} such that $0 \in M$ and $\mathcal{E} \cap (M + \bar{e}) \subset E$ and such that the following property is satisfied. For each $X = \text{co}\{x_1, \dots, x_{k+1}\} \subset M$ there exists $\beta > 0$ such that

$$(5.36) \quad \bar{e} + \alpha X \subset \mathcal{E}, \quad \alpha \in [0, \beta),$$

and (5.4) and (5.5) are satisfied. Furthermore, if $k = 0$ then (5.6) holds.

Condition 5.5. Condition 5.4 is satisfied, $\mathcal{X}_0, \dots, \mathcal{X}_j$ are Banach spaces and the following property is satisfied. For each $X = \text{co}\{x_1, \dots, x_{k+2}\} \subset M$ there exists $\beta > 0$ satisfying (5.36) and such that

$$(5.37) \quad \alpha \rightarrow \Phi\left(\bar{e} + \sum_{i=1}^{k+2} \alpha_i x_i\right) : \mathbb{P}^{k+2}(\beta) \rightarrow \mathcal{Y}$$

is twice \tilde{F} -differentiable at $\alpha = 0$.

THEOREM 5.4. *Suppose \bar{e} solves OP. If Condition 5.4 is satisfied then there exists $l \in \mathcal{X}^*$ satisfying (5.7)–(5.11). If, furthermore, Condition 5.5 is satisfied, then for each $y \in \mathcal{D}$ there exists $l \in \mathcal{X}^*$ satisfying (5.7)–(5.11) and*

$$(5.38) \quad l(D^2\Phi(\bar{e}; y)) \geq 0.$$

Proof. Arguments similar to those used to prove Proposition 5.5 show that the conditions on M and (5.36) imply (5.3). Thus, the first part of the theorem follows from Theorem 5.1. To prove the second part of the theorem let $X = \text{co}\{x_1, \dots, x_{k+1}, y\} \subset M$, where $y \in \mathcal{D}$, and note that (5.36) with this choice of X is equivalent to (5.12). Thus, Condition 5.2 is satisfied with (5.13) deleted. Since also (5.33)–(5.35) (with $M' = M$) are satisfied, Proposition 5.5 implies that Condition 5.3

must hold with $M' = M$. By Proposition 5.3, Condition 5.3 is satisfied with M replaced by cone M . Since Condition 5.3 implies Condition 5.2, it follows from Theorem 5.2 and Proposition 5.1 (with $R = \text{cone } M$ and $Q = \{0\}$) that there exists $l \in \mathcal{X}^{*k}$ satisfying (5.17)–(5.20), (5.11) and (5.38). Finally, Proposition 5.2 implies that l satisfying (5.7)–(5.11) must also satisfy (5.19) and (5.20). Thus, these last two conditions have been omitted. \square

Remark 5.8. Remark 5.5 takes on added importance in the context of Theorem 5.4. This is because when Γ is a ray, l no longer depends on $y \in \mathcal{D}$. In this case Theorem 5.4 can be strengthened by deleting the phrase “for each $y \in \mathcal{D}$ ” and replacing (5.38) by

$$(5.39) \quad l(D^2\Phi(\bar{e}; y)) \geq 0, \quad y \in \mathcal{D}.$$

Remark 5.9. Theorem 5.4 generalizes Warga [35, Thm. 2.3]. To obtain his result, specialize OP by setting $j = 0$, $\mathcal{X}_0 = \mathbb{R}$ and $Z_0 = \mathbb{R}_-$. The hypotheses of Theorem 5.4 are weaker than those of [35, Thm. 2.3] in several important ways. In [35], $\mathcal{E} \cap (M + \bar{e}) \subset E$ is replaced by $\mathcal{E} \cap (M + \bar{e}) = E$, \mathbb{P} is replaced by \mathbb{B} in (5.37) (see Remark 5.1), the map in (5.37) is assumed to be twice continuously differentiable in a neighborhood of the origin and an additional normality-like condition is assumed. It is shown in [13] that this normality condition implies that Γ (specialized to the problem of [35]) is a ray.

We now prove Theorem 5.2. For brevity, we assume $k > 0$; the case $k = 0$ follows from similar arguments. For $I \subset \{0, \dots, j\}$ let $\Phi_I \triangleq (\bar{\phi}_0, \dots, \bar{\phi}_j, \psi)$, where $\bar{\phi}_i \triangleq \phi_i$, $i \in I$, and $\phi_i \triangleq 0$, $i \in \{0, \dots, j\}/I$. We will show that Condition 4.1 is satisfied with \bar{e} and \mathcal{X} as specified, $M = M'$, $n = 2$, $m_i = 0$ for $i \in I_N$, $m_i = 1$ for $i \in I_{A0}$, $Y_{i1} = D\phi_i(\bar{e}; y)$ for $i \in I_{A0}$, and $V(x) = D\Phi_{I_0}(\bar{e}; x) + \frac{1}{2}D^2\Phi_{I_0}(\bar{e}; y)$. Since $I' = I'_0$ and $I'' = I''_0$, (5.19) and (5.20) follow from (4.10) and (4.11). Also, (4.12) and (5.19) imply (5.21).

We now show that Condition 5.2 implies Condition 4.1. Note that from (5.14) and Theorem A.2 (with $\nu = 2$) it follows that $V: M' \rightarrow \mathcal{X}'$ is affine. Thus (see Condition 4.1 for notation), V_ψ is affine and V_i is Z_i -convex, $i \in I''_0$. Let $X = \text{co}\{x_1, \dots, x_{k+1}\} \subset M'$ be a k -simplex and define $f: \mathbb{P}^{k+2}(\beta) \rightarrow \mathcal{X}'$ by

$$f(\alpha) = \Phi_{I_0}(\bar{e} + \alpha_1 y + \sum_{i=2}^{k+2} \alpha_i x_{i-1}).$$

Also, let $\beta_0 > 0$ satisfy $\beta_0 + \beta_0^2 < \beta$. From (5.14), Proposition A.1.III, Remark A.1 and Theorem A.2, it follows that, for $\alpha \in [0, \beta_0)$, $\mu \in \Delta^k$ and $x = \sum_{i=1}^{k+1} \mu_i x_i$,

$$(5.40) \quad \begin{aligned} \Phi_{I_0}(\bar{e} + \alpha y + \alpha^2 x) &= f(\alpha, \alpha^2 \mu_1, \dots, \alpha^2 \mu_{k+1}) \\ &= f(0) + \alpha f_{\alpha_1}(0) + \frac{1}{2} \alpha^2 \left[f_{\alpha_1 \alpha_2}(0) + 2 \sum_{i=1}^{k+1} \mu_i f_{\alpha_1 \alpha_{i+1}}(0) \right] + R(\alpha, \mu) \\ &= \Phi_{I_0}(\bar{e}) + \alpha D\Phi_{I_0}(\bar{e}; y) + \alpha^2 V(x) + R(\alpha, \mu), \end{aligned}$$

where $\alpha^{-2}R(\alpha, \mu) \xrightarrow{\alpha \rightarrow 0^+} 0$ uniformly for $\mu \in \Delta^k$. From (5.14) it follows that the mapping

$$\alpha \rightarrow \phi_i\left(\bar{e} + \alpha_1 y + \sum_{i=2}^{k+2} \alpha_i x_{i-1}\right): \mathbb{P}^{k+2}(\beta) \rightarrow \mathcal{X}_i$$

is continuous at $\alpha = 0$, $i \in I_N$, and \tilde{F} -differentiable at $\alpha = 0$, $i \in I'_0/I_N$. Thus, Proposition A.1.I implies

$$(5.41) \quad \phi_i(\bar{e} + \alpha y + \alpha^2 x) - \phi_i(\bar{e}) \xrightarrow{\alpha \rightarrow 0^+} 0$$

uniformly for $\mu \in \Delta^k$, $i \in I_N$, and Proposition A.1.II yields

$$(5.42) \quad \alpha^{-1}[\phi_i(\bar{e} + \alpha y + \alpha^2 x) - \phi_i(\bar{e}) - \alpha D\phi_i(e; y)] \xrightarrow{\alpha \rightarrow 0^+} 0$$

uniformly for $\mu \in \Delta^k$, $i \in I'_0/I_N$. For N_0, \dots, N_j, η and τ as specified in Condition 4.1, it is now easy to see from (5.40)–(5.42) that there exists $\alpha \in (0, \tau)$ such that (4.4), (4.5) (without need for Z_i) and (4.7) are satisfied with $\zeta(x) = \bar{e} + \alpha y + \alpha^2 x$. Finally, (4.6) is a consequence of (5.14).

To prove Theorem 5.1, let $n = 1$, $m_i = 0$ for $i \in I_N$, $m_i = 1$ for $i \in I_{A_0}$, and $V(x) = D\Phi_{I_{A_0}}(\bar{e}; x)$. Since $I' = I_N$ and $I'' = I_{A_0}$, (5.9) and (5.10) follow from (4.10) and (4.11). Also, (5.11) follows from (4.12) and (5.9). The remainder of the proof follows from arguments that should by now be clear. \square

We now point out several ways in which the results of this section can be generalized. First, it is possible to take advantage of the presence of Z_i in (4.4) and (4.5) by replacing continuity and differentiability conditions such as (5.51) and (5.52) by conditions involving semicontinuity and semidifferentiability. This approach appears in [20] for first-order necessary conditions but apparently has not been extended to higher-order necessary conditions.

Another approach to generalizing Theorems 5.1 and 5.2 is based on the concept of a conical approximation. This involves the existence of a map θ depending on X such that, instead of (5.3), (5.12) and (5.13), the following conditions hold:

$$(5.3)^* \quad \theta(\bar{e} + \alpha X) \subset E, \quad \alpha \in [0, \beta),$$

$$(5.12)^* \quad \theta(\bar{e} + \alpha_1 y + \alpha_2 X) \subset \mathcal{E}, \quad (\alpha_1, \alpha_2) \in \mathbb{P}^2(\beta),$$

$$(5.13)^* \quad \theta(\bar{e} + \alpha y + \alpha^2 X) \subset E, \quad (\alpha, \alpha^2) \in \mathbb{P}^2(\beta).$$

Moreover, in (5.5) and (5.14), $\Phi \circ \theta$ plays the role of Φ . Additional assumptions such as the following are then required: if X and \hat{X} have corresponding maps θ and $\hat{\theta}$, then

$$(5.43) \quad D\Phi \circ \theta(\bar{e}; x) = D\Phi \circ \hat{\theta}(\bar{e}; x), \quad x \in X \cap \hat{X}.$$

A closely related approach appears in [7, Def. B.1.3]. See [28] for a related notion of conical approximation in the context of second-order necessary conditions for a nonlinear programming problem.

6. Applications to nonlinear programming. In this section we further illustrate the use of Theorem 4.1 by deriving first-, second- and third-order necessary conditions for a nonlinear programming problem. The third-order conditions are new while the first- and second-order conditions sharpen results from the previous literature. To keep the conditions reasonably simple, the hypotheses of this section are considerably stronger than those of § 5.

To derive n th-order necessary conditions, the strengthened assumptions for OP are:

$$(6.1) \quad \mathcal{E} \text{ is a Banach space,}$$

$$(6.2) \quad \mathcal{Z}_i = \mathbb{R}, Z_i = \bar{\mathbb{R}}_-, i \in \{0, \dots, j\},$$

$$(6.3) \quad \phi_0, \dots, \phi_j \text{ and } \psi \text{ are } C^n.$$

A mapping from \mathcal{E} into \mathbb{R}^r is C^n if it is n -times Fréchet differentiable and its first-through n th-order derivatives are continuous. Let NP denote OP in the presence of (6.1)–(6.3). As before, let $j=0$ and $k=0$ denote the absence of (1.2)' and (1.3), respectively. Define the (Lagrangian) function $L: \mathcal{E} \times \mathbb{R}^{1+j+k} \rightarrow \mathbb{R}$ by

$$(6.4) \quad L(e, l) = \begin{cases} \sum_{i=0}^j l_i \phi_i(e) + l_\psi^T \psi(e), & l = (l_0, \dots, l_j, l_\psi), \quad k > 0, \\ \sum_{i=0}^j l_i \phi_i(e), & l = (l_0, \dots, l_j), \quad k = 0. \end{cases}$$

Notation for the Fréchet derivative and partial Fréchet derivative, such as in

$$L_{ee}(e, l)(y, \hat{y}) = \sum_{i=0}^j l_i \phi_i''(e)(y, \hat{y}) + l_\psi^T \psi''(e)(y, \hat{y}),$$

is made precise in the Appendix. We require some notation pertaining to a feasible element \bar{e} . Recall that if $j > 0$ then $I_N = \{i \in \{1, \dots, j\}: \phi_i(\bar{e}) < 0\}$ and, because of (6.2), $I_{A0} = \{0\} \cup \{i \in \{1, \dots, j\}: \phi_i(\bar{e}) = 0\}$. If $j = 0$ then $I_N = \emptyset$ and $I_{A0} = \{0\}$. Let the set of variations M be chosen to satisfy

$$(6.5) \quad 0 \in M \subset E - \bar{e}, \quad M \text{ is convex.}$$

For the following definitions let $\psi = 0$ when $k = 0$. If $n = 1$ define

$$(6.6) \quad \mathcal{D}_1 \triangleq \{y \in M: \phi_i'(\bar{e})(y) \leq 0, i \in I_{A0}, \psi'(\bar{e})(y) = 0\},$$

$$(6.7) \quad \mathcal{J}_1(y) \triangleq \{i \in I_{A0}: \phi_i'(\bar{e})(y) < 0\},$$

where $y \in \mathcal{D}_1$, and

$$(6.8) \quad \mathcal{J}_1 \triangleq \bigcup_{y \in \mathcal{D}_1} \mathcal{J}_1(y).$$

If $n = 2$, then let

$$(6.9) \quad \begin{aligned} \mathcal{D}_2(y) &\triangleq \{\hat{y} \in M: \tfrac{1}{2} \phi_i''(\bar{e})(y)^2 + \phi_i'(\bar{e})(\hat{y}) \leq 0, \\ &\quad i \in I_{A0} / \mathcal{J}_1(y), \tfrac{1}{2} \psi''(\bar{e})(y)^2 + \psi'(\bar{e})(\hat{y}) = 0\}, \end{aligned}$$

where $y \in \mathcal{D}_1$,

$$(6.10) \quad \mathcal{D}_2 \triangleq \{(y, \hat{y}) \in M^2: y \in \mathcal{D}_1, \hat{y} \in \mathcal{D}_2(y)\}$$

and

$$(6.11) \quad \mathcal{J}_2(y, \hat{y}) \triangleq \{i \in I_{A0} / \mathcal{J}_1(y): \tfrac{1}{2} \phi_i''(\bar{e})(y)^2 + \phi_i'(\bar{e})(\hat{y}) < 0\},$$

where $\hat{y} \in \mathcal{D}_2(y)$.

THEOREM 6.1. *Suppose \bar{e} solves NP. I. If $n = 1$ then there exists $l \in \mathbb{R}^{1+j+k}$ satisfying*

$$(6.12) \quad |l| = 1,$$

$$(6.13) \quad l_i \geq 0, \quad i \in \{0, \dots, j\},$$

$$(6.14) \quad l_i = 0, \quad i \in I_N,$$

$$(6.15) \quad L_e(\bar{e}, l)(x) \geq 0, \quad x \in M.$$

II. If $n = 2$, then for each $y \in \mathcal{D}_1$ there exists $l \in \mathbb{R}^{1+j+k}$ satisfying (6.12)–(6.15) and

$$(6.16) \quad L_{ee}(\bar{e}, l)(y)^2 \geq 0.$$

III. If $n = 3$, then for each $(y, \hat{y}) \in \mathcal{D}_2$ there exists $l \in \mathbb{R}^{1+j+k}$ satisfying (6.12)–(6.15) and

$$(6.17) \quad l_i = 0, \quad i \in \mathcal{I}_2(y, \hat{y}),$$

$$(6.18) \quad L_{ee}(\bar{e}, l)(y, \hat{y}) + \frac{1}{6}L_{eee}(\bar{e}, l)(y)^3 \geq 0.$$

Remark 6.1. From Proposition 5.2 it follows that l satisfying (6.12)–(6.15) must also satisfy

$$(6.19) \quad l_i = 0, \quad i \in \mathcal{I}_1.$$

Remark 6.2. If $0 \in \text{int } M$ then (6.15) is clearly equivalent to the condition

$$(6.20) \quad L_e(\bar{e}, l) = 0.$$

Part I of Theorem 6.1 with $\mathcal{E} = \mathbb{R}^r$ is well known. When $E = \mathcal{E}$ (and thus $M = \mathcal{E}$) see, e.g., [25, Thm. 1]; when $0 \notin \text{int } M$ see, e.g., [7, Thm. 2.3.12]. Second-order necessary conditions similar to those of part II can also be found in the literature. For example, part II follows from [4, Thm. 3.2] when $E = \mathcal{E} = \mathbb{R}^r$. There the condition $l_i = 0, i \in \mathcal{I}_1(y)$, is included. Remark 6.1 shows, however, that this adds no new information. After some manipulation, part II with $\mathcal{E} = \mathbb{R}^r$ and $j = 0$ can be obtained from [28, Thm. 6]. The only third-order necessary conditions from the literature which appear to be related to part III are [14, Thm. 2.5], [16, Thm. 5.1] and [22, Thm. 2.5]. Because of differing hypotheses, however, these results cannot be compared directly.

Note that l depends on $y \in \mathcal{D}_1$ in part II and on $(y, \hat{y}) \in \mathcal{D}_2$ in part III. Specific examples of second-order necessary conditions where this dependence actually occurs have been given in [4] and [22]. Note that for both parts II and III, l must belong to the set

$$(6.21) \quad \hat{\Gamma} \triangleq \{l \in \mathbb{R}^{1+j+k} : l \text{ satisfies (6.12)–(6.15)}\}.$$

Using $\hat{\Gamma}$, Theorem 6.1 can be written more compactly. This equivalent version of the theorem will help establish further connections with results from the literature.

COROLLARY 6.1. Suppose \bar{e} solves NP. I. If $n = 1$, then

$$(6.22) \quad \hat{\Gamma} \neq \emptyset.$$

II. If $n = 2$, then

$$(6.23) \quad \text{for each } y \in \mathcal{D}_1 \text{ there exists } l \in \hat{\Gamma} \text{ satisfying (6.16).}$$

III. If $n = 3$, then

$$(6.24) \quad \text{for each } (y, \hat{y}) \in \mathcal{D}_2 \text{ there exists } l \in \hat{\Gamma} \text{ satisfying (6.17) and (6.18).}$$

Remark 6.3. As shown in Remark 5.10, (6.23) can be expressed in the equivalent form

$$(6.25) \quad \max_{l \in \hat{\Gamma}} L_{ee}(\bar{e}, l)(y)^2 \geq 0, \quad y \in \mathcal{D}_1,$$

which is similar to [18, Thm. 6]. His result is valid for infinite-dimensional equality constraints when $\psi'(\bar{e})$ has full range. A similar result was also obtained in [23].

Remark 6.4. Condition (6.19) clarifies some second-order necessary conditions

from the prior literature which involve a constraint qualification. In [26, p. 29] and [24, p. 102], for example, the element l satisfying the first-order necessary conditions (6.12)–(6.14) and (6.20) (since $M = \mathcal{E}$) is used to determine a set of critical directions

$$(6.26) \quad \begin{aligned} \mathcal{D}_1(l) &\triangleq \{y \in \mathcal{E} : \phi'_i(\bar{e})(y) = 0 \text{ if } i \in I_{A_0} \text{ and } l_i > 0, \\ &\phi'_i(\bar{e})(y) \leq 0 \text{ if } i \in I_{A_0} \text{ and } l_i = 0, \\ &\psi'(\bar{e})(y) = 0\}. \end{aligned}$$

This definition gives the impression that $\mathcal{D}_1(l)$ depends on l . However, using (6.19) it is easy to show that $\mathcal{D}_1(l) = \mathcal{D}_1$ for all l satisfying (6.12)–(6.14) and (6.20). Thus $\mathcal{D}_1(l)$ is independent of l .

Remark 6.5. Sometimes NP satisfies a constraint qualification which implies that (6.23) can be replaced by a stronger condition such as

There exists $l \in \hat{\Gamma}$ such that

$$(6.27) \quad L_{ee}(\bar{e}, l)(y)^2 \geq 0, \quad y \in \mathcal{D}_1,$$

or

$$(6.28) \quad L_{ee}(\bar{e}, l)(y)^2 \geq 0, \quad l \in \hat{\Gamma}, \quad y \in \mathcal{D}_1.$$

See [3], [4] for details of specific constraint qualifications in this context. Note that (6.28) implies (6.27) and (6.27) implies (6.23). If $\hat{\Gamma}$ is a singleton then (6.23), (6.27) and (6.28) are equivalent.

We now present an example illustrating the use of Theorem 6.1. Let $j = 3$, $k = 0$, $M = E = \mathcal{E} = \mathbb{R}^2$, $\bar{e} = (1, 0)$, $\phi_0(t_1, t_2) = t_1$, $\phi_1(t_1, t_2) = -(1 - t_1)^3 + t_2$, $\phi_2(t_1, t_2) = -t_1$ and $\phi_3(t_1, t_2) = -t_2$. This example is often discussed in connection with the Kuhn–Tucker constraint qualification (see, e.g., [9, p. 20]) which fails to hold at \bar{e} . Note that $I_N = \{2\}$, $I_{A_0} = \{0, 1, 3\}$, $\phi'_0(\bar{e}) = [1 \ 0]$ and $\phi'_1(\bar{e}) = -\phi'_3(\bar{e}) = [0 \ 1]$. Thus, part I is satisfied uniquely with $l = (0, \frac{1}{2}, 0, \frac{1}{2})$. Because $\mathcal{D}_1 = \mathbb{R}_- \times \{0\}$ and $\phi''_1(\bar{e}) = \phi''_3(\bar{e}) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, l also satisfies (6.16) for all $y \in \mathcal{D}_1$. Thus, part II is satisfied. Define $y = (-1, 0)$ and, since $\mathcal{D}_2(y) = \mathbb{R} \times \{0\}$, let $\hat{y} = (0, 0)$. Note that $\mathcal{J}_2(y, \hat{y}) = \emptyset$. Since $\phi'''_3(\bar{e}) = 0$, $(\phi_1)_{t_1 t_1}(\bar{e}) = 6$ and all other components of $\phi'''_1(\bar{e})$ are zero, (6.18) implies that $l_1 \leq 0$, which contradicts $l_1 = \frac{1}{2}$. Thus, part III is not satisfied and \bar{e} is not optimal.

If in this example ϕ_0 is redefined to be $\phi_0(t_1, t_2) = -t_1$, then \bar{e} is optimal and the necessary conditions are satisfied with l as given above. These examples show that even when $l_0 = 0$ (and thus ϕ_0 is absent from the Lagrangian) the higher-order necessary conditions yield useful information. The reason for this is that ϕ_0 still plays a role in the necessary conditions because of its appearance in the definition of \mathcal{D}_1 .

We now prove Theorem 6.1. Parts I and II follow most easily from Theorem 5.4, but Theorem 4.1 can also be used without great difficulty. When $n = 1$, it is easy to see that (6.1)–(6.3) and (6.5) imply Condition 5.4 with $\mathcal{X} = \mathcal{E}$. Note that: (5.8) and (6.2) imply (6.13); (5.9) implies (6.14); and (5.11) implies (6.15). Note that (5.10) can be ignored since it yields no useful information. When $n = 2$, (6.1)–(6.3) imply Condition 5.5 with \mathcal{X} as just defined. It remains only to note that (5.38) implies (6.16).

Part III follows from Theorem 4.1. The arguments used in the proof of Proposition 5.5 show that M can be replaced by cone M . We define $n = 3$, $I' = I_N \cup \mathcal{J}_1(y) \cup \mathcal{J}_2(y, \hat{y})$, $I'' = \{0, \dots, j\}/I'$, $m_i = 0$ for $i \in I_N$, $m_i = 1$ for $i \in \mathcal{J}_1(y)$, $m_i = 2$ for $i \in I'' \cup \mathcal{J}_2(y, \hat{y})$, $Y_{i1} = \phi'_i(\bar{e})(y)$ for $i \in I_{A_0}$, $Y_{i2} = \frac{1}{2}\phi''_i(\bar{e})(y)^2 + \phi'_i(\bar{e})(\hat{y})$ for $i \in I'' \cup \mathcal{J}_2(y, \hat{y})$ and $V(x) = \frac{1}{6}\Phi'''_{I''}(\bar{e})(y)^3 + \Phi''_{I''}(\bar{e})(y, \hat{y}) + \Phi'_{I''}(\bar{e})(x)$. (See the proof of Theorem 5.2 for the definition of $\Phi_{I''}$.)

Arguments similar to those used to prove Proposition 5.1 show that (4.12) yields (6.15) and (6.18). Conditions (6.14) and (6.17) follow from (4.10). The arguments required to show that Condition 4.1 is satisfied are a simple extension of those used in the proof of Theorem 5.2. We note only that $\zeta(x) = \bar{e} + \alpha y + \alpha^2 \hat{y} + \alpha^3 x$ and, because of (6.3), the only result needed from the Appendix is a weakened version of Proposition A.1.

7. Concluding remarks. It has been shown that various higher-order necessary conditions can be obtained systematically from a single result. Specific cases examined in detail include first- and second-order necessary conditions involving directional differentials and first-, second- and third-order necessary conditions for a nonlinear programming problem. With only minor extensions of the arguments used to obtain these results it is possible in both cases to obtain necessary conditions of arbitrary order.

Theorem 5.4, which generalizes [35, Thm. 2.3], contributes in several ways to generalizing the results of [34], [35] and [36]. First, since Theorem 5.4 involves finite- and infinite-dimensional inequality constraints, it allows the handling of both endpoint and state-space inequality constraints. Secondly, the cost criterion ϕ_0 in OP may be nonscalar (e.g., Pareto-type). And thirdly, Theorem 5.4 does not require a normality condition which is difficult to verify in practice. The results of [34] and [35] have been extended in this direction in [13]. Other second-order necessary conditions from the literature requiring such an assumption (e.g., [6] and [15]) can also benefit from this generalization. Finally, the third- and higher-order necessary conditions mentioned above lead to necessary conditions in optimal control theory of still higher order. It is hoped that these conditions will be useful in treating singular optimal control problems for which there is an extensive body of literature containing specialized higher-order necessary conditions (see, e.g., [1], [2], [12], [20], [21] and the references therein).

While this paper was being revised a theory of higher-order optimality conditions appeared in [22]. Although a direct comparison of their results to the present paper is rather complex because of differing assumptions, one interesting aspect of their development appears to be more general. In Condition 4.1 it is assumed that $\{Y_{ir}\}_{r=1}^{m_i} \subset Z_i$ so that Y as given by (4.13) is an element of Z . Their conditions require that $m_i = n$ for all i yet allow for a more general situation in which $Y_{i1}, \dots, Y_{i, m_i}$ satisfy

$$Y_{ir} \in Z_i + \sum_{q=0}^{r-1} \mathbb{R} Y_{iq}, \quad r = 1, \dots, m_i,$$

where $Y_{i0} = \phi_i(\bar{e})$.

Appendix. This section collects together differentiation results needed for the proofs of Theorems 5.1, 5.2 and 6.1. Some of them concern the notion of \tilde{F} -differentiability given in Definitions 5.1 and 5.2 and extend well-known results from the literature. Also included is a restatement of a result from [33].

Let \mathcal{X} be a Banach space with norm $|\cdot|$, \mathcal{Y} be a topological vector space, $A \subset \mathcal{X}$ and $f: A \rightarrow \mathcal{Y}$. When \mathcal{Y} is a Banach space higher-order \tilde{F} -derivatives $f^{(n)}(\bar{x}) \in \mathcal{B}(\mathcal{X}; \mathcal{B}(\mathcal{X}; \dots, \mathcal{B}(\mathcal{X}; \mathcal{Y})) \dots)$, $n \geq 2$, may be defined inductively in the manner of Definition 5.2. Following others (e.g., [11, p. 192]) we view $f^{(n)}(\bar{x})$, $n \geq 2$, as an element of $\mathcal{B}_n(\mathcal{X}; \mathcal{Y})$. Specifically, the values of $f^{(n)}(\bar{x})$ on \mathcal{X}^n are given by $f^{(n)}(\bar{x})(h_1, \dots, h_n) \triangleq (\dots ((f^{(n)}(\bar{x}))(h_1))(h_2)) \dots (h_n)$. For notational simplicity $f^{(n)}(\bar{x})(h)^n \triangleq f^{(n)}(\bar{x})(h, \dots, h)$. The following is an extension of Taylor's Theorem. Its proof is a slight modification of the proof of [11, Thm. 3.6.2].

THEOREM A.1. Let \mathcal{X} and \mathcal{Y} be Banach spaces with the norm on \mathcal{X} denoted by $|\cdot|$, $A \subset \mathcal{X}$ and $f: A \rightarrow \mathcal{Y}$. If f is m -times \tilde{F} -differentiable at $\bar{x} \in A$, where $m \in \mathbb{N}$, and A' is a convex set satisfying $\bar{x} \in A' \subset A$ then

$$(A.1) \quad \lim_{\substack{x \rightarrow \bar{x} \\ x \in A'/\bar{x}}} |x - \bar{x}|^{-m} \left[f(x) - f(\bar{x}) - \sum_{i=1}^m (i!)^{-1} f^{(i)}(\bar{x})(x - \bar{x})^i \right] = 0.$$

Partial \tilde{F} -derivatives are introduced as follows (see [33, p. 168], and [30, pp. 58–59]).

DEFINITION A.1. Let $\mathcal{X}_1, \dots, \mathcal{X}_r$ be Banach spaces, \mathcal{Y} be a topological vector space, $A \subset \mathcal{X} \triangleq \mathcal{X}_1 \times \dots \times \mathcal{X}_r$, $f: A \rightarrow \mathcal{Y}$ and $\bar{x} \triangleq (\bar{x}_1, \dots, \bar{x}_r) \in A$. For $i \in \{1, \dots, r\}$ define the set $A_i(\bar{x}) \triangleq \{x_i \in \mathcal{X}_i: (\bar{x}_1, \dots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \dots, \bar{x}_r) \in A\}$ and the mapping $x_i \rightarrow f_i(x_i): A_i(\bar{x}) \rightarrow \mathcal{Y}$ by $f_i(x_i) \triangleq f(\bar{x}_1, \dots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \dots, \bar{x}_r)$. If $f_i(x_i)$ is \tilde{F} -differentiable at \bar{x}_i then $f_{x_i}(\bar{x}) \triangleq f'_i(\bar{x}_i) \in \mathcal{B}(\mathcal{X}_i; \mathcal{Y})$ is the partial \tilde{F} -derivative of f with respect to x_i at \bar{x} .

Note that ([33, p. 169]) if f is \tilde{F} -differentiable at \bar{x} , then $f_{x_i}(\bar{x})$ exists for all $i \in \{1, \dots, r\}$ and

$$(A.2) \quad f'(\bar{x})(h) = \sum_{i=1}^r f_{x_i}(\bar{x})(h_i), \quad h \triangleq (h_1, \dots, h_r) \in \mathcal{X}_1 \times \dots \times \mathcal{X}_r.$$

If \mathcal{Y} is a Banach space then for $n \in \mathbb{N}$ and $\nu_1, \dots, \nu_n \in \{1, \dots, r\}$ the higher-order \tilde{F} -partial derivative $f_{x_{\nu_1} \dots x_{\nu_n}}(\bar{x}) \in \mathcal{B}(\mathcal{X}_{\nu_n}, \dots, \mathcal{X}_{\nu_1}; \mathcal{Y})$ can be defined inductively. If $\nu_1 = \dots = \nu_n = \bar{\nu}$, then $f_{(x_{\bar{\nu}})^n}(\bar{x}) \triangleq f_{x_{\nu_1} \dots x_{\nu_n}}(\bar{x})$. If $f^{(n)}(\bar{x})$ exists, then the following relation is valid [11, p. 197]

$$(A.3) \quad f^{(n)}(\bar{x})(h)^n = \sum f_{x_{\nu_1} \dots x_{\nu_n}}(\bar{x})(h_{\nu_n}, \dots, h_{\nu_1}), \quad h \triangleq (h_1, \dots, h_r) \in \mathcal{X}_1 \times \dots \times \mathcal{X}_r,$$

where the summation is over all $(\nu_1, \dots, \nu_n) \in \{1, \dots, r\}^n$.

The following proposition concerns a one-parameter expansion of a mapping defined on $\mathbb{P}^n(\beta)$. The result, which follows from Theorem A.1, shows that the remainder term satisfies a uniform convergence condition.

PROPOSITION A.1. Let $\kappa, n \in \mathbb{N}$, $\beta > 0$, \mathcal{Y} be a topological vector space and $f: \mathbb{P}^{\kappa+n}(\beta) \rightarrow \mathcal{Y}$. For $\beta_0 > 0$ satisfying $\sum_{i=1}^n \beta_0^i \leq \beta$ define $F: [0, \beta_0] \times \Delta^\kappa \rightarrow \mathcal{Y}$ by

$$(A.4) \quad F(\alpha, \mu) = \begin{cases} f(\alpha \mu_1, \dots, \alpha \mu_{\kappa+1}), & n = 1, \\ f(\alpha, \alpha^2, \dots, \alpha^{n-1}, \alpha^n \mu_1, \dots, \alpha^n \mu_{\kappa+1}), & n \geq 2. \end{cases}$$

I. If f is continuous at $\alpha = 0$ then

$$(A.5) \quad F(\alpha, \mu) - f(0) \xrightarrow{\alpha \rightarrow 0^+} 0$$

uniformly for $\mu \in \Delta^\kappa$.

II. If f is \tilde{F} -differentiable at $\alpha = 0$, then

$$(A.6) \quad \alpha^{-1}[F(\alpha, \mu) - f(0) - \alpha F_\alpha(0, \mu)] \xrightarrow{\alpha \rightarrow 0^+} 0$$

uniformly for $\mu \in \Delta^\kappa$.

III. If \mathcal{Y} is a Banach space, $n \geq 2$ and f is m -times \tilde{F} -differentiable at $\alpha = 0$, where $m \in \{2, \dots, n\}$, then

$$(A.7) \quad \alpha^{-m} \left[F(\alpha, \mu) - f(0) - \sum_{i=1}^m (i!)^{-1} \alpha^i F_{\alpha^i}(0, \mu) \right] \xrightarrow{\alpha \rightarrow 0^+} 0$$

uniformly for $\mu \in \Delta^\kappa$.

Proof. We prove part III only, since similar arguments apply in the other cases. Since $f^{(m)}(0)$ exists, Theorem A.1 can be applied with $\bar{x} = 0 \in A' = A = \mathbb{P}^{\kappa+n}(\beta)$ to obtain

$$(A.8) \quad \lim_{\substack{\alpha \rightarrow 0 \\ \alpha \in \mathbb{P}^{\kappa+n}(\beta) \setminus \{0\}}} |\alpha|^{-m} \left[f(\alpha) - f(0) - \sum_{i=1}^m (i!)^{-1} f^{(i)}(0)(\alpha)^i \right] = 0.$$

Specialize (A.8) by replacing α by $(\alpha, \alpha^2, \dots, \alpha^{n-1}, \alpha^n \mu_1, \dots, \alpha^n \mu_{\kappa+1})$, where $\alpha \in (0, \beta_0)$ and $\mu \in \Delta^\kappa$. Then (A.8) implies

$$(A.9) \quad \lim_{\alpha \rightarrow 0^+} \alpha^{-m} [F(\alpha, \mu) - f(0) - \theta_m(\alpha, \mu)] = 0$$

uniformly for all $\mu \in \Delta^\kappa$, where

$$(A.10) \quad \theta_m(\alpha, \mu) \triangleq \sum_{i=1}^m (i!)^{-1} f^{(i)}(0)(\alpha, \alpha^2, \dots, \alpha^{n-1}, \alpha^n \mu_1, \dots, \alpha^n \mu_{\kappa+1})^i.$$

Using the chain rule [33, p. 172] to express $F_{\alpha^i}(0, \mu)$ in terms of the partial \tilde{F} -derivatives of f and applying (A.3) to (A.10) it can be shown that

$$(A.11) \quad \lim_{\alpha \rightarrow 0^+} \alpha^{-m} \left[\theta_m(\alpha, \mu) - \sum_{i=1}^m (i!)^{-1} \alpha^i F_{\alpha^i}(0, \mu) \right] = 0$$

uniformly for all $\mu \in \Delta^\kappa$. The desired result (A.7) now follows from (A.9) and (A.11). \square

Remark A.1. The \tilde{F} -partial derivative $F_{\alpha^i}(0, \mu)$ can be written explicitly in terms of the \tilde{F} -partial derivatives of f . For example:

$$\begin{aligned} n = 1: \quad F_\alpha(0, \mu) &= \sum_{i=1}^{\kappa+1} \mu_i f_{\alpha_i}(0); \\ n = 2: \quad F_\alpha(0, \mu) &= f_{\alpha_1}(0), \\ F_{\alpha\alpha}(0, \mu) &= f_{\alpha_1\alpha_1}(0) + 2 \sum_{i=1}^{\kappa+1} \mu_i f_{\alpha_{i+1}}(0); \\ n = 3: \quad F_\alpha(0, \mu) &= f_{\alpha_1}(0), \\ F_{\alpha\alpha}(0, \mu) &= f_{\alpha_1\alpha_1}(0) + 2f_{\alpha_2}(0), \\ F_{\alpha\alpha\alpha}(0, \mu) &= f_{\alpha_1\alpha_1\alpha_1}(0) + 6f_{\alpha_1\alpha_2}(0) + 6 \sum_{i=1}^{\kappa+1} \mu_i f_{\alpha_{i+2}}(0). \end{aligned}$$

The following result follows from [33, Thm. II.3.3].

THEOREM A.2. Let \mathcal{X} be a vector space, \mathcal{Y} be a topological vector space, $A \subset \mathcal{X}$, $F: A \rightarrow \mathcal{Y}$, $\bar{x} \in A$, $\nu \in \mathbb{N}$ and $X \triangleq \text{co}\{x_1, \dots, x_\nu\} \subset \mathcal{X}$. Suppose there exists $\beta > 0$ such that $\bar{x} + \alpha X \subset A$, $\alpha \in [0, \beta)$, and define $f: \mathbb{P}^\nu(\beta) \rightarrow \mathcal{Y}$ by $f(\alpha) = F(\bar{x} + \sum_{i=1}^\nu \alpha_i x_i)$. If $f'(0)$ exists then $DF(\bar{x}; x)$ exists for all $x \in \text{cone } X$ and the mapping $x \rightarrow DF(\bar{x}; x): \text{cone } X \rightarrow \mathcal{Y}$ is positively homogeneous and affine.

Note that $f_{\alpha_i}(0) = DF(\bar{x}; x_i)$. Also, if $f''(0)$ exists then $f_{\alpha_i\alpha_j}(0) = D^2F(\bar{x}; x_i, x_j)$.

Acknowledgment. I wish to thank Elmer G. Gilbert for providing numerous helpful suggestions throughout the course of this work.

REFERENCES

- [1] A. A. AGRACHEV AND R. V. GAMKRELIDZE, *A second-order optimality principle for a time-optimal problem*, Math. USSR-Sb., 29 (1976), pp. 547–576.
- [2] D. J. BELL AND D. H. JACOBSON, *Singular Optimal Control Problems*, Academic Press, London, 1975.
- [3] A. BEN-ISRAEL, A. BEN-TAL AND S. ZLOBEC, *Optimality in Nonlinear Programming*, John Wiley, New York, 1981.
- [4] A. BEN-TAL, *Second-order and related extremality conditions in nonlinear programming*, J. Optim. Theory Appl., 31 (1980), pp. 143–165.
- [5] 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.
- [6] D. S. BERNSTEIN AND E. G. GILBERT, *Optimal periodic control: The π test revisited*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 673–684.
- [7] M. CANON, G. CULLUM AND E. POLAK, *Theory of Optimal Control and Mathematical Programming*, McGraw-Hill, New York, 1970.
- [8] A. J. DUBOVICKII AND A. A. MILYUTIN, *Second variations in extremal problems with constraints*, Soviet Math. Dokl., 6 (1965), pp. 12–16.
- [9] A. FIACCO AND G. MCCORMICK, *Nonlinear Programming*, John Wiley, New York, 1968.
- [10] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.
- [11] T. M. FLETT, *Differential Analysis*, Cambridge Univ. Press, Cambridge, 1980.
- [12] R. GABASOV AND F. M. KIRILLOVA, *High-order necessary conditions for optimality*, this Journal, 10 (1972), pp. 127–168.
- [13] E. G. GILBERT AND D. S. BERNSTEIN, *Second-order necessary conditions in optimal control: accessory-problem results without normality*, J. Optim. Theory Appl., 41 (1983), pp. 75–106.
- [14] B. GOLLAN, *High-order necessary conditions for an abstract optimization problem*, Math. Programming Stud., 14 (1981), pp. 69–76.
- [15] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1964.
- [16] K. H. HOFFMANN AND H. J. KORNSTAEDT, *High-order necessary conditions in abstract mathematical programming*, J. Optim. Theory Appl., 26 (1978), pp. 533–568.
- [17] R. B. HOLMES, *Geometric Functional Analysis and its Applications*, Springer-Verlag, New York, 1975.
- [18] 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.
- [19] V. L. KLEE, *Separation and support properties of convex sets—A survey*, in Control Theory and the Calculus of Variations, A. V. Balakrishnan, ed., Academic Press, New York, 1969.
- [20] H. W. KNOBLOCH, *Higher-order Necessary Conditions in Optimal Control Theory*, Springer-Verlag, New York, 1981.
- [21] A. J. KRENER, *The high-order maximal principle and its application to singular extremals*, this Journal, 15 (1977), pp. 256–293.
- [22] F. LEMPIO AND J. ZOWE, *High-order optimality conditions*, in Modern Applied Mathematics—Optimization and Operations Research, B. Korte, ed., North-Holland, Amsterdam, 1982, pp. 147–193.
- [23] E. S. LEVITIN, A. A. MILJUTIN AND N. P. OSMOLOVSKII, *On conditions for a local minimum in a problem with constraints*, in Mathematical Economics and Functional Analysis, B. S. Mitjagin, ed., Nauka, Moscow, 1974, pp. 139–202. (In Russian.)
- [24] H. MAURER AND J. ZOWE, *First- and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.
- [25] G. P. MCCORMICK, *Second-order conditions for constrained minima*, SIAM J. Appl. Math., 15 (1967), pp. 641–652.
- [26] ———, *Optimality criteria in nonlinear programming*, in Nonlinear Programming, SIAM-AMS Proceedings, Vol. 9, R. W. Cottle and C. E. Lemke, eds., American Mathematical Society, Providence, RI, 1976, pp. 27–38.
- [27] E. J. MCSHANE, *Sufficient conditions for a weak relative minimum in the problem of Bolza*, Trans. Amer. Math. Soc., 52 (1942), pp. 344–379.
- [28] E. J. MESSERLI AND E. POLAK, *On second-order necessary conditions for optimality*, this Journal, 7 (1969), pp. 272–291.
- [29] L. W. NEUSTADT, *A general theory of extremals*, J. Comput. System Sci., 3 (1969), pp. 57–92.
- [30] ———, *Optimization—A Theory of Necessary Conditions*, Princeton Univ. Press, Princeton, NJ, 1976.
- [31] F. A. VALENTINE, *Convex Sets*, McGraw-Hill, New York, 1964.

- [32] C. VIRSAN, *Necessary conditions of extremality of high order*, *Revue Roumaine de Mathématiques Pures et Appliquées*, 28 (1973), pp. 591–611.
- [33] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [34] ———, *A second-order Lagrangian condition for restricted control problems*, *J. Optim. Theory Appl.*, 24 (1978), pp. 475–483.
- [35] ———, *A second-order condition that strengthens Pontryagin's maximum principle*, *J. Differential Equations*, 28 (1978), pp. 284–307.
- [36] ———, *A hybrid relaxed-Lagrangian second-order condition for minimum*, *Differential Games and Control Theory*, Proc. 3rd Kingston Conference, Part A, Marcel Dekker, New York, 1979, pp. 77–94.

STABILITY IN MATHEMATICAL PROGRAMMING WITH NONDIFFERENTIABLE DATA*

ALFRED AUSLENDER†

Abstract. We study the variation under perturbation of isolated local minimizers of a nonlinear and nondifferentiable optimization problem. For this we extend to the Lipschitzian case a fundamental result concerning regular points. Then we introduce the notion of lower second-order directional derivative, from which we obtain a second-order sufficiency theorem. These two results are finally used for obtaining bounds for the variations of some classes of isolated minimizers.

Key words. mathematical programming, stability theory, second-order directional derivative, sufficient conditions, locally Lipschitzian functions

Introduction. Let \mathbb{R}^N be the usual vector space of N -tuples with the usual inner product denoted by (\cdot, \cdot) , and let m, p be positive integers. We denote by $\langle 1, p \rangle$ the set of integers included in $[1, p]$. Let W be an open set in \mathbb{R}^q and let $f_i, g_j, i \in \langle 0, m \rangle, j \in \langle 1, p \rangle$ be real-valued locally Lipschitzian functions defined on $\mathbb{R}^N \times W$. Let

$$C(w) = \{x \in \mathbb{R}^N : f_i(x, w) \leq 0, \forall i \in \langle 1, m \rangle, g_j(x, w) = 0 \forall j \in \langle 1, p \rangle\}.$$

The purpose of this paper is to study the behavior of some classes of stationary points of the optimization problem

$$P(w): \quad \min f_0(x, w) \text{ subject to } x \in C(w),$$

when the parameter w belongs to a neighborhood of a point $\bar{w} \in W$. More precisely we want to generalize for nondifferentiable data some results obtained by Robinson in [22]. It follows that the object of this paper is not the same as in the other studies in stability theory with nondifferentiable data: Rockafellar [24], [25], Golan [13] and others. The first part of this paper deals with regular points as they were defined by Robinson [21] for the differentiable case and by Ioffe in [17]. The main result of this part is the following:

A point $\bar{x} \in C(\bar{w})$ is regular if an extended Mangasarian-Fromovitz condition is satisfied at the point.

In order to obtain stability results about stationary points in the differentiable case it is necessary to use second-order sufficient conditions.

Then § 2 is devoted to second-order sufficiency theorems for the locally Lipschitzian case. This leads, in particular, to the notion of a lower second-order directional derivative for locally Lipschitzian functions, a notion which seems to be interesting when used with lower- C^2 functions, a new class of functions introduced by Rockafellar [23].

Second-order sufficient conditions for certain classes of nondifferentiable functions were given recently by several authors, for example, Ioffe [18], Fletcher and Watson [12], Demjanov and Malozemov [10], Ben-Tal and Zowe [5], Spingarn [26] and Chaney [6], [7]. All these authors with the exception of Chaney, study only certain particular classes of locally Lipschitzian functions.

Finally, as an application of §§ 1, 2 we give in § 3 bounds for the distance between local minimizers of problem $P(w)$ and the local minimizers of the original problem.

* Received by the editors June 29, 1981, and in revised form October 10, 1982.

† Département de Mathématiques Appliquées, Université de Clermont II, B.P. 45, 63170 Aubiere, France.

1. Regular points for locally Lipschitzian functions. Let E_{f_j} be the set in $\mathbb{R}^N \times W$ of points where f_j is differentiable. From Rademacher's theorem the complement of E_{f_j} is a set of measure zero. Let us denote by $\partial_x^* f_j(x, w)$ the partial generalized gradient of f_j at (x, w) introduced by Hiriart-Urruty in [16]

$$\partial_x^* f_j(x, w) = \text{co} \left\{ \lim_{(x_i, w_i) \rightarrow (x, w)} \nabla_x f_j(x_i, w_i) \mid (x_i, w_i) \in E_{f_j} \right\}.$$

Here $\nabla_x f_j(x_i, w_i)$ denotes the gradient of the function $f_j(\cdot, w_i)$ at x_i and $\text{co}(A)$ the convex hull of A . This nonempty compact convex set may not coincide with the generalized gradient of $f_j(\cdot, w)$ at x introduced by Clarke and which we denote by $\partial_x f_j(x, w)$

$$\partial_x f_j(x, w) = \text{co} \left\{ \lim_{x_i \rightarrow x} \nabla_x f_j(x_i, w) \mid x_i \in E_{f_j}(\cdot, w) \right\}.$$

In this formula $E_{f_j}(\cdot, w)$ is the set of points in \mathbb{R}^N where $f_j(\cdot, w)$ is differentiable.

The reason we sometimes use $\partial_x^* f_j(\cdot, \cdot)$ instead of $\partial_x f_j(\cdot, \cdot)$ is that the former is a multi-valued map which is upper semi-continuous in both variables, which is not always the case for $\partial_x f_j(\cdot, \cdot)$. Let us remark that we always have

$$(1.0) \quad \partial_x f_j(\cdot, \cdot) \subset \partial_x^* f_j(\cdot, \cdot), \quad \partial_x g_i(\cdot, \cdot) \subset \partial_x^* g_i(\cdot, \cdot) \quad \forall i, j.$$

This is a consequence of the following inequality:

$$\limsup_{\substack{x_i \rightarrow x \\ \lambda \rightarrow 0^+}} \frac{f_j(x_i + \lambda v, w) - f_j(x_i, w)}{\lambda} \leq \limsup_{\substack{x_i \rightarrow x \\ w_i \rightarrow w \\ \lambda \rightarrow 0^+}} \frac{f_j(x_i + \lambda v, w_i + \lambda 0) - f_j(x_i, w_i)}{\lambda}$$

and of the fundamental theorem of Clarke ([9, Prop. 5, see also (2.0) below).

Now we give a condition that generalizes the Mangasarian-Fromovitz condition to the nondifferentiable case and which is close to the condition given by Hiriart-Urruty in [16].

DEFINITION 1.1. Let $\bar{x} \in C(\bar{w})$ and let $I(\bar{x}, \bar{w}) = \{i \in \langle 1, m \rangle : f_i(\bar{x}, \bar{w}) = 0\}$. The point \bar{x} is said to satisfy the *extended Mangasarian-Fromovitz condition for the set* $C(\bar{w})$ if:

1) For each set $\{c_i, d_j \mid i \in I(\bar{x}, \bar{w}), j \in \langle 1, p \rangle\}$ with $c_i \in \partial_x^* f_i(\bar{x}, \bar{w})$, $d_j \in \partial_x^* g_j(\bar{x}, \bar{w})$ there exists a vector h such that

$$(c_i, h) < 0 \quad \forall i \in I(\bar{x}, \bar{w}), \quad (d_j, h) = 0 \quad \forall j \in \langle 1, p \rangle.$$

2) Each set $\{d_j \mid j \in \langle 1, p \rangle\}$ with $d_j \in \partial_x^* g_j(\bar{x}, \bar{w})$ is composed of linearly independent vectors.

The following definition of regularity is an extension of that given by Robinson in the differentiable case, and given by Ioffe in the nonperturbed case.

DEFINITION 1.2. The point $\bar{x} \in C(\bar{w})$ is called *regular at* \bar{w} if there exists an open neighborhood V of \bar{x} , an open neighborhood W^* of \bar{w} ($W^* \subset W$) and a constant c such that for each $x \in V$, each $w \in W^*$ we have

$$(1.1) \quad d_{C(w)}(x) \leq c \max (f_i^+(x, w), |g_j(x, w)| \mid i \in \langle 1, m \rangle, j \in \langle 1, p \rangle).$$

In this formula $d_{C(w)}(x)$ denotes the distance of x to $C(w)$ and $f_i^+ = \max (f_i, 0)$.

THEOREM 1.1. Suppose that $\bar{x} \in C(\bar{w})$ satisfies the extended Mangasarian-Fromovitz condition for the set $C(\bar{w})$ and that $C(w)$ is nonempty for each $w \in W_0$, where W_0 is an open neighborhood of \bar{w} in W . Then \bar{x} is regular at \bar{w} .

Proof. a) Let $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m)$ with $\alpha_i = \pm 1$, set $I = I(\bar{x}, \bar{w})$ and

$$T_\alpha(x, w) = \text{co}(\partial_x^* f_i(x, w), \alpha_j \partial_x^* g_j(x, w), i \in I, j \in \langle 1, p \rangle).$$

We claim that for each $\alpha \notin T_\alpha(\bar{x}, \bar{w})$. Indeed in the contrary case there would exist $\alpha, \lambda_i \geq 0, c_i \in \partial_x^* f_i(\bar{x}, \bar{w})$ with $i \in I, \rho_j \geq 0, d_j \in \partial_x^* g_j(\bar{x}, \bar{w}), j \in \langle 1, p \rangle$ such that

$$0 = \sum_{i \in I(\bar{x}, \bar{w})} \lambda_i c_i + \sum_{j=1}^p \alpha_j \rho_j d_j \quad \text{with} \quad \sum_{i \in I} \lambda_i + \sum_{j=1}^p \rho_j = 1,$$

which contradicts the extended Mangasarian–Fromovitz condition. Let $B(\bar{x}, \bar{\rho})$ be the closed ball centered at \bar{x} with radius $\bar{\rho}$. Since the convex hull and the union of a finite number of multi-valued maps which are upper semi-continuous at a point is again upper semi-continuous at this point $\bigcup_\alpha T_\alpha$ is upper semi-continuous at (\bar{x}, \bar{w}) , and there exists $\bar{\rho} > 0$ such that $B(\bar{w}, \bar{\rho}) \subset W_0$ and such that

$$0 \notin \bigcup_\alpha T_\alpha(x, w) \quad \forall x \in B(\bar{x}, \bar{\rho}) \quad \forall w \in B(\bar{w}, \bar{\rho}).$$

Consequently there exists $c > 0$ such that

$$(1.2) \quad \|y\| \geq \frac{1}{c} \quad \forall y \in T_\alpha(x, w) \quad \forall x \in B(\bar{x}, \bar{\rho}) \quad \forall w \in B(\bar{w}, \bar{\rho}) \quad \forall \alpha.$$

Note that by $\|\cdot\|$, we mean the usual Euclidean norm.

b) Now we set

$$(1.3) \quad F(x, w) = \max(f_i^+(x, w), |g_j(x, w)| \mid i \in \langle 1, m \rangle, j \in \langle 1, p \rangle)$$

so that $F(x, w)$ is always ≥ 0 . Moreover, since the functions f_i are continuous we can also choose $\bar{\rho}$ such that for $x \in B(\bar{x}, \bar{\rho}), w \in B(\bar{w}, \bar{\rho})$ we have

$$(1.4) \quad \max_{i \notin I} f_i(x, w) < 0, \\ F(x, w) = \max(f_i^+(x, w), |g_j(x, w)| \mid i \in I, j \in \langle 1, p \rangle).$$

Now for $\rho \in [0, \bar{\rho}]$, $w \in B(\bar{w}, \bar{\rho})$ set

$$H(\rho, w) = \max c\{F(x, w) \mid x \in B(\bar{x}, \rho)\}.$$

Since $F(\bar{x}, \bar{w}) = 0$ we have $H(0, \bar{w}) = 0$. Now by the “maximum theorem” H is continuous. It follows then that there exists $\rho^* < \bar{\rho}$ such that

$$(1.5) \quad H(\rho, w) \leq \frac{\bar{\rho}}{2} \quad \forall \rho \in [0, \rho^*] \quad \forall w \in B(\bar{w}, \rho^*).$$

c) Now we claim that the theorem is true with $V = B(\bar{x}, \rho^*/4)$ and $W^* = B(\bar{w}, \rho^*)$. Indeed, in the contrary case, there would exist $u \in B(\bar{x}, \rho^*/4), w \in B(\bar{w}, \rho^*)$ such that

$$d_{C(w)}(u) > cF(u, w),$$

and there would exist $t \in]1, \frac{3}{2}[$ such that

$$d_{C(w)}(u) > tF(u, w)c.$$

Let $\gamma = tF(u, w)c$. Since $F(u, w) \geq 0$, $d_{C(w)}(u) > 0$ and then $F(u, w) > 0$, so that it follows that $\gamma > 0$. Moreover we have from (1.5) that

$$(1.6) \quad \gamma \leq \frac{t\bar{\rho}}{2} < \frac{3}{4}\bar{\rho}.$$

Then since $\gamma/t = F(u, w)c$ and since $F(v, w) \geq 0$ we have

$$F(u, w) \leq \inf (F(v, w) | v \in \mathbb{R}^N) + \frac{\gamma}{tc}.$$

From [11, Chapt. I, Thm. 6.1 and (6.11)] there then exists $u_\gamma(w)$ such that

$$(1.7) \quad \|u - u_\gamma(w)\| \leq \gamma,$$

$$(1.8) \quad u_\gamma(w) \text{ minimizes the functional } v \rightarrow F(v, w) + \frac{1}{tc} \|v - u_\gamma(w)\| \text{ on } \mathbb{R}^N.$$

From (1.6) and (1.7) it follows that $\|u_\gamma(w) - \bar{x}\| < \bar{\rho}$. Since $d_{C(w)}(u) > \gamma$ it follows from (1.7) that $u_\gamma(w) \notin C(w)$, and then

$$F(u_\gamma(w), w) > 0.$$

Set $J(w) = \{j \in \langle 1, p \rangle : |g_j(u_\gamma(w), w)| = F(u_\gamma(w), w)\}$. This set of indices may be empty. Now let

$$\varepsilon_j(w) = \begin{cases} 1 & \text{if } g_j(u_\gamma(w), w) > 0, \\ -1 & \text{if } g_j(u_\gamma(w), w) < 0. \end{cases}$$

Since $F(u_\gamma(w), w) > 0$ then there exists a neighborhood $X(w)$ of $u_\gamma(w)$ such that for each $x \in X(w)$ we have

$$F(x, w) = \max \{f_i(x, w), g_j(x, w)\varepsilon_j(w) | i \in I, j \in J(w)\}$$

from which it follows, when using (1.0), that

$$\partial_x F(u_\gamma(w), w) \in \bigcup_{\alpha} T_{\alpha}(u_\gamma(w), w).$$

Now from relation (1.8) it follows that there exists $y_\gamma(w)$ such that

$$\|y_\gamma(w)\| \leq 1, \quad \frac{1}{tc} y_\gamma(w) \in \partial_x F(u_\gamma(w), w).$$

Then from (1.2) we have

$$\frac{1}{c} > \frac{1}{tc} \geq \left\| \frac{1}{tc} y_\gamma(w) \right\| \geq \frac{1}{c}$$

which is a contradiction.

Remark 1.1. In the differentiable case Theorem 1.1 was given by Robinson in [21]. When w is a fixed parameter a sufficient condition to obtain regularity was given in [17] by Ioffe. But even in the differentiable case Ioffe did not obtain a regularity theorem under the classical Mangasarian–Fromovitz condition. Regularity theorems were only obtained as corollaries in the linear case (Hoffman estimates) or for equality constraints. In any event, his way of proving his main theorem was quite different than the one used by Robinson. The proof given by Ioffe was based upon the variational principle of Ekeland, and this idea is used again in our proof.

Remark 1.2. Let $L(x, \lambda, \mu, w) = f_0(x, w) + \sum_{i \in I(\bar{x}, \bar{w})} \lambda_i f_i(x, w) + \sum_{j=1}^m \mu_j g_j(x, w)$ and let $\partial_x L(x, \lambda, \mu, w)$ denote the generalized gradient of $L(\cdot, \lambda, \mu, w)$ at x . Now if

$\bar{x} \in C(\bar{w})$ satisfies the extended Mangasarian–Fromovitz condition at \bar{w} for the set $C(\bar{w})$ and if \bar{x} is a local minimum of $f_0(\cdot, \bar{w})$ on $C(\bar{w})$ then by [16, Thm. 4.2] the set $\Omega(\bar{x}, \bar{w})$, defined by

$$\Omega(\bar{x}, \bar{w}) = \{(\lambda_i, \mu_i): 0 \in \partial_x L(\bar{x}, \lambda, \mu, \bar{w}), \lambda_i \geq 0 \forall i \in I(\bar{x}, \bar{w})\},$$

is a nonempty set.

Remark 1.3. One can ask if it is really necessary to assume in Theorem 1.1 that $C(w) \neq \emptyset$ for all w in a neighborhood W_0 . Could not this be shown to follow from the extended Mangasarian–Fromovitz condition? In [3, Thm. 2.1] this was proved for vertical perturbations

$$f_j(x, w) = f_j(x) + w_j, \quad g_i(x, w) = g_i(x) + w_{i+m}$$

with the additional assumption that the functions g_i are continuously differentiable. In any event, the question remains open for the general case.

2. Second-order sufficient conditions.

2.1. Preliminaries. When f is a real-valued locally Lipschitzian function defined on \mathbb{R}^N , recall that f is almost everywhere differentiable, that $\partial f(x)$ is the generalized gradient of f at x defined as the convex hull of the set of limits of the form $\lim \nabla f(x + h_i)$, where $h_i \rightarrow 0$ as $i \rightarrow +\infty$, and that the generalized directional derivative $f^0(x; v)$ defined by

$$(2.0) \quad f^0(x; v) = \limsup_{\substack{\lambda \rightarrow 0^+ \\ h \rightarrow 0}} \frac{f(x + h + \lambda v) - f(x + h)}{\lambda}$$

satisfies

$$(2.1) \quad f^0(x; v) = \max \{ \langle v, z \rangle \mid z \in \partial f(x) \}.$$

For the following we shall also use the upper and the lower Dini directional derivatives

$$D'_+ f(x; v) = \limsup_{\lambda \rightarrow 0^+} \frac{f(x + \lambda v) - f(x)}{\lambda},$$

$$D'_- f(x; v) = \liminf_{\lambda \rightarrow 0^+} \frac{f(x + \lambda v) - f(x)}{\lambda}.$$

Since f is locally Lipschitzian these quantities are finite. If $D'_+ f(x; v) = D'_- f(x; v)$ then the common value denoted by $f'(x; v)$ is the usual directional derivative of f at x in direction v . Let φ be a function defined on $\mathbb{R}^N \setminus \{0\}$ with values in \mathbb{R} . For $d \neq 0$ one can define

$$\varphi_1(d) = \liminf_{\substack{d \\ h \rightarrow 0}} \varphi(h), \quad \varphi_2(d) = \limsup_{\substack{d \\ h \rightarrow 0}} \varphi(h)$$

as follows:

$$\liminf_{\substack{d \\ h \rightarrow 0}} \varphi(h) = \sup_{\varepsilon > 0} \left(\inf \left\{ \varphi(h) \mid \left\| \frac{h}{\|h\|} - \frac{d}{\|d\|} \right\| \leq \varepsilon, 0 < \|h\| \leq \varepsilon \right\} \right),$$

$$\limsup_{\substack{d \\ h \rightarrow 0}} \varphi(h) = \inf_{\varepsilon > 0} \left(\sup \left\{ \varphi(h) \mid \left\| \frac{h}{\|h\|} - \frac{d}{\|d\|} \right\| \leq \varepsilon, 0 < \|h\| \leq \varepsilon \right\} \right).$$

Let us now introduce the notion of a lower second-order directional derivative $f''_-(x; d; d)$ of f at x in direction $d \neq 0$ by setting

$$(2.2) \quad f''_-(x; d; d) = 2 \liminf_{\substack{d \\ h \rightarrow 0}} \frac{1}{\|h\|} \left[\frac{f(x+h) - f(x)}{\|h\|} - D'_-f\left(x; \frac{h}{\|h\|}\right) \right].$$

Remark 2.1. $f''_-(x; \cdot; \cdot)$ is lower-semicontinuous and

$$f''_-(x; d; d) = f''_-(x; \lambda d; \lambda d) \quad \forall \lambda \neq 0 \quad \forall d \neq 0.$$

Remark 2.2. If f is twice differentiable at x then

$$f''_-(x; d; d) = \left(\nabla^2 f(x) \frac{d}{\|d\|}, \frac{d}{\|d\|} \right),$$

where $\nabla^2 f(x)$ is the Hessian matrix of f at x .

In § 2.3 we shall give some additional properties of this notion in the case of lower- C^2 functions, but before that let us use this notion for obtaining second-order sufficient conditions.

2.2 Second-order sufficient conditions. Let C be a nonempty closed set in \mathbb{R}^N and P the optimization problem

$$P: \inf (f(x) | x \in C).$$

DEFINITION 2.1. A point $\bar{x} \in C$ is said to be an *isolated local minimum with order i* ($i = 1$ or 2) of problem P if there exists a real $m > 0$ and a neighborhood V of \bar{x} such that

$$(2.3) \quad f(x) > f(\bar{x}) + \frac{1}{2}m\|x - \bar{x}\|^i \quad \forall x \in V \cap C, \quad x \neq \bar{x}.$$

For the following we shall assume that $\bar{x} \in C$; then let us denote by $T(C; \bar{x})$ the usual tangent cone of C at \bar{x} , that is

$$T(C; \bar{x}) = \{d \in \mathbb{R}^N : \exists \rho_n \downarrow 0, d_n \rightarrow d \text{ with } \bar{x} + \rho_n d_n \in C \text{ for all } n\}.$$

Recall that $T(C; \bar{x}) = \mathbb{R}^N$ if $C = \mathbb{R}^N$. In order to state the results, set

$$(2.4) \quad L_+(\bar{x}) = \{d \neq 0 : D'_+f(\bar{x}; d) \leq 0\}, \quad L(\bar{x}) = \{d \neq 0 : D'_-f(\bar{x}; d) \leq 0\},$$

$$(2.5) \quad K_+(\bar{x}) = T(C; \bar{x}) \cap L_+(\bar{x}), \quad K(\bar{x}) = T(C; \bar{x}) \cap L(\bar{x})$$

and set $\delta(\cdot | C)$ to be the usual indicator function of C

$$\delta(u | C) = \begin{cases} 0 & \text{if } u \in C, \\ +\infty & \text{if } u \notin C. \end{cases}$$

For functions whose directional derivatives exist, the sets $K_+(\bar{x})$ and $K(\bar{x})$ coincide and the following proposition then gives a characterization of isolated local minimums with order i of problem P .

PROPOSITION 2.1. a) Let \bar{x} be an isolated local minimum with order i of problem P . If $i = 1$ then $K_+(\bar{x})$ is empty, else if $i = 2$ we have

$$(2.6) \quad f^*(\bar{x}; d) > 0 \quad \forall d \in \mathbb{R}^N,$$

where f^* is defined by

$$(2.7) \quad f^*(\bar{x}; d) = 2 \liminf_{\substack{d \\ h \rightarrow 0}} \left[\frac{1}{\|h\|} \left[\frac{f(\bar{x}+h) - f(\bar{x})}{\|h\|} \right] + \delta(\bar{x}+h | C) \right].$$

b) Conversely if $K(\bar{x})$ is empty then \bar{x} is an isolated local minimum with order 1 of problem P. If $K(\bar{x})$ is nonempty and if we have

$$(2.8) \quad f^*(\bar{x}; d) > 0 \quad \forall d \in K(\bar{x}),$$

then \bar{x} is an isolated local minimum with order 2 of problem P.

Proof. A) a) Suppose that \bar{x} is an isolated local minimum with order 1 and that $K_+(\bar{x})$ is nonempty. Then there would exist $d \in L_+(\bar{x})$ with $d \neq 0$, and $\rho_n \downarrow 0$, $d_n \rightarrow d$ with $\bar{x} + \rho_n d_n \in C$ for all n . Since (2.3) is satisfied we have for n sufficiently large

$$\frac{f(\bar{x} + \rho_n d_n) - f(\bar{x} + \rho_n d)}{\rho_n \|d_n\|} + \frac{f(\bar{x} + \rho_n d) - f(\bar{x})}{\rho_n \|d_n\|} \geq \frac{m}{2}.$$

Since f is locally Lipschitzian, as $n \rightarrow \infty$, we obtain

$$D'_+ f(\bar{x}; d) \geq \frac{m}{2} \|d\| > 0,$$

which contradicts (2.4).

b) If \bar{x} is an isolated local minimum with order 2 then (2.6) is obtained immediately from (2.3).

B) Suppose now that (2.3) is not satisfied for some i ($i = 1$ or 2). Then there would exist a sequence of positive reals $\{m_j\}$ converging to 0, a sequence $\{x_j\}$ converging to \bar{x} with $x_j \neq \bar{x}$ such that

$$(2.9) \quad f(x_j) \leq f(\bar{x}) + \frac{1}{2} m_j \|x_j - \bar{x}\|^i, \quad x_j \in C.$$

Let $v_j = (x_j - \bar{x})/\|x_j - \bar{x}\|$; without loss of generality, we can suppose that the sequence $\{v_j\}$ converges to some v , with $\|v\| = 1$. Let $\alpha_j = \|x_j - \bar{x}\|$ it follows from (2.9) that

$$(2.10) \quad \bar{x} + \alpha_j v_j \in C, \quad \alpha_j \rightarrow 0^+,$$

$$(2.11) \quad \frac{f(\bar{x} + \alpha_j v_j) - f(\bar{x} + \alpha_j v)}{\alpha_j} + \frac{f(\bar{x} + \alpha_j v) - f(\bar{x})}{\alpha_j} \leq \frac{1}{2} m_j \alpha_j^{i-1}.$$

From (2.10) it follows that $v \in T(C; \bar{x})$ and from (2.11), since f is locally Lipschitzian, as $j \rightarrow \infty$ we obtain

$$D'_- f(\bar{x}; v) \leq 0.$$

Finally, v belongs to $K(\bar{x})$. This implies in particular, that (2.3) is satisfied for $i = 1$ when $K(\bar{x})$ is empty. Now suppose that $K(\bar{x})$ is nonempty and that (2.3) is not satisfied for $i = 2$. Then from (2.9) we obtain

$$(2.12) \quad \frac{f(x_j) - f(\bar{x})}{\|x_j - \bar{x}\|^2} + \delta(x_j | C) \leq \frac{1}{2} m_j.$$

Passing to the limit as $j \rightarrow \infty$ it follows that

$$f^*(x; v) \leq 0,$$

which contradicts (2.8).

In general it is not easy to compute $f^*(\bar{x}; d)$. In order to obtain (2.8) one can try to obtain lower bounds of $f^*(\bar{x}; \cdot)$ which are strictly positive on $K(\bar{x})$. Consider, for example, the unconstrained case $C = \mathbb{R}^N$; then $T(C, \bar{x}) = \mathbb{R}^N$, $K(\bar{x}) = L(\bar{x})$, $\delta(\cdot | C) = 0$ and an obvious necessary condition for a point x to be a local minimum is that

$$(2.13) \quad D'f(x; d) \geq 0 \quad \forall d \in \mathbb{R}^N.$$

Points that satisfy (2.13) will be called stationary points for f . Let \bar{x} be such a point, then it follows from (2.13) that

$$f^*(\bar{x}; d) \geq f''_-(\bar{x}; d; d) \quad \forall d \in \mathbb{R}^N.$$

Then we easily obtain:

COROLLARY 2.2. *Let $C = \mathbb{R}^N$. Suppose that \bar{x} is a stationary point, that $L(\bar{x})$ is nonempty and that*

$$(2.14) \quad f''_-(\bar{x}; d; d) > 0 \quad \forall d \in L(\bar{x}).$$

Then \bar{x} is an isolated local minimum with order 2 of problem P.

Remark 2.3. When f is twice continuously differentiable at \bar{x} Corollary 2.2 is a generalization of the standard second-order sufficiency theorem.

To show that assumption (2.14) is "suitable" we shall prove now that under the conditions given by Demjanov and Malozemov [10] for the discrete minimax case (2.13) and (2.14) are satisfied. Let us first recall that a locally Lipschitzian function $f: \mathbb{R}^N \rightarrow \mathbb{R}$ is subdifferentially regular if for every $x \in \mathbb{R}^N$, $v \in \mathbb{R}^N$ the ordinary directional derivative $f'(x; v)$ exists and satisfies

$$f'(x; v) = f^0(x; v).$$

For such functions, introduced by Clarke [8], (2.13) is equivalent to

$$0 \in \partial f(x)$$

and for such points we have

$$L(x) = \{d \neq 0: f'(x; d) = 0\}.$$

Let now t_i , $i \in \langle 1, m \rangle$ be real-valued functions defined on \mathbb{R}^N and twice continuously differentiable, and let

$$t(x) = \max (t_i(x) | i \in \langle 1, m \rangle).$$

Set $I(x) = \{i: t_i(x) = t(x)\}$; then it is well known that t is subdifferentially regular and that

$$t'(x; d) = \max ((\nabla t_i(x), d) | i \in I(x)).$$

Furthermore:

COROLLARY 2.3. *Suppose (assumptions of Demjanov and Malozemov [10]) that $0 \in \partial t(\bar{x})$, that $L(\bar{x})$ is nonempty and that for some $\gamma > 0$, $\rho > 0$ we have*

$$(2.15) \quad \min_{d \in G_\gamma} \max_{i \in I(\bar{x}; d)} (\nabla^2 t_i(\bar{x})d, d) \geq \rho,$$

where

$$I(\bar{x}; d) = \{i \in I(\bar{x}): (\nabla t_i(\bar{x}), d) = t'(\bar{x}; d)\},$$

$$G_\gamma = \{d: \|d\| = 1, 0 \leq t'(\bar{x}; d) \leq \gamma\}.$$

Then (2.14) is satisfied.

Proof. Since the functions t_j are twice continuously differentiable we have for each j

$$(2.16) \quad t_j(\bar{x} + h) = t_j(\bar{x}) + (\nabla t_j(\bar{x}), h) + \frac{1}{2}(\nabla^2 t_j(\bar{x})h, h) + \|h\|^2 \varepsilon_j(\bar{x}, h),$$

where $\lim_{h \rightarrow 0} \varepsilon_j(\bar{x}, h) = 0$.

Let $I(\bar{x}; h; h) = \{i \in I(\bar{x}; h) : (\nabla^2 t_i(\bar{x})h, h) = \max_{j \in I(\bar{x}; h)} (\nabla^2 t_j(\bar{x})h, h)\}$. Since $t(\bar{x} + h) \geq \max_{j \in I(\bar{x}; h; h)} (t_j(\bar{x} + h))$ and since $t_j(\bar{x}) + (\nabla t_j(\bar{x}), h) + \frac{1}{2}(\nabla^2 t_j(\bar{x})h, h)$ is constant for $j \in I(\bar{x}; h; h)$, it follows from (2.16) that for $j \in I(\bar{x}; h; h)$ we have

$$(2.17) \quad \frac{1}{\|h\|} \left(\frac{t(\bar{x} + h) - t(\bar{x})}{\|h\|} - t' \left(\bar{x}; \frac{h}{\|h\|} \right) \right) \geq \frac{1}{2} \left(\nabla^2 t_j(\bar{x}) \frac{h}{\|h\|}, \frac{h}{\|h\|} \right) + \min_{j \in \langle 1, m \rangle} \varepsilon_j(\bar{x}, h).$$

Since $L(\bar{x}) \neq \emptyset$ let now $d \neq 0$ such that

$$t'(\bar{x}; d) = 0.$$

Since $0 \in \partial t(\bar{x})$, $t'(\bar{x}; h)$ is ≥ 0 for each h ; then since $t'(\bar{x}; \cdot)$ is continuous there exists $\varepsilon > 0$ such that

$$0 < \|h\| \leq \varepsilon, \left\| \frac{h}{\|h\|} - \frac{d}{\|d\|} \right\| \leq \varepsilon \Rightarrow 0 \leq t' \left(\bar{x}; \frac{h}{\|h\|} \right) \leq \gamma.$$

From (2.15) it follows that

$$\left(\nabla^2 t_j(\bar{x}) \frac{h}{\|h\|}, \frac{h}{\|h\|} \right) \geq \rho \quad \forall j \in I(\bar{x}; h; h)$$

and (2.17) becomes

$$\frac{1}{\|h\|} \left(\frac{t(\bar{x} + h) - t(\bar{x})}{\|h\|} - t' \left(\bar{x}; \frac{h}{\|h\|} \right) \right) \geq \frac{\rho}{2} + \min_{j \in \langle 1, m \rangle} \varepsilon_j(\bar{x}; h),$$

which implies

$$t''_-(\bar{x}; d; d) \geq \rho > 0.$$

In fact we can give an improved version of corollary 2.2.

COROLLARY 2.4. Let $C = \mathbb{R}^N$. Suppose that g is another real-valued locally Lipschitzian function for which

$$(2.18) \quad g(\bar{x}) = f(\bar{x}), \quad g(x) \leq f(x) \quad \forall x \in V,$$

where V is a neighborhood of \bar{x} . Suppose also that $L(\bar{x})$ is nonempty and that

$$(2.19) \quad D'_-g(\bar{x}; d) \geq 0 \quad \forall d \in \mathbb{R}^N,$$

$$(2.20) \quad g''_-(\bar{x}; d; d) > 0 \quad \forall d \in L(\bar{x}).$$

Then \bar{x} is an isolated local minimum with order 2 of problem P.

Proof. From (2.18) and (2.19) we obtain

$$f^*(x; d) \geq g''_-(x; d; d).$$

Then the result follows when using Proposition 2.1 and (2.20).

Remark 2.4. We shall see in § 2.3 that Corollary 2.4 is really an improved version of Corollary 2.2, but this can be seen immediately by means of the following example.

Let $f(x, y) = \max \{x^2 + y^2, 4x - x^2 - y^2\}$. We have

$$(0, 0) \in \partial f(0, 0).$$

Let C be the circle defined by

$$C = \{x : t_i(x) \leq 0 \ \forall i \in \langle 1, p \rangle, s_j(x) = 0 \ \forall j \in \langle 1, m \rangle\}.$$

Outside this circle we have $f(x, y) = x^2 + y^2$, while inside the circle we have $f(x, y) = 4x - x^2 - y^2$. Let $d = (0, 1)$. Then with $z^* = (0, 0)$, $f'(z^*; d) = 0$. Pick $z_k = (x_k, y_k)$ in the interior of the circle so that $\{z_k\}$ converges to z^* in direction d . Now we have

$$\frac{1}{\|z_k\|} \left[\frac{f(z_k) - f(z^*)}{\|z_k\|} - f' \left(z^*; \frac{z_k}{\|z_k\|} \right) \right] = \frac{1}{\|z_k\|^2} [4x_k - x_k^2 - y_k^2 - 4x_k] = -1.$$

Hence (2.14) is not satisfied but (2.18), (2.19) and (2.20) are satisfied for $g(x, y) = x^2 + y^2$.

Let us consider now the usual constrained case where C is given by

$$C = \{x : t_i(x) \leq 0 \ \forall i \in \langle 1, p \rangle, s_j(x) = 0 \ \forall j \in \langle 1, m \rangle\}$$

and suppose that t_i, s_j are real-valued locally Lipschitzian functions defined on \mathbb{R}^N . Let $I(\bar{x}) = \{i \in \langle 1, p \rangle : t_i(\bar{x}) = 0\}$. For given real $\lambda_i \geq 0, i \in I(\bar{x}), \mu_j, j \in \langle 1, m \rangle$ set

$$J(\bar{x}) = \{i \in I(\bar{x}) : \lambda_i > 0\}, L_{\lambda, \mu}(x) = f(x) + \sum_{i \in J(\bar{x})} \lambda_i t_i(x) + \sum_{j=1}^m \mu_j s_j(x).$$

COROLLARY 2.5. *Suppose that $K(\bar{x})$ is nonempty and that there exist (λ, μ) such that*

$$(2.21) \quad D' L_{\lambda, \mu}(\bar{x}; d) \geq 0 \quad \forall d \in \mathbb{R}^N,$$

$$(2.22) \quad (L_{\lambda, \mu})''(\bar{x}; d; d) > 0 \quad \forall d \in K(\bar{x}).$$

Then \bar{x} is an isolated local minimum with order 2 of problem P.

Proof. By definition of $L_{\lambda, \mu}$ we have

$$L_{\lambda, \mu}(\bar{x}) = f(\bar{x}), \quad L_{\lambda, \mu}(x) \leq f(x) \quad \forall x \in C.$$

Then by (2.21) we obtain

$$\left[\frac{1}{\|h\|} \left[\frac{f(\bar{x} + h) - f(\bar{x})}{\|h\|} + \delta(\bar{x} + h | C) \right] \right] \geq \frac{1}{\|h\|} \left[\frac{L_{\lambda, \mu}(\bar{x} + h) - L_{\lambda, \mu}(\bar{x})}{\|h\|} - D' L_{\lambda, \mu}(\bar{x}; \frac{h}{\|h\|}) \right]$$

from which it follows, when passing to the limit, that

$$f^*(\bar{x}; d) \geq (L_{\lambda, \mu})''(\bar{x}; d; d) > 0 \quad \forall d \in K(\bar{x}),$$

and the theorem is proved by using Proposition 2.1.

Remark 2.5. When the data functions are twice continuously differentiable it is obvious that Corollary 2.5 coincides with Hestenes [15, Thm. 10.3], from which we can obtain easily the usual classical second-order sufficiency theorem (see for example Han and Mangasarian [14]).

Second-order sufficient conditions for certain classes of nondifferentiable functions were given recently by several authors: Ioffe [18], Fletcher and Watson [12], Demjanov and Malozemov [10], Ben-Tal and Zowe [5], Spingarn [26] and Chaney [6], [7]. All these authors with the exception of Chaney study only certain particular kinds of locally Lipschitzian functions. In [10] Demjanov and Malozemov are interested in the discrete minimax problem. In [5] Ben-Tal and Zowe are concerned with

three particularly important topics: l_1 -approximation, the minimization of the exact penalty function and the minimization of the classical penalty function. In [12] Fletcher and Watson consider the problem

$$\text{minimize } f(x) = \Phi(x) + \|r(x)\|_A$$

subject to $\|c(x)\|_B \leq h$, where f, r, c are twice continuously differentiable functions and $\|\cdot\|_A, \|\cdot\|_B$ are any norms on \mathbb{R}^N . In [18] Ioffe minimizes functions of the form

$$f(x) = g(G(x)),$$

where g is a sublinear function and G is assumed to be twice continuously differentiable. In [26] Spingarn proves another kind of theorem: if \bar{x} is a local minimizer and ∂f^{-1} is Lipschitz continuous at $(0, \bar{x})$ then one has a relation like (2.3). In Chaney's paper [6], if we consider for example the unconstrained case, condition (2.3) is satisfied under assumptions other than (2.13) and (2.14). The point of view adopted here is not the same. Corollary 2.2 is centered on the notion of lower second-order directional derivatives, and this corollary is useful if it is possible to easily obtain lower bounds for f'' .

2.3. Some additional properties for lower- C^2 functions. If we restrict ourselves to lower- C^2 functions more can be said about the lower second-order directional derivative and about sufficient conditions. These functions were introduced by Rockafellar in [23] as follows:

DEFINITION 2.2. A real-valued function f defined on \mathbb{R}^N is *lower- C^2* if for each point $\bar{x} \in \mathbb{R}^N$ there is for some open neighborhood X of \bar{x} a representation

$$(2.23) \quad f(x) = \max_{s \in S} F(x, s) \quad \text{for all } x \in X,$$

where S is a compact topological space and $F: X \times S \rightarrow \mathbb{R}$ is a function which has partial derivative and about sufficient conditions. These functions were introduced by continuous not just in x but jointly in $(x, s) \in X \times S$.

Now we recall the fundamental theorem given by Rockafellar [23, Thm. 6].

THEOREM 2.6. For a locally Lipschitzian function $f: \mathbb{R}^N \rightarrow \mathbb{R}$ the following properties are equivalent:

- a) f is lower- C^2
- b) ∂f is strictly hypomonotone, that is:

$$(2.24) \quad \liminf_{\substack{x' \rightarrow x, y' \in \partial f(x') \\ x'' \rightarrow x, y \in \partial f(x'')}} \frac{(x' - x'', y' - y'')}{\|x' - x''\|^2} > -\infty \quad \text{for all } x.$$

c) For every $\bar{x} \in \mathbb{R}^N$ there is a convex neighborhood X of \bar{x} on which f has a representation

$$(2.25) \quad f = g - h \text{ on } X \text{ with } g \text{ convex and finite, } h \text{ quadratic convex.}$$

d) For every $\bar{x} \in \mathbb{R}^N$ there is a neighborhood X of \bar{x} and a representation of f as in (2.23) with S a compact topological space, $F(x, s)$ quadratic in x and continuous in s .

Remark 2.6. From (2.24) and (2.25) it follows that these functions coincide with those introduced by Malivert in [19].

Let us now remark that, when f is a lower- C^2 function, f is sub-differentially regular and then let us introduce the quantities:

$$D_-''f(x; d; d) = \liminf_{\substack{d \\ h \rightarrow 0}} \frac{f'(x+h; h/\|h\|) - f'(x; h/\|h\|)}{\|h\|},$$

$$D_+''f(x; d; d) = \limsup_{\substack{d \\ h \rightarrow 0}} \frac{f'(x+h; h/\|h\|) - f'(x; h/\|h\|)}{\|h\|},$$

$$f''_+(x; d; d) = 2 \limsup_{\substack{d \\ h \rightarrow 0}} \frac{1}{\|h\|} \left[\frac{f(x+h) - f(x)}{\|h\|} - f'\left(x; \frac{h}{\|h\|}\right) \right].$$

Now remark with Rockafellar [23] that formula (2.25) is crucial since it implies that local properties of convex functions will be carried over to general lower- C^2 functions. Hence we have:

PROPOSITION 2.7. *Suppose that f is a lower- C^2 function. Then*

$$(2.26) \quad -\infty < D_-''f \leq f''_- \leq f''_+ \leq D_+''f.$$

Proof. Following (2.24), we have

$$-\infty < D_-''f(x; d; d) \quad \forall x \quad \forall d \neq 0.$$

Now, following (2.25), we have

$$\begin{aligned} f(x+h) - f(x) &= \int_0^{\|h\|} f'\left(x + \frac{\theta h}{\|h\|}; \frac{h}{\|h\|}\right) d\theta \\ &= \int_0^{\|h\|} f'\left(x + \frac{\theta h}{\|h\|}; \frac{h}{\|h\|}\right) - f'\left(x; \frac{h}{\|h\|}\right) d\theta + \|h\| f'\left(x; \frac{h}{\|h\|}\right) \end{aligned}$$

so that

$$\begin{aligned} \frac{1}{\|h\|} \left[\frac{f(x+h) - f(x)}{\|h\|} - f'\left(x; \frac{h}{\|h\|}\right) \right] &= \frac{1}{\|h\|^2} \int_0^{\|h\|} \left[f'\left(x + \frac{\theta h}{\|h\|}; \frac{h}{\|h\|}\right) - f'\left(x; \frac{h}{\|h\|}\right) \right] d\theta \\ &= \int_0^1 u \frac{f'(x + u\|h\|h/\|h\|; h/\|h\|) - f'(x; h/\|h\|)}{u\|h\|} du, \end{aligned}$$

which obviously implies (2.26).

Proposition 2.7 shows that f''_- is finite and gives a lower bound for f''_- which can be easier to compute in some cases. Now recall from Mignot's theorem [20] and representation (2.25) that there exists a set \bar{E}_f in X of measure zero such that ∂f is differentiable at each point of $\bar{E}_f^c = X \setminus \bar{E}_f$ in the following sense:

∂f is differentiable at x if $\partial f(x) = \{\nabla f(x)\}$ and there is a linear transformation denoted by $D''f(x)$ such that

$$\|\partial f(z) - \nabla f(x) - D''f(x)(z-x)\| = o(\|z-x\|)$$

or, in other words,

$$\begin{aligned} \forall \varepsilon > 0 \exists \delta: \forall z \text{ with } \|z-x\| \leq \delta \quad \forall z^* \in \partial f(z) \\ \|z^* - \nabla f(x) - D''f(x)(z-x)\| \leq \varepsilon \|z-x\|. \end{aligned}$$

From this property and from Proposition 2.7 follows:

COROLLARY 2.8. *Suppose that f is a lower- C^2 function. Then we have*

$$f''_-(x; d; d) = \left(D''f(x) \frac{d}{\|d\|}, \frac{d}{\|d\|} \right) \quad \forall d \neq 0 \quad \forall x \in \bar{E}_f^c.$$

Finally we shall now prove a sufficiency proposition for lower- C^2 functions which can also be obtained by using Chaney [6, Corollary 2.18]. This proposition requires “knowledge” of the function $F(\cdot, \cdot)$ which appears in the definition.

PROPOSITION 2.9. *Suppose that f is lower- C^2 and is defined by (2.23). Let $I(\bar{x}) = \{s \in S: f(\bar{x}) = F(\bar{x}, s)\}$. Suppose that there exist indices $s_i, i \in \langle 1, m \rangle$ in $I(\bar{x})$ and reals a_i such that*

$$\sum_{i=1}^m a_i \nabla_x F(\bar{x}, s_i) = 0, \quad \sum_{i=1}^m a_i = 1, \quad a_i \geq 0 \quad \forall i \in \langle 1, m \rangle.$$

Suppose also that $L(\bar{x})$ is nonempty and that

$$d \neq 0, f'(\bar{x}; d) = 0 \Rightarrow \sum_{i=1}^m a_i (\nabla_{xx}^2 F(\bar{x}, s_i) d, d) > 0.$$

Then, for $C = \mathbb{R}^N$, \bar{x} is an isolated local minimum with order 2 of problem P.

Proof. Use Corollary 2.4. Choose the auxiliary function $g(\cdot)$ to be $\sum_{i=1}^m a_i F(\cdot, s_i)$.

3. Stability of perturbed systems. In this section the notation is the same as in § 1. Let $r > 0$ and set

$$C_r(w) = C(w) \cap B(\bar{x}, r),$$

where $B(\bar{x}, r)$ is the closed ball centered at \bar{x} with radius r . Also let

$$(3.0) \quad \alpha_r(w) = \min \{f_0(x, w) \mid x \in C_r(w)\}, \quad M_r(w) = \{x \in C_r(w): f_0(x, w) = \alpha_r(w)\}.$$

The following theorem now gives, as was announced in the introduction, bounds for the distance between the local minimizers of perturbed problems and the local minimizers of the original problem. This theorem is a generalization of Alt [2, Thm. 4.6] given for differentiable data.

THEOREM 3.1. *Let $\bar{x} \in C(\bar{w})$. Suppose that \bar{x} satisfies the extended Mangasarian–Fromovitz condition for the set $C(\bar{w})$ and that for each $s \in (0, \infty]$ there exists a real $t(s) > 0$ such that $C_s(w)$ is nonempty for each $w \in B(\bar{w}, t(s))$. Suppose furthermore that \bar{x} is an isolated local minimum with order i ($i = 1$ or 2) of problem $P(\bar{w})$. Then there exist constants r, r_1, L strictly positive such that*

$$(3.1) \quad \|x - \bar{x}\|^i \leq L \|w - \bar{w}\| \quad \forall x \in M_r(w) \quad \forall w \in B(\bar{w}, r_1).$$

Proof. This proof is similar to Alt’s proof given for differentiable data.

A) Since \bar{x} is an isolated local minimum with order i of problem $P(\bar{w})$, there exists $s > 0, \varepsilon > 0$ such that

$$(3.2) \quad f_0(x, \bar{w}) > f_0(\bar{x}, \bar{w}) + \frac{1}{2}\varepsilon \|x - \bar{x}\|^i \quad \forall x \in C_s(\bar{w}), \quad x \neq \bar{x}.$$

Since $\bar{x} \in C(\bar{w})$ satisfies the extended Mangasarian–Fromovitz condition for the set $C(\bar{w})$, \bar{x} belongs to $C_s(\bar{w})$ and satisfies the same condition but for the set $C_s(\bar{w})$.

Furthermore, since $C_s(w) \neq \emptyset$ for $w \in B(\bar{w}, t(s))$ it follows from Theorem 1.1 that there exist $r \in]0, \min(s, t(s))]$ and a constant c such that

$$C_s(w) \neq \emptyset \quad \forall w \in B(\bar{w}, r),$$

$$d_{C_s(w)}(x) \leq c \max_{i,j} (f_i^+(x, w), |g_j(x, w)|) \quad \forall x \in B(\bar{x}, r) \quad \forall w \in B(\bar{w}, r).$$

Let $l(r) = \min(r, t(r))$. For each $w, w' \in B(\bar{w}, l(r))$, $C_r(w')$ is nonempty and for each $x \in C_r(w')$ this inequality is equivalent to:

$$d_{C_s(w)}(x) \leq c \max_{i,j} (f_i^+(x, w) - f_i^+(x, w'), |g_j(x, w)| - |g_j(x, w')|).$$

Now for each $x \in B(\bar{x}, r)$ and each $w \in B(\bar{w}, l(r))$ let $x_w(x)$ be such that

$$(3.3) \quad x_w(x) \in C_s(w), \quad \|x - x_w(x)\| = d_{C_s(w)}(x).$$

Since the functions f_i and g_j are locally Lipschitzian it follows that there exist $L_0 > 0, L_1 > 0$ such that

$$(3.4) \quad \|x - x_w(x)\| \leq L_0 \|w - w'\| \quad \forall x \in C_r(w') \quad \forall w, w' \in B(\bar{w}, l(r)),$$

$$|f_0(x, w') - f_0(x_w(x), w)| \leq L_1 (\|x - x_w(x)\| + \|w - w'\|)$$

$$\forall x \in C_r(w') \quad \forall w, w' \in B(\bar{w}, l(r)),$$

from which it follows that if we set $\delta = (1 + L_0)L_1$ then

$$(3.5) \quad |f_0(x, w') - f_0(x_w(x), w)| \leq \delta \|w - w'\| \quad \forall x \in C_r(w') \quad \forall w, w' \in B(\bar{w}, l(r)).$$

B) Now let $r_1 = \min(l(r), l(r)/L_0)$; if $x = \bar{x}$ and $w' = \bar{w}$ it follows from (3.3) and (3.4) that for $w \in B(\bar{w}, r_1)$ we have

$$x_w(\bar{x}) \in C_r(w), \quad \|\bar{x} - x_w(\bar{x})\| = d_{C_r(w)}(\bar{x}),$$

and then from (3.5) it follows that

$$(3.6) \quad \alpha_r(w) \leq f_0(x_w(\bar{x}), w) \leq \alpha_r(\bar{w}) + \delta \|w - \bar{w}\| \quad \forall w \in B(\bar{w}, r_1).$$

C) If we set $w = \bar{w}$ we obtain, from (3.2) and (3.5), for each $w' \in B(\bar{w}, r_1)$ and for each $x \in C_r(w')$

$$(3.7) \quad f_0(\bar{x}, \bar{w}) + \frac{1}{2}\varepsilon \|x_w(x) - \bar{x}\|^i \leq f_0(x_w(x), \bar{w}) \leq f_0(x, w') + \delta \|w' - \bar{w}\|.$$

Since $\alpha_r(\bar{w}) = f_0(\bar{x}, \bar{w})$, taking $x \in M_r(w')$ then we obtain

$$(3.8) \quad \alpha_r(\bar{w}) - \alpha_r(w') \leq \delta \|w' - \bar{w}\| \quad \forall w' \in B(\bar{w}, r_1),$$

$$(3.9) \quad \alpha_r(w') \geq \alpha_r(\bar{w}) + \frac{1}{2}\varepsilon \|x_w(x) - \bar{x}\|^i - \delta \|w' - \bar{w}\| \quad \forall w' \in B(\bar{w}, r_1).$$

From (3.6) and (3.8) we obtain

$$(3.10) \quad |\alpha_r(\bar{w}) - \alpha_r(w)| \leq \delta \|w - \bar{w}\| \quad \forall w \in B(\bar{w}, r_1).$$

Now, since we have

$$\|x - \bar{x}\| \leq \|x - x_w(x)\| + \|x_w(x) - \bar{x}\|,$$

$$\|x - \bar{x}\|^2 \leq \|x - x_w(x)\|^2 + 2\|x - x_w(x)\| \|x_w(x) - \bar{x}\| + \|\bar{x} - x_w(x)\|^2,$$

then it follows from (3.4) and (3.9) that there exists a constant L_2 such that

$$\alpha_r(w) \geq \alpha_r(\bar{w}) + \frac{1}{2}\varepsilon \|x - \bar{x}\|^i - (L_2 + \delta)\|w - \bar{w}\| \quad \forall w \in B(\bar{w}, r_1) \quad \forall x \in M_r(w),$$

and (3.1) follows from (3.10).

Remark 3.1. Let us return to Remark 1.3 with vertical perturbations and assume that \bar{x} satisfies the extended Mangasarian–Fromovitz condition for the set $C(\bar{w})$. Then since $\bar{x} \in C(\bar{w})$, for each $s \in (0, \infty]$ the point \bar{x} satisfies the extended Mangasarian–Fromovitz condition for the set $C_s(\bar{w})$ and it follows from [3, Thm. 2.1] that there exists $t(s) > 0$ such that $C_s(w)$ is nonempty for each $w \in B(\bar{w}, t(s))$.

Acknowledgments. The author wishes to thank the two referees for many important suggestions and help on this paper. In particular, the introduction of the tangent cone in § 2, the improvement of Corollary 2.2 by Corollary 2.4 and the example in Remark 2.4 which were suggested by one of the referees.

REFERENCES

- [1] A. D. ALEXANDROFF, *Almost everywhere existence of the second differential of a convex function and some properties of convex surfaces connected with it*, Leningrad State Univ. Ann. Math., Ser. 6 (1939), pp. 3–35. (In Russian.)
- [2] W. ALT, *Lipschitzian perturbations of infinite optimization problems*, Proc. Second International Symposium Concerning Stability in Mathematical Programming, A. V. Fiacco, ed., Marcel Dekker, New York, to appear.
- [3] A. AUSLENDER, *Differentiable stability in nonconvex and nondifferentiable programming*, Mathematical Programming Study, 10 (1979), pp. 29–41.
- [4] A. BEN-TAL, *Second-order and related extremality conditions in nonlinear programming*, J. Optim. Theory Appl., 31 (1980), pp. 143–167.
- [5] A. BEN-TAL AND J. ZOWE, *Discrete l_1 approximation and related nonlinear nondifferentiable problems*, contributed paper, Mathematical Programming Symposium, Confolant, 1981.
- [6] R. W. CHANEY, *Second-order sufficiency conditions for nondifferentiable programming problems*, this Journal, 20 (1982), pp. 20–33.
- [7] ———, *A general sufficiency theorem for nonsmooth nonlinear programming*, Tech. Rep., Western Washington University, Bellingham, WA, 1982.
- [8] F. H. CLARKE, *Generalized gradients of Lipschitz functionals*, Madison M.R.C. Tech. Rept., University of Wisconsin, 1976.
- [9] ———, *A new approach to Lagrange multipliers*, Math. Oper. Res., 1 (1976), pp. 165–174.
- [10] V. F. DEMJANOV AND V. N. MALOZEMOV, *Introduction to Minimax Theory*, John Wiley, New York, 1974.
- [11] I. EKELAND AND R. TEMAN, *Analyse convexe et problèmes variationnels*, Dunod, Paris, 1974.
- [12] R. FLETCHER AND G. A. WATSON, *First and second-order conditions for a class of nondifferentiable optimization problems*, Math. Programming, 18 (1980), pp. 286–291.
- [13] B. GOLAN, *Perturbation theory for abstract optimization problems*, J. Optim. Theory Appl., 35 (1981), pp. 417–442.
- [14] S. P. HAN AND O. L. MANGASARIAN, *Exact penalty functions in nonlinear programming*, Math. Programming, 17 (1979), pp. 251–259.
- [15] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
- [16] J. B. HIRIART-URRUTY, *Refinements of necessary optimality conditions in nondifferentiable programming I*, Appl. Math. Opt., 5 (1979), pp. 63–82.
- [17] A. D. IOFFE, *Regular points of Lipschitz functions*, Trans. Amer. Math. Soc., 251 (1979), pp. 61–69.
- [18] ———, *Necessary and sufficient conditions for a local minimum*, this Journal, 17 (1979), pp. 245–289.
- [19] C. MALIVERT, *Méthode de descente sur un fermé non convexe, analyse non convexe*, Bull. Soc. Mat. France, 60 (1979), pp. 113–124.
- [20] F. MIGNOT, *Contrôle dans les équations variationnelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130–185.
- [21] S. M. ROBINSON, *Stability theory for systems of inequalities, Part II: Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–512.
- [22] ———, *Generalized equations and their solutions, Part II: Applications to nonlinear programming*, Tech. Rep. 2048, Mathematics Research Center, University of Wisconsin, Madison, 1980.
- [23] R. T. ROCKAFELLAR, *Favorable classes of Lipschitz continuous functions in subgradient optimization*, Tech. Rept., I.A.S.A., 1980.
- [24] ———, *Lagrange multipliers and subderivatives of optimal value functions in nonlinear programming*, Math. Programming Study, 15, to appear.

- [25] ———, *Proximal subgradients, marginal values and augmented Lagrangians in nonconvex optimization*, Math. Oper. Res., to appear.
- [26] J. E. SPINGARN, *Submonotone mappings and the proximal point algorithm*, Tech. Rept., School of Mathematics, Georgia Inst. Technology, Atlanta, 1981.

FINITE DIMENSIONAL COMPENSATORS FOR PARABOLIC DISTRIBUTED SYSTEMS WITH UNBOUNDED CONTROL AND OBSERVATION*

RUTH F. CURTAIN†

Abstract. It is proved that for a class of parabolic distributed parameter systems with unbounded control and observation there exists a finite dimensional compensator using dynamic output feedback. Finite dimensional means here that the dynamics relating the input to the output is finite dimensional. A constructive design algorithm is presented and several examples are considered.

Key words. parabolic systems, compensators, unbounded control, dynamic output feedback

1. Introduction. The important problem of stabilizing infinite dimensional systems has received much attention in the literature. Although state feedback stabilization is an interesting theoretical problem, in infinite dimensions one can never observe the whole state and so it is necessary to stabilize by output feedback. *Static* output feedback via either distributed or boundary control and/or observations has been considered by Nambu [19] and Triggiani [30], [31] for second order parabolic systems. Stabilization by *dynamic* output feedback (often called compensation in the literature) has been considered by Curtain [6], Fujii [15] and Nambu [20] for classes of parabolic systems, including boundary control and observations at points or on the boundary. All these approaches share the common disadvantage that the stabilization scheme is *infinite-dimensional*.

A major advance was made by Schumacher [26], [27], when he gave a theory for designing finite dimensional compensators for a large class of systems, including parabolic and delay systems. However, in his theory it must be assumed that the control and observation operators are bounded, which for distributed systems means that point and boundary action are excluded.

It is the purpose of this paper to develop a theory for finite dimensional compensator design for distributed parameter systems, where the control and observation may be implemented pointwise or at the boundary of the domain. Although this paper borrows much from Schumacher [26], it is not clear how one could extend his design to unbounded control operators B , as his design is given in terms of eigenfunctions of $A + BF$. What is presented here is an alternative compensator design which does extend to the unbounded case, and proves to be a simpler approach for the bounded case as well [5].

The design is outlined in § 2 for the bounded case to clarify the connections with Schumacher's work and to clarify the ideas underlying the sequel. In § 3 a theoretical existence theorem is proved under technical assumptions reminiscent of earlier work [6], [7]. These assumptions are fairly general and guarantee the existence of a finite dimensional compensator, but, as in [27], there is no upper bound on the order. To ensure an implementable compensator design, Schumacher gave a test for the stability of the compensator in terms of zeros of a Weinstein-Aronzajn determinant and we extend this result in § 4, but here we are forced to assume that the control is bounded. One way out of this dilemma is to reformulate the problem so as to obtain a bounded

* Received by the editors March 30, 1982, and in revised form November 20, 1982. This paper was written while the author was visiting the Mathematics Department of the University of Melbourne, Australia.

† Rijksuniversiteit Groningen, Mathematisch Instituut, Postbus 800, 9700 Av Groningen, The Netherlands.

B as in [11], [21] and this is followed up in § 5, together with some examples of a diffusion system with point or boundary observation and control for both Neumann and Dirichlet conditions. In the conclusions in § 6 the scope of this approach is discussed in more detail together with comparisons to related recent work in [9], [10], [25].

2. Motivation of the compensator design. In view of the complicated technical assumptions needed to establish a rigorous theory, it seems useful to motivate the compensator design by considering a very special class of systems:

$$(2.1) \quad \dot{z} = Az + Bu, \quad y = Cz.$$

We assume that A is a self-adjoint operator on a real, separable Hilbert space Z and, furthermore, that A has compact resolvent. This implies that A has a point spectrum and, for simplicity, we suppose that the eigenvalues $\{\lambda_i\}$ are simple and $\lambda_1 > \lambda_2 > \lambda_3 > \dots$. The corresponding eigenfunctions then generate a complete orthonormal basis $\{e_i\}_{i=1}^\infty$ for Z . Using this basis for Z , we have that $A = \text{diag}(\lambda_1, \lambda_2, \dots)$ and A generates the analytic semigroup $T_t = \text{diag}(e^{\lambda_1 t}, e^{\lambda_2 t}, \dots)$ (see [7, p. 46]). We assume that $B \in \mathcal{L}(R^m, Z)$ and $C \in \mathcal{L}(Z, R^k)$.

Now the problem is to design a finite dimensional compensator for (2.1) to stabilize the system

$$(2.2) \quad \dot{w} = Mw + Ly, \quad u = Qw.$$

Combining (2.1) and (2.2) we obtain the extended operator A_e on $Z \oplus W$, $W \cong R^n$ for some $n > 0$.

$$(2.3) \quad A_e = \begin{pmatrix} A & BQ \\ LC & M \end{pmatrix}.$$

If the unstable eigenvalues of A are $\lambda_1, \dots, \lambda_r$, then it is known [29], [25], how to choose $F_0 \in \mathcal{L}(R^m, R^r)$ and $G_0 \in \mathcal{L}(R^r, R^k)$ so that $A + B(F_0 0)$ and $A + \begin{pmatrix} G_0 \\ 0 \end{pmatrix} C$ are stable with eigenvalues $\lambda_{r+1}, \lambda_{r+2}, \dots$, determined and r arbitrarily assignable ones. Let $F = (F_0 0)$ and $G = \begin{pmatrix} G_0 \\ 0 \end{pmatrix}$ and define

$$Z^r = \text{span}\{e_1, \dots, e_r\},$$

and let R be the isomorphism between Z^n and R^n .

For our compensator we let $W = R^n$ for $n \geq r$ and choose $Q = FR^{-1}$, $L = -RG$ and $M = R(A + \Pi_n BF + GC)R^{-1}$, where Π_n is the projection from Z to Z^n . Then Q , L and M are well defined matrices of appropriate sizes, and are calculable in terms of $\lambda_1, \dots, \lambda_n$, F_0 , G_0 and $\Pi_n B$ and C/Z^n . We proceed to show that the corresponding extended matrix A_e on $Z \oplus R^n$ is asymptotically stable for n sufficiently large. Now

$$(2.4) \quad A_e = \begin{pmatrix} A & BFR^{-1} \\ -RG C & R(A + \Pi_n BF + GC)R^{-1} \end{pmatrix} = H\tilde{A}_e H^{-1},$$

where

$$(2.5) \quad H = \begin{pmatrix} I & 0 \\ R\Pi_n & -T^{-1} \end{pmatrix}, \quad H^{-1} = \begin{pmatrix} I & 0 \\ TR\Pi_n & -T \end{pmatrix}$$

and

$$(2.6) \quad T: w \rightarrow \begin{pmatrix} R^{-1}w \\ w \end{pmatrix} \text{ maps } W = R^n \text{ to } M,$$

the n -dimensional subspace of $Z \oplus R^n$ defined by

$$(2.7) \quad M = \left\{ \begin{pmatrix} x \\ Rx \end{pmatrix} \middle| x \in Z^n \right\}.$$

Thus T and H are isomorphisms and since $BF\Pi_n = BF$,

$$(2.8) \quad \tilde{A}_e = \begin{pmatrix} A + BF & -BFR^{-1}T^{-1} \\ 0 & TR(A + GC)R^{-1}T^{-1} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ TRG(C - C\Pi_n) & 0 \end{pmatrix}.$$

The spectrum of

$$\begin{pmatrix} A + BF & -BFR^{-1}T^{-1} \\ 0 & TR(A + GC)R^{-1}T^{-1} \end{pmatrix}$$

is $\sigma(A + BF) \cup \sigma(TR(A + GC)R^{-1}T^{-1})$ and we have seen that this can be made to be in $\text{Re } \lambda \leq \lambda_{r+1}$ for arbitrary r . \tilde{A}_e is a degenerate perturbation of this operator since the range of G is finite-dimensional and so its spectrum is also discrete. Bounded perturbations of analytic semigroups are also analytic and so the spectrum of the generator determines the stability of the semigroup, and we have as an estimate for the semigroup generated by \tilde{A}_e ,

$$\|\tilde{T}_t^e\| \leq m e^{\lambda_{r+1}t + m\|G(C - C\Pi_n)\|t}$$

and m does not depend on n . So for $\|C - C\Pi_n\|$ sufficiently small, or n sufficiently large, \tilde{T}_t^e will be asymptotically stable. The stability can also be directly checked by calculating the eigenvalues of \tilde{A}_e from the Weinstein–Aronszajn method [16, p. 244].

The above approach is applicable to a large class of systems, for example the parabolic and delay systems considered by Schumacher in [26] and [27]. In fact, the only difference lies in our different choices of the A_e operator. He uses the isomorphism

$$\tilde{H} = \begin{pmatrix} I & -R^{-1} \\ 0 & T \end{pmatrix}$$

to show that his A_e is similar to one of a simpler form and he later perturbs the G operator. Under the \tilde{H} isomorphism, our A_e is similar to

$$\begin{pmatrix} A + GC & 0 \\ -TRGC & TR(A + \Pi_n BF)R^{-1}T^{-1} \end{pmatrix} + \begin{pmatrix} 0 & (B - \Pi_n B)FR^{-1}T^{-1} \\ 0 & 0 \end{pmatrix},$$

which indicates the lack of bias of our design with respect to either the B or C components. It is difficult to compare these two approaches from a theoretical point of view, as neither gives an upper bound for n , the compensator order; it is more appropriately done via numerical studies e.g. [5]. Our interest here and the motivation for deriving an alternative approach is to extend the theory to allow for unbounded B and C operators. The theory of Schumacher assumed that B and C were bounded, and in particular relied on eigenfunctions of $A + BF$, which for unbounded B seems difficult to interpret.

We proceed in § 3 to show how A_e in (2.4) can be interpreted for unbounded B and C for a class of parabolic type systems.

3. A theoretical existence result. We shall use the mathematical formulation for unbounded B and C operators outlined in [7, Chapt. 8], and it shall be necessary to develop some further results. Consider the following abstract systems

$$(3.1) \quad z(t) = T_t z_0 + \int_0^t T_{t-s} B u(s) ds,$$

$$(3.2) \quad y(t) = C S_t x_0,$$

where T_t is a strongly continuous semigroup with infinitesimal generator A on the Banach space Z , $z_0 \in Z$; $B: U \rightarrow Z$ is a linear map on the Banach space U and $u \in L^p[0, t_1; U]$. S_t is a strongly continuous semigroup on the Banach space X , $x_0 \in X$ and $C: X \rightarrow Y$ is a linear map onto the Banach space Y . B and C will not be bounded in general, but satisfy the following hypotheses:

H1($B, T_b, Z, U, \tilde{W}, g_1$). There exists a Banach space \tilde{W} with Z dense in \tilde{W} such that:

- (a) $\tilde{W} \supset Z, \tilde{W} \supset R(B)$;
- (b) $B \in \mathcal{L}(U, \tilde{W})$;
- (c) $T_t \in \mathcal{L}(\tilde{W}, Z), t > 0$;
- (d) $\|T_t w\|_Z \leq g_1(t) \|w\|_{\tilde{W}} \forall w \in \tilde{W}$ for $t \in (0, t_1)$, where $g_1 \in L^q(0, t_1), 1 < q < \infty$.

H2($C, S_b, X, Y, \tilde{W}, g_2$). There exists a Banach space \tilde{W} dense in X such that:

- (a) $\tilde{W} \subset X, \tilde{W} \subset D(C)$;
- (b) $C \in \mathcal{L}(\tilde{W}, Y)$;
- (c) $S_t \in \mathcal{L}(X, \tilde{W}), t < 0$;
- (d) $\|S_t x\|_{\tilde{W}} \leq g_2(t) \|x\|_X \forall x \in X$ for $t \in (0, t_1)$, where $g_2 \in L^r(0, t_1), 1 < r < \infty$.

Whenever we write $\|B\|$ and $\|C\|$, the norms refer to those in $\mathcal{L}(U, \tilde{W})$ and $\mathcal{L}(\tilde{W}, Y)$, respectively.

We remark that the semigroup property of T_t implies that the (d) of H1 holds for all $t > 0$, where on $[t_1, \infty)$, $g_1(t) \leq m g_1(t_1) e^{w(t-t_1)} = m_1 e^{wt}$. Similarly for S_t with $g(t) = m_2 e^{-wt}$ on $[t_1, \infty)$ where the exponents refer to the growth constant of the respective semigroup.

As we shall see in the examples in § 5, the B operator will be typically identifiable as AD for some $D \in \mathcal{L}(U, Z)$ and the operator C will be A -bounded. The assumptions H1 and H2 are in fact more general and allow us to deduce more about successive perturbations and growth rates of semigroups (Lemma 3.2) and the spectrum determined growth assumption (Lemma 3.3).

We shall be concerned with two types of perturbed semigroups, which arise naturally in the case of state feedback $u = Fz(t)$ for an $F \in \mathcal{L}(Z, U)$ and in the case of output injection $u = Gy(t)$ for a $G \in \mathcal{L}(Y, Z)$. The first case leads to the integral equation

$$(3.3) \quad V_t z = T_t z + \int_0^t T_{t-s} B F V_s z ds, \quad z \in Z$$

and the second to the equation

$$(3.4) \quad U_t x = S_t x + \int_0^t U_{t-s} G C S_s x ds, \quad x \in X.$$

Equations (3.3) and (3.4) are duals if we suppose that X is reflexive and we identify $Z = X^*, T_t = S_t^*, Y^* = U, B = C^*$ and $F = G^*$. If we let $\tilde{W}^* = \tilde{W}$ then we see that $H2(C, S_b, X, Y, \tilde{W}, g_2) \Leftrightarrow H1(B, T_b, Z, U, \tilde{W}, g_1)$. We shall use this duality in the sequel to deduce results about C perturbations from corresponding ones about B perturbations.

The existence of solutions of (3.3) and (3.4) was established in [7] essentially using the following results on Volterra integral equations.

LEMMA 3.1 [7, pp. 224–226]. *Consider the integral equation*

$$(3.5) \quad f(t) = h(t) + \int_0^t k(t-s)f(s) ds,$$

where $h \in L^P(0, T)$, $k \in L^1(0, T)$ and are both positive. Equation (3.5) has a unique solution $f \in L^P(0, T)$ and

$$(3.6) \quad \|f\|_{L^P(0, T)} \leq \frac{e^{\alpha T}}{1 - M_\alpha} \|h\|_{L^P[0, T]},$$

where α is chosen so that $M_\alpha = \int_0^T e^{-\alpha s} k(s) ds < 1$. If instead of equality in (3.5) we have inequality, then the estimate (3.6) remains valid.

We now proceed to extend the results on V_t proved in [7].

LEMMA 3.2. *If $H1(B, T, Z, U, \tilde{W}, g_1)$ is satisfied, then:*

(a) *Equation (3.3) has a unique solution V_t which is a strongly continuous semigroup on Z .*

(b) *The growth constant of V_t is bounded by*

$$(3.7) \quad \alpha = w_1 + \|B\| \|F\| (\|g_1\|_{L^q(0, t_1)} + m_1),$$

where w_1 is the growth constant of T_t .

(c) *There exists $\bar{g}_1 \in L^q(0, t_1)$ such that $H1(B, V, Z, U, \tilde{W}, \bar{g}_1)$ is satisfied.*

(d) *If $H2(C, T, Z, Y, \tilde{W}, g_2)$ is satisfied with $g_1 g_2 \in L^1(0, t_1)$, then there exists a $\bar{g}_2 \in L^r(0, t_1)$ such that $H2(C, V, Z, Y, \tilde{W}, \bar{g}_2)$ is satisfied.*

Proof. (a) This assertion is proved in [7] by first establishing a solution of (3.3) in $L^P[(0, t_2); \mathcal{L}(Z)]$ analogously to the proof of Lemma 3.1. That the solution V_t is a strongly continuous semigroup is then proved directly using (3.3).

(b) We now proceed to show that (3.3) has a solution $L^P[(0, T); \mathcal{L}(Z)]$ for all T by showing that α in Lemma 3.1 can be chosen to be independent of T . The corresponding $k(t) = \|B\| \|F\| g_1(t)$ where $g_1(t)$ satisfies (d) of H1. This follows from taking estimates of (3.3)

$$\|V_t z\| \leq \|T_t z\| + \int_0^t g_1(t-s) \|B\| \|F\| \|V_s z\| ds$$

and identifying $f(t) = \|V_t z\|$, $h(t) = \|T_t z\|$. Now

$$\begin{aligned} M_\alpha &= \int_0^T e^{-\alpha s} k(s) ds \\ &= \|B\| \|F\| \left[\int_0^{t_1} g_1(s) e^{-\alpha s} ds + \int_{t_1}^T g_1(s) e^{-\alpha s} ds \right] \\ &\leq \|B\| \|F\| \left(\frac{2}{\alpha} \|g_1\|_{L^q(0, t_1)} + \int_{t_1}^T m_1 e^{(w_1 - \alpha)s} ds \right) \quad \text{by (d) of H1} \\ &\leq \|B\| \|F\| \left(\frac{2\|g_1\|}{\alpha} + \frac{2m_1}{\alpha - w_1} \right) \quad \text{choosing } \alpha > w_1. \end{aligned}$$

So

$$(3.8) \quad M_\alpha \leq \frac{2\|B\| \|F\|}{\alpha - w_1} (\|g_1\| + m_1).$$

A suitable choice for α is given by

$$(3.9) \quad \alpha = w_1 + 2\|B\| \|F\| (\|g_1\| + m_1).$$

This ensures an $M_\alpha < 1$ and independent of T and so (3.3) has a unique solution in $L^p[0, T; \mathcal{L}(Z)]$ for all finite T . You can prove that it is a strongly continuous semigroup on $[0, T]$ as in [7]. We now proceed to estimate $\|V_t z\|$ from (3.3).

$$\begin{aligned} \|V_t z\| &\leq \|T_t z\| + \int_0^t \|T_{t-s} B F V_s z\| ds \\ &\leq m e^{w_1 t} \|z\| + \int_0^t g_1(t-s) \|B\| \|F\| \|V_s\| \|z\| ds, \end{aligned}$$

so for all $t > 0$, we have

$$\begin{aligned} \|e^{-\alpha t} V_t\| &\leq m e^{(w_1 - \alpha)t} + \int_0^t e^{-\alpha(t-s)} g_1(t-s) \|B\| \|F\| \|e^{-\alpha s} V_s\| ds \\ &\leq m e^{(w_1 - \alpha)t} + M_\alpha \sup_{0 \leq s \leq t} \|e^{-\alpha s} V_s\| \quad \text{from (3.8).} \end{aligned}$$

Now $\alpha > w_1$, and $M_\alpha < 1$ is independent of t and so

$$\sup_{0 \leq s \leq t} \|e^{-\alpha s} V_s\| \leq \frac{m}{1 - M_\alpha}$$

or, in other words,

$$(3.10) \quad \|V_t\| \leq \frac{m}{1 - M_\alpha} e^{\alpha t}$$

and α given by (3.9) represents a bound for the growth constant of V_t .

(c) From (3.3) and (d) of H1 we obtain the estimate for $w \in \tilde{W}$

$$\|V_t w\|_Z \leq g_1(t) \|w\|_{\tilde{W}} + \int_0^t g_1(t-s) \|B\| \|F\| \|V_s w\|_Z ds.$$

Now for $0 \leq t \leq t_1$, letting $f(t) = \|V_t w\|_Z$, we have

$$\begin{aligned} f(t) &\leq g_1(t) \|w\|_{\tilde{W}} + \|B\| \|F\| \|g_1\|_q \|f\|_{L^1[0, t_1]} \\ &\leq g_1(t) \|w\|_{\tilde{W}} + \|B\| \|F\| \|g_1\|_q \frac{e^{\alpha t_1}}{1 - M_\alpha} \|g_1\|_1 \|w\|_{\tilde{W}} \quad \text{from Lemma 3.1.} \end{aligned}$$

Thus we can define $\bar{g}_1(t) = g_1(t) + \|B\| \|F\| \|g_1\|_q (e^{\alpha t_1} / (1 - M_\alpha)) \|g_1\|_1$ on $(0, t_1)$.

(d) From (3.3) (d) of H1 and (d) of H2, we obtain the following estimate

$$\begin{aligned} \|V_t z\|_{\tilde{W}} &\leq g_2(t) \|z\|_Z + \int_0^t g_2\left(\frac{t-s}{2}\right) g_1\left(\frac{t-s}{2}\right) \|B\| \|F\| \|V_s z\|_Z ds \\ &\leq g_2(t) \|z\|_Z + \int_0^t g_2\left(\frac{t-s}{2}\right) g_1\left(\frac{t-s}{2}\right) K e^{\alpha s} \|z\|_Z ds \quad \text{by (b)} \\ &\leq \bar{g}_2(t) \|z\|_Z, \end{aligned}$$

where

$$\begin{aligned}\bar{g}_2(t) &= g_2(t) + K \int_0^{t/2} g_1(u) g_2(u) e^{\alpha(t-2u)} du \\ &\leq g_2(t) + K e^{\alpha t} \int_0^{t/2} g_1(u) g_2(u) du\end{aligned}$$

and so if $g_1 g_2 \in L^1(0, t_1)$, $\bar{g}_2 \in L^q(0, t_1)$.

Using the duality referred to after (3.4), one can deduce obvious dual versions of this lemma for C perturbations.

The stability estimate obtained in (3.7) is not very sharp; if B is bounded we obtain an upper estimate of $w_1 + m_1 \|B\| \|F\|$, $m = m_1$. In fact a better way of examining the stability of perturbations of a semigroup is to examine the spectrum of its infinitesimal generator. As is well known [29] a necessary and sufficient condition for this is

A1. *The spectrum determined growth assumption.* The following condition:

$$\sup \operatorname{Re} \sigma(A) = \lim_{t \rightarrow \infty} \frac{\log \|T_t\|}{t}$$

holds for analytic semigroups as was recently shown in [8], also for our class of unbounded perturbations.

LEMMA 3.3 [8]. *Let T_t be a strongly continuous semigroup with infinitesimal generator A and suppose that V_t is the perturbed semigroup by an $F \in \mathcal{L}(Z, U)$ under $H1(B, T_b, Z, U, \tilde{W}, g_1)$.*

(a) *For $\operatorname{Re} \lambda$ greater than the maximum growth constant of T_t and V_t , $R(\lambda, A)B$ is bounded and*

$$R(\lambda, A_1)z = R(\lambda, A)z + R(\lambda, A)BR(\lambda, A_1)z, \quad z \in Z,$$

where A_1 is the infinitesimal generator of V_t .

(b) *If T_t is analytic, then V_t is analytic and V_t satisfies A1.*

The dual result for C perturbations under $H2(C, T_b, Z, U, \tilde{W}, g_2)$ also holds.

We now prove a generalization of [25, Prop. 4.7] to the unbounded case.

LEMMA 3.4. *Let S_t be a strongly continuous semigroup on a Banach space X with generator A_0 and suppose that the semigroup V_t is defined by (3.3) under assumption $H1(B, T_b, Z, U, \tilde{W}, g_1)$ with $F \in \mathcal{L}(Z, U)$ and that V_t has the generator A_1 .*

Consider the following integral equation on $Z \oplus X$

$$(3.11) \quad U_t h = \tilde{T}_t h + \int_0^t \tilde{T}_{t-s} \begin{pmatrix} 0 & BD \\ 0 & 0 \end{pmatrix} U_s h ds,$$

where $\tilde{T}_t = \begin{pmatrix} V_t & 0 \\ 0 & S_t \end{pmatrix}$ is the strongly continuous semigroup generated by $\begin{pmatrix} A_1 & 0 \\ 0 & A_0 \end{pmatrix}$.

Then (3.11) has a unique solution $U_t \in \mathcal{L}(Z \oplus X)$ which is a strongly continuous semigroup on $Z \oplus X$ for all $D \in \mathcal{L}(X, U)$.

If the growth rates of V_t and S_t are bounded by α and w_2 respectively, then the growth rate of U_t is bounded by $\mu = \max(\alpha, w_2)$.

Proof. The growth rate of \tilde{T}_t is clearly bounded by μ . From Lemma 3.2(c), V_t satisfies $H1(B, V_b, Z, U, \tilde{W}, \bar{g}_1)$ and it is readily verified that $H1(\tilde{B}, \tilde{T}_b, Z \oplus X, U, \tilde{W} \oplus X, \tilde{g})$ holds for $\tilde{g}(t) = k\bar{g}_1(t)$ on $(0, t_1)$ for some $k > 0$. So by Lemma 3.2(a), (3.11) has a unique solution U_t which is a strongly continuous semigroup. Let

$$U_t = \begin{pmatrix} U_t^{11} & U_t^{12} \\ U_t^{21} & U_t^{22} \end{pmatrix}$$

with respect to $Z \oplus X$ and substitute in (3.11), giving

$$\begin{pmatrix} U_t^1 & U_t^{12} \\ U_t^{21} & U_t^2 \end{pmatrix} h = \begin{pmatrix} V_t & 0 \\ 0 & S_t \end{pmatrix} h + \int_0^t \begin{pmatrix} V_{t-s} & 0 \\ 0 & S_{t-s} \end{pmatrix} \begin{pmatrix} 0 & BD \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_s^1 & U_s^{12} \\ U_s^{21} & U_s^2 \end{pmatrix} h \, ds,$$

and equating components yields

$$\begin{aligned} U_t^1 &= V_t + \int_0^t V_{t-s} BD U_s^{21} \, ds = V_t, & U_t^{21} &= 0, \\ U_t^2 &= S_t, & U_t^{12} &= \int_0^t V_{t-s} BD U_s^2 \, ds = \int_0^t V_{t-s} BDS_s \, ds. \end{aligned}$$

Now

$$\begin{aligned} \left\| U_t \begin{pmatrix} z \\ x \end{pmatrix} \right\| &\leq \max \{ \|V_t z + U_t^{12} x\|, \|S_t x\| \} \\ &\leq \max \{ m e^{\alpha t} \|z\| + \|U_t^{12} x\|, m_2 e^{w_2 t} \|x\| \} \quad \text{by assumption.} \end{aligned}$$

Now

$$\begin{aligned} \|U_t^{12} x\| &\leq \int_0^t \|V_{s/2} V_{s/2} BDS_{t-s} x\| \, ds \\ &\leq \int_0^t m e^{\alpha s/2} \bar{g}_1\left(\frac{s}{2}\right) \|B\| \|D\| \|S_{t-s} x\| \, ds \quad \text{by Lemma 3.2(c)} \\ &\leq m m_2 \|B\| \|D\| \int_0^{t/2} \bar{g}_1(s) e^{w_2(t-2s)+\alpha s} \, ds \|x\| \\ &\leq m m_2 \|B\| \|D\| \left(c_1 \|\bar{g}_1\|_{L^q(0,t_1)} + \int_{t_1}^{t/2} \bar{m}_1 e^{w_2 t 2s(\alpha-w_2)} \, ds \right) \|x\| \quad \text{for } t > 2t_1 \\ &\leq (k_1 e^{w_2 t} + k_2 e^{\alpha t}) \|x\| \quad \text{for all } t. \end{aligned}$$

So the growth rate is bounded by $\mu = \max(\alpha, w_2)$.

We now consider stabilizability by unbounded feedback.

DEFINITION 3.5. We say that (3.1) under $H1(B, T_b, Z, U, \tilde{W}, g_1)$ is *stabilizable* if there exists an $F \in \mathcal{L}(U, Z)$ such that the strongly continuous semigroup V_t defined by (3.3) is exponentially stable, i.e., $\|V_t\| \leq m e^{-\alpha t}$ for some $m, \alpha > 0$.

A key assumption in feedback stabilizability is the following assumption on the generator:

A2. The spectrum decomposition assumption. For every $\delta > 0$, define $\sigma_u(A) = \sigma(A) \cap \{\lambda : \operatorname{Re} \lambda \geq -\delta\}$, $\sigma_s(A) = \sigma(A) \cap \{\lambda : \operatorname{Re} \lambda < -\delta\}$.

We assume that $\sigma_u(A)$ is bounded and separated from $\sigma_s(A)$, so that a simple rectifiable closed curve Γ can be drawn so as to enclose an open set containing $\sigma_u(A)$ in its interior and $\sigma_s(A)$ in its exterior.

This induces a natural state space decomposition

$$Z = Z^u \oplus Z^s, \quad \text{where } Z^u = \Pi Z, \quad Z^s = (I - \Pi)Z$$

and $\Pi = (1/2\pi i) \int_{\Gamma} (\lambda I - A)^{-1} d\lambda$ is a bounded projection in Z . $A^u = A|Z^u$ is bounded, $A^s = A|Z^s$, and $\sigma(A^s) = \sigma_s(A)$, $\sigma(A^u) = \sigma_u(A)$. Furthermore, Π reduces T_b , by which we mean that Π and $(I - \Pi)$ commute with A and T_b , and $T_t^u = \Pi T_t$ is the semigroup generated by A^u ($T_t^u = e^{A^u t}$). $T_t^s = (I - \Pi)T_t$ is the semigroup generated by A^s .

The major result in stabilizability by bounded feedback of Triggiani in [29] can be extended to the unbounded case ([16], [7]). As the latter results are incomplete, we include a simpler proof here.

THEOREM 3.6. *Suppose that A satisfies assumption A2, A^s satisfies assumption A1, and suppose that $H1(B, T_b, Z, U, \tilde{W}, g_1)$ is satisfied. Then if ΠB is bounded and $(A^u, \Pi B)$ is stabilizable on Z^u , (3.1) is stabilizable by the feedback $u = BFz$ for some $F \in \mathcal{L}(U, Z)$. (By Π here we mean the extension of the projection $\Pi: Z \rightarrow Z^u$ to \tilde{W} .)*

Proof. We note first that if ΠB is bounded and $F = F\Pi$, then $V_t^u = \Pi V_t$ is a strongly continuous semigroup on Z^u defined by

$$(3.12) \quad V_t^u = T_t^u + \int_0^t T_{t-s}^u \Pi B F V_s^u ds = \exp(A^u + \Pi B F)t.$$

Since $(A^u, \Pi B)$ is stabilizable there exists an $F_0 \in \mathcal{L}(U, Z^u)$, such that V_t^u is stable, with $F = (F_0 \ 0)$. We are now in the situation of Lemma 3.4:

$$\begin{pmatrix} A^u + \Pi B F_0 & 0 \\ 0 & A^s \end{pmatrix}$$

generates a strongly continuous semigroup, as does its perturbation by

$$\begin{pmatrix} 0 & 0 \\ (I - \Pi)B F_0 & 0 \end{pmatrix}$$

and $\Pi = (1/2\pi i) \int_{\Gamma} (\lambda I - A)^{-1} d\lambda$ is a bounded projection in Z . $A^u = A|Z^u$ is bounded, T_t^s and that of V_t . Since A^s satisfies A1, the growth rate of T_t^s is $< -\delta$ and so V_t is exponentially stable with $F = (F_0 \ 0)$.

It is tempting to try to replace assumption A1 on A_s with the requirement that A have compact resolvent and that A generate an analytic semigroup. Then by Lemma 3.3, we know that the stability of V_t is determined by the spectrum of its infinitesimal generator. The problem here is that we do not have an explicit representation for the generator at this stage; we need further assumptions concerning B .

We now merely state the dual results for detectability as the proofs are analogous to those of Theorem 3.6.

DEFINITION 3.7. We say that $y = CT_t z_0$ under $H2(C, T_b, Z, Y, \tilde{W}, g_2)$ is *detectable* if there exists a $C \in \mathcal{L}(Z, Y)$, so that the strongly continuous semigroup U_t defined by (3.4) with $S_t = T_t$ is exponentially stable.

THEOREM 3.8. *Suppose that A satisfies assumption A2, A^s satisfies assumption A1 and that $H2(C, T_b, Z, Y, \tilde{W}, g_2)$ is satisfied. Then if $C^u = C|Z^u$ is bounded and (A^u, C^u) is detectable on Z^u , $y = CT_t z_0$ is detectable via a $G \in \mathcal{L}(Y, Z)$.*

Proof. If $G_0 \in \mathcal{L}(Y, Z^u)$ makes $A^u + G_0 C^u$ stable, then we choose $G = \begin{pmatrix} G_0 \\ 0 \end{pmatrix}$. The resulting decay rate is bounded by the maximum of $-\delta$ and that of $\exp((A^u + G_0 C^u)t)$. We note that $U_t^u = U_t|Z^u = \exp(A^u + G_0 C^u)t$.

We are now in a position to establish some results about our general extended operator for our compensated system. We consider (3.1) and (3.2) with $S_t = T_t$ and $X = Z$ under assumptions $H1(B, T_b, Z, U, \tilde{W}, g_1)$ and $H2(C, T_b, Z, Y, \tilde{W}, g_2)$ and consider the auxiliary system

$$(3.13) \quad \dot{w} = Mw + Ly, \quad u = Qw$$

with state space $W \cong R^n$ and $M \in \mathcal{L}(W)$, $L \in \mathcal{L}(Y, W)$, $Q \in \mathcal{L}(W, U)$.

The extended system operator is then given "formally" by

$$A_e = \begin{pmatrix} A & BD \\ LC & M \end{pmatrix},$$

although in general A_e will not be a generator of a strongly continuous semigroup, as B maps out of Z . Firstly we note that as a consequence of Lemma 3.4 $\begin{pmatrix} A & 0 \\ 0 & M \end{pmatrix}$ generates the strongly continuous semigroup

$$\tilde{T}_t = \begin{pmatrix} T_t & 0 \\ 0 & e^{Mt} \end{pmatrix} \quad \text{on } Z \oplus W.$$

It is easily seen that $H1(\begin{pmatrix} B \\ 0 \end{pmatrix}, \tilde{T}_t, Z \oplus W, U, \tilde{W} \oplus W, \tilde{g}_1)$ will be satisfied for a $\tilde{g}_1(t) = Kg_1(t)$ on $[0, t_1)$ for a suitable constant K and w the growth constant of \tilde{T}_t . So under the perturbation $\begin{pmatrix} B \\ 0 \end{pmatrix}(Q \ 0)$, we obtain a well defined semigroup \tilde{V}_t on $Z \oplus W$ given by

$$(3.14) \quad \tilde{V}_t h = \tilde{T}_t h + \int_0^t \tilde{T}_{t-s} \begin{pmatrix} 0 & BQ \\ 0 & 0 \end{pmatrix} \tilde{V}_s h \, ds.$$

From Lemma 3.2(c), it follows that there exists a suitable \tilde{g}_1 such that $H1(\begin{pmatrix} B \\ 0 \end{pmatrix}, \tilde{V}_t, Z \oplus W, U, \tilde{W} \oplus W, g_1)$ is satisfied. If furthermore, $g_1 g_2 \in L^1(0, t_1)$ then by Lemma 3.2(d), there exists a suitable \tilde{g}_2 such that $H2(\begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix}, \tilde{V}_t, Z \oplus W, Y, \tilde{W} \oplus W, \tilde{g}_2)$ is satisfied. This means that the perturbation of \tilde{V}_t by $\begin{pmatrix} 0 & 0 \\ LC & 0 \end{pmatrix}$ also generates a strongly continuous semigroup T_t^e on $Z \oplus W$, which is defined as the unique solution of

$$(3.15) \quad T_t^e h = \tilde{V}_t h + \int_0^t T_{t-s}^e \begin{pmatrix} 0 & 0 \\ LC & 0 \end{pmatrix} \tilde{V}_s h \, ds$$

and the infinitesimal generator of T_t^e is the desired "interpretation" of A_e . We collect our results in the following lemma.

LEMMA 3.9. *Under the assumptions $H1(B, T_t, Z, U, \tilde{W}, g_1)$ and $H2(C, T_t, Z, Y, \tilde{W}, g_2)$, if $g_1 g_2 \in L^1(0, t_1)$, then the following extended system is well-defined on the extended state space $Z \oplus W$, ($W \cong R^n$):*

$$(3.16) \quad \begin{aligned} z(t) &= T_t z_0 + \int_0^t T_{t-s} B u(s) \, ds, & y(t) &= C z(t), \\ \dot{w} &= M w + L y, & u &= Q w \end{aligned}$$

for all $M \in \mathcal{L}(W)$, $L \in \mathcal{L}(Y, W)$, $Q \in \mathcal{L}(W, U)$, and

$$(3.17) \quad z_e(t) = \begin{pmatrix} z \\ w \end{pmatrix} = T_t^e \begin{pmatrix} z_0 \\ w_0 \end{pmatrix},$$

where T_t^e is defined by (3.15).

We define the useful isomorphisms first introduced by Schumacher in [26]. Let Z^n be an n -dimensional subspace of Z and let R be the isomorphism between Z^n and $W \cong R^n$.

Then $M = \{ \begin{pmatrix} x \\ R x \end{pmatrix} | x \in Z^n \}$ is an n -dimensional subspace of $Z \oplus W$ which induces the isomorphism

$$T: w \rightarrow \begin{pmatrix} R^{-1} w \\ w \end{pmatrix} \quad \text{between } W \text{ and } M.$$

It is readily verified that H is an isomorphism on $Z \oplus W$

$$H = \begin{pmatrix} I & 0 \\ R\Pi_n & -T^{-1} \end{pmatrix}, \quad H^{-1} = \begin{pmatrix} I & 0 \\ TR\Pi_n & -T \end{pmatrix},$$

where Π_n is the projection from Z onto Z^n , (or its extension to \tilde{W}), and $\Pi_n^{-1}x = \begin{pmatrix} x \\ 0 \end{pmatrix}$ for $x \in Z^n$.

If A has a discrete spectrum, then by increasing δ in the decomposition in A2, we obtain a sequence of finite dimensional subspaces $\{Z^n\}$ of Z , with corresponding projection operator Π_n ; we choose n to correspond to the dimension of the subspace Z^n .

LEMMA 3.10. Suppose that $H_1(B, T, Z, U, \tilde{W}, g_1)$ and $H_2(C, T, Z, Y, \tilde{W}, g_2)$ are satisfied and that $g_1 g_2 \in L^1(0, t_1)$. Supposing that A has a discrete spectrum and that $\Pi_n B, C|Z^n$ are bounded, we choose $n \geq p$ and let

$$Q = F_0 \Pi_p R^{-1}, \quad L = -R \Pi_n^{-1} G_0, \quad M = R(A + \Pi_n B F_0 \Pi_p + \Pi_n^{-1} G_0 C) r^{-1},$$

where $F_0 \in \mathcal{L}(Z^p, U)$, $G_0 \in \mathcal{L}(Y, Z^p)$, then the extended semigroup \tilde{T}_t^e is well defined on $Z \oplus W$ and is similar to the semigroup T_t^e on $Z \oplus M$ via

$$(3.18) \quad T_t^e = H \tilde{T}_t^e H^{-1}.$$

\tilde{T}_t^e is defined by the following integral equations:

$$(3.19) \quad \tilde{T}_t^e h = T_t^1 h + \int_0^t \tilde{T}_{t-s}^e \begin{pmatrix} 0 & 0 \\ TR \Pi_n^{-1} G_0 (C - C \Pi_n) & 0 \end{pmatrix} T_s^1 h ds, \quad h \in Z \oplus M,$$

$$(3.20) \quad T_t^1 = \begin{pmatrix} V_t & T_t^{12} \\ 0 & TR U_t R^{-1} T^{-1} \end{pmatrix},$$

$$(3.21) \quad T_t^{12} h = - \int_0^t V_{t-s} B F_0 \Pi_p U_s R^{-1} T^{-1} h ds, \quad h \in M$$

and V_t and U_t are defined by (3.3) and (3.4) with $F = F_0 \Pi_p$, $G = \Pi_p G_0$, $S_t = T_t$.

Proof. That T_t^e is well defined follows from Lemma 3.9 and the fact that Q, L and M map into the right spaces: $A: Z^n \rightarrow Z^n \forall n$. We now consider $\tilde{T}_t^e = TR U_t R^{-1} T^{-1}$ will be a well defined semigroup, if the restriction of U_t to Z^n is well defined semigroup. Now U_t is given by

$$U_t x = T_t x + \int_0^t U_{t-s} \Pi_n^{-1} G_0 C T_s x ds.$$

Thus

$$U_t \Pi_n x = T_n \Pi_n x + \int_0^t U_{t-s} \Pi_n \Pi_n^{-1} G_0 C T_s \Pi_n x ds,$$

and this says that $U_t \Pi_n = \exp \{A^n + \Pi_n^{-1} G_0 C|Z^n\} t$, a finite dimensional semigroup ($T_t \Pi_n = \exp \{A^n t\}$). Thus

$$\tilde{V}_t = \begin{pmatrix} V_t & 0 \\ 0 & TR U_t R^{-1} T^{-1} \end{pmatrix}$$

is a well defined strongly continuous semigroup, with

$$\tilde{B} = \begin{pmatrix} 0 & -B F_0 \Pi_p R^{-1} T^{-1} \\ 0 & 0 \end{pmatrix}, \quad \tilde{g}_1(t) = K \tilde{g}_1(t) \quad \text{on } (0, t_1),$$

where $\bar{g}_1(t)$ is as in Lemma 3.2(c), we see that $H1(\tilde{B}, \tilde{V}, Z \oplus W, U, \tilde{W} \oplus W, \tilde{g}_1)$ is satisfied and so T_t^1 defined by (3.20), (3.21) is a well defined semigroup on $Z \oplus W$ by Lemma 3.2(a). \tilde{T}_t^e will be well-defined if we can find a suitable $\tilde{g}_2(t)$ such that $H2(\tilde{C}, T_t^1, Z \oplus W, Y, \tilde{W} \oplus W, \tilde{g}_2)$ is satisfied for

$$\tilde{C} = \begin{pmatrix} 0 & 0 \\ TR\Pi_n^{-1}G_0(C - C\Pi_n) & 0 \end{pmatrix}.$$

Consider

$$\begin{aligned} \left\| T_t^1 \begin{pmatrix} z \\ y \end{pmatrix} \right\|_{W \oplus W} &= \max \{ \|V_t z + T_t^{12} y\|_W, \|TRU_t R^{-1} T^{-1} y\|_W \} \quad \text{from (3.20)} \\ &\leq \max \{ \|V_t z\|_W + \|T_t^{12} y\|_W, \bar{K} e^{\mu t} \|y\|_W \} \\ &\leq \max \{ \bar{g}_2(t) \|z\|_W + \|T_t^{12} y\|_W, \bar{K} e^{\mu t} \|y\|_W \} \end{aligned}$$

since $H2(C, V, Z, Y, \tilde{W}, \bar{g}_2)$ is satisfied by Lemma 3.2(d). Now

$$\begin{aligned} \|T_t^{12} y\|_W &\leq \int_0^t \|V_{t-s} B F_0 \Pi_p U_s R^{-1} T^{-1} y\|_W ds \\ &\leq \int_0^t \bar{g}_2 \left(\frac{t-s}{2} \right) \|V_{(t-s)/2} B F_0 \Pi_p U_s R^{-1} T^{-1} y\|_Z ds \quad \text{by } H2(C, V, Z, Y, \tilde{W}, \bar{g}_2) \\ &\leq \int_0^t \bar{g}_2 \left(\frac{t-s}{2} \right) \bar{g}_1 \left(\frac{t-s}{2} \right) \|B\| C e^{\mu s} \|y\|_W ds, \\ &\quad \text{by } H1(B, V, Z, U, \tilde{W}, \bar{g}_1), \text{ Lemma 3.2(c).} \end{aligned}$$

Now from Lemma 3.2(c), (d), we see that we can take \bar{g}_1 and \bar{g}_2 to be

$$\bar{g}_1(t) = g_1(t) + K_1 e^{\alpha t}, \quad \bar{g}_2(t) = g_2(t) + K_2 e^{\alpha t}.$$

Thus $\bar{g}_1 \bar{g}_2 \in L^1(0, t_1)$ and the required \tilde{g}_2 exists.

So both T_t^e and \tilde{T}_t^e are well-defined strongly continuous semigroups, and if B and C were also bounded, we would be able to write the generators A_e and \tilde{A}_e

$$(3.22) \quad A_e = \begin{pmatrix} A & BFR^{-1} \\ -RGC & R(A + \Pi_n BF + GC)R^{-1} \end{pmatrix},$$

$$(3.23) \quad \tilde{A}_e = \begin{pmatrix} A + BF & -BFRT^{-1} \\ TRG(C - C\Pi_n) & TR(A + GC)R^{-1}T^{-1} \end{pmatrix}$$

with $F = F_0 \Pi_p$, $G = \Pi_p G_0$. Using (3.22) and (3.23) it is trivial to verify that

$$A_e = H \tilde{A}_e H^{-1}$$

from which would follow (3.18).

However, we are not justified in writing the infinitesimal generators by (3.22), (3.23) at this stage, which means that (3.18) must be verified using the integral formulas, which is incredibly tedious, and is therefore omitted. We stress that we have proved the important fact that every intermediate semigroup is well defined by its integral equations.

Combining the results of this section we state our existence result concerning finite dimensional compensators which generalizes § 2.

THEOREM 3.11. *We consider the system (3.1) and (3.2) with $S_t = T_t$, $X = Z$. Under the following assumptions, there exists a finite dimensional compensator:*

(a) $H1(B, T_t, Z, U, \tilde{W}, g_1)$ and $H2(C, T_t, Z, Y, \tilde{W}, g_2)$ are satisfied with $g_1 g_2 \in L^1(0, t)$.

(b) A satisfies the spectrum decomposition assumption A2 and A^s satisfies the spectrum determined growth assumption A1.

(c) U and Y are finite dimensional and $\Pi_p B, C|Z^p$ are bounded.

(d) $\lim_{n \rightarrow \infty} \|C - C\Pi_n\|_{\mathcal{L}(W, Y)} = 0$.

(e1) A has a discrete spectrum and $(AP, \Pi_p B, C|Z^p)$ is minimal for all p . (We shall call this modal minimality.)

(e2) For a $p > 0$, Z^p is finite dimensional and A^s is exponentially stable and $(AP, \Pi_p B, C|Z^p)$ is minimal.

Proof. Under both sets of assumptions we choose a $\delta > 0$, so that $Z^u = Z^p$ and A^s is exponentially stable. Then from (d) and (e1) or (e2), we can find an $F_0 \in \mathcal{L}(Z^p, U)$ and $G_0 \in \mathcal{L}(Y, Z^p)$ so that $A^p + BF_0\Pi_p$ and $A^p + G_0C|Z^p$ are exponentially stable. Under assumptions (a)–(d) we may apply Theorems 3.6 and 3.8 to deduce that V_t given by (3.3) and U_t given by (3.4) with $S_t = T_t$, $F = F_0\Pi_p$ and $G = \begin{pmatrix} G_0 \\ 0 \end{pmatrix} := \Pi_p^{-1}G_0$ are exponentially stable with decay rate δ . Note that the gain operators G and F are now fixed, and depend only on p .

We now construct the compensator (3.16) as in Lemma 3.10 choosing $W = R^n$, $n \geq p$. We examine the stability of \tilde{T}_t^e , as it is similar to T_t^e by (5.18). Now \tilde{T}_t^e is a perturbation of T_t^1 defined by (3.19) and by Lemma 3.4, we see that the growth rate of T_t^1 is the maximum of that of V_t and $TRU_tR^{-1}T^{-1}$, which is just $-\delta$. We can now apply Lemma 3.2(b) to ascertain the effect of the perturbation

$$\begin{pmatrix} 0 & 0 \\ TR\Pi_n^{-1}G_0(C - C\Pi_n) & 0 \end{pmatrix} \text{ on } T_t^1.$$

The new growth rate is bounded by

$$-\delta + K\|G_0\|\|C - C\Pi_n\|_{L(W, Y)},$$

where K depends on the norm of $\tilde{g}_2(t)$ used in Lemma 3.10 and thus depends on p . Since the norms of T and R are independent of n in our construction, we see that $\tilde{g}_2(t)$ and K are also independent of n ; they do depend on p , or course. With p fixed, assumption (d) shows us that for a sufficiently large n , the new growth rate will be negative. If assumption (e1) holds, then by initially choosing p large enough, we can improve the stability by then increasing n .

The above result is constructive, but we have no a priori way of knowing the values of p and n needed to achieve stability. One might try to check the stability of T_t^e by calculating the eigenvalues of its infinitesimal generator as done by Schumacher in [26] for the bounded case. Some initial results in this direction are:

LEMMA 3.12. *Under assumptions (a), (c), (e1) or (e2) of Theorem 3.11, if A generates an analytic semigroup, then A^s and T_t^e satisfy the spectrum determined growth assumption.*

Proof. (a) A^s is a projection of A which generates a strongly continuous semigroup and so A^s also generates an analytic semigroup.

(b) Let

$$\tilde{T}_t = \begin{pmatrix} T_t & 0 \\ 0 & e^{Mt} \end{pmatrix}, \quad \tilde{B} = \begin{pmatrix} B \\ 0 \end{pmatrix},$$

Q, M as in Lemma 3.10. Now since $\begin{pmatrix} A & 0 \\ 0 & M \end{pmatrix}$ will generate an analytic semigroup $H1(\tilde{B}, \tilde{T}, Z \oplus W, U, \tilde{W}, Cg_1)$ is satisfied for some constant C and so by Lemma 3.3(b), \tilde{V}_t is analytic. \tilde{V}_t is the perturbation of \tilde{T}_t under $\begin{pmatrix} 0 & BQ \\ 0 & 0 \end{pmatrix}$. Now there exists a suitable \tilde{g}_2 such that $H2((C0), \tilde{V}, Z \oplus W, Y, \tilde{W} \oplus W, \tilde{g}_2)$ is satisfied since $g_1 g_2 \in L^1(0, t_1)$ and so by the dual to Lemma 3.3(b), T_t^e defined by (3.15) satisfies the spectrum growth assumption.

LEMMA 3.13. *Under the assumptions of Theorem 3.11, if A has compact resolvent, then so does the infinitesimal generator of T_t^e, A_e .*

Proof. If \tilde{A} is the infinitesimal generator of \tilde{V}_t defined by (3.14) then by Lemma 3.3(a), for $\operatorname{Re} \lambda$ sufficiently large, $R(\lambda, \begin{pmatrix} A & 0 \\ 0 & M \end{pmatrix}) \begin{pmatrix} 0 & BQ \\ 0 & 0 \end{pmatrix}$ is bounded and

$$(3.24) \quad R(\lambda, \tilde{A})x = R\left(\lambda, \begin{pmatrix} A & 0 \\ 0 & M \end{pmatrix}\right)x + R\left(\lambda, \begin{pmatrix} A & 0 \\ 0 & M \end{pmatrix}\right) \begin{pmatrix} 0 & BQ \\ 0 & 0 \end{pmatrix} R(\lambda, \tilde{A})x,$$

where Q and M are as in Lemma 3.10. Now $R(\lambda, \begin{pmatrix} A & 0 \\ 0 & M \end{pmatrix})$ is compact, since $R(\lambda, A)$ is, and W is finite dimensional. The range of the operator in (3.24) is finite dimensional since $Q = F_0 \Pi_p R^{-1}$. Thus $R(\lambda, \tilde{A})$ is compact.

T_t^e is the perturbation of \tilde{V}_t by $\begin{pmatrix} 0 & 0 \\ LC & 0 \end{pmatrix}$ with the finite range operator $L = -\Pi_n^{-1} G_0$. By the dual of lemma 3.3(a), for sufficiently large λ , $\begin{pmatrix} 0 & 0 \\ LC & 0 \end{pmatrix} R(\lambda, \tilde{A})$ is bounded and $R(\lambda, A_e)$ is compact since

$$(3.25) \quad R(\lambda, A_e)x = R(\lambda, \tilde{A})x + R(\lambda, A_e) \begin{pmatrix} 0 & 0 \\ LC & 0 \end{pmatrix} R(\lambda, \tilde{A})x.$$

We remark finally if A has compact resolvent, then $\sigma(A)$ is discrete and A2 is automatically satisfied.

4. Verifiable criteria for stability. In order to test the stability of a finite dimensional compensator designed according to Theorem 3.11, we need to have an explicit representation for the infinitesimal generator of the extended semigroup T_t^e or \tilde{T}_t^e . Although the rather weak assumptions H1 and H2 guarantee that the perturbed semigroups exist, we do not know what the resulting infinitesimal generator looks like. To overcome this problem we introduce some (provisional) stronger assumptions.

Assumptions A3.

- (4.1) A has compact resolvent.
- (4.2) A generates an analytic semigroup.
- (4.3) C is A -bounded, that is $D(C) \supset D(A)$ and there exist constants $m_1 \geq 0, m_2 \geq 0$ such that

$$\|Cz\|_Y \leq m_1 \|z\| + m_2 \|Az\| \quad \text{for } z \in D(A).$$

- (4.4) U and Y are finite dimensional and B is bounded. (That $C|Z^p$ is bounded follows from (4.3) and the fact that $A|Z^p$ is bounded.)

Now B bounded implies H1, but (4.3) does not necessarily imply H2, although the two conditions are very close in spirit. Nonetheless the construction of the compensator outlined in Theorem 3.11 can be carried out under the above assumptions, noting that A -bounded perturbations of analytic semigroups generate strongly continuous semigroups. In fact A -degenerate perturbations (A -bounded and finite range) of analytic semigroups generate analytic semigroups [33]. Thus from (4.2), we see that $A, A + BF$ and $A + GC$ all generate analytic semigroups. Since W is finite

dimensional, for any $Q \in \mathcal{L}(W)$, $A_1 = \begin{pmatrix} A & 0 \\ 0 & M \end{pmatrix}$ generates an analytic semigroup and since $\begin{pmatrix} 0 & 0 \\ -RGC & 0 \end{pmatrix}$ is A_1 -degenerate, $\begin{pmatrix} A & 0 \\ -RGC & M \end{pmatrix}$ generates an analytic semigroup. Finally, $\begin{pmatrix} 0 & BFR^{-1} \\ 0 & 0 \end{pmatrix}$ is a bounded, finite-rank perturbation and so A_e generates an analytic semigroup, where

$$(4.5) \quad A_e = \begin{pmatrix} A & BFR^{-1} \\ -RGC & R(A + \Pi_n BF + GC)R^{-1} \end{pmatrix}.$$

Thus A_e given by (4.5) is the infinitesimal generator of T_t^e and $D(A_e) = D(A) \oplus W$. Moreover, A_e satisfies the spectrum determined growth assumption. Similarly, the infinitesimal generator \tilde{A}_e of \tilde{T}_t^e is given by

$$(4.6) \quad \tilde{A}_e = \begin{pmatrix} A + BF & -BFR^{-1}T^{-1} \\ TRG(C - C\Pi_n) & TR(A + GC)R^{-1}T^{-1} \end{pmatrix}$$

with $D(\tilde{A}_e) = D(A) \oplus M$.

Thus under Assumptions A3, we can determine the stability of our compensator by determining $\sigma(\tilde{A}_e)$. As in Schumacher [26] we can appeal to the perturbation result of Weinstein-Aronzajn in Kato [16], since \tilde{A}_e generates a strongly continuous semigroup, the degenerate case ($\sigma(\tilde{A}_e) = C$) is excluded and since \tilde{A}_e is an A_0 -degenerate perturbation of $A_0 = \begin{pmatrix} A & 0 \\ 0 & RAR^{-1} \end{pmatrix}$, we may conclude that $\sigma(\tilde{A}_e)$ is discrete and can be determined via the formula

$$(4.7) \quad \nu(\lambda, \tilde{A}_e) = \nu(\lambda, A_1) + \bar{\nu}(\lambda, f) \quad \text{for } \lambda \in C,$$

where

$$A_1 = \begin{pmatrix} A + BF & -BFR^{-1}T^{-1} \\ 0 & TR(A + GC)R^{-1}T^{-1} \end{pmatrix},$$

and the multiplicity functions $\nu(\lambda, A)$ and $\bar{\nu}(\lambda, f)$ are defined by

$$(4.8) \quad \nu(\lambda, A) = \begin{cases} 0 & \text{if } \lambda \in \rho(A), \\ \text{dimension of the eigenspace} & \text{if } \lambda \in \sigma(A), \end{cases}$$

$$(4.9) \quad \bar{\nu}(\lambda, f) = \begin{cases} k & \text{if } \lambda \text{ is a zero of } f \text{ of order } k, \\ -k & \text{if } \lambda \text{ is a pole of } f \text{ of order } k, \\ 0 & \text{otherwise,} \end{cases}$$

where f is the following $W-A$ determinant

$$(4.10) \quad f(\lambda) = \det(I + Q(A_1 - \lambda I)^{-1}) = \det((\tilde{A}_e - \lambda I)(A_1 - \lambda I)^{-1}),$$

where

$$Q = \begin{pmatrix} 0 & 0 \\ TRG(G - C\Pi_n) & 0 \end{pmatrix}$$

is the A_1 -degenerate perturbation. We note that $\sigma(A_1) = \sigma(A + BF) \cup \sigma(TR(A + GC)R^{-1}T^{-1}) = \sigma(A^p + \Pi_p BF_0) \cup \sigma_s(A) \cup \sigma(A^n + \Pi_n^{-1}G_0C|Z_n)$ and $f(\lambda)$ reduces to the following determinant of an $n \times n$ matrix

$$(4.11) \quad \begin{aligned} f(\lambda) &= \det[I + G(C - C\Pi_n)(\lambda - A - BF)^{-1}BF(\lambda - A^n - \Pi_n^{-1}GC|Z_n)^{-1}] \\ &= \det[I + (C - \Pi_n)(\lambda - A - BF)^{-1}BF(\lambda - A^n - \Pi_n^{-1}GC|Z_n)^{-1}G] \end{aligned}$$

by a result on determinants of degenerate operators [16, p. 244], since the dual of Lemma 3.3(a) guarantees that $(C - C\Pi_n)(\lambda - A - BF)^{-1}$ is bounded. Equation (4.11) can be evaluated as an infinite series in Fourier coefficients with respect to the

eigenvectors of A . It is interesting to compare this with the analogous formula for $\tilde{f}(\lambda)$ obtained by Schumacher in [26]:

$$(4.12) \quad \tilde{f}(\lambda) = \det(I - C(\lambda - A)^{-1}\hat{G}),$$

which may seem simpler at first glance, but it is not, because \hat{G} must be calculated. In (4.10) all the operators are either known a priori or can be chosen, like F_0 and G_0 .

In this section we have shown that under assumptions A3, H2(C, T_b, Z, Y, W, g_2) and if either (A, B, C) is modal minimal or (e2) of Theorem 3.11 is satisfied, then we can construct a finite dimensional compensator, whose stability we can check by using (4.11) to determine the eigenvalues of the extended system operator.

We remark that (4.3) is usually satisfied together with H2, as the conditions are very close. Equations (4.1) and (4.2) are quite reasonable and H1 or H2 usually mean that A is analytic. If (A, B) is approximately controllable and (C, A) is initially observable then $(A^p, \Pi_p B, C|Z^p)$ will be minimal for all $p > 0$. Alternatively, (e2) often holds in applications. The only restrictive assumption is that B be bounded and in § 5 we proceed to circumvent this.

5. Examples. The conditions in § 3 are fairly general and are satisfied by all the examples considered in [6], so there do exist finite dimensional compensators for sufficiently large n . These examples also satisfy all the extra assumptions needed in § 4, except that B is not bounded for boundary control. So here the W-A theory is not applicable, and we have no proven method of knowing when n is sufficiently large. A way around this is to reformulate the problem using different function spaces, as was done by Fattorini in [11] and by Pohjolainen in [22] and we shall outline the main results we shall need.

Let Z be a real separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_0$ and norm $\|\cdot\|_0$. Suppose that $-A$ is an unbounded, positive self-adjoint operator in H_0 , such that $\langle -Ax, x \rangle_0 \geq \alpha \|x\|_0^2$ for some $\alpha > 0$. Then for all real r , $(-A)^r$ is defined on $D_r = D((-A)^r)$ and D_r is dense in Z . This induces a Hilbert space H_r , via the inner product

$$(5.1) \quad \langle x, y \rangle_r = \langle (-A)^r x, (-A)^r y \rangle_0, \quad x, y \in D_r.$$

If $r > 0$, $D_r = H_r$ is complete with respect to the induced norm and if $r < 0$, we define H_r to be the completion of H_0 with respect to the norm.

The restriction of $-A$ to H_{r+1} , $r \geq 0$ is denoted by $-A_r \in \mathcal{L}(H_{r+1}, H_r)$. $-A_r$ is also self-adjoint and ≥ 0 for all $r \geq 0$. For $r < 0$, we define $-A_r$ to be the self-adjoint extension of $-A$ to H_r . Again $-A_r \in \mathcal{L}(H_{r+1}, H_r)$ and is ≥ 0 . A_r generates an analytic semigroup $T_r(t)$ on H_r for all r and $T_r(t)$ coincides with the restriction on extension in H_r of T_b , the analytic semigroup generated by A . We have $\|T_r(t)\|_r \leq M_r e^{\beta t}$ for all r , where $0 > \beta > -\alpha$.

We consider the following class of partial differential equations on the real separate Hilbert space Z

$$(5.2) \quad \dot{z} = A_0 z, \quad z(0) = z_0 \in Z, \quad S_b z(t) = \Gamma u(t), \quad t > 0,$$

where A_0 is a formal differential operator in Z and $S_b: D(S_b) \rightarrow E$ is a linear boundary operator with $D(S_b) \supset D(A_0)$, $u(t) \in U \cong \mathbb{R}^n$ is the control and we assume that $u(t)$ is C^1 and $\Gamma \in \mathcal{L}(U, E)$.

We let $D(A) = \{x \in D(A_0) | S_b x = 0\}$ and $Ax = A_0 x$ for $x \in D(A)$ and suppose that A satisfies all the aforementioned assumptions on Z , and we define H_r , etc. Finally we suppose that there exists an $r > 0$ and $D \in \mathcal{L}(U, H_r)$ such that

$$(5.3) \quad S_b(D_u) = \Gamma u \quad \text{and} \quad AD_u = 0.$$

Then it can be shown that (5.2) has a unique mild solution in Z if and only if the following transformed equation has a unique mild solution in H_{r-1} :

$$(5.4) \quad \dot{z} = A_{r-1}Z - A_{r-1}Du, \quad z(0) = z_0 \in H_{r-1}.$$

We remark that we do need to assume that u is C^1 but this is always the case in our applications and that this enables us to pose our problem on H_{r-1} with $B = -A_{r-1}D \in \mathcal{L}(U, H_{r-1})$, a bounded B operator. This solution of (5.4) is

$$(5.5) \quad z(t) = T_{r-1}(t)z_0 - \int_0^t A_{r-1}T_{r-1}(t-s)Du(s) ds \in D(A_{r-1}) = H_r,$$

since $T(t)$ is analytic. We now consider some examples adapted from [7] and [22].

Example 5.5. Control and observation at points or on the boundary under Neumann boundary conditions.

$$(5.6) \quad \frac{\partial z}{\partial t} = \frac{1}{\pi^2} \frac{\partial^2 z}{\partial x^2}, \quad \frac{\partial z}{\partial x} = 0, \quad \text{at } x = 0, x = 1.$$

$$(5.7) \quad \left[\frac{\partial z}{\partial x} \right]_{\xi_1}^{\xi_1^+} = -u(t), \quad [z(x, t)]_{\xi_1^+}^{\xi_1^+} = 0, \quad 0 \leq \xi_1 \leq 1,$$

$$(5.8) \quad y = z(\xi_2, t), \quad 0 \leq \xi_2 \leq 1.$$

We remark that ξ_1 and ξ_2 may also be on the boundary in this example, as in the case considered by Fujii in [15]. In contrast to [15], we obtain a finite dimensional compensator. Now from [6], we see that with state space $Z = L^2(0, 1)$ and

$$(5.9) \quad A = \frac{1}{\pi^2} \frac{d^2}{dx^2}, \quad D(A) = \{h \in Z; h_x, h_{xx} \in Z \text{ and } h_x(0) = 0 = h_x(1)\}$$

if we choose $\tilde{W} = \mathcal{H}^{1/2+\varepsilon}(0, 1)$, the usual L^2 -Sobolev space [18], then $C \in \mathcal{L}(\tilde{W}, R)$ and the semigroup T_t generated by A satisfies

$$(5.10) \quad \|T_t z\|_{\tilde{W}} \leq \frac{a}{t^{1/4+\varepsilon}} \|z\|_Z.$$

Thus $H_2(C, T, L^2(0, 1), R, \mathcal{H}^{1/2+\varepsilon}(0, 1), g)$ is satisfied with $g = a/(t^{1/4+\varepsilon})$. B can be interpreted as heat injection at the point ξ_1 by recognizing that $B = -\delta = C^*$ in Z , and $T_t = T_t^*$, we see that $H_1(B, T, L^2(0, 1), R, \tilde{W}^*, g)$ is satisfied, and furthermore $g_2 \in L^1(0, t_1)$. So assumption (a) of Theorem 3.11 is satisfied. A satisfies the assumptions of Lemma 3.12 and 3.13 and thus assumption (b) of Theorem 3.11. Assumption (c) is obvious and to examine (d) we note that the eigenvalues of A are $\lambda_n = -n^2$, $n = 0, 1, \dots$ and the eigenfunctions are $1, \phi_n = \sqrt{2} \cos n\pi x$; $n = 1, 2, \dots$. Thus

$$Z^p = \text{span} \{\phi_n; n = 0, 1, \dots, p-1\},$$

$$A^p \cong \text{diag}[0, 1, \dots, -(p-1)^2],$$

$$\Pi_p B \approx (1, \sqrt{2} \cos \pi \xi_1, \dots, \sqrt{2} \cos \pi(p-1)\xi_1),$$

$$C|Z^p = (1, \sqrt{2} \cos \pi \xi_2, \dots, \sqrt{2} \cos \pi(p-1)\xi_2).$$

So $(A^p, \Pi_p B, C|Z^p)$ will be minimal for all p for irrational ξ_1 or ξ_2 or even with $\xi_1 = 0$ or 1 (boundary action). Thus all the assumptions of Theorem 3.11 are easily satisfied and there exists a finite dimensional compensator on $L^2(0, 1)$ for sufficiently large p .

In order to choose p we must reformulate (5.6) so as to obtain a bounded B . To do this we let $w = e^{-t}z$ to obtain the equivalent system

$$(5.11) \quad \frac{\partial w}{\partial t} = \frac{1}{\pi^2} \frac{\partial^2 w}{\partial x^2} - w, \quad \frac{\partial w}{\partial x} = 0 \quad \text{at } x = 0, 1,$$

$$(5.12) \quad \left[\frac{\partial w}{\partial x} \right]_{\xi_1}^{\xi_1^+} = -\bar{u}(t), \quad [w(x, t)]_{\xi_1}^{\xi_1^+} = 0, \quad 0 \leq \xi_1 \leq 1,$$

$$(5.13) \quad \bar{y}(t) = w(\xi_2, t),$$

where $\bar{u}(t) = e^{-t}u(t)$; $\bar{y}(t) = e^{-t}y(t)$.

Then $\bar{A} = (1/\pi^2) d^2/dx^2 - 1$, with $D(\bar{A}) = D(A)$, is self-adjoint with the complete eigenset $-1 - n^2$, ϕ_n is as before and the H_r spaces are well defined with

$$\|f\|_r^2 = \sum_{n=0}^{\infty} (n^2 + 1)^{2r} |\langle f, \phi_n \rangle_0|^2,$$

where $\langle \cdot, \cdot \rangle$ is the inner product in $L^2(0, 1)$. For the case $\xi_1 = 0$, an appropriate D is given by

$$Du = \frac{e}{1-e^2} (e^x + e^2 e^{-x})$$

and for other values of ξ_1 , see [22] or [5].

The Fourier coefficients are $O(1/(1+n^2))$. So $\bar{A}D \in \mathcal{L}(R, H_{-\alpha})$ for any $\alpha > 0$. We consider the following system on the state space $\bar{Z} = H_{-1/2+\varepsilon}$ for any $\varepsilon, 0 < \varepsilon < \frac{1}{2}$:

$$(5.14) \quad \frac{\partial z}{\partial t} = \tilde{A}z + Bu, \quad \frac{\partial z}{\partial x} = 0 \quad \text{at } x = 0, 1,$$

$$(5.15) \quad y(t) = z(\xi_2, t)$$

where $\tilde{A} = \bar{A}_{-1/2+\varepsilon} + 1$ and $B = \bar{A}_{-1/2+\varepsilon} D \in \mathcal{L}(R, H_{-1/2+\varepsilon})$. \tilde{A} now satisfies our assumptions (4.1), (4.2), (4.4) and from [14] we see that C is \tilde{A} -bounded on $H_{-1/2+\varepsilon}$, since for $h \in D(\tilde{A})$

$$(5.16) \quad |Ch| \leq \text{const} \|\tilde{A}h\|_{-1/2} \leq \text{const} \|\tilde{A}h\|_{-1/2+\varepsilon}.$$

It is easily verified that $H_2(C, T_{-1/2+\varepsilon}(t), H_{-1/2+\varepsilon}, R, W, g_2)$ is satisfied with $W = H_{1/2}$ and $g_2(t) = \text{const}/t^{1-\varepsilon}$. So by Theorem 3.11, there exists a finite dimensional compensator on $H_{-1/2+\varepsilon} \oplus R^p$, and its stability is determined by the eigenvalues of \tilde{A}^ε via the W-A formula (4.11). In [5] this construction was carried out for several values of ξ_1 and ξ_2 , with similar results; namely, it was possible to stabilize the system (5.6)–(5.8) by means of a second order compensator: the gains were of order 2 and the decay rate was -1 . The theory in § 4 tells us that the combined system is exponentially stable on $H_{-1/2+\varepsilon} \oplus R^2$ with decay rate -1 . However, T_t^ε is also defined on $L^2(0, 1) \oplus R^2$ and its infinitesimal generator, A_ε , will be the restriction of the generator on $H_{-1/2+\varepsilon} \oplus R^2$ to $L^2(0, 1) \oplus R^2$. By Lemma 3.13, we know that $\sigma(A^\varepsilon)$ is discrete on $L^2(0, 1) \oplus R^2$ and so it remains the same and Lemma 3.12 ensures that the stability of T_t^ε on $L^2(0, 1) \oplus R^2$ is determined by $\sigma(A_\varepsilon)$. Thus the combined system is exponentially stable on $L^2(0, 1) \oplus R^2$ with decay rate -1 .

The following example can be developed in an analogous way and similar numerical calculations for compensators can also be found in [5].

Example 5.2. Control and observation in points under Dirichlet boundary conditions.

$$\begin{aligned}
 \frac{\partial z}{\partial t} &= \frac{1}{\pi^2} \frac{\partial^2 z}{\partial x^2} + 4z, \\
 z(0, t) &= 0 = z(1, t), \\
 \left[\frac{\partial z}{\partial x}(x, t) \right]_{\xi_1^-}^{\xi_1^+} &= -u(t) \quad [z(x, t)]_{\xi_1^-}^{\xi_1^+} = 0 \quad 0 < \xi_1 < 1, \\
 y &= z(\xi_2, t) \quad 0 \leq \xi_2 \leq 1.
 \end{aligned}
 \tag{5.17}$$

We choose the same state space $Z = L^2(0, 1)$ and define A to be

$$A = \frac{1}{\pi^2} \frac{d^2}{dx^2} + 4, \quad D(A) = \{h \in Z; h_x, h_{xx} \in Z \text{ and } h(0) = 0 = h(1)\}.
 \tag{5.18}$$

Then as in Example 5.1, $H1(\delta, T, Z, R, \tilde{W}, g)$ and $H2(C, T, Z, R, \tilde{W}, g)$ are satisfied with $\tilde{W}^* = W = \mathcal{H}^{1/2+\varepsilon}(0, 1)$ and $g = ae^{t/t^{1/4+\varepsilon}}$. T_t is the semigroup generated by A given by (5.18). A has the complete eigenset $\lambda_n = (4 - n^2)$, $\phi_n = \sqrt{2} \sin n\pi x$; $n = 1, 2, \dots$ and

$$\begin{aligned}
 Z^p &= \text{span} \{\phi_1, \dots, \phi_p\}, \\
 A^p &\simeq \text{diag} [3, 0, -5, \dots, (4 - p^2)], \\
 \Pi_p B &\simeq (\sqrt{2} \sin \pi \xi_1, \sqrt{2} \sin 2\pi \xi_1, \dots, \sqrt{2} \sin p\pi \xi_1), \\
 C|Z^p &\simeq (\sqrt{2} \sin \pi \xi_2, \dots, \sqrt{2} \sin p\pi \xi_2).
 \end{aligned}$$

If we choose $\xi_1 = \frac{1}{3}$, $\xi_2 = \frac{3}{4}$, say, then we see that $(A^p, \Pi_p B, C|Z^p)$ is minimal for $p = 2$, but not for $p = 3$. So (e1) of Theorem 3.11 is not satisfied, and we check (e2); this is satisfied, since in [22], we have the estimates

$$\|C\|_{\mathcal{L}(W, R)} = \|c\|_{H_{-1/2}} \quad \text{with } c = \sum_{i=1}^{\infty} \sqrt{2} \sin \frac{3\pi i}{4} \sqrt{2} \sin \pi i x.$$

Thus

$$\|C - C\Pi_n\|_{\mathcal{L}(W, R)}^2 = \text{const} \sum_{i=n+1}^{\infty} \frac{1}{i^2} \left(\sin \frac{3\pi i}{4} \right)^2 \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

So again by Theorem 3.11, there exists a finite dimensional compensator using $p = 2$ and sufficiently large n . To choose n , however, (5.17) must be reformulated to obtain a bounded B . This can be done on $H_{-1/2+\varepsilon}$, as in Example 5.1, and one can finally show that a constructed compensator will be stable on the original $L^2(0, 1) \oplus R^n$.

Finally we consider the more difficult case of boundary control under Dirichlet conditions.

Example 5.3. Boundary control under Dirichlet boundary conditions. We consider Example 5.2 with the boundary control

$$z(0, t) = -u_1(t).$$

In [6] it is shown that this problem can be formulated with $B = \delta'$. B is the dual of the observation operator $\bar{C}h = h_x(0)$ and thus one can show that $H_1(B, T, L^2(0, 1), \bar{W}, R, \bar{g})$ is satisfied with

$$\bar{W} = \mathcal{H}^{3/2+\varepsilon}(0, 1)^* \quad \text{and} \quad \bar{g} = \frac{\text{const } e^{3t}}{t^{3/4+\varepsilon/2}} \quad \text{from [6].}$$

We see then that with point observations as in Examples 5.1 and 5.2 we will not have $\bar{g}g \in L^1(0, t_1)$, so Theorem 3.11 is only justified for bounded observations, for example,

$$Ch = \frac{1}{2\mu} \int_{\xi_2-\mu}^{\xi_2+\mu} h(x) dx.$$

In [22] it is shown that one can formulate the problem with a bounded B on the state space $H_{-3/4-\varepsilon}$, but here again the point observations will not be A -bounded. So we can design finite dimensional compensators for boundary control under Dirichlet conditions, but at the expense of using bounded observations. Unfortunately we have to report that even for the case of bounded observations, we encountered numerical problems in designing a compensator for this example [5], [6].

6. Conclusions. The main result of this paper Theorem 3.11 is the existence of a finite dimensional compensator for a class of parabolic systems, generated by an operator A satisfying various assumptions. In § 4 assumptions A3 require that A have compact resolvent and generate an analytic semigroup and that C be A -bounded and B be bounded, which apart from the bounded B assumption are quite reasonable. In § 5 it was shown how one can get around the unbounded B assumption for self-adjoint A . The question remains what one can achieve for the nonself-adjoint case and this hinges around assumption (d) of Theorem 3.11. For the self-adjoint case we know that the eigenfunctions of A are complete and orthogonal and this together with C being A -bounded satisfies (d). For the general nonself-adjoint case one often has completeness of the eigenfunctions, but not orthogonality. An analysis of the proof of Theorem 3.11 reveals that if one can find a subset of the eigenfunctions, which approximate C in the $L(W, Y)$ norm closely enough, then the result holds. So the existence theorem is quite general, but the implementation is a story apart and that is the reason that we have elaborated on this for self-adjoint systems in § 5.

Although the emphasis in this paper has been on parabolic systems, the compensator design is also applicable to more general distributed systems, for example, the “flexible” distributed systems, which have been proposed as prototypes for large space flexible structures. These are not pure hyperbolic systems as the eigenvalues do not asymptote vertically; see [9] for an example which satisfies our assumptions here.

Since this paper was written, another method has been proposed to accommodate boundary control in [9] and [10]. This results in a compensator of order one greater than the scheme here, but the implementation of the control is via a smoother map and it may produce better results for Dirichlet boundary conditions (cf. Example 5.3).

It is interesting to note that we have been able to design a finite dimensional compensator for the heat equation with boundary control and/or observations, contrary to the expectations of Fujii in [15].

It has recently come to our attention that in the report [25] by Sakawa an almost identical scheme to ours was proposed for *self-adjoint* systems with *bounded* B and C . The notation is different and the work was done quite independently but the construction used is essentially the same.

Acknowledgments. Finally I would like to thank the referees for their painstaking care in refereeing this paper.

REFERENCES

- [1] M.J. BALAS, *Feedback control of flexible systems*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 673–679.
- [2] ———, *Modal control of certain flexible dynamic systems*, this Journal, 16 (1978), pp. 450–462.
- [3] ———, *Feedback control of linear diffusion processes*, Int. J. Control, 29 (1979), pp. 523–533.
- [4] M.K.P. BHAT, *Regulator theory for evolution systems*, Ph.D thesis, Univ. Toronto, Toronto, 1979.
- [5] J. BONTSEMA, *Finite dimensional compensators for parabolic systems*, Tech. Rep., Rijksuniversiteit Groningen, 1982.
- [6] R.F. CURTAIN, *finite-dimensional compensator design for parabolic distributed systems with point sensors and boundary input*, IEEE Trans. Automat. Control, AC-26 (1982), pp. 98–104.
- [7] R.F. CURTAIN AND A.J. PRITCHARD, *Infinite-dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sciences 8, Springer Verlag, Berlin, 1978.
- [8] R.F. CURTAIN, *Stability of infinite-dimensional systems by spectrum analysis: The spectrum determined growth assumption*, Systems and Control Letters, 2 (1982), pp. 106–109.
- [9] ———, *Finite-dimensional compensators for some hyperbolic systems with boundary control*, presented at the Workshop on Control Theory for Distributed Parameter Systems, Vorau, Austria, July 11–17, 1982.
- [10] ———, *Stabilization of boundary control distributed systems via integral dynamic output feedback of a finite-dimensional compensator*, presented at the 5th International Conference on Analysis and Optimization of Systems, INRIA, Versailles, France, Dec. 14–17, 1982 (to be published by Springer Lecture Notes).
- [11] H.O. FATTORINI, *Boundary control systems*, this Journal, 6 (1968), pp. 348–385.
- [12] R. GRESSANG AND G. LAMONT, *Observers for systems characterized by semigroups*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 523–528.
- [13] E. HILLE AND R.S. PHILLIPS, *Functional Analysis and Semigroups*, AMS Colloquium Publications 31, American Mathematical Society, Providence, RI, 1957.
- [14] N. FUJII AND M. HIRAI, *A finite dimensional observer for a class of distributed parameter systems*, Int. J. Control, 32 (1980), pp. 951–962.
- [15] N. FUJII, *Feedback stabilization of distributed parameter systems by a functional observer*, this Journal, 18 (1980), pp. 108–120.
- [16] T. KATO, *Perturbation Theory for Linear Operators*, Springer Verlag, Berlin, 1966.
- [17] H. KWAKERNAAK AND R. SILVAN, *Linear Optimal Control Systems*, Wiley, Interscience, New York, 1972.
- [18] J.L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications I*, Springer Verlag, Berlin, 1972.
- [19] T. NAMBU, *Feedback stabilization of distributed parameter systems of parabolic type*, J. Differential Equations, 33 (1979), pp. 167–188.
- [20] ———, *Feedback stabilization of diffusion equations by a functional observer*, J. Differential Equations, 43 (1982), pp. 257–280.
- [21] P.A. ORNER AND A.M. FOSTER, *A design procedure for a class of distributed parameter control systems*, Trans. ASME (1971), pp. 86–93.
- [22] S. POHJOLAINEN, *Robust multivariable controller for distributed parameter systems*, Ph.D. Thesis, Tampere University of Technology Publ., 9, 1980.
- [23] A.J. PRITCHARD AND J. ZABCZYK, *Stability and stabilizability of infinite-dimensional systems*, SIAM Rev., 23 (1981), pp. 25–52.
- [24] Y. SAKAWA AND T. MARSUSHITA, *Feedback stabilization of a class of distributed systems and construction of a state estimator*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 748–753.
- [25] Y. SAKAWA, *Feedback stabilization of linear diffusion systems*, this Journal, 21 (1983), pp. 667–676.
- [26] J. M. SCHUMACHER, *A direct approach to compensator design for distributed systems*, this Journal, 21 (1983), pp. 823–836.
- [27] ———, *Dynamic feedback in finite- and infinite-dimensional linear systems*, M.C. Tracts No. 143, Mathematisch Centrum, Amsterdam, 1982.
- [28] M. SLEMROD, *Asymptotic behaviour of C-semigroups as determined by the spectrum of the generator*, Indiana Univ. Math. J., 25 (1976), pp. 783–791.
- [29] R. TRIGGIANI, *On the stabilization problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383–403.
- [30] ———, *On Nambu's boundary stabilizability problem for diffusion processes*, J. Differential Equations, 33 (1979), pp. 189–200.
- [31] ———, *Boundary feedback stabilizability of parabolic equations*, J. Appl. Math. Optim., 6 (1980), pp. 201–220.

- [32] K. YOSIDA, *Functional Analysis*, Springer-Verlag, 1965.
- [33] J. ZABCZYK, *On decomposition of generators*, this Journal, 16 (1978), pp. 523–534.
- [34] ———, *Stabilization of Boundary Control Systems*, Lecture Notes in Control and Information Sciences, 14, Springer-Verlag, New York, 1979.

ASYMPTOTIC CONVERGENCE ANALYSIS OF THE PROXIMAL POINT ALGORITHM*

FERNANDO JAVIER LUQUE†

Abstract. The asymptotic convergence of the proximal point algorithm (PPA), for the solution of equations of type $0 \in Tz$, where T is a multivalued maximal monotone operator in a real Hilbert space, is analyzed. When $0 \in Tz$ has a nonempty solution set \bar{Z} , convergence rates are shown to depend on how rapidly T^{-1} grows away from \bar{Z} in a neighbourhood of 0. When this growth is bounded by a power function with exponent s , then for a sequence $\{z^k\}$ generated by the PPA, $\{z^k - \bar{Z}\}$ converges to zero, like $o(k^{-s/2})$, linearly, superlinearly, or in a finite number of steps according to whether $s \in (0, 1)$, $s = 1$, $s \in (1, +\infty)$, or $s = +\infty$.

Key words. monotone maps, algorithms, asymptotic convergence, multiplier methods, convex programming.

1. Introduction. Let H be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$, and induced norm $|\cdot|$, where for all $z \in H$, $|z| = \langle z, z \rangle^{1/2}$. Let us consider a multivalued mapping $T: H \rightarrow 2^H$. Its domain $D(T)$ is defined by

$$D(T) = \{z \in H: Tz \neq \emptyset\},$$

its range by

$$R(T) = \bigcup \{Tz: z \in H\}$$

and its graph by

$$G(T) = \{(z, w) \in H \times H: w \in Tz\}.$$

The inverse point-to-set mapping T^{-1} is defined by $T^{-1}w = \{z \in H: w \in Tz\}$ if $w \in R(T)$ and $T^{-1}w = \emptyset$ otherwise. It is an elementary fact that $D(T^{-1}) = R(T)$, $R(T^{-1}) = D(T)$, and $G(T^{-1}) = \{(w, z) \in H \times H: (z, w) \in G(T)\}$.

Such a mapping T is said to be a monotone operator if and only if

$$\forall z, z' \in D(T), \quad \forall w \in Tz, \quad \forall w' \in Tz' \quad \langle z - z', w - w' \rangle \geq 0.$$

If, in addition, its graph is not properly contained in the graph of any other monotone operator, then T is maximal monotone. For a detailed treatment of the theory and applications of such mappings, the reader may consult the works by Brézis (1973), Browder (1976), Pascali and Sburlan (1978) and the references cited therein.

A fundamental problem is to find a vector $z \in H$ such that $0 \in Tz$. Some of the most important problems in the area of convex programming and related fields can be cast into this general framework.

If f is a closed proper convex function, then $T = \partial f$ is maximal monotone (Moreau (1965)). Thus solving the equation $0 \in Tz$ is equivalent to minimizing the convex function f , since f attains its minimum at \bar{z} if and only if $0 \in \partial f(\bar{z})$.

* Received by the editors September 21, 1981, and in revised form November 2, 1982. This research is part of the author's Ph.D. thesis written under the supervision of Professor D. P. Bertsekas of the Massachusetts Institute of Technology. It was supported by the ITP Foundation of Madrid and the National Science Foundation under grant 79-20834.

† Laboratory for Information and Decision Systems and Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

Let H_1, H_2 be real Hilbert spaces and let K be a closed proper saddle function on $H = H_1 \times H_2$, which is convex in the first argument and concave in the second, and let the subdifferential of K at $(x, y) \in H_1 \times H_2$, $\partial K(x, y)$, be defined as the set of vectors $(u, v) \in H_1 \times H_2$ satisfying

$$\forall (x', y') \in H_1 \times H_2 \quad K(x', y) - \langle x' - x, u \rangle \geq K(x, y) \geq K(x, y') - \langle y' - y, v \rangle,$$

then the multifunction

$$T(x, y) = \{(u, -v) \in H_1 \times H_2 : (u, v) \in \partial K(x, y)\}$$

is a maximal monotone operator (Rockafellar (1970a)). The solutions (\bar{x}, \bar{y}) of the equation $(0, 0) \in T(x, y)$ are the saddle points of K .

A variational inequality problem is to find a vector $\bar{z} \in C$ satisfying

$$\exists \bar{w} \in A\bar{z} : \quad \forall v \in C \quad \langle a - \bar{w}, \bar{z} - v \rangle \geq 0,$$

where $C \subseteq H$ is a nonempty closed convex set, $A : H \rightarrow 2^H$ is a multivalued monotone mapping with $D(A) = C$, and a is a given vector in H . Equivalently, it can be expressed by: Find a vector $\bar{z} \in C$ such that

$$a \in A\bar{z} + N_C(\bar{z}),$$

where $N_C(z)$ is the normal cone to C at z . Its expression valid for all $u \in H$ is (Rockafellar (1970b, p. 15))

$$N_C(u) = \{x \in H : \langle x, u - v \rangle \geq 0 \text{ for all } v \in C\}.$$

When C is a cone and C° denotes its polar, the variational inequality problem above is reduced to the complementarity problem of finding a vector $\bar{z} \in C$ such that

$$\exists \bar{w} \in A\bar{z} : \quad a - \bar{w} \in C^\circ, \quad \langle a - \bar{w}, \bar{z} \rangle = 0.$$

These last two problems can be reduced to solving $0 \in Tz$ for the operator T defined by (Rockafellar (1976a))

$$Tz = \begin{cases} -a + Az + N_C(z), & z \in C, \\ \emptyset, & z \notin C. \end{cases}$$

Conditions for the maximal monotonicity of such operators, T were given by Rockafellar (1970c, Thm. 5). Further results are contained in papers by Rockafellar (1978), (1980) and McLinden (1980).

We will now introduce the Proximal Point Algorithm (PPA). Most of the notation has been borrowed from Rockafellar (1976a).

Minty (1962) proved that if T is a maximal monotone operator and c is a positive constant, for any $u \in H$ there is a unique z such that $u \in (I + cT)z$. The operator $P = (I + cT)^{-1}$ (the proximal mapping associated with cT in the terminology of Moreau (1965)) is thus single-valued from all of H into H . The monotonicity of T is a necessary and sufficient condition for the nonexpansiveness of P wherever P is defined (Brézis (1973, Prop. 2.1, p. 21)). Thus

$$\forall z, z' \in H \quad |Pz - Pz'| \leq |z - z'|.$$

The PPA generates, for any starting point $z^0 \in H$, a sequence $\{z^k\}$ according to the rule

$$z^{k+1} \cong P_k z^k$$

where $P_k = (I + c_k T)^{-1}$ and $\{c_k\}$ is some sequence of positive real numbers.

The criterion for the approximate computation of z^{k+1} used in this analysis will be

$$(A_r) \quad |z^{k+1} - P_k z^k| \leq \varepsilon_k \min \{1, |z^{k+1} - z^k|^r\}$$

where r and ε_k satisfy

$$r \geq 0, \quad \forall k \quad \varepsilon_k \geq 0, \quad \sum_{k=0}^{\infty} \varepsilon_k < +\infty.$$

It has been shown by Rockafellar (1976a, Thm. 1) that when $0 \in Tz$ has at least one solution, the condition $|z^{k+1} - P_k z^k| \leq \varepsilon_k$ is a sufficient condition for $|z^{k+1} - z^k| \rightarrow 0$. Therefore when the PPA is implemented with criterion (A_r) $r \geq 1$, there exists some $k' \in \mathbb{Z}_+$ such that for all $k \geq k'$, $|z^{k+1} - z^k|^r \leq |z^{k+1} - z^k| < 1$, and thus the larger r is, the more accurate the computation of z^{k+1} will be. In previous papers dealing with the PPA (Rockafellar (1976a), (1978), (1980)), the value of r was always taken equal to 1, but as will be shown below, one takes r strictly greater than one in order to achieve superlinear convergence of order greater than one.

As shown by Rockafellar (1976a, Prop. 3), the estimate $|z^{k+1} - P_k z^k| \leq c_k \text{dist}(0, S_k z^{k+1})$ holds for all k , where $S_k z = Tz + c_k^{-1}(z - z^k)$. Therefore criterion (A_r) is implied by

$$(A'_r) \quad \text{dist}(0, S_k z^{k+1}) \leq \frac{\varepsilon_k}{c_k} \min \{1, |z^{k+1} - z^k|^r\}.$$

The set of solutions (possibly empty) of the equation $0 \in Tz$, will be denoted by $\bar{Z} = \{z \in H : 0 \in Tz\}$. When T is maximal monotone, for every $u \in H$, $T^{-1}u = \{z \in H : u \in Tz\}$ is a, possibly empty, closed, convex set (Minty (1964, Thm. 1)). Therefore $\bar{Z} = T^{-1}0$ is closed and convex. If \bar{Z} is nonempty, the vector in \bar{Z} closest to z will be denoted by \bar{z} . We will use the notation

$$|z - \bar{Z}| = \min \{|z - z'| : z' \in \bar{Z}\} = |z - \bar{z}|.$$

Our analysis will focus on the convergence properties of the sequence $\{|z^k - \bar{Z}|\}$ corresponding to any sequence $\{z^k\}$ generated by the PPA.

In addition to the proximal mappings $P_k = (I + c_k T)^{-1}$ where $c_k > 0$ and T is a maximal monotone operator, use will also be made of the mappings Q_k defined by

$$Q_k = I - P_k.$$

Clearly $0 \in Tz \Leftrightarrow P_k z = z \Leftrightarrow Q_k z = 0$.

Rockafellar (1976a, Prop. 1) proved the following facts:

$$(1.1) \quad \forall k \geq 0, \quad \forall z \in H \quad c_k^{-1} Q_k z^k \in TP_k z^k,$$

$$(1.2) \quad \forall k \geq 0, \quad \forall z, z' \in H \quad |P_k z - P_k z'|^2 + |Q_k z - Q_k z'|^2 \leq |z - z'|^2.$$

In the same paper, the following theorem was also proved.

THEOREM 1.1 (Rockafellar (1976a, Thm. 1)). *Let $\bar{Z} = T^{-1}0 \neq \emptyset$, and let $\{z^k\}$ be any sequence generated by the PPA with stopping criterion (A_r) , $r \geq 0$, and a sequence of positive numbers $\{c_k\}$ such that $\liminf_{k \rightarrow \infty} c_k > 0$. Then $\{z^k\}$ is bounded and converges in the weak topology to a unique point $z^\infty \in \bar{Z}$. Also*

$$(1.3) \quad 0 = \lim_{k \rightarrow \infty} |c_k^{-1} Q_k z^k| = \lim_{k \rightarrow \infty} |Q_k z^k| = \lim_{k \rightarrow \infty} |z^{k+1} - z^k|.$$

This paper addresses the issue of the speed of convergence of the PPA, in both its exact and approximate versions. Under various hypotheses, we show linear convergence, superlinear convergence, convergence in a finite number of steps and convergence in one step. A condition that implies sublinear convergence is given, and an estimate of the convergence rate in this case is also provided. The results previously available on the speed of convergence of the PPA have been reported by Rockafellar (1976a) for the case in which $\bar{Z} = \{\bar{z}\}$. If T^{-1} is Lipschitz continuous at 0 with modulus $a \geq 0$, then the approximate algorithm with $\{c_k\}$ nondecreasing converges linearly at a rate bounded by $a/(a^2 + c_\infty^2)^{1/2}$, which becomes superlinear if $\{c_k\}$ is unbounded (Rockafellar (1976a, Thm. 2, p. 883)). If $0 \in \text{int } T\bar{z}$, the exact algorithm with $\liminf_{k \rightarrow \infty} c_k > 0$ converges in a finite number of steps, while the approximate one with $\{c_k\}$ nondecreasing achieves superlinear convergence without requiring that $\{c_k\}$ be unbounded (Rockafellar (1976a, Thm. 3, p. 888)).

The hypotheses used are of a geometric nature and concern the growth properties of the multivalued mapping T^{-1} in a neighbourhood of 0, away from the solution set. The general form of these growth conditions is

$$\exists \delta > 0: \quad \forall w \in B(0, \delta), \quad \forall z \in T^{-1}w \quad |z - \bar{z}| \leq f(|w|),$$

where $B(0, \delta) = \{x \in H: |x| < \delta\}$, and $f: [0, +\infty) \rightarrow [0, +\infty)$ is a continuous function such that $f(0) = 0$. This type of assumption was suggested to the author by Professor D. Bertsekas of the Massachusetts Institute of Technology. The characterizations of Rockafellar (1976a) discussed above are special cases of this one.

When f is linear with slope $a > 0$, we are able to guarantee linear convergence at a rate bounded by $a/(a^2 + c_\infty^2)^{1/2}$. This is valid for both the exact and the approximate versions of the PPA with $r \geq 1$ in criterion (A_r) . By means of an example we show that this bound is tight. The extension to general solution sets \bar{Z} , and the proof of tightness of the bound, are new.

As shown by Rockafellar (1976b), the quadratic method of multipliers for convex programming is a realization of the PPA in which $T = -\partial g$, g being the essential objective function of the ordinary dual program. Taking this into account, our Theorem 2.1 allows some extensions of the circumstances under which the quadratic method of multipliers achieves linear convergence as reported in Kort and Bertsekas (1973), (1976) and Rockafellar (1976b).

When f is a power function with exponent $s \geq 1$, we show in Theorem 3.1 that superlinear convergence of order at least s is obtained for the exact algorithm. For the approximate implementation, with criterion (A_r) and $r \geq 1$, the order of convergence is at least $\min\{r, s\}$. This result is entirely new. A comparison is made with results on the superlinear convergence of the quadratic method of multipliers reported by Kort and Bertsekas (1973), (1976).

If f is flat in some neighbourhood of 0 in $[0, +\infty)$, i.e., if there is some $\delta > 0$ such that $f(x) = 0$ for all $x \in [0, \delta)$, then the exact algorithm converges in a finite number of steps. The approximate version of the algorithm, with stopping criterion (A_r) , $r \geq 1$, achieves superlinear convergence of order r at least. A sufficient condition for the convergence of the exact algorithm in a single step is also given.

When T^{-1} is such that its growth exceeds any linear bounding in any neighbourhood, however small, of zero, it is proved in Theorem 4.1 that then the PPA will converge sublinearly when the penalty sequence $\{c_k\}$ remains bounded. This result is valid for both the exact and approximate versions. To the best of our knowledge this is the first result dealing with sublinear convergence of the PPA.

Finally, if f is a power function with exponent $s \in (0, 1)$, we give a conservative estimate of the speed of convergence. It is shown that $|z^k - \bar{Z}|$ decreases to zero faster than $k^{-s/2}$. This result is also new.

This section ends with a proposition on the global convergence of the PPA which will be used repeatedly in what follows.

PROPOSITION 1.2. *Let $\bar{Z} \neq \emptyset$, and let $\{z^k\}$ be any sequence generated by the PPA with stopping criterion (A_r) , $r \geq 0$, and a sequence of positive numbers $\{c_k\}$ such that $\liminf_{k \rightarrow \infty} c_k > 0$. Let us also assume that*

$$(1.4) \quad \exists \delta > 0: \quad \forall w \in B(0, \delta), \quad \forall z \in T^{-1}w \quad |z - \bar{Z}| \leq f(|w|),$$

where $B(0, \delta) = \{z \in H: |z| < \delta\}$ and $f: [0, +\infty) \rightarrow [0, +\infty)$ is such that $f(0) = 0$ and is upper semicontinuous at 0 (in this case equivalent to continuity at 0). Then $|z^k - \bar{Z}| \rightarrow 0$.

Proof. By Theorem 1.1, (1.3), we have $|c_k^{-1}Q_k z^k| \rightarrow 0$. There exists, then, some $k_1 \in \mathbb{Z}_+$ such that $|c_k^{-1}Q_k z^k| < \delta$ for all $k \geq k_1$. By (1.1) and assumption (1.4),

$$\forall k \geq k_1 \quad |P_k z^k - \bar{Z}| \leq f(|c_k^{-1}Q_k z^k|).$$

Using the continuity of f ,

$$\limsup_{k \rightarrow \infty} |P_k z^k - \bar{Z}| \leq \lim_{k \rightarrow \infty} f(|c_k^{-1}Q_k z^k|) = 0,$$

from which it follows that $|P_k z^k - \bar{Z}| \rightarrow 0$.

Also the definition of $|z - \bar{Z}|$, the triangle inequality and criterion (A_r) yield for all k

$$\begin{aligned} |z^{k+1} - \bar{Z}| &\leq |z^{k+1} - \overline{P_k z^k}| \leq |z^{k+1} - P_k z^k| + |P_k z^k - \overline{P_k z^k}| \\ &\leq \varepsilon_k \min\{1, |z^{k+1} - z^k|^r\} + |P_k z^k - \bar{Z}|. \end{aligned}$$

Since $\varepsilon_k \rightarrow 0$, the result follows.

Remark. Condition (1.4) is not necessary. To see this consider in $l^2(\mathbb{N})$ the quadratic function $q(x) = \sum_{i=0}^{\infty} x_i^2 / (i+1)$. The PPA for $T = \partial q$ converges strongly to the unique solution $x = 0$ (Kryanev (1973)). Nonetheless, the graph of T is as flat as we may want in any neighbourhood of zero, and such an f does not exist.

2. Linear convergence. In this section a sufficient condition for the linear convergence of $\{|z^k - \bar{Z}|\}$ to zero when the PPA is operated in an approximate manner is provided. The upper bound on the rate of convergence is shown to be tight. Implications for the quadratic method of multipliers are pointed out, and comparisons with previous results are made. The main result of this section is embodied in the following theorem.

THEOREM 2.1. *Let $\bar{Z} \neq \emptyset$, and let $\{z^k\}$ be any sequence generated by the PPA with stopping criterion (A_r) , $r \geq 1$, and a nondecreasing sequence of positive numbers $\{c_k\}$ such that $0 < c_k \uparrow c_\infty \leq +\infty$. Let us also assume that*

$$(2.1) \quad \exists a > 0, \quad \exists \delta > 0: \quad \forall w \in B(0, \delta), \quad \forall z \in T^{-1}w \quad |z - \bar{Z}| \leq a|w|.$$

Then $|z^k - \bar{Z}| \rightarrow 0$ linearly with a rate bounded from above by $a/(a^2 + c_\infty^2)^{1/2} < 1$. If $c_\infty = +\infty$, the convergence is superlinear.

Proof. The hypothesis implies that of Theorem 1.1, thus its conclusion is in force. By (1.3) there exists some $k_1 \in \mathbb{Z}_+$ such that $|c_k^{-1}Q_k z^k| < \delta$ for all $k \geq k_1$. Using formula (1.1) and assumption (2.1),

$$(2.2) \quad \forall k \geq k_1 \quad |P_k z^k - \bar{Z}| \leq \frac{a}{c_k} |Q_k z^k|.$$

Equation (1.2) with $z = z^k$, $z' = \bar{z}^k \in \bar{Z}$ (thus $P_k z' = z'$, $Q_k z' = 0$), together with the fact $|P_k z^k - \bar{z}^k| \geq |P_k z^k - \bar{Z}|$, yields

$$(2.3) \quad \forall k \geq 0 \quad |Q_k z^k|^2 \leq |z^k - \bar{Z}|^2 - |P_k z^k - \bar{Z}|^2.$$

Using (2.3) to eliminate $|Q_k z^k|$ in (2.2) and rearranging, we obtain

$$\forall k \geq k_1 \quad |P_k z^k - \bar{Z}|^2 \frac{c_k^2 + a^2}{a^2} \leq |z^k - \bar{Z}|^2.$$

Introducing $\mu_k = a/(a^2 + c_k^2)^{1/2} < 1$ we obtain

$$(2.4) \quad \forall k \geq k_1 \quad |P_k z^k - \bar{Z}| \leq \mu_k |z^k - \bar{Z}|.$$

From (2.3) we have

$$(2.5) \quad \forall k \geq 0 \quad |Q_k z^k| \leq |z^k - \bar{Z}|.$$

The triangle inequality gives

$$\forall k \geq 0 \quad |z^k - \overline{P_k z^k}| \leq |z^k - \bar{z}^k| + |\bar{z}^k - \overline{P_k z^k}|.$$

Projection onto a nonempty closed convex set (\bar{Z} in our case) is a nonexpansive operation (in fact it is a proximal mapping; see Moreau (1965, p. 279)). Thus $|z^k - \overline{P_k z^k}| \leq |z^k - P_k z^k|$, and using (2.5),

$$(2.6) \quad \forall k \geq 0 \quad |z^k - \overline{P_k z^k}| \leq 2|z^k - \bar{Z}|.$$

By Theorem 1.1, $|z^{k+1} - z^k| \rightarrow 0$, and therefore there exists an index $k_2 \in \mathbb{Z}_+$ such that for all $k \geq k_2$ $|z^{k+1} - z^k| < 1$. If $r \geq 1$, then for all $k \geq k_2$ $|z^{k+1} - z^k|^r \leq |z^{k+1} - z^k|$, and criterion (A_r) can be used to obtain the estimate

$$\begin{aligned} \forall k \geq k_2 \quad |z^{k+1} - \overline{P_k z^k}| &\leq |z^{k+1} - P_k z^k| + |P_k z^k - \overline{P_k z^k}| \\ &\leq \varepsilon_k |z^{k+1} - z^k| + |P_k z^k - \bar{Z}| \\ &\leq \varepsilon_k |z^{k+1} - \overline{P_k z^k}| + \varepsilon_k |z^k - \overline{P_k z^k}| + |P_k z^k - \bar{Z}|. \end{aligned}$$

But $|z^{k+1} - \overline{P_k z^k}| \geq |z^{k+1} - \bar{Z}|$. By criterion (A_r) , $\varepsilon_k \rightarrow 0$; thus there is some $k_3 \in \mathbb{Z}_+$ that for all $k \geq k_2$ $|z^{k+1} - z^k| < 1$. If $r \geq 1$, then for all $k \geq k_2$ $|z^{k+1} - z^k|^r \leq |z^{k+1} - z^k|$, obtain

$$(2.7) \quad \forall k \geq \tilde{k} \quad |P_k z^k - \bar{Z}| \geq (1 - \varepsilon_k) |z^{k+1} - \bar{Z}| - 2\varepsilon_k |z^k - \bar{Z}|.$$

Let $\bar{k} = \max\{k_1, \tilde{k}\}$. Combining (2.4) and (2.7),

$$\forall k \geq \bar{k} \quad |z^{k+1} - \bar{Z}| \leq \frac{\mu_k + 2\varepsilon_k}{1 - \varepsilon_k} |z^k - \bar{Z}|.$$

Thus the rate of linear convergence β satisfies

$$(2.8) \quad \beta \leq \limsup_{k \rightarrow \infty} \frac{|z^{k+1} - \bar{Z}|}{|z^k - \bar{Z}|} \leq \lim_{k \rightarrow \infty} \frac{\mu_k + 2\varepsilon_k}{1 - \varepsilon_k} = \frac{a}{(a^2 + c_\infty^2)^{1/2}} < 1.$$

Example. We will show by means of an example that the bound for the rate of linear convergence obtained in Theorem 2.1 is achieved.

Let us consider in $H = \mathbb{R}^2$ the linear transformation $A : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, where A is given by the matrix

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

(Its effect is to rotate vectors counterclockwise by an angle of $\pi/2$.) Let us consider the quadratic form $\langle z, Az \rangle$. It is obvious that $\langle z, Az \rangle = 0$ for all $z \in \mathbb{R}^2$. The mapping $T: \mathbb{R}^2 \rightarrow 2^{\mathbb{R}^2}$, given by $Tz = \{Az\}$, is monotone because

$$\forall z, z' \in \mathbb{R}^2 \quad \langle z - z', Az - Az' \rangle = \langle z - z', A(z - z') \rangle = 0.$$

Clearly T is single valued and is continuous in \mathbb{R}^2 . Therefore it is maximal monotone (Pascali and Sburlan (1978, Cor. 2.3, p. 106)). T is not the subdifferential mapping of any proper lower semicontinuous convex function $f: \mathbb{R}^2 \rightarrow (-\infty, +\infty]$, as A is not selfadjoint (Rockafellar (1970b, p. 240)).

It is easy to see that $|Az| = |z|$ for all $z \in \mathbb{R}^2$. Therefore $\bar{Z} = \{0\}$, the constants a, δ appearing in assumption (2.4) are $a = 1, \delta = +\infty$, and the inequality $|z - \bar{Z}| \leq a|w|$ becomes $|z| = |w|$, valid for all $w \in \mathbb{R}^2$ and all $z \in T^{-1}w$, i.e., $z = A^{-1}w$.

When the PPA is implemented in its exact form, i.e., $\varepsilon_k \equiv 0$, it becomes $z^{k+1} = (I + c_k T)^{-1} z^k$, and in our case we obtain $|z^{k+1}| = |(I + c_k A)^{-1} z^k|$.

Elementary computations show that

$$(I + c_k A)^{-1} = \frac{1}{1 + c_k^2} \begin{pmatrix} 1 & c_k \\ -c_k & 1 \end{pmatrix},$$

and also that

$$|z^{k+1}| = |(I + c_k A)^{-1} z^k| = \frac{|z^k|}{\sqrt{1 + c_k^2}}.$$

Therefore the convergence rate is

$$\beta = \lim_{k \rightarrow \infty} \frac{|z^{k+1}|}{|z^k|} = \lim_{k \rightarrow \infty} \frac{1}{\sqrt{1 + c_k^2}} = \frac{1}{\sqrt{1 + c_\infty^2}}.$$

Since in this example $a = 1$, we have $\beta = a/(a^2 + c_\infty^2)^{1/2}$, and the bound is achieved.

Let us consider the following convex programming problem:

$$(2.9) \quad \begin{aligned} & \min f_0(x) \\ & \text{s.t. } f_i(x) \leq 0, \quad i = 1, 2, \dots, m, \quad x \in C, \end{aligned}$$

where C is a nonempty closed convex subset of \mathbb{R}^n and $f_i: C \rightarrow \mathbb{R}$ is a lower semicontinuous convex function for $i = 0, 1, \dots, m$.

Its ordinary dual problem is

$$(2.10) \quad \max g_0(y) \quad \text{s.t. } y \geq 0$$

where $g_0: \mathbb{R}_+^m \rightarrow \bar{\mathbb{R}}$ is the concave function defined by

$$g_0(y) = \inf_{x \in C} \{f_0(x) + y_1 f_1(x) + \dots + y_m f_m(x)\}.$$

The quadratic method of multipliers involves a sequence of minimizations of the augmented Lagrangian function

$$L(x, y, c) = \begin{cases} f_0(x) + \frac{1}{2c} \sum_{i=1}^m [(\max\{0, y_i + c f_i(x)\})^2 - y_i^2], & c > 0, \quad x \in C, \\ +\infty, & x \notin C. \end{cases}$$

The method of multipliers can thus be expressed as

$$(2.11) \quad \begin{aligned} x^k &= \arg \min_x L(x, y^k, c_k), \\ y_i^{k+1} &= \max \{0, y_i^k + c_k f_i(x^k)\}, \quad i = 1, 2, \dots, m. \end{aligned}$$

Rockafellar (1976b) has shown that the quadratic method of multipliers (2.11) for the solution of (2.9) is a realization of the PPA, in which $T = -\partial g$, where g is the essential objective function of the dual problem (2.10) defined by

$$g(y) = \begin{cases} g_0(y) & \text{if } y \in \mathbb{R}_+^m, \\ -\infty & \text{if } y \notin \mathbb{R}_+^m, \end{cases}$$

assuming that g is proper, i.e. $\sup g > -\infty$, so that T is maximal monotone. He assumes that T^{-1} is Lipschitz continuous at the origin—which implies that there is only one Lagrange multiplier vector \bar{y} —and that the minimization to determine x^k in (2.11) is carried out only approximately with stopping criterion.

$$L(x^{k+1}, y^k, c_k) - \inf_x L(x, y^k, c_k) \leq (\varepsilon_k^2 / 2c_k) |y^{k+1} - y^k|^2,$$

where y^{k+1} is given as in (2.11), $\varepsilon_k \geq 0$ for all k , and $\sum_{k=0}^{\infty} \varepsilon_k < +\infty$. He concludes that the sequences $\{y^k\}$ generated by the algorithm converge towards \bar{y} linearly, at a rate bounded as in (2.8). He also shows that

$$|y^{k+1} - P_k y^k|^2 / 2c_k \leq L(x^{k+1}, y^k, c_k) - \inf_x L(x, y^k, c_k),$$

and thus the stopping criterion implies (A_r) with $r = 1$.

Kort and Bertsekas (1973), (1976) have also studied the convergence for this method of multipliers. In their analysis they assume:

(i) Problem (2.9) has a nonempty compact solution set \bar{X} , and a nonempty compact set of Lagrange multipliers \bar{Y} .

(ii) f_0 is strongly convex with modulus $\mu > 0$. This implies that $\bar{X} = \{\bar{x}\}$.

(iii) The following growth condition on the dual function:

$$\exists b > 0, \quad \exists \delta > 0: \quad \forall y \in B(\bar{Y}, \delta) \quad g(y) \leq \bar{g} - \frac{1}{b} |y - \bar{Y}|^2,$$

where

$$\bar{g} = \max_y g, \quad B(\bar{Y}, \delta) = \{y \in \mathbb{R}^m: |y - \bar{Y}| < \delta\},$$

and

$$|y - \bar{Y}| = \min \{|y - y'|: y' \in \bar{Y}\},$$

which is well defined, as \bar{Y} is closed and convex.

The strong convexity assumption (ii) allows them to develop an implementable criterion which implies the following:

$$(2.12) \quad |y^{k+1} - P_k y^k|^2 \leq \frac{\eta_k}{2\mu} |y^{k+1} - y^k|^2,$$

where η_k is a prespecified sequence such that $\eta_k < 2\mu$ for all k large enough. When minimization of the augmented Lagrangian is carried out exactly, (ii) need not be assumed. The growth condition (iii) implies (2.1) with $a = b$. Linear convergence is

guaranteed if $\bar{\eta} < 4\mu\bar{c}/b$ where $\bar{\eta} = \limsup_{k \rightarrow \infty} \eta_k$, $\bar{c} = \limsup_{k \rightarrow \infty} c_k$. If $\eta_k \rightarrow 0$, the rate of linear convergence is bounded (as in the case of exact implementation of (2.11)) by $a/(a + \bar{c})$.

When interpreted in the framework of the method of multipliers, Theorem 2.1 gives a sufficient condition for its linear convergence under still weaker assumptions than those discussed above. First, both \bar{X} and \bar{Y} are required to be only nonempty (they will always be closed and convex by the lower semicontinuity and convexity of the functions f_i , $i = 0, 1, \dots, m$), and no assumption is made on their compactness. Secondly, the strong convexity assumption on f_0 is not made.

3. Superlinear convergence and convergence in a finite number of iterations. The $(Q-)$ order of convergence (Ortega and Rheinboldt (1970)) of $\{z^k - \bar{Z}\}$, assuming that $|z^k - \bar{Z}| \neq 0$ for all k , is the supremum t of the numbers $\tau \geq 1$ such that

$$\limsup_{k \rightarrow \infty} \frac{|z^{k+1} - \bar{Z}|}{|z^k - \bar{Z}|^\tau} < +\infty.$$

THEOREM 3.1. *Let $\bar{Z} \neq \emptyset$, and let $\{z^k\}$ be any sequence generated by the PPA with stopping criterion (A_r) , $r \geq 1$, and a nondecreasing sequence of positive numbers $\{c_k\}$, such that $0 < c_k \uparrow c_\infty \leq +\infty$. Let us also assume that*

$$(3.1) \quad \exists a > 0, \quad \exists s \geq 1, \quad \exists \delta > 0: \quad \forall w \in B(0, \delta), \quad \forall z \in T^{-1}w \quad |z - \bar{Z}| \leq a|w|^s.$$

Then $|z^k - \bar{Z}| \rightarrow 0$, and its $(Q-)$ order of convergence satisfies $t \geq \min\{r, s\}$.

Proof. The hypothesis of the theorem subsumes that of Theorem 1.1, and therefore $|c_k^{-1}Q_k z^k| \rightarrow 0$. By (1.1) and assumption (3.1), there is some $k_1 \in \mathbb{Z}_+$ such that

$$(3.2) \quad \forall k \geq k_1 \quad |P_k z^k - \bar{Z}| \leq \frac{a}{c_k^s} |Q_k z^k|^s.$$

Using (2.3) to eliminate $|Q_k z^k|$ in (3.2), we obtain

$$\forall k \geq k_1 \quad |P_k z^k - \bar{Z}|^2 + \frac{c_k^2}{a^{2/s}} |P_k z^k - \bar{Z}|^{2/s} \leq |z^k - \bar{Z}|^2,$$

from which

$$(3.3) \quad \forall k \geq k_1 \quad |P_k z^k - \bar{Z}| \leq \frac{a|z^k - \bar{Z}|^s}{(c_k^2 + a^{2/s}|P_k z^k - \bar{Z}|^{2(s-1)/s})^{s/2}}.$$

The triangle inequality and criterion (A_r) yield the following estimate for all k :

$$\begin{aligned} |z^{k+1} - \overline{P_k z^k}| &\leq |z^{k+1} - P_k z^k| + |P_k z^k - \overline{P_k z^k}| \\ &\leq \varepsilon_k |z^{k+1} - z^k|^r + |P_k z^k - \bar{Z}| \\ &\leq \varepsilon_k |z^{k+1} - z^k|^{r-1} (|z^{k+1} - \overline{P_k z^k}| + |z^k - \overline{P_k z^k}|) + |P_k z^k - \bar{Z}|. \end{aligned}$$

Rearranging and using the fact that $|z^k - \overline{P_k z^k}| \leq 2|z^k - \bar{Z}|$ (cf. (2.6)), we have that

$$\begin{aligned} \forall k \geq 0 \quad |P_k z^k - \bar{Z}| &\geq (1 - \varepsilon_k |z^{k+1} - z^k|^{r-1}) |z^{k+1} - \overline{P_k z^k}| \\ &\quad - 2\varepsilon_k |z^{k+1} - z^k|^{r-1} |z^k - \bar{Z}|. \end{aligned}$$

By Theorem 1.1, $|z^{k+1} - z^k| \rightarrow 0$, and therefore there is some $k_2 \in \mathbb{Z}_+$ such that $|z^{k+1} - z^k| < 1$, and $|z^{k+1} - z^k|^{r-1} \leq 1$ for all $k \geq k_2$ as $r \geq 1$. Also, by criterion (A_r) , $\varepsilon_k \rightarrow 0$, and thus there is some $k_3 \in \mathbb{Z}_+$ such that $\varepsilon_k < 1$ for all $k \geq k_3$. Hence for all $k \geq \tilde{k} = \max\{k_2, k_3\}$, $1 - \varepsilon_k |z^{k+1} - z^k|^{r-1} \geq 1 - \varepsilon_k > 0$, and being $|z^{k+1} - P_k z^k| \geq |z^{k+1} - \bar{Z}|$,

$$(3.4) \quad \forall k \geq \tilde{k} \quad |P_k z^k - \bar{Z}| \geq (1 - \varepsilon_k) |z^{k+1} - \bar{Z}| - 2\varepsilon_k |z^{k+1} - z^k|^{r-1} |z^k - \bar{Z}|.$$

From (2.5), the triangle inequality and the fact that $r \geq 1$,

$$\begin{aligned} |z^k - \bar{Z}| &\geq |Q_k z^k| = |z^k - P_k z^k| \geq |z^k - z^{k+1}| - |z^{k+1} - P_k z^k| \\ &\geq |z^k - z^{k+1}| (1 - \varepsilon_k |z^k - z^{k+1}|^{r-1}), \end{aligned}$$

which can be transformed into the following estimate valid for all $k \geq \tilde{k}$:

$$(3.5) \quad |z^k - z^{k+1}| \leq \frac{|z^k - \bar{Z}|}{1 - \varepsilon_k}.$$

Combining (3.4) and (3.5) gives

$$(3.6) \quad \forall k \geq \tilde{k} \quad |P_k z^k - \bar{Z}| \geq (1 - \varepsilon_k) |z^{k+1} - \bar{Z}| - \frac{2\varepsilon_k}{(1 - \varepsilon_k)^{r-1}} |z^k - \bar{Z}|^r.$$

Let $\bar{k} = \max\{k_1, \tilde{k}\} = \max\{k_1, k_2, k_3\}$. The combination of (3.3) and (3.6) yields for all $k \geq \bar{k}$

$$(3.7) \quad |z^{k+1} - \bar{Z}| \leq \frac{a |z^k - \bar{Z}|^s}{(1 - \varepsilon_k)(c_k^2 + a^{2/s} |P_k z^k - \bar{Z}|^{2(s-1)/s})^{s/2}} + \frac{2\varepsilon_k}{(1 - \varepsilon_k)^r} |z^k - \bar{Z}|^r.$$

Assumption (3.1) implies the hypothesis of Proposition (1.2) with $f(|\cdot|) = a|\cdot|^s$, and thus $|z^k - \bar{Z}| \rightarrow 0$. Also, by criterion (A_r) , $\varepsilon_k \rightarrow 0$, and therefore from (3.7) it clearly follows that the (Q_-) order of convergence of $\{z^k - \bar{Z}\}$ is at least $\min\{r, s\} \geq 1$.

Remark. An alternative proof can be obtained by using (2.5) instead of (2.3) (to eliminate $|Q_k z^k|$ in (3.2)), and then (3.6) to obtain

$$\forall k \geq \bar{k} \quad |z^{k+1} - \bar{Z}| \leq \frac{a}{(1 - \varepsilon_k)c_k^s} |z^k - \bar{Z}|^s + \frac{2\varepsilon_k}{(1 - \varepsilon_k)^r} |z^k - \bar{Z}|^r.$$

The proof chosen has the advantage of clearly showing the connection in (3.7) with the linear convergence case (take $s = 1$ to obtain (2.8)).

In the context of the quadratic method of multipliers, Kort and Bertsekas (1973), (1976) have also specified conditions for the superlinear convergence to zero of the sequences $\{|y^k - \bar{Y}|\}$. The assumptions under which this result is obtained include (i) and (ii) as in § 2 above for the case of inexact minimization of the augmented Lagrangian, while (iii) takes the form

$$\exists b > 0, \quad \exists q \in (1, 2), \quad \exists \delta > 0: \quad \forall y \in B(Y, \delta) \quad g(y) \leq \bar{g} - \frac{1}{b} |\bar{y} - \bar{Y}|^q.$$

With the help of the subgradient inequality for the concave function g ,

$$\forall y^* \in \partial g(y), \quad \forall \bar{y} \in \bar{Y} \quad \bar{g} \leq g(y) + \langle \bar{y} - y, y^* \rangle \leq g(y) + |y - y^*| |y^*|,$$

the assumption above becomes

$$\exists b > 0, \quad \exists q \in (1, 2), \quad \exists \delta > 0: \quad \forall y \in B(\bar{Y}, \delta), \quad \forall y^* \in \partial g(y) \quad |y - \bar{Y}| \leq b |y^*|^{1/(q-1)}.$$

Clearly when $q \in (1, 2)$, $s = 1/(q-1) \in (1, +\infty)$, thus obtaining a growth condition on ∂g^{-1} analogous to the assumption (3.1) used in the proof of our theorem. When the algorithm is implemented in exact form (the strong convexity of f_0 is not needed in this case), the $(Q-)$ order of convergence is at least $1/(q-1)$, which coincides with our result (3.7). When the algorithm is implemented only approximately (see (2.12)), the $(Q-)$ order of convergence obtained is $2/q$ (Kort and Bertsekas (1976, Prop. 7, p. 286)). Taking into account that $1/(q-1) = s$, this order becomes $2s/(1+s)$ in our notation and satisfies

$$\forall s > 1 \quad 1 < \frac{2s}{1+s} < s.$$

In order to achieve the same order of convergence as with the exact algorithm, the sequence η_k in (2.12) has to be replaced by $\min \{\hat{\eta}_k, c|y^{k+1} - y^k|^2\}$, where $\hat{\eta}_k \rightarrow 0$, $c > 0$, and $a \geq s-1$ (Kort and Bertsekas (1976, Cor. 7.1, p. 288)). With this modification the actual criterion for the approximate implementation implies

$$|y^{k+1} - P_k y^k|^2 \leq c|y^{k+1} - y^k|^{s+1}.$$

This is less stringent than criterion (A_r) with $r \geq s$ which implies that for all k large enough (after $|y^{k+1} - y^k| < 1$)

$$|z^{k+1} - P_k z^k| \leq \varepsilon_k |z^{k+1} - z^k|^s, \quad \sum_{k=0}^{\infty} \varepsilon_k < +\infty.$$

The difference in orders of convergence might be accounted for by the following:

- a) the presence in the method of multipliers of subgradient inequalities which are not available for a general monotone operator;
- b) the assumptions made on f_0 , \bar{X} and \bar{Y} .

We analyze now the conditions under which finite convergence is obtained.

THEOREM 3.2. *Let $\bar{Z} \neq \emptyset$, and let $\{z^k\}$ be any sequence generated by the PPA either in exact form ($\varepsilon_k \equiv 0$), or with stopping criterion (A_r) , with $r = 0$ or $r \geq 1$, and a sequence of positive numbers $\{c_k\}$, such that $\liminf_{k \rightarrow \infty} c_k > 0$. Let us also assume that*

$$(3.8) \quad \exists \delta > 0: \quad \forall w \in B(0, \delta), \quad \forall z \in T^{-1}w \quad z \in \bar{Z}.$$

Then for all k large enough

$$(3.9) \quad |z^{k+1} - \bar{Z}| \leq \frac{\varepsilon_k}{(1 - \varepsilon_k)^r} |z^k - \bar{Z}|^r.$$

If the PPA is operated in exact form ($\varepsilon_k \equiv 0$), convergence is achieved in a finite number of iterations. Otherwise, if $r \geq 1$, superlinear convergence of order at least r is guaranteed.

Proof. Theorem 1.1 applies, and by (1.3), $|c_k^{-1} Q_k z^k| \rightarrow 0$, so there is some $k_1 \in \mathbb{Z}_+$ such that $|c_k^{-1} Q_k z^k| < \delta$ for all $k \geq k_1$. By (1.1) and assumption (3.8),

$$(3.10) \quad \forall k \geq k_1 \quad |P_k z^k - \bar{Z}| = 0.$$

Equation (1.3) implies that $|z^{k+1} - z^k| \rightarrow 0$, so there is some $k_2 \in \mathbb{Z}_+$ such that $|z^{k+1} - z^k| \leq 1$ for all $k \geq k_2$, and the inequality $\min \{1, |z^{k+1} - z^k|^r\} \leq |z^{k+1} - z^k|^r$ is valid for all $k \geq k_2$ for $r = 0$ or $r \geq 1$. Letting $\tilde{k} = \max \{k_1, k_2\}$, the triangle inequality, criterion (A_r) and (3.10) yield

$$(3.11) \quad \begin{aligned} \forall k \geq \tilde{k} \quad |z^{k+1} - \bar{Z}| &\leq |z^{k+1} - \overline{P_k z^k}| \leq |z^{k+1} - P_k z^k| + |P_k z^k - \overline{P_k z^k}| \\ &\leq \varepsilon_k \min \{1, |z^{k+1} - z^k|^r\} \leq \varepsilon_k |z^{k+1} - z^k|^r, \end{aligned}$$

valid for $r = 0$ or $r \geq 1$.

By criterion (A_r) , $\varepsilon_k \rightarrow 0$, thus there is some $k_3 \in \mathbb{Z}_+$ such that $\varepsilon_k < 1$ for all $k \geq k_3$. When $r \geq 1$ and $k \geq \bar{k} = \max \{k_1, k_2, k_3\} \geq \bar{k}$, (3.5) holds, and (3.11) can be transformed into

$$(3.12) \quad \forall k \geq \bar{k} \quad |z^{k+1} - \bar{Z}| \leq \frac{\varepsilon_k}{(1 - \varepsilon_k)^r} |z^k - \bar{Z}|^r.$$

The theorem follows because it is clear that (3.10), (3.11) and (3.12) are equivalent to (3.9) when the PPA is implemented in exact form ($\varepsilon_k = 0$), with (A_r) and $r = 0$, and with (A_r) and $r \geq 1$ respectively.

Remark. A condition for the convergence of the exact PPA in a single step can be easily obtained as follows. By (1.1) $c_0^{-1}Q_0z^0 \in TP_0z^0$, so if $|c_0^{-1}Q_0z^0| < \delta$ then $z^1 = P_0z^0 \in \bar{Z}$. Q_0 is the proximal mapping for the maximal monotone operator $(c_0T)^{-1}$, and thus it is nonexpansive. Hence for any $z, z' \in H$, $|Q_0z - Q_0z'| \leq |z - z'|$. We know that if $z' \in \bar{Z}$, then $Q_0z' = 0$. Let us choose $z = z^0$, $z' = z^0 \in \bar{Z}$. Then the estimate $|Q_0z^0| \leq |z^0 - \bar{Z}|$ is obtained. Thus a sufficient condition for $|c_0^{-1}Q_0z^0| < \delta$ is $c_0 > |z^0 - \bar{Z}|/\delta$. A condition of this type appeared for the first time in Bertsekas (1975).

Rockafellar (1976a, Thm. 3, p. 888) showed the finite convergence of the PPA under the assumption that $0 \in \text{int } T\bar{z}$ for some $\bar{z} \in H$. This assumption implies that \bar{z} is the unique solution of $0 \in Tz$. On the other hand, our result applies in the general case in which \bar{Z} need not be a singleton or even compact.

Viewed in the context of the quadratic method of multipliers, Theorem 3.2 guarantees finite convergence without the need of making compactness assumptions on \bar{X} (Bertsekas (1975)) or uniqueness of the Lagrange multipliers, i.e. $\bar{Y} = \{\bar{y}\}$ (Rockafellar (1976b)).

The generalization of Rockafellar's criterion $0 \in \text{int } T\bar{z}$, for some $\bar{z} \in H$, to a general nonempty \bar{Z} would be

$$(3.13) \quad \exists \delta > 0: \quad B(0, \delta) \subseteq T\bar{Z}.$$

Instead we have used (cf. (3.8))

$$(3.14) \quad \exists \delta > 0: \quad T^{-1}B(0, \delta) \subseteq \bar{Z},$$

which is the obvious limiting case of (3.1) when $s \rightarrow \infty$ and $\delta < 1$ (this last condition can be arranged by taking some $\delta' < \min \{1, \delta\}$).

It is interesting to explore the relationship between (3.13) and (3.14). From our analysis (see Prop. 3.5 below), it follows not only that (3.13) implies (3.14) but also that \bar{Z} is bounded. On the other hand there are instances in which (3.14) holds but (3.13) does not. For example, if \bar{Z} is unbounded, as it happens for $H = \mathbb{R}$, when the graph of T is given by $G(T) = \mathbb{R}_- \times \{0\} \cup \{0\} \times [0, 1] \cup \mathbb{R}_+ \times \{1\}$. To show this relationship we will first prove two technical lemmas.

LEMMA 3.3. *Let T be a maximal monotone operator such that $\bar{Z} = T^{-1}0$ is nonempty.*

Then $Tz \subseteq N_{\bar{Z}}(z)$ for all $z \in H$, where $N_{\bar{Z}}(z)$ denotes the normal cone to \bar{Z} at z . In particular if $z \in \text{int } \bar{Z}$, the interior of \bar{Z} in the strong topology of H , $Tz = \{0\}$.

Proof. For all $z \in H$, the cone normal to \bar{Z} at z is given by (Rockafellar (1970b, p. 15))

$$(3.15) \quad N_{\bar{Z}}(z) = \{x \in H: \forall u \in \bar{Z} \langle z - u, x \rangle \geq 0\}.$$

If $z \notin D(T)$, then $Tz = \emptyset$ and the inclusion $Tz \subseteq N_{\bar{Z}}(z)$ is trivial. Let $z \in D(T)$, and $w \in Tz$. Then the monotonicity of T implies

$$\forall z' \in \bar{Z} \quad \langle z - z', w \rangle \geq 0$$

and it follows that $w \in N_{\bar{Z}}(z)$. If $z \in \text{int } \bar{Z}$ then $N_{\bar{Z}}(z) = \{0\} \supseteq Tz \neq \emptyset$, so $Tz = \{0\}$.

LEMMA 3.4. *Let C be a nonempty closed convex and bounded subset of a real Hilbert space H . Let $N_C(z)$ denote the normal cone to C at z .*

If $z \notin C$, then $N_C(z)$ has a nonempty interior in the strong topology which is also a convex cone.

Proof. Since C is nonempty closed and convex, for any $z \in H$ there is a unique vector $\bar{z} \in C$ which is closest at z . This vector is characterized by Luenberger (1969, Thm. 1, p. 69) $\langle z - \bar{z}, \bar{z} - u \rangle \geq 0$ for all $u \in C$. But $\langle z - \bar{z}, \bar{z} - u \rangle = \langle z - \bar{z}, z - u \rangle - |z - \bar{z}|^2$ which clearly shows that for all $u \in C$ $\langle z - \bar{z}, z - u \rangle \geq |z - \bar{z}|^2 \geq 0$, and therefore $z - \bar{z} \in N_C(z)$ (see (3.15)).

It will now be shown that if C is bounded and $z \notin C$, then $z - \bar{z} \in \text{int } N_C(z)$. Let us suppose that $z - \bar{z} \notin \text{int } N_C(z)$. Then for any $\delta > 0$ there is some $v \in B(0, \delta)$ such that $z - \bar{z} + v \notin N_C(z)$. By the definition of $N_C(z)$, this implies that there is some vector $p \in C$ such that $\langle z - p, z - \bar{z} + v \rangle < 0$, or $\langle z - p, z - \bar{z} \rangle < \langle p - z, v \rangle$. But $\langle z - p, z - \bar{z} \rangle = \langle z - \bar{z}, z - \bar{z} \rangle + \langle \bar{z} - p, z - \bar{z} \rangle \geq |z - \bar{z}|^2 > 0$, because $p \in C$ and \bar{z} is the projection of z onto C .

Using successively the Cauchy-Bunyakovsky and triangle inequalities, and boundedness of C (i.e., there exists $M \in \mathbb{R}$ such that $C \subseteq B(0, M)$), we get

$$\begin{aligned} 0 &< |z - \bar{z}|^2 < \langle p - z, v \rangle \leq |p - z||v| \\ &\leq (|p| + |z|)|v| \leq (M + |z|)|v|. \end{aligned}$$

Thus $|v| > |z - \bar{z}|^2 / (M + |z|) > 0$. Let us choose $0 < \delta < |z - \bar{z}|^2 / (M + |z|)$ to obtain a contradiction with $v \in B(0, \delta)$, and therefore $z - \bar{z} \in \text{int } N_C(z)$. It is easy to prove that if K is a convex cone so is $\text{int } K$, and therefore $\text{int } N_C(z)$ is a convex cone.

PROPOSITION 3.5. *Let T and \bar{Z} be as above. Then $0 \in \text{int } T\bar{Z}$ implies that \bar{Z} is bounded. Moreover, there is some $\delta > 0$ such that for all $w \in B(0, \delta)$, $z \in T^{-1}w \Rightarrow z \in \bar{Z}$. In particular, if $w \in B(0, \delta) \setminus \{0\}$ and $z \in T^{-1}w$ then $z \in \partial\bar{Z} = \bar{Z} \setminus \text{int } \bar{Z}$, or more suggestively, $T^{-1}(B(0, \delta) \setminus \{0\}) \subseteq \partial\bar{Z}$.*

Proof. \bar{Z} is closed, therefore it contains its boundary $\partial\bar{Z}$, and $\bar{Z} = \text{int } \bar{Z} \cup \partial\bar{Z}$. We also have by Lemma 3.3

$$T\bar{Z} = \cup \{Tz : z \in \bar{Z}\} = T \text{int } \bar{Z} \cup T \partial\bar{Z} = \{0\} \cup T \partial\bar{Z} = T \partial\bar{Z}.$$

By the hypothesis, there is some $\delta > 0$ such that $B(0, \delta) \subseteq T\bar{Z}$, and thus $B(0, \delta) \subseteq T \partial\bar{Z}$.

Let $N_{\bar{Z}}(z)$ be the cone normal to \bar{Z} at z . The hypothesis, Lemma 3.3, and the fact that $N_{\bar{Z}}(z)$ is a cone imply

$$(3.16) \quad B(0, \delta) \subseteq \cup \{Tz : z \in \partial\bar{Z}\} \subseteq \cup \{N_{\bar{Z}}(z) : z \in \partial\bar{Z}\} = H.$$

Let ψ denote the indicator function of \bar{Z} . For all z in \bar{Z} , $N_{\bar{Z}}(z) = \partial\psi(z)$, thus $R(\partial\psi) = H$, and ψ^* , the support function of \bar{Z} , is everywhere finite. The boundedness of \bar{Z} follows from a result of Rockafellar (1966, Thm. 5B, p. 57).

To prove the second part of the proposition, let us assume that for some $z \in D(T) \setminus \bar{Z}$ there is some $w \in Tz$ such that $|w| < \delta$. Since \bar{Z} is convex and bounded, by Lemma 3.4, for any $z \notin \bar{Z}$, the interior of $N_{\bar{Z}}(z)$ is a nonempty convex cone. Let

$p \in \text{int } N_{\bar{Z}}(z) \cap B(0, \delta - |w|) \neq \emptyset$. Clearly, $0 < |p| < \delta - |w|$, and

$$\forall u \in \bar{Z} \quad \langle z - u, p + w \rangle = \langle z - u, p \rangle + \langle z - u, w \rangle \geq 0,$$

because $p \in \text{int } N_{\bar{Z}}(z) \subseteq N_{\bar{Z}}(z)$, and $w \in Tz \subseteq N_{\bar{Z}}(z)$. The triangle inequality yields $|p + w| \leq |p| + |w| < \delta - |w| + |w| = \delta$, and $p + w \in B(0, \delta)$. By (3.16), there is some $z' \in \partial \bar{Z}$ such that $p + w \in Tz' \subseteq N_{\bar{Z}}(z')$. The monotonicity of T implies $0 \leq \langle z - z', w - (p + w) \rangle = -\langle z - z', p \rangle$. But $p \in N_{\bar{Z}}(z)$ and $z' \in \bar{Z}$ imply that $\langle z - z', p \rangle \geq 0$, thus $\langle z - z', p \rangle = 0$. As $p \in \text{int } N_{\bar{Z}}(z)$, there is some $\tau > 0$ such that $B(p, \tau) \subseteq N_{\bar{Z}}(z)$. For any $\nu \in (0, \tau)$, $p + \nu(z' - z)/|z' - z| \in N_{\bar{Z}}(z)$. By the definition of $N_{\bar{Z}}(z)$ given in (3.15), this implies that

$$\left\langle z - z', p + \nu \frac{z' - z}{|z' - z|} \right\rangle \geq 0.$$

Since $\nu > 0$ and $\langle z - z', p \rangle = 0$ we obtain $0 \leq \langle z - z', z' - z \rangle < 0$ a contradiction. Therefore we cannot assume that for some $z \in D(T) \setminus \bar{Z}$ there exists some $w \in Tz$ with $|w| < \delta$. It follows that $|w| < \delta$ implies $z \in \bar{Z}$.

4. Sublinear convergence. This section starts with a partial converse to Theorem 2.1.

THEOREM 4.1. *Let $\bar{Z} \neq \emptyset$, and let $\{z^k\}$ be any sequence generated by the PPA with stopping criterion (A_r) , with $r \geq 1$, and a nondecreasing sequence of positive numbers $\{c_k\}$, such that $0 < c_k \uparrow c_\infty < +\infty$. Let us also assume that*

$$(4.1) \quad \forall a > 0, \quad \exists \delta > 0: \quad \forall w \in B(0, \delta), \quad \forall z \in T^{-1}w \quad |z - \bar{Z}| \geq a|w|.$$

Then if $\{z^k\}$ does not converge to \bar{Z} in a finite number of steps (i.e., $z^k \notin \bar{Z}$ for all k)

$$\liminf_{k \rightarrow \infty} \frac{|z^{k+1} - \bar{Z}|}{|z^k - \bar{Z}|} = 1,$$

and $\{z^k - \bar{Z}\}$ cannot converge to zero faster than sublinearly.

Proof. Let us choose some fixed $a > 0$. Theorem 1.1 applies and by (1.3) $|c_k^{-1} Q_k z^k| \rightarrow 0$. Therefore there is some $k_1 \in \mathbb{Z}_+$ such that $|c_k^{-1} Q_k z^k| < \delta$ for all $k \geq k_1$. By (1.1) and assumption (4.1),

$$(4.2) \quad \forall k \geq k_1 \quad |P_k z^k - \bar{Z}| \geq \frac{a}{c_k} |Q_k z^k|.$$

By the triangle inequality

$$(4.3) \quad |Q_k z^k| = |z^k - P_k z^k| \geq |z^k - z^{k+1}| - |z^{k+1} - P_k z^k|.$$

The triangle inequality, and the fact that projection onto a nonempty closed convex set is a nonexpansive mapping (Moreau (1965, p. 279)) yield

$$(4.4) \quad \begin{aligned} |P_k z^k - \bar{Z}| &\leq |P_k z^k - z^{k+1}| + |z^{k+1} - \overline{z^{k+1}}| + |\overline{z^{k+1}} - \overline{P_k z^k}| \\ &\leq 2|P_k z^k - z^{k+1}| + |z^{k+1} - \bar{Z}|. \end{aligned}$$

Using (4.3) and (4.4), we see that (4.2) can be transformed into

$$(4.5) \quad \forall k \geq k_1 \quad (2c_k + a)|P_k z^k - z^{k+1}| + c_k |z^{k+1} - \bar{Z}| \geq a|z^k - z^{k+1}|.$$

By (1.3) there is some $k_2 \in \mathbb{Z}_+$ such that $|z^{k+1} - z^k|^r \leq |z^{k+1} - z^k| < 1$ for all $k \geq k_2$. Hence criterion (A_r) yields

$$(4.6) \quad \forall k \geq k_2 \quad |P_k z^k - z^{k+1}| \leq \varepsilon_k \min \{1, |z^{k+1} - z^k|^r\} \leq \varepsilon_k |z^{k+1} - z^k|.$$

We also have that

$$(4.7) \quad |z^k - z^{k+1}| \geq |z^k - \bar{z}^{k+1}| - |z^{k+1} - \bar{z}^{k+1}| \geq |z^k - \bar{z}| - |z^{k+1} - \bar{z}|.$$

By combining (4.5)–(4.7) we obtain for all $k \geq \tilde{k} = \max \{k_1, k_2\}$

$$(c_k + a)|z^{k+1} - \bar{z}| \geq a|z^k - \bar{z}| - (2c_k + a)\varepsilon_k |z^{k+1} - z^k|.$$

Criterion (A_r) implies that $\varepsilon_k \rightarrow 0$. Thus there is some $k_3 \in \mathbb{Z}_+$ such that $\varepsilon_k < 1$ for all $k \geq k_3$. If $k \geq \bar{k} = \max \{k_3, \tilde{k}\}$, the above inequality and (3.5) are valid, and using the latter to substitute for $|z^{k+1} - z^k|$ in the former, we arrive at

$$\forall k \geq \bar{k} \quad (c_k + a)|z^{k+1} - \bar{z}| \geq a|z^k - \bar{z}| - \frac{(2c_k + a)\varepsilon_k}{1 - \varepsilon_k} |z^k - \bar{z}|.$$

From this expression, taking into account that $\varepsilon_k \rightarrow 0$, we obtain

$$\liminf_{k \rightarrow \infty} \frac{|z^{k+1} - \bar{z}|}{|z^k - \bar{z}|} \geq \lim_{k \rightarrow \infty} \frac{a}{a + c_k} = \frac{a}{a + c_\infty}.$$

Since a can be arbitrarily large, the theorem follows.

The preceding theorem provides an essentially negative result. In the next theorem we try to quantify the speed of convergence. Since the convergence may be sublinear, we will have to look for an estimate of the form $|z^k - \bar{z}| = O(k^{-\sigma})$ for some $\sigma > 0$.

THEOREM 4.2. *Let $\bar{Z} \neq \emptyset$, and let $\{z^k\}$ be any sequence generated by the PPA in exact form with a nondecreasing sequence of positive numbers $\{c_k\}$, such that $0 < c_k \uparrow c_\infty < +\infty$. Let us also assume that*

$$(4.8) \quad \exists a > 0, \quad \exists s \in (0, 1), \quad \exists \delta > 0: \quad \forall w \in B(0, \delta), \quad \forall z \in T^{-1}w \quad |z - \bar{z}| \leq a|w|^s.$$

Then $|z^k - \bar{z}| \rightarrow 0$ as $o(k^{-s/2})$, i.e., $\lim_{k \rightarrow \infty} |z^k - \bar{z}|^{2/s} k = 0$.

Proof. By Theorem 1.1, $|c_k^{-1} Q_k z^k| \rightarrow 0$, thus there exists some $k_1 \in \mathbb{Z}_+$ such that $|c_k^{-1} Q_k z^k| < \delta$ for all $k \geq k_1$. Also by (1.1) $c_k^{-1} Q_k z^k \in TP_k z^k = Tz^{k+1}$. Using these facts and assumption (4.8) gives

$$\forall k \geq k_1 \quad |z^{k+1} - \bar{z}| \leq \frac{a}{c_k^s} |Q_k z^k|^s.$$

Using (2.3) eliminate $|Q_k z^k|$ and rearranging, we come to

$$\forall k \geq k_1 \quad |z^{k+1} - \bar{z}|^2 + \frac{c_k^2}{a^{2/s}} |z^{k+1} - \bar{z}|^{2/s} \leq |z^k - \bar{z}|^2,$$

from which we obtain the following inequality for all $n \geq k_1$:

$$\sum_{k=k_1}^n |z^{k+1} - \bar{z}|^2 + \sum_{k=k_1}^n \frac{c_k^2}{a^{2/s}} |z^{k+1} - \bar{z}|^{2/s} \leq \sum_{k=k_1}^n |z^k - \bar{z}|^2,$$

which reduces to

$$\forall n \geq k_1 \quad |z^{n+1} - \bar{z}|^2 + \sum_{k=k_1}^n \frac{c_k^2}{a^{2/s}} |z^{k+1} - \bar{z}|^{2/s} \leq |z^{k_1} - \bar{z}|^2.$$

Taking the limit as $n \rightarrow \infty$, $|z^{n+1} - \bar{Z}| \rightarrow 0$, by Proposition 1.2

$$\sum_{k=k_1}^{\infty} \frac{c_k^2}{a^{2/s}} |z^{k+1} - \bar{Z}|^{2/s} \leq |z^{k_1} - \bar{Z}|^2 < +\infty.$$

For the series to converge, its terms have to decrease to zero faster than the terms of the harmonic series, thus

$$\limsup_{k \rightarrow \infty} c_k^2 |z^{k+1} - \bar{Z}|^{2/s} k = 0.$$

Obviously, any speed of decrease can be obtained by making $c_k \uparrow \infty$ fast enough. If $c_k \uparrow c_\infty < +\infty$ then it follows that $|z^k - \bar{Z}| = o(k^{-s/2})$.

Remark. This estimate seems conservative, at least when $s \uparrow 1$ because for $s = 1$ linear convergence is achieved and then $|z^k - \bar{Z}| = o(k^{-\sigma})$ for all $\sigma > 0$ (Ortega and Rheinboldt (1970)).

REFERENCES

- [1] D. P. BERTSEKAS (1975), *Necessary and sufficient conditions for a penalty method to be exact*, Math. Programming, 9, pp. 87–99.
- [2] H. BRÉZIS (1973), *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland, Amsterdam.
- [3] F. E. BROWDER (1976), *Nonlinear Operators and Nonlinear Equations of Evolution in Banach Spaces*, Proc. Symposia in Pure Mathematics, 18 part 2, American Mathematical Society, Providence, RI.
- [4] B. W. KORT AND D. P. BERTSEKAS (1973), *Multiplier methods for convex programming*, in Proceedings 1973 IEEE Conference on Decision and Control, San Diego, CA, December 1973, pp. 428–432.
- [5] ——— (1976), *Combined primal-dual and penalty methods for convex programming*, this Journal, 14, pp. 268–294.
- [6] A. V. KRYANEV (1973), *The solution of incorrectly posed problems by methods of successive approximations*, Soviet Math. Dokl., 14, pp. 673–676.
- [7] D. G. LUENBERGER (1969), *Optimization by Vector Space Methods*, John Wiley, New York.
- [8] L. McLINDEN (1980), *The complementarity problem for maximal monotone multifunctions*, in Variational Inequalities and Complementarity Problems: Theory and Applications, R. W. Cottle, F. Giannessi and J.-L. Lions, eds., John Wiley, Chichester, UK, Chapter 17, pp. 251–270.
- [9] G. J. MINTY (1962) *Monotone (nonlinear) operators in Hilbert space*, Duke Math. J., 29, 341–362.
- [10] ——— (1964), *On the solvability of nonlinear functional equations of “monotonic” type*, Pacific J. Math., 14, pp. 249–255.
- [11] J.-J. MOREAU (1965), *Proximité et dualité dans un espace Hilbertien*, Bull. Soc. Math. France, 93, pp. 273–299.
- [12] J. M. ORTEGA AND W. C. RHEINOLDT (1970), *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York.
- [13] D. PASCALI AND S. SBURLAN (1978), *Nonlinear Mappings of Monotone Type*, Sijthoff & Noordhoff, Alphen aan den Rijn, the Netherlands.
- [14] R. T. ROCKAFELLAR (1966), *Level sets and continuity of conjugate convex functions*, Trans. Amer. Math. Soc., 123, pp. 46–63.
- [14a] ——— (1970a), *Monotone operators associated with saddle functions and minimax problems*, in Nonlinear Functional Analysis, F. E. Browder, ed., Proceedings of Symposia in Pure Mathematics, 18 part 1, American Mathematical Society, Providence, RI, pp. 241–250.
- [15] ——— (1970b), *Convex Analysis*, Princeton Univ. Press, Princeton, NJ.
- [16] ——— (1970c), *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc., 149, pp. 75–88.
- [17] ——— (1976a), *Monotone operators and the proximal point algorithm*, this Journal, 4, pp. 877–898.
- [18] ——— (1976b), *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1, pp. 97–116.
- [19] ——— (1978), *Monotone operators and augmented Lagrangian methods in nonlinear programming*, in Nonlinear Programming 3, O. L. Mangasarian, R. R. Meyer and S. M. Robinson, eds., Academic Press, New York, pp. 1–25.

- [20] ——— (1980), *Lagrange multipliers and variational inequalities*, in *Variational Inequalities and Complementarity Problems: Theory and Applications*, R. W. Cottle, F. Giannessi and J.-L. Lions, eds., John Wiley, Chichester, UK, Chapter 20, pp. 303–322.

EXTENDED JACOBI SUFFICIENCY CRITERION FOR OPTIMAL CONTROL*

VERA ZEIDAN†

Abstract. In this paper we present sufficient conditions for strong local minimality in optimal control. We do not require the dynamic f to be linear nor the control set U to be open. Moreover, these conditions extend a known sufficiency criterion in the calculus of variations involving the Jacobi condition.

Key words. sufficient conditions, strong local minimality, optimal control problem, extended Jacobi condition

1. Introduction. Let f, g and l^0 be given functions:

$$f: [a, b] \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n, \quad g: [a, b] \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}, \quad l^0: \mathbb{R}^n \rightarrow \mathbb{R},$$

let U be a subset of \mathbb{R}^m , and let A be a point of \mathbb{R}^n . The optimal control problem is defined to be:

$$\text{minimize } l^0(x(b)) + \int_a^b g(t, x(t), u(t)) dt$$

over all absolutely continuous functions x from $[a, b]$ to \mathbb{R}^n with derivative \dot{x} (almost everywhere), and over all measurable functions u from $[a, b]$ to \mathbb{R}^m satisfying:

$$\dot{x}(t) = f(t, x(t), u(t)) \text{ a.e.}, \quad x(a) = A, \quad u(t) \in U \text{ a.e.}$$

The Hamiltonian H of this problem is defined as follows:

$$H(t, x, p) = \sup \{ \langle p, f(t, x, u) \rangle - g(t, x, u) : u \in U \}.$$

There are two main sufficiency criteria existing in the literature for optimal control problems. The first requires the Hamiltonian H to be concave in x and convex in p [9]. The second criterion involves the Hamilton–Jacobi inequality [1], [4], [5] and [10]. Also, there are other criteria which apply in simplified contexts. For instance, when the control function \hat{u} is assumed to lie in the interior of the control set U , or when the differential equation is linear in x and u .

In this paper we develop a new sufficiency criterion for autonomous optimal control problems where the Hamiltonian is not necessarily concave-convex. Our sufficient conditions are obtained by considering the optimal control problem as being a generalized Bolza problem, and then by applying the sufficiency criterion developed in [11]. Thus, our conditions extend known sufficient conditions in the calculus of variations involving the Jacobi condition.

If we consider the special case of optimal control problems treated by Mayne in [6], that is the case when the control $\hat{u}(t)$ belongs to the interior of the control set U , then we see that our conditions reduce to his result in [6, Thm. 3.2].

2. Statement of the sufficiency theorem. Consider the autonomous optimal control problem:

$$(C) \quad \text{minimize } J(x, u) = l^0(x(b)) + \int_a^b g(x(t), u(t)) dt$$

* Received by the editors November 5, 1982, and in revised form February 15, 1983. This research was supported by a grant from Fonds FCAC.

† Centre de Recherche de Mathématiques Appliquées, Université de Montréal, Montréal, Québec, Canada H3C 3J7.

subject to:

$$(2.1) \quad \dot{x}(t) = f(x(t), u(t)) \text{ a.e.,}$$

$$(2.2) \quad x(a) = A,$$

$$(2.3) \quad u(t) \in U \text{ a.e.,}$$

where $u: [a, b] \rightarrow \mathbb{R}^m$ is measurable, and $x: [a, b] \rightarrow \mathbb{R}^n$ is absolutely continuous. Such a function x is called an arc.

The Hamiltonian corresponding to the Problem (C) is

$$(2.4) \quad H(x, p) = \sup \{ \langle p, f(x, u) \rangle - g(x, u) : u \in U \}.$$

DEFINITION. Let x be an arc from $[a, b]$ to \mathbb{R}^n and u be a measurable function from $[a, b]$ to \mathbb{R}^m . The pair (x, u) is *admissible* if it satisfies (2.1)–(2.3).

DEFINITION. An admissible pair (\hat{x}, \hat{u}) is a *strong local minimum* for (C) if there exists a positive number γ such that (\hat{x}, \hat{u}) minimizes $J(x, u)$ over all admissible pairs (x, u) satisfying

$$|x(t) - \hat{x}(t)| < \gamma \quad \text{for all } t \in [a, b].$$

Let \hat{x} be a given arc. The following hypothesis will be made:

- (H) There exists a positive number ε such that $f(x, u)$ and $g(x, u)$ are C^2 on $\{(x, u) \in \mathbb{R}^n \times U : |x - \hat{x}(t)| < \varepsilon \text{ for some } t \in [a, b]\}$ and $l^0(x)$ is C^2 on $\{x : |x - \hat{x}(b)| < \varepsilon\}$.

Suppose we are given arcs \hat{x}, \hat{p} from $[a, b]$ to \mathbb{R}^n and a positive number β . For $\hat{z} = (\hat{x}, \hat{p})$ we define

$$N_\beta(\hat{z}) = \{z \in \mathbb{R}^n \times \mathbb{R}^n : |z - \hat{z}(t)| < \beta \text{ for some } t \in [a, b]\}.$$

DEFINITION. The Hamiltonian H is said to be C^{1+} near $\hat{z} = (\hat{x}, \hat{p})$ if there exists a positive number β such that $H(\cdot, \cdot)$ is C^1 with locally Lipschitz gradient on $N_\beta(\hat{z})$.

If the Hamiltonian $H(z)$ is C^{1+} near \hat{z} then $\nabla H(\cdot)$ is Lipschitz near \hat{z} . Hence, by [2], the generalized Jacobian $\partial \nabla H(\cdot)$ exists and it is defined at a point z as being the convex hull of all matrices of the form

$$M = \lim_{i \rightarrow \infty} \{\nabla^2 H(z_i)\},$$

where $z_i \rightarrow z$ and the usual Jacobian $\nabla^2 H(z_i)$ exists for each i .

DEFINITION. Suppose that the Hamiltonian H is C^{1+} near a given arc $\hat{z} = (\hat{x}, \hat{p})$. We say that the *extended Jacobi condition* is satisfied at \hat{z} if there exists a Lipschitz matrix function $Q(\cdot)$ on $[a, b]$ such that, for all t in $[a, b]$, $Q(t)$ is symmetric and satisfies

$$\eta(t) - Q(t)\gamma(t)Q(t) + Q(t)\beta(t) + \delta(t)Q(t) - \alpha(t) > 0$$

for all t in $[a, b]$, for all matrices

$$\begin{pmatrix} \alpha(t) & \delta(t) \\ \beta(t) & \gamma(t) \end{pmatrix} \in \partial \nabla H(\hat{z}(t))$$

and for all $\eta(t) \in \partial Q(t)$.

Remark. If the Hamiltonian H is C^2 near \hat{z} and if H is concave in x , then the extended Jacobi condition is satisfied. In fact, in this case, the function $Q = 0$ satisfies this condition, where the strict inequality is replaced by inequality. But, using the embedding theorem for differential equations, we conclude that there exists a function Q_0 (which is even C^1) satisfying the extended Jacobi condition.

THEOREM. Let (\hat{x}, \hat{u}) be an admissible pair for (C). Assume that Hypothesis (H) holds for some $\varepsilon > 0$, and

- (1) U is a nonempty convex compact polyhedron in \mathbb{R}^m ;
- (2) there exists an arc \hat{p} from $[a, b]$ to \mathbb{R}^n satisfying

$$-\dot{\hat{p}}(t) = f_x(\hat{x}(t), \hat{u}(t))^T \hat{p}(t) - g_x(\hat{x}(t), \hat{u}(t)) \quad \text{a.e.,}$$

with

$$\hat{p}(b) = -l_x^0(\hat{x}(b));$$

- (3) for all $t \in [a, b]$ and for all $u \in U$ such that $u \neq \hat{u}(t)$,

$$\langle \hat{p}(t), f(\hat{x}(t), \hat{u}(t)) \rangle - g(\hat{x}(t), \hat{u}(t)) > \langle \hat{p}(t), f(\hat{x}(t), u) \rangle - g(\hat{x}(t), u);$$

- (4) $g_{uu}(\hat{x}(t), \hat{u}(t)) - D_u((f_u(\hat{x}(t), u))^T \hat{p}(t))|_{u=\hat{u}(t)} > 0$ for all $t \in [a, b]$;

- (5) the extended Jacobi condition is satisfied by a matrix function Q such that

$$Q(b) < l_{xx}^0(\hat{x}(b)).$$

Then the pair (\hat{x}, \hat{u}) provides a strong local minimum for (C).

Remark. Condition (2) is the adjoint equation and condition (3) is a strengthening of the maximum principle of Pontryagin.

Remark. As we shall soon see, conditions (1), (3) and (4) of the theorem imply that the Hamiltonian H is C^{1+} near $\hat{z} = (\hat{x}, \hat{p})$. Hence, the use of the generalized Jacobian $\partial \nabla H(\hat{z}(t))$ in condition (5) is justified.

Remark. In the classical setting, it is shown in [11] that the extended Jacobi condition reduces to the Jacobi condition.

3. Proof of the theorem. The proof of the sufficiency theorem is based on converting the optimal control problem (C) to a generalized Bolza problem (P_C) and then applying the sufficiency criterion given in [11, Thm. 2].

Define

$$(3.1) \quad l(x_1, x_2) = \psi_{\{A\}}(x_1) + l^0(x_2),$$

where $\psi_{\{A\}}(\cdot)$ is the indicator function of the set $\{A\}$, and

$$(3.2) \quad L(x, v) = \inf \{g(x, u) : v = f(x, u), u \in U\}.$$

Then the functions l and L are extended real-valued.

Now, consider the generalized Bolza problem corresponding to (C):

$$(P_C) \quad \text{minimize} \quad J_C(x) = l(x(a), x(b)) + \int_a^b L(x(t), \dot{x}(t)) dt$$

over all arcs x from $[a, b]$ to \mathbb{R}^n , where the functions l and L are defined by (3.1) and (3.2), respectively.

Problems (P_C) and (C) have the same Hamiltonian. In fact, (3.1) and (3.2) imply that

$$\tilde{H}(x, p) = \sup \{ \langle p, v \rangle - L(x, v) : v \in \mathbb{R}^n \} = \sup \{ \langle p, f(x, u) \rangle - g(x, u) : u \in U \} = H(x, p).$$

Condition (3) and (3.2) yield

$$(3.3) \quad L(\hat{x}(t), \dot{\hat{x}}(t)) = g(\hat{x}(t), \hat{u}(t)) \quad \text{a.e.,}$$

and hence $J_C(\hat{x})$ is finite. From the compactness of U and [8, Thm. 6] we conclude that L , defined in (3.2), is measurable and that (\hat{x}, \hat{u}) is a strong local minimum for (C) whenever \hat{x} provides a strong local minimum for (P_C).

To prove the optimality of \hat{x} for the generalized Bolza problem (P_C) we will apply [11, Thm. (2)].

We first observe that the compactness of U and condition (3) imply that \hat{u} is continuous on $[a, b]$. Hence, from (2.1) it follows that \hat{x} is C^1 and therefore, the arc \hat{p} in condition (2) is C^1 and (3.3) holds for all t in $[a, b]$.

Define

$$A(t, v) = \{u \in U : \dot{\hat{x}}(t) + v = f(\hat{x}(t), u)\}.$$

Let $t \in [a, b]$ and $v \in \mathbb{R}^n$. If $A(t, v)$ is not empty, (3.2) and the compactness of U yield that there exists $u \in A(t, v)$ such that

$$L(\hat{x}(t), \dot{\hat{x}}(t) + v) = g(\hat{x}(t), u).$$

Then, from (3.3) and condition (3) it follows that

$$\begin{aligned} L(\hat{x}(t), \dot{\hat{x}}(t) + v) - L(\hat{x}(t), \dot{\hat{x}}(t)) &= g(\hat{x}(t), u) - g(\hat{x}(t), \hat{u}(t)) \\ &\geq \langle \hat{p}(t), f(\hat{x}(t), u) - f(\hat{x}(t), \hat{u}(t)) \rangle = \langle \hat{p}(t), v \rangle. \end{aligned}$$

On the other hand, if $A(t, v)$ is empty, then

$$L(\hat{x}(t), \dot{\hat{x}}(t) + v) = +\infty,$$

and hence the above inequality remains valid. Thus [11, Thm. 2(a)] is satisfied.

To show that our Hamiltonian H is C^{1+} near \hat{z} we need to establish the following result.

LEMMA 3.1. *Assume Hypothesis (H) and conditions (1), (3) and (4) of the theorem. Then there exists a positive number β such that, for all z in $N_\beta(\hat{z})$, the supremum in (2.4) is attained at an unique point $u(z)$. Furthermore, the function $u(\cdot)$ is Lipschitz and satisfies*

$$u(\hat{z}(t)) = \hat{u}(t) \quad \text{for all } t \in [a, b].$$

Proof. Condition (4) and Hypothesis (H) imply that there exists a positive number δ ($\delta \leq \varepsilon$) such that for all $(x, p) \in N_\delta(\hat{z})$ and for all $u \in \{u \in U : |u - \hat{u}(t)| < \delta\}$ we have

$$(3.4) \quad g_{uu}(x, u) - D_u((f_u(x, u))^T p) > 0.$$

The compactness of U and the continuity of the functions $f(x, \cdot)$ and $g(x, \cdot)$ yield that the supremum in (2.4) is always attained for any $z \in N_\varepsilon(\hat{z})$. However, arguing by contradiction, (3.4) and condition (3) imply that there exist a positive number γ ($\gamma \leq \delta$) and a unique function $u(x, p)$ such that, for all $(x, p) \in N_\gamma(\hat{z})$,

$$(3.5) \quad H(x, p) = \sup \{\langle p, f(x, u) \rangle - g(x, u) : u \in U\} = \langle p, f(x, u(x, p)) \rangle - g(x, u(x, p)).$$

Since $u(x, p)$ is unique (3.5) implies that $u(\cdot, \cdot)$ is in fact continuous. Moreover, by condition (3) we have

$$u(\hat{x}(t), \hat{p}(t)) = \hat{u}(t).$$

It remains to show that $u(\cdot, \cdot)$ is Lipschitz. From (3.5) it follows that, for $(x, p) \in N_\gamma(\hat{z})$,

$$(3.6) \quad 0 \in g_u(x, u(x, p)) - \langle p, f_u(x, u(x, p)) \rangle + \partial\psi_U(u(x, p)),$$

where $\partial\psi_U(u_0)$ is the normal cone to U at u_0 , that is

$$\partial\psi_U(u_0) = \begin{cases} \{y \in \mathbb{R}^m : \langle y, u - u_0 \rangle \leq 0 \text{ for } u \in U\} & \text{if } u_0 \in U, \\ \emptyset & \text{if } u_0 \notin U. \end{cases}$$

Let $\lambda_1(z), \dots, \lambda_m(z)$ be the eigenvalues of the matrix function

$$g_{uu}(x, u(z)) - D_u((f_u(x, u(z)))^T p),$$

and define

$$\lambda(z) = \min_{i=1}^m \{\lambda_i(z)\}.$$

Using (3.4), we conclude that, for all $z \in N_\gamma(\hat{z})$ and for all $h \in \mathbb{R}^m$,

$$h \cdot [g_{uu}(x, u(z)) - D_u((f_u(x, u(z)))^T p)]h \geq \lambda(z) \|h\|^2.$$

Since the function $\lambda_i(\cdot)$ is continuous for each i , so is the function $\lambda(\cdot)$. Hence, there exist positive numbers μ and β ($\beta \leq \delta$) such that, for all $z \in N_\beta(\hat{z})$ and for all $h \in \mathbb{R}^m$,

$$(3.7) \quad h \cdot [g_{uu}(x, u(z)) - D_u((f_u(x, u(z)))^T p)]h \geq \mu \|h\|^2.$$

We have that U is a convex compact polyhedron, that Hypothesis (H) holds, and that relations (3.6) and (3.7) are satisfied for all $z \in N_\beta(\hat{z})$. Then, by [7, Thm. 4.2 and Corollary 4.3] we conclude the following:

There exists a positive number λ such that for each $y_0 \in N_\beta(\hat{z})$ we can find a neighborhood $N(y_0)$ of y_0 with

$$|u(z) - u(y_0)| \leq \lambda |z - y_0| \quad \text{for all } z \in N(y_0),$$

which proves that $u(\cdot)$ is uniformly point-wise Lipschitz. As we have shown in the proof of [11, Thm. 3], the uniformly point-wise Lipschitz condition implies in this case that $u(\cdot)$ is Lipschitz. \square

By Lemma 3.1, there exists a positive number β ($\beta \leq \varepsilon$) and a unique Lipschitz function $u(\cdot)$ on $N_\beta(\hat{z})$ such that

$$H(x, p) = \langle p, f(x, u(z)) \rangle - g(x, u(z)) \quad \text{and} \quad u(\hat{z}(t)) = \hat{u}(t).$$

Moreover, from [2, Thm. (2.1)] it follows that $H(\cdot, \cdot)$ is C^1 on $N_\beta(\hat{z})$ with gradient

$$(3.8) \quad \nabla H(x, p) = (pf_x(x, u(z)) - g_x(x, u(z)), f(x, u(z))).$$

Thus, (3.8) implies that $H(\cdot, \cdot)$ is C^{1+} near \hat{z} . The admissibility of (\hat{x}, \hat{u}) , condition (2) and (3.8) imply that $\hat{z} = (\hat{x}, \hat{p})$ satisfies the Hamiltonian equations, i.e. [11, Thm. 2(b)]. From the last remark given in [11], [11, Thm. (2)] remains valid if [11, Thm. 2(c)] is replaced by the extended Jacobi condition. Hence, our proof here is complete when we show that [11, condition (d) of Theorem 2] is satisfied.

Since $l^0(\cdot)$ is C^2 near $\hat{x}(b)$, and $\hat{p}(b) = -l_x^0(\hat{x}(b))$, then for $x \in \{x: |x - \hat{x}(b)| < \varepsilon\}$ we have

$$(3.9) \quad \begin{aligned} l^0(x) &= l^0(\hat{x}(b)) - \langle \hat{p}(b), x - \hat{x}(b) \rangle + \frac{1}{2} \langle x - \hat{x}(b), l_{xx}^0(\hat{x}(b))(x - \hat{x}(b)) \rangle \\ &\quad + \|x - \hat{x}(b)\| o(x - \hat{x}(b)), \end{aligned}$$

where

$$\lim_{x \rightarrow \hat{x}(b)} \frac{o(x - \hat{x}(b))}{\|x - \hat{x}(b)\|} = 0.$$

Let $\lambda = \min_{i=1}^n \{\lambda_i\}$, where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of the matrix $l_{xx}^0(\hat{x}(b)) - Q(b)$. By Theorem 2.1 (5), we deduce that, for all $d \in \mathbb{R}^n$,

$$\langle d, (l_{xx}^0(\hat{x}(b)) - Q(b))d \rangle \geq \lambda \|d\|^2.$$

Choose a positive number α such that for all d satisfying $\|d\| < \alpha$ we have

$$\frac{\lambda}{2} + \frac{o(d)}{\|d\|} \geq 0.$$

Thus, using (3.9), we obtain for all such d

$$\begin{aligned} l^0(\hat{x}(b) + d) - l^0(\hat{x}(b)) + \langle \hat{p}(b), d \rangle - \frac{1}{2} \langle d, Q(b)d \rangle &= \frac{1}{2} \langle d, (l_{xx}^0(\hat{x}(b)) - Q(b))d \rangle + \|d\|o(d) \\ &\geq \|d\|^2 \left(\frac{\lambda}{2} + \frac{o(d)}{\|d\|} \right) \geq 0. \end{aligned}$$

Therefore, the function $l(\cdot, \cdot)$, defined by (3.1), satisfies [11, condition (d), Theorem 2]. Q.E.D.

Remark. Consider the nonautonomous optimal control problem. If we assume that $\hat{u}(t)$ is in the interior of U , then the Hamiltonian $H(t, z)$ is C^2 in z . For this special case, our theorem here remains valid, where the assumption that U is a convex polyhedron (which was needed to prove that H is C^{1+}) can be omitted. If, in addition, we assume that Q is C^1 , our conditions here reduce to the result of Mayne [6, Thm. 3.2]. Moreover, our result improves on [6], since our notion of local solution (optimal with respect to arcs in an L^∞ -ball about $\hat{x}(\cdot)$) is less restrictive than that in [6] (optimal with respect to controls in an L^1 -ball about $\hat{u}(\cdot)$).

Remark. For the case where both boundary values are fixed ($x(a) = A$, $x(b) = B$), Theorem 2.1 remains valid when the boundary conditions

$$\hat{p}(b) = -l_x^0(\hat{x}(b)) \quad \text{and} \quad Q(b) < l_{xx}^0(\hat{x}(b))$$

are removed.

4. Example. Consider the optimal control problem:

$$(C') \quad \text{minimize} \quad \int_0^1 \left(2u_1^2 + 2u_2^2 - \frac{x^3}{24} \right) dt$$

subject to

$$\dot{x}(t) = u_1(t)x^2(t) + u_2^3(t)x^4(t),$$

$$x(0) = x(1) = 1,$$

$$(u_1(t), u_2(t)) \in U = [-1, 0] \times [-1, 0] \quad \text{for all } t \in [0, 1].$$

We have

$$f(x, u) = u_1x^2 + u_2^3x^4, \quad g(x, u) = 2u_1^2 + 2u_2^2 - \frac{x^3}{24}$$

and $U = [-1, 0] \times [-1, 0]$, which is a convex compact polyhedron in \mathbb{R}^2 .

Take $\hat{x}(t) = 1$, $\hat{u}(t) = (0, 0)$ and $\hat{p}(t) = (1 - t)/8$. The pair (\hat{x}, \hat{u}) is admissible for our problem, and, for all $t \in [0, 1]$,

$$-\dot{\hat{p}}(t) = \frac{1}{8} = f_x(\hat{x}(t), \hat{u}(t))\hat{p}(t) - g_x(\hat{x}(t), \hat{u}(t)).$$

Now, observe that the function

$$pf(x, u) - g(x, u) = p(u_1x^2 + u_2^3x^4) - 2u_1^2 - 2u_2^2 + \frac{x^3}{24}$$

is strictly concave in u if (x, p) are such that

$$(4.1) \quad -3u_2px^4 + 2 > 0 \quad \text{for all } u_2 \in [-1, 0].$$

Since $\hat{x}(t) = 1$ and $\hat{p}(t) = (1-t)/8$ satisfy (4.1), we can find a neighborhood $N_\varepsilon(\hat{z})$ of $\hat{z} = (\hat{x}, \hat{p})$ such that (4.1) is satisfied for all $(x, p) \in N_\varepsilon(\hat{z})$. For all such $z = (x, p)$, we have that the maximum of $pf(x, \cdot) - g(x, \cdot)$ taken over U is attained at

$$(4.2) \quad u(z) = \begin{cases} (0, 0) & \text{if } px^2 \geq 0, \\ \left(\frac{px^2}{4}, 0\right) & \text{if } -4 \leq px^2 \leq 0, \\ (-1, 0) & \text{if } px^2 \leq -4. \end{cases}$$

Hence, from (2.4) it follows that

$$(4.3) \quad H(x, p) = \begin{cases} \frac{x^3}{24} & \text{if } px^2 \geq 0, \\ \frac{p^2x^4}{8} + \frac{x^3}{24} & \text{if } -4 \leq px^2 \leq 0, \\ -px^2 - 2 - \frac{x^3}{24} & \text{if } px^2 \leq -4. \end{cases}$$

Equation (4.2) implies that \hat{x} , \hat{p} and \hat{u} satisfy Theorem 2.1 (3). Also, we have

$$g_{uu}(\hat{x}(t), \hat{u}(t)) - f_{uu}(\hat{x}(t), \hat{u}(t))\hat{p}(t) = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix} > 0.$$

To show that the pair (\hat{x}, \hat{u}) is optimal for (C'), by the previous remark, it suffices to apply our theorem where the boundary conditions are removed. Thus, it remains to prove that the extended Jacobi condition holds.

From (4.3) it follows that, for $z \in N_\varepsilon(\hat{z})$, $H(\cdot)$ is C^1 with gradient

$$(4.4) \quad \nabla H(x, p) = \begin{cases} \left(\frac{x^2}{8}, 0\right) & \text{if } px^2 \geq 0, \\ \left(\frac{p^2x^3}{2} + \frac{x^2}{8}, \frac{x^4p}{4}\right) & \text{if } -4 \leq px^2 \leq 0, \\ \left(-2px + \frac{x^2}{8}, x^2\right) & \text{if } px^2 \leq -4. \end{cases}$$

Equation (4.4) implies

$$\begin{aligned} H_{xx}(\hat{x}(t), \hat{p}(t)) &= \frac{1}{4}, & H_{xp}(\hat{x}(t), \hat{p}(t)) &= H_{px}(\hat{x}(t), \hat{p}(t)) = 0, \\ H_{pp}(\hat{x}(t), \hat{p}(t)) &= \begin{cases} 0 & \text{if } t \neq 1, \\ \text{does not exist} & \text{if } t = 1. \end{cases} \end{aligned}$$

In this case our Hamiltonian is not concave in x . Also, H is not C^2 near \hat{z} , but from Lemma 3.1, the Hamiltonian is C^{1+} near \hat{z} .

From [3, § 1], the following generalized gradient inclusion holds:

$$\partial \nabla H(z) \subset A(z) = \begin{pmatrix} \partial_x H_x(z) & \partial_p H_x(z) \\ \partial_x H_p(z) & \partial_p H_p(z) \end{pmatrix}.$$

Thus, to check the extended Jacobi condition for the elements of $\partial \nabla H(\hat{z}(t))$ it suffices to check this condition for the elements of $A(\hat{z}(t))$.

From (4.4) we deduce that $\partial_p H_p(\hat{z}(1)) = [0, \frac{1}{4}]$, and hence

$$A(\hat{z}(t)) = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \partial_p H_p(\hat{z}(t)) \end{pmatrix},$$

where

$$\partial_p H_p(\hat{z}(t)) = \begin{cases} 0 & \text{if } t \neq 1, \\ [0, \frac{1}{4}] & \text{if } t = 1. \end{cases}$$

Take $Q_0(t) = t$. Then for $(\frac{\alpha}{\beta} \frac{\delta}{\gamma}) \in A(\hat{z}(t))$ we have

$$\dot{Q}_0(t) - \gamma Q_0^2(t) + Q_0(t)\beta + \delta Q_0(t) - \alpha = \frac{3}{4} - \gamma t^2.$$

Since $\gamma \in [0, \frac{1}{4}]$ and $t \in [0, 1]$, we obtain

$$\frac{3}{4} - \gamma t^2 \geq \frac{1}{2} > 0.$$

Thus, \hat{z} satisfies the extended Jacobi condition, and therefore, by the theorem, (\hat{x}, \hat{u}) provides a strong local minimum for (C').

Remark. Since our control function $\hat{u} = (0, 0)$ is on the boundary of the control set U , the sufficiency theorem of [6] cannot be applied to our problem.

REFERENCES

- [1] V. G. BOLTYANSKII, *Sufficient conditions for optimality and the justification of the dynamic programming method*, this Journal, 4 (1966), pp. 326–361.
- [2] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
- [3] ———, *Generalized gradients of Lipschitz functionals*, Tech. Rep. 1687, Mathematics Research Center, Madison, WI, 1976; Adv. in Math., 40 (1981), pp. 52–67.
- [4] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.
- [5] V. F. KROTOV, *Methods for solving variational problems on the basis of sufficient conditions for an absolute minimum*: I, II, III, Automat. Remote Control (1962), pp. 1473–1484; (1963), pp. 539–553; (1964), pp. 924–933.
- [6] D. Q. MAYNE, *Sufficient conditions for a control to be a strong minimum*, J. Opt. Theory Appl., 21 (1977), pp. 339–351.
- [7] S. M. ROBINSON, *Generalized equations and their solutions: Part II: Applications to nonlinear programming*, preprint, University of Wisconsin-Madison, Madison, 1980.
- [8] R. T. ROCKAFELLAR, *Optimal arcs and the minimum value function in problems of Lagrange*, Trans. Amer. Math. Soc., 180 (1973), pp. 53–83.
- [9] A. SEIERSTAD AND K. SYDSAETER, *Sufficient conditions in optimal control theory*, Inter. Econo. Rev. J., 18 (1977), pp. 367–391.
- [10] R. B. VINTER AND R. M. LEWIS, *A verification theorem which provides a necessary and sufficient condition for optimality*, IEEE Trans. Autom. Control, AC-25 (1980), pp. 84–89.
- [11] V. ZEIDAN, *Sufficient conditions for the generalized problem of Bolza*, Trans. Amer. Math. Soc., to appear.

DIFFERENCE EQUATION STATE APPROXIMATIONS FOR NONLINEAR HEREDITARY CONTROL PROBLEMS*

I. G. ROSEN†

Abstract. Discrete approximation schemes for the solution of nonlinear hereditary control problems are constructed. The methods involve approximation by a sequence of optimal control problems in which the original infinite dimensional state equation has been approximated by a finite dimensional discrete difference equation. Convergence of the state approximations is argued using linear semigroup theory and is then used to demonstrate that solutions to the approximating optimal control problems in some sense approximate solutions to the original control problem. Two schemes, one based upon piecewise constant approximation and the other involving spline functions, are discussed. Numerical results are presented and used to compare the schemes to other available approximation methods for the solution of hereditary control problems.

Key words. approximation, nonlinear, delay systems, hereditary systems, control

1. Introduction. The purpose of this paper is two-fold. It first serves to describe how the abstract approximation framework developed for the integration of linear functional differential equation (FDE) initial value problems in [25] can be extended so as to be applicable to certain nonlinear problems as well. Secondly, the application of the resulting approximation schemes to the generation of approximate solutions to optimal control problems in which the dynamics of the underlying system are governed by nonlinear FDE is discussed.

The approach we take is not new. We consider the nonlinear FDE in an equivalent form, i.e., as an implicit abstract evolution equation in an infinite dimensional Hilbert space Z . We then construct a sequence of finite dimensional approximating discrete difference equations by approximating the solution semigroup of operators (and its infinitesimal generator) defined by the linear part of the equation using piecewise constant or spline based subspaces of Z . Linear semigroup theory and discrete analogues of the Trotter-Kato theorem [18] and the well-known Gronwall inequality are then used to argue convergence. Approximate solutions to the optimal control problem are generated by considering a sequence of approximating optimal control problems in each of which the infinite dimensional FDE state equation has been approximated in the spirit of the discussions above. Using the fact that the state approximations converge, we are then able to demonstrate that solutions to the approximating optimal control problems (which can be solved by conventional methods) in some sense approximate solutions to the original control problem. The application of our approximation framework to the integration of nonlinear FDE is based largely upon results which first appeared in [24]. The idea of approximating the infinite dimensional optimal control problem by a sequence of finite dimensional discrete optimal control problems closely parallels the approach taken in [26] where similar methods are used to obtain

* Received by the editors June 28, 1982, and in revised form February 7, 1983. This work was supported in part by the Air Force Office of Scientific Research under contract AFOSR 76-3092D, in part by the National Science Foundation under grant NSF-MCS-7905774-02; in part by the U.S. Army Research Office under contract ARO-DAAG29-79-C-0161; and in part by the Bowdoin College Faculty Research Fund. Additional support was provided by the National Aeronautics and Space Administration under NASA contracts NAS1-15810 and NAS1-16394 while the author was in residence at the Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, Virginia 23665.

† Department of Mathematics, Bowdoin College, Brunswick, Maine 04011. Present address: The Charles Stark Draper Laboratory, Cambridge, Massachusetts 02139.

approximate solutions to parameter identification problems involving FDE state equations.

Banks and Burns [3], [4] were among the first to propose the idea of approximating hereditary control problems by a sequence of finite dimensional approximating control problems. The semidiscrete methods for problems with linear state equations which they developed were later extended by Banks [2] so as to be applicable to nonlinear problems as well. Using similar approaches, Reber [21] and Rockey [23] developed fully discrete schemes for the approximation of FDE which they then applied to the solution of control problems. Reber developed first order convergent schemes for linear nonautonomous equations. In the constant coefficient case, his work becomes a special case of the more general theory to be presented below. In [23], the linear or nonlinear FDE is first recast as an equivalent Volterra integral equation in L_2 , and is then discretized using piecewise constant or spline subspaces. An algebraic system for the Fourier coefficients of the solution results, which is then solved using standard methods. Recently, within the context of the framework developed in [3], Gibson [13] and Kunisch [20] have formulated semidiscrete approximation schemes which yield approximating closed loop solutions to the linear quadratic control problem with hereditary system dynamics.

The discussion of our results below closely parallels the presentation in [2]. The treatment in [2] relies heavily upon the linear theory developed in [3] and [4] by considering the nonlinearities (which are assumed to satisfy local Lipschitz and affine growth conditions) to be a perturbation of the linear part of the equation. Since the basis for our approximation schemes involves the approximation of the solution semigroup $e^{\mathcal{A}t}$ using rational function approximations to the exponential and finite dimensional approximation of \mathcal{A} , we too depend heavily upon the linear theory, and hence consider precisely the same class of equations which are studied in [2]. An unfortunate consequence, however, is that this precludes the inclusion of nonlinearities in the discrete delay terms (i.e. terms of the form $x(t-r)$). This is in contrast to the work of Kappel and Schappacher [17] and Kappel [16] and the recent paper by Daniel [10] in which the nonlinearities in the equation are handled more directly and which do permit discrete delay terms to enter into the equation in a nonlinear fashion. The convergence arguments for the approximation schemes developed in [16] and [17] are based largely on ideas from nonlinear semigroup theory and approximation results analogous to those used in the linear case. Daniel on the other hand, avoids the semigroup approach entirely and relies instead, directly upon the dissipative properties of the nonlinear operators arising in the abstract formulation of the FDE in order to argue the convergence of spline based semidiscrete approximation schemes. These results are obtained, however, at the expense of requiring somewhat stronger assumptions (global Lipschitz and additional smoothness) on the nonlinearities in the equation and the placement of additional restrictions on the class of admissible controls.

We conclude this section with an outline of the rest of the paper and a brief description of our notation. In § 2 we define the nonlinear FDE with which we shall be concerned and state the hypotheses it must satisfy in order for us to carry out our analysis. We also state fundamental existence and uniqueness results and describe the equivalent formulation of the FDE as an abstract evolution equation in the Hilbert space Z . In § 3 we first recall the abstract approximation results for linear equations discussed in [25]. We then extend them so that they are applicable to the nonlinear equation as well and state and prove the fundamental convergence result. In § 4 we briefly describe the details involved in the construction of actual schemes to which our general convergence results apply. We also outline two specific schemes, one using

piecewise constant functions and the other using splines. Section 5 contains the results pertaining to the application of the approximation schemes to the solution of optimal control problems while in § 6 we demonstrate the feasibility of our methods by presenting several numerical examples.

The notation we use, is, for the most part, standard. The superscripts on the Lebesgue spaces $L_p^n(a, b)$, the space of functions with p continuous derivatives $C_p^n(a, b)$ and the Sobolev spaces $H_p^n(a, b)$ denote that they consist of functions (or equivalence classes of functions) defined on (a, b) with range in R^n . The symbol L_∞^{loc} is used to denote the class of functions which are locally essentially bounded. The space of continuous functions from an interval (a, b) with range in the abstract space Z is denoted by $\mathcal{C}([a, b], Z)$. We assume that this space is endowed with the usual supremum norm. For a linear operator \mathcal{A} and a complex number λ contained in the resolvent set of \mathcal{A} we denote the resolvent of \mathcal{A} at λ by $R(\lambda; \mathcal{A})$.

2. Nonlinear hereditary control systems and their abstract formulation. In this paper we consider nonlinear hereditary control systems which are governed by functional differential state equations of retarded type of the form

$$(2.1) \quad \dot{x}(t) = Lx_t + f(t, x(t), x_t, u(t)), \quad t \in [0, T],$$

with initial conditions given by

$$(2.2) \quad x(0) = \eta, \quad x_0 = \phi$$

where $\eta \in R^n$, $\phi \in L_2^n(-r, 0)$ and x_t denotes the function on $[-r, 0]$ defined by $x_t(\theta) = x(t + \theta)$, $-r \leq \theta \leq 0$. The linear part of the equation, given by the linear operator $L: L_2^n(-r, 0) \rightarrow R^n$ will be assumed to be of the form

$$L\phi = \sum_{j=0}^{\nu} A_j \phi(-\tau_j) + \int_{-r}^0 A(\theta) \phi(\theta) d\theta$$

where the A_j are $n \times n$ matrices, $A(\cdot)$ is a square integrable $n \times n$ matrix valued function defined on the interval $(-r, 0)$ and $0 = \tau_0 < \tau_1 < \tau_2 < \dots < \tau_\nu = r$.

In addition, we assume that the nonlinear perturbation term $f: R^1 \times R^n \times L_2^n(-r, 0) \times R^m \rightarrow R^n$ satisfies the following hypotheses:

(H1) The mapping $(t, \eta, \phi, v) \rightarrow f(t, \eta, \phi, v)$ is continuous on $R^1 \times R^n \times L_2^n(-r, 0) \times R^m$.

(H2) For any bounded subset \mathcal{D} of $R^n \times L_2^n(-r, 0)$ there exist $m_i = m_i(\mathcal{D})$, $m_i \in L_\infty^{\text{loc}}$, $i = 1, 2$, such that for $v \in R^m$, $t \in R^1$ and $(\eta, \phi), (\xi, \psi) \in \mathcal{D}$ one has

$$|f(t, \eta, \phi, v) - f(t, \xi, \psi, v)| \leq \{m_1(t) + m_2(t)|v|\} \{|\eta - \xi| + |\phi - \psi|\}.$$

(H3) There exists a continuous $n \times m$ matrix valued mapping $t \rightarrow B(t)$ such that $f(t, 0, 0, v) = B(t)v$ for all $t \in R^1$ and $v \in R^m$. In addition there exist functions $\hat{m}_i \in L_\infty^{\text{loc}}$, $i = 1, 2$, such that

$$|f(t, \eta, \phi, v)| \leq \{\hat{m}_1(t) + \hat{m}_2(t)|v|\} \{|\eta| + |\phi|\} + |B(t)||v|$$

for all $t \in R^1$, $v \in R^m$ and all $(\eta, \phi) \in R^n \times L_2^n(-r, 0)$ with $|(\eta, \phi)|^2 = |\eta|^2 + |\phi|^2$ sufficiently large.

(H4) There exists a continuous function $g: R^1 \times R^n \times L_2^n(-r, 0) \rightarrow R^1$ such that

$$|f(t, \eta, \phi, v) - f(t, \eta, \phi, w)| \leq g(t, \eta, \phi)|v - w|$$

for all $t \in R^1$, $\eta \in R^n$, $\phi \in L_2^n(-r, 0)$ and $v, w \in R^m$.

Hypotheses (H2) and (H3) together yield the following growth condition satisfied by f :

(G) There exist functions $\tilde{m}_1, \tilde{m}_2 \in L_\infty^{\text{loc}}$ such that

$$|f(t, \eta, \phi, v)| \in \{\tilde{m}_1(t) + \tilde{m}_2(t)|v|\} \{|\eta| + |\phi|\} + |B(t)||v|$$

for all $t \in R^1$, $\eta \in R^n$, $\phi \in L_2^n(-r, 0)$ and $v \in R^m$.

A solution $x(t) = x(t; \eta, \phi, u)$ to (2.1), (2.2) is defined to be a function $x \in L_2^n(-r, T)$ such that the mapping $t \rightarrow x(t)$ is absolutely continuous on $(0, T)$, (2.1) is satisfied almost everywhere on $(0, T)$ and for which $x(0) = \eta$, $x_0 = \phi$. Using standard arguments, the following existence, uniqueness and continuous dependence result for solutions to the initial value problem (2.1), (2.2) can be established.

THEOREM 2.1. *Under hypotheses (H1)–(H4), given $u \in L_2^m(0, T)$ and $(\eta, \phi) \in R^n \times H_1^n(-r, 0)$ with $\eta = \phi(0)$, there exists a unique solution to the initial value problem (2.1), (2.2) on $[0, T]$. Moreover, the mapping $(\phi, u) \rightarrow (x(t; \phi(0), \phi, u), x_t(\phi(0), \phi, u))$ from $H_1^n(-r, 0) \times L_2^m(0, T)$ into $R^n \times L_2^n(-r, 0)$ where x is the unique solution to (2.1), (2.2) corresponding to $u \in L_2^m(0, T)$ and $x(0) = \phi(0)$, $x_0 = \phi$, is continuous with respect to the topology on $H_1^n(-r, 0) \times L_2^m(0, T)$ induced by the supremum norm on $H_1^n(-r, 0)$ and the standard L_2 norm on $L_2^m(0, T)$.*

Fundamental to the development of our approximation schemes below will be the equivalence which exists between the FDE initial value problem (2.1), (2.2) above and an abstract evolution equation set in the Hilbert space $Z = R^n \times L_2^n(-r, 0)$ with inner product $\langle \cdot, \cdot \rangle_Z = \langle \cdot, \cdot \rangle_{R^n} + \langle \cdot, \cdot \rangle_{L_2^n(-r, 0)}$. For each $t \geq 0$ let $S(t): Z \rightarrow Z$ denote the solution operator for the associated linear homogeneous initial value problem corresponding to (2.1), (2.2). That is, for $(\eta, \phi) \in Z$ we have

$$S(t)(\eta, \phi) = (x(t), x_t)$$

where x is the unique solution to (2.1), (2.2) with $f \equiv 0$. Based upon existence, uniqueness and continuous dependence results for the linear homogeneous problem (see [3], [4], [25]), one may conclude that $\{S(t): t \geq 0\}$ represents a parameterized family of well defined bounded linear transformations forming a \mathcal{C}_0 -semigroup of operators on Z . The infinitesimal generator of $\{S(t): t \geq 0\}$, \mathcal{A} , and its domain of definition $D(\mathcal{A})$ may be computed and are given by

$$D(\mathcal{A}) = \{(\eta, \phi) \in Z: \eta = \phi(0), \phi \in H_1^n(-r, 0)\}, \quad \mathcal{A}(\phi(0), \phi) = (L\phi, \dot{\phi}).$$

If we define the inner product $\langle \cdot, \cdot \rangle_g$ on Z by

$$\langle (\eta, \phi), (\xi, \psi) \rangle_g = \eta^T \xi + \int_{-r}^0 \phi(\theta)^T \psi(\theta) g(\theta) d\theta$$

where

$$g(\theta) = \begin{cases} 1, & -r \leq \theta < \tau_{\nu-1}, \\ 2, & -\tau_{\nu-1} \leq \theta < \tau_{\nu-2}, \\ \vdots & \vdots \\ \nu, & \tau_1 \leq \theta \leq 0, \end{cases}$$

then it clearly follows that for $(\eta, \phi) \in Z$

$$|(\eta, \phi)|_Z \leq |(\eta, \phi)|_g \leq \sqrt{\nu} |(\eta, \phi)|_Z.$$

Furthermore, it can be shown that the operator \mathcal{A} satisfies the following dissipative inequality with respect to the g inner product:

$$(2.3) \quad \langle \mathcal{A}z_0, z_0 \rangle_g \leq \omega \langle z_0, z_0 \rangle_g$$

with

$$\omega = \frac{\nu+1}{2} + |A_0| + \frac{1}{2} \sum_{i=1}^{\nu} |A_i|^2 + \frac{1}{2} \int_{-r}^0 |A(\theta)|^2 d\theta,$$

and hence $\mathcal{A} \in G(\sqrt{\nu}, \omega)$ —that is, the semigroup of operators $\{S(t): t \geq 0\}$ satisfies the exponential bound given by

$$|S(t)| \leq \sqrt{\nu} e^{\omega t}.$$

Let $\pi_1: Z \rightarrow R^n$ and $\pi_2: Z \rightarrow L_2^n(-r, 0)$ denote the two coordinate projections of Z onto R^n and $L_2^n(-r, 0)$ respectively. That is, for $(\eta, \phi) \in Z$, we have

$$\pi_1(\eta, \phi) = \eta, \quad \pi_2(\eta, \phi) = \phi.$$

Let the mapping $F: R^1 \times Z \times R^m \rightarrow Z$ be defined by

$$F(t, z, v) = (f(t, \pi_1 z, \pi_2 z, v), 0).$$

Hypotheses (H1)–(H4) imposed upon f naturally imply that the mapping F defined above will have the following properties:

(P1) For any $z \in \mathcal{C}([0, T], Z)$ and $u \in L_2^m(0, T)$, the mapping $t \rightarrow |F(t, z(t), u(t))|$ is in $L_2^1(0, T)$.

(P2) For any bounded subset \mathcal{D} of Z , there exist M_1, M_2 (depending on \mathcal{D}) in L_∞^{loc} such that $|F(t, z, v) - F(t, w, v)| \leq \{M_1(t) + M_2(t)|v|\}|z - w|$ for all $z, w \in \mathcal{D}$, $t \in R^1$ and $v \in R^m$.

For $z_0 \in Z$ and $u \in L_2^m(0, T)$, let the mapping $z: [0, T] \rightarrow Z$ be defined implicitly by the following expression

$$(2.4) \quad z(t) = z(t; z_0, u) = S(t)z_0 + \int_0^t S(t-\sigma)F(\sigma, z(\sigma), u(\sigma)) d\sigma.$$

Using hypotheses (H1)–(H4) and properties (P1) and (P2) above, together with standard arguments involving Picard iterates and the Gronwall inequality, Banks [2] is able to establish the following lemma.

LEMMA 2.1. *Under hypotheses (H1)–(H4), (2.4) above defines for each $z_0 \in Z$ and $u \in L_2^m(0, T)$ a unique function $t \rightarrow z(t; z_0, u) \in \mathcal{C}([0, T], Z)$. Moreover, the mapping $(\phi(0), \phi, u) \rightarrow z(t, (\phi(0), \phi), u)$ is continuous on $D(\mathcal{A}) \times L_2^m(0, T)$ with respect to the $Z \times L_2$ and $R^n \times C^n \times L_2$ topologies.*

Finally, using the above results, the equivalence which we desire between the FDE initial value problem (2.1), (2.2) and an abstract evolution equation set in Z , in particular the system given by (2.4), can be established. The details of the proof can be found in [2].

THEOREM 2.2. *For f satisfying hypotheses (H1)–(H4), $z_0 = (\phi(0), \phi) \in D(\mathcal{A})$ and $u \in L_2^m(0, T)$ we have*

$$z(t; z_0, u) = (x(t; \phi(0), \phi, u), x_t(\phi(0), \phi, u)), \quad t \in [0, T],$$

where $z(t; z_0, u)$ is the unique solution to (2.4) guaranteed to exist by Lemma 2.1, and $x(t; \phi(0), \phi, u)$ is the unique solution to the FDE initial value problem (2.1), (2.2) guaranteed to exist by Theorem 2.1.

3. An abstract approximation framework. In this section we develop an abstract approximation framework under which approximation schemes applicable to the abstract evolution equation given by (2.4) can be constructed. In addition, we establish conditions which are sufficient to conclude convergence of schemes constructed within the framework. The approach we take is based upon, and an extension of, the discrete approximation framework for the integration of linear FDE initial value problems described in [25]. Indeed our schemes will be based upon the approximation of the semigroup of operators $\{S(t): t \geq 0\}$ defined on Z by a sequence of discrete semigroups (see [18]) which are defined on finite dimensional approximating subspaces of Z and which are constructed using rational function approximations to the exponential and finite dimensional approximations to the infinitesimal generator \mathcal{A} of $\{S(t): t \geq 0\}$. The fundamental convergence results for these constructions are given in Theorem 3.1 to follow and are used extensively throughout our discussions below.

For each $N = 1, 2, \dots$ let Z_N be a finite dimensional subspace of Z of dimension k_N and let $P_N: Z \rightarrow Z_N$ be the associated orthogonal (not necessarily with respect to the standard inner product on Z) projection of Z onto Z_N . Define $\mathcal{A}_N: Z_N \rightarrow Z_N$ to be a bounded linear operator on Z_N and let $S_N(t) = e^{\mathcal{A}_N t}$ for all $t \geq 0$.

THEOREM 3.1. *Suppose*

- (1) $P_N z \rightarrow z$ as $N \rightarrow \infty$ for each $z \in Z$.
- (2) *There exist constants M, β , independent of N for which $\mathcal{A}, \mathcal{A}_N \in G(M, \beta)$, $N = 1, 2, \dots$, (i.e. $|S(t)| \leq M e^{\beta t}$, $|S_N(t)| \leq M e^{\beta t}$, $N = 1, 2, \dots$).*
- (3) *There exists $D_1 \subset D(\mathcal{A})$, a dense subset of Z for which*

$$|\mathcal{A}_N P_N z - \mathcal{A} z| \rightarrow 0 \quad \text{as } N \rightarrow \infty \quad \text{for each } z \in D_1.$$

- (4) *There exist $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda > \beta$ and D_2 a dense subset of Z for which $R(\lambda; \mathcal{A}) D_2 \subseteq D_1$.*

- (5) *$C(z)$ is a rational function of the complex variable z for which*

- (a) $|C(z) - e^z| = O(|z|^{q+1})$ as $|z| \rightarrow 0$ with $q > 0$;
- (b) *If $C(z) = n(z)/d(z)$ then $\operatorname{degree} C(z) = \operatorname{degree} n(z) - \operatorname{degree} d(z) \leq q + 1$;*
- (c) *$C(z)$ has no poles in $\{z \in \mathbb{C}: \operatorname{Re} z \leq 0\}$.*

Then the operators $C((r/N)\mathcal{A}_N) = n((r/N)\mathcal{A}_N) d((r/N)\mathcal{A}_N)^{-1}$ exist for all N sufficiently large. If, in addition, for ρ_N , that positive integer for which $\rho_N(r/N) \leq T < (\rho_N + 1)r/N$, we have that the infinite collection of operators on Z_N , $\{C((r/N)\mathcal{A}_N)^k\}_{k=0}^{\rho_N}$ are uniformly bounded with respect to N then

$$(3.1) \quad \left| C\left(\frac{r}{N}\mathcal{A}_N\right)^k P_N z - S(t_k^N) z \right| \rightarrow 0$$

as $N \rightarrow \infty$ for each $z \in Z$ uniformly in k , $k = 0, 1, 2, \dots, \rho_N$, where $t_k^N = k(r/N)$, $k = 0, 1, 2, \dots, \rho_N$.

Theorem 3.1 is based primarily upon a result due to Hersh and Kato [14] and is in fact a fully discrete analogue of the well-known Trotter–Kato results which are commonly used in establishing the convergence of semidiscrete approximations to semigroups of operators (see [18]). That it is possible to actually construct schemes (i.e. $Z_N, P_N, \mathcal{A}_N, C(z)$) which satisfy the hypotheses of Theorem 3.1 is exhibited in § 4.

Remark 3.1. As a corollary to Theorem 3.1, it is possible to estimate the rate of convergence in (3.1). Indeed, if for $z \in \mathcal{S}$, a particular subset of Z which is defined in [25, Thm. 4.17] we have that the convergence in hypothesis (3) is $O(r/N)^p$ for some $p > 0$, then the rate of convergence in (3.1) will be $O(r/N)^p + O(r/N)^q = O(r/N)^l$ for $z \in \mathcal{S}$ where $l = \min(p, q)$.

Before we can proceed to apply the results of Theorem 3.1 in the development of approximation schemes for the nonlinear system (2.4), we must first consider the linear nonhomogeneous problem. We shall require the following result from [25]. For $f \in L_2^n(0, T)$ and $z_0 \in Z$, let $z \in \mathcal{C}([0, T], Z)$ be given by

$$z(t) = S(t)z_0 + \int_0^t S(t-\sigma)(f(\sigma), 0) d\sigma$$

and let $\{z_k^N\}_{k=0}^{\rho_N} \subset Z_N$ be given by

$$z_k^N = C\left(\frac{r}{N}\mathcal{A}_N\right)^k P_N z_0 + \frac{r}{N} \sum_{j=1}^k C\left(\frac{r}{N}\mathcal{A}_N\right)^{k-j} D\left(\lambda \frac{r}{N}\mathcal{A}_N\right) P_N(f_j^N, 0)$$

where $f_j^N = (N/r) \int_{(j-1)r/N}^{j(r/N)} f(\sigma) d\sigma$, $D(z)$ is a rational function of the complex variable z and $0 \leq \lambda \leq 1$.

THEOREM 3.2. *Suppose that Z_N , P_N , \mathcal{A}_N , $C(z)$ satisfy the hypotheses of Theorem 3.1. Suppose further that*

- (1) *The infinite collection of operators on Z_N , $\{C((r/N)\mathcal{A}_N)^k\}_{k=0}^{\rho_N}$ are uniformly bounded with respect to N and*
- (2) *The operators $D(\lambda(r/N)\mathcal{A}_N)$ exist for all N sufficiently large and satisfy $|D(\lambda(r/N)\mathcal{A}_N)P_N z - z| \rightarrow 0$ as $N \rightarrow \infty$ for each $z \in Z$.*

Then

$$|z_k^N - z(t_k^N)| \rightarrow 0$$

as $N \rightarrow \infty$ for each $z_0 \in Z$ uniformly in k , $k = 0, 1, 2, \dots, \rho_N$ and uniformly in f for f in bounded subsets of $L_2^n(0, T)$.

Several of our arguments below rely upon an application of the following lemma. The result given in Lemma 3.1 is a discrete analogue of the well-known generalized Gronwall differential inequality. The proof, which is not difficult, has been omitted but can be found in [24].

LEMMA 3.1. *Suppose that $\{\alpha_j\}_{j=0}^\infty$ and $\{\beta_j\}_{j=0}^\infty$ are sequences of nonnegative real numbers and that $\{\phi_j\}_{j=0}^\infty$ is a sequence of real numbers which satisfy*

$$\phi_n \leq \alpha_n + \sum_{k=0}^{n-1} \beta_k \phi_k, \quad n = 1, 2, \dots$$

Then we have that

$$\phi_n \leq \alpha_n + \sum_{j=0}^{n-1} \beta_j \alpha_j \left(\exp \left(\sum_{k=j+1}^{n-1} \beta_k \right) \right), \quad n = 1, 2, \dots$$

If in addition $\alpha_j = \alpha \geq 0$, $j = 0, 1, 2, \dots$, then

$$\phi_n \leq \alpha \exp \left(\sum_{k=0}^{n-1} \beta_k \right), \quad n = 1, 2, \dots$$

For Z_N , P_N , \mathcal{A}_N , $C(z)$, $D(z)$ and ρ_N as described above, $z_0 \in Z$, $u \in L_2^m(0, T)$ and f satisfying hypotheses (H1)–(H4), we define the collection $\{z_k^N\}_{k=0}^{\rho_N} \subset Z_N$ by

$$(3.2) \quad z_k^N = z_k^N(z_0, u) = C\left(\frac{r}{N}\mathcal{A}_N\right)^k P_N z_0 + \frac{r}{N} \sum_{j=1}^k C\left(\frac{r}{N}\mathcal{A}_N\right)^{k-j} D\left(\lambda \frac{r}{N}\mathcal{A}_N\right) P_N F_j^N, \\ k = 0, 1, 2, \dots, \rho_N,$$

where

$$F_i^N \equiv \frac{N}{r} \int_{(i-1)r/N}^{i(r/N)} F(\sigma, z_{i-1}^N, u(\sigma)) d\sigma = \left(\frac{N}{r} \int_{(i-1)r/N}^{i(r/N)} f(\sigma, \pi_1 z_{i-1}^N, \pi_2 z_{i-1}^N, u(\sigma)) d\sigma, 0 \right),$$

$$i = 1, 2, \dots, \rho_N.$$

LEMMA 3.2. For $Z_N, P_N, \mathcal{A}_N, C(z), D(z)$ and ρ_N satisfying the hypotheses of Theorem 3.2 and $u \in \mathcal{E}$, a bounded subset of $L_2^m(0, T)$, the collection $\{z_k^N\}_{k=0}^{\rho_N}$ defined by (3.2) above are bounded in $(\times_{k=0}^{\rho_N} Z, \|\cdot\|_\infty^N)$ uniformly in N for all N sufficiently large and uniformly in $u \in \mathcal{E}$ where $\|\{z_k^N\}_{k=0}^{\rho_N}\|_\infty^N \equiv \max_{0 \leq k \leq \rho_N} \|z_k^N\|_Z$.

Proof. Let K_0 denote the uniform bound on the operators $\{C((r/N)\mathcal{A}_N)^k\}_{k=0}^{\rho_N}$ for all N sufficiently large, which is assumed to exist in hypothesis (1) of Theorem 3.2, and let K_1 denote the uniform bound on the operators $D(\lambda(r/N)\mathcal{A}_N)$ for all N sufficiently large, whose existence can be argued using hypothesis (2) of Theorem 3.2 and the uniform boundedness principle. Then, for $k = 0, 1, 2, \dots, \rho_N$,

$$\begin{aligned} |z_k^N| &\leq K_0 |z_0| + \frac{r}{N} K_0 K_1 \sum_{j=1}^k |P_N F_j^N| \\ &\leq K_0 |z_0| + K_0 K_1 \sum_{j=1}^k \int_{(j-1)r/N}^{j(r/N)} |f(\sigma, \pi_1 z_{j-1}^N, \pi_2 z_{j-1}^N, u(\sigma))| d\sigma. \end{aligned}$$

Applying the growth condition (G) satisfied by f , we find

$$\begin{aligned} |z_k^N| &\leq K_0 |z_0| + K_0 K_1 \sum_{j=1}^k \int_{(j-1)r/N}^{j(r/N)} ((\tilde{m}_1(\sigma) + \tilde{m}_2(\sigma)|u(\sigma)|) \\ &\quad \cdot \{|\pi_1 z_{j-1}^N| + |\pi_2 z_{j-1}^N|\} + |B(\sigma)||u(\sigma)|) d\sigma \\ &\leq K_0 |z_0| + K_0 K_1 |B|_{L_2^{q \times m}(0, T)} \|u\|_{L_2^m(0, T)} \\ &\quad + K_0 K_1 \sum_{j=1}^{\rho_N} \{|\pi_1 z_{j-1}^N| + |\pi_2 z_{j-1}^N|\} \int_{(j-1)r/N}^{j(r/N)} (\tilde{m}_1(\sigma) + \tilde{m}_2(\sigma)|u(\sigma)|) d\sigma \\ &\leq K_0 |z_0| + K_0 K_1 |B|_{L_2^{q \times m}(0, T)} \|u\|_{L_2^m(0, T)} \\ &\quad + \sqrt{2} K_0 K_1 \sum_{j=1}^{\rho_N} |z_{j-1}^N| \int_{(j-1)r/N}^{j(r/N)} (\tilde{m}_1(\sigma) + \tilde{m}_2(\sigma)|u(\sigma)|) d\sigma, \end{aligned}$$

and hence by Lemma 3.1

$$\begin{aligned} |z_k^N| &\leq (K_0 |z_0| + K_0 K_1 |B|_{L_2^{q \times m}(0, T)} \|u\|_{L_2^m(0, T)}) \\ &\quad \cdot \exp \left(\sqrt{2} K_0 K_1 \int_0^T (\tilde{m}_1(\sigma) + \tilde{m}_2(\sigma)|u(\sigma)|) d\sigma \right) \\ &\leq (K_0 |z_0| + K_0 K_1 |B|_{L_2^{q \times m}(0, T)} \|u\|_{L_2^m(0, T)}) \\ &\quad \cdot \exp (\sqrt{2} K_0 K_1 (\|\tilde{m}_1\|_{L_\infty} T + \|\tilde{m}_2\|_{L_\infty} T^{1/2} \|u\|_{L_2^m(0, T)})). \end{aligned}$$

THEOREM 3.3. For $Z_N, P_N, \mathcal{A}_N, C(z), D(z)$ and ρ_N satisfying the hypotheses of Theorem 3.2, $u \in \mathcal{E}$, a bounded subset of $L_2^m(0, T)$, $\{z_k^N\}_{k=0}^{\rho_N}$ given by (3.2), and z given by (2.4) we have

$$|z_k^N(z_0, u) - z(t_k^N; z_0, u)| \rightarrow 0$$

as $N \rightarrow \infty$ for each $z_0 \in Z$ uniformly in $k, k = 0, 1, 2, \dots, \rho_N$ and uniformly in u for $u \in \mathcal{E}$.

Proof. For $z \in \mathcal{C}([0, T], Z)$, the unique solution to (2.4) guaranteed to exist by Lemma 2.1, and $t \in [0, T]$ define the function h by

$$h(t) = h(t, u) = f(t, \pi_1 z(t), \pi_2 z(t), u(t)).$$

Then

$$z(t) = S(t)z_0 + \int_0^t S(t-\sigma)(h(\sigma), 0) d\sigma$$

and using hypotheses (H1)–(H4) it is easily verified that for $u \in \mathcal{E}$, h lies in a bounded subset of $L_2^n(0, T)$. If we define $\{\tilde{z}_k^N\}_{k=0}^{\rho_N} \subset Z_N$ by

$$\tilde{z}_k^N = C\left(\frac{r}{N}\mathcal{A}_N\right)^k P_{NZ_0} + \frac{r}{N} \sum_{j=1}^k C\left(\frac{r}{N}\mathcal{A}_N\right)^{k-j} D\left(\lambda \frac{r}{N}\mathcal{A}_N\right) P_N(h_j^N, 0)$$

where $h_i^N = (N/r) \int_{t_{i-1}^N}^{t_i^N} h(\sigma) d\sigma$, then it follows that

$$\begin{aligned} |z_k^N - z(t_k^N)| &\leq |z_k^N - \tilde{z}_k^N| + |\tilde{z}_k^N - z(t_k^N)| \\ &\leq |\tilde{z}_k^N - z(t_k^N)| + \sum_{j=1}^k K_0 K_1 \int_{t_{j-1}^N}^{t_j^N} |f(\sigma, \pi_1 z_{j-1}^N, \pi_2 z_{j-1}^N, u(\sigma)) \\ &\quad - f(\sigma, \pi_1 z(\sigma), \pi_2 z(\sigma), u(\sigma))| d\sigma \\ &= |\tilde{z}_k^N - z(t_k^N)| + \sum_{j=0}^{k-1} K_0 K_1 \int_{t_j^N}^{t_{j+1}^N} |F(\sigma, z_j^N, u(\sigma)) - F(\sigma, z(\sigma), u(\sigma))| d\sigma \end{aligned}$$

where K_0 and K_1 are as they were defined in the proof of Lemma 3.2. Since $\{z(t; z_0, u): t \in [0, T], u \in \mathcal{E}\}$ lies in a bounded subset of Z (see [2]), as does $\{z_k^N(z_0, u), k = 0, 1, 2, \dots, \rho_N, u \in \mathcal{E}\}$ uniformly in N for all N sufficiently large, property (P2) implies

$$\begin{aligned} |z_k^N - z(t_k^N)| &\leq |\tilde{z}_k^N - z(t_k^N)| + \sum_{j=0}^{k-1} K_0 K_1 \int_{t_j^N}^{t_{j+1}^N} (M_1(\sigma) + M_2(\sigma)|u(\sigma)|) |z_j^N - z(\sigma)| d\sigma \\ (3.3) \quad &\leq |\tilde{z}_k^N - z(t_k^N)| + K_0 K_1 \sum_{j=0}^{k-1} \int_{t_j^N}^{t_{j+1}^N} (M_1(\sigma) + M_2(\sigma)|u(\sigma)|) |z(t_j^N) - z(\sigma)| d\sigma \\ &\quad + \sum_{j=0}^{k-1} K_0 K_1 \int_{t_j^N}^{t_{j+1}^N} (M_1(\sigma) + M_2(\sigma)|u(\sigma)|) d\sigma |z_j^N - z(t_j^N)|. \end{aligned}$$

Let $\varepsilon > 0$ be given. Theorem 3.2 implies that $|\tilde{z}_k^N - z(t_k^N)| < \varepsilon$ for all N sufficiently large uniformly in k , $k = 0, 1, 2, \dots, \rho_N$ and uniformly in u for $u \in \mathcal{E}$. Furthermore by [3, Thm. 3.2], the operator $\mathcal{F}: L_2^n(0, T) \rightarrow \mathcal{C}([0, T], Z)$ defined by

$$\mathcal{F}(f)(t) = S(t)z_0 + \int_0^t S(t-\sigma)(f(\sigma), 0) d\sigma$$

is a compact affine operator. Since $u \in \mathcal{E}$ a bounded subset of $L_2^m(0, T)$ implies $h(\cdot, u)$ lies in a bounded subset of $L_2^n(0, T)$, it follows that $\{z(\cdot; z_0, u): u \in \mathcal{E}\}$ is a relatively compact subset of $\mathcal{C}([0, T], Z)$. Therefore, the mappings $t \rightarrow z(t; z_0, u)$, $u \in \mathcal{E}$, are uniformly equicontinuous on $[0, T]$ and $|z(t_k^N) - z(\sigma)| < \varepsilon$, $\sigma \in [t_k^N, t_{k+1}^N]$ for all N sufficiently large uniformly in k , $k = 0, 1, 2, \dots, \rho_N$, and uniformly in u , $u \in \mathcal{E}$. The

above arguments together with the inequalities given by (3.3) imply

$$\begin{aligned}
 |z_k^N - z(t_k^N)| &\leq \varepsilon \left(1 + K_0 K_1 \int_0^T (M_1(\sigma) + M_2(\sigma)|u(\sigma)|) d\sigma \right) \\
 &\quad + \sum_{j=0}^{k-1} K_0 K_1 \int_{t_j^N}^{t_{j+1}^N} (M_1(\sigma) + M_2(\sigma)|u(\sigma)|) d\sigma |z_j^N - z(t_j^N)| \\
 (3.4) \quad &\leq \varepsilon (1 + K_0 K_1 T |M_1|_{L_\infty} + K_0 K_1 |M_2|_{L_\infty} T^{1/2} |u|_{L_2}) \\
 &\quad + \sum_{j=0}^{k-1} K_0 K_1 \int_{t_j^N}^{t_{j+1}^N} (M_1(\sigma) + M_2(\sigma)|u(\sigma)|) d\sigma |z_j^N - z(t_j^N)|
 \end{aligned}$$

for all N sufficiently large. If we now apply Lemma 3.1 to (3.4), it then follows that

$$\begin{aligned}
 |z_k^N - z(t_k^N)| &\leq \varepsilon (1 + \gamma) \exp \left(K_0 K_1 \sum_{j=0}^{k-1} \int_{t_j^N}^{t_{j+1}^N} (M_1(\sigma) + M_2(\sigma)|u(\sigma)|) d\sigma \right) \\
 &\leq \varepsilon (1 + \gamma) \exp \left(K_0 K_1 \int_0^T (M_1(\sigma) + M_2(\sigma)|u(\sigma)|) d\sigma \right) \\
 &\leq \varepsilon (1 + \gamma) \exp(\gamma)
 \end{aligned}$$

for all N sufficiently large, where

$$\gamma = K_0 K_1 T |M_1|_{L_\infty} + K_0 K_1 |M_2|_{L_\infty} T^{1/2} |u|_{L_2},$$

and the theorem is proven.

COROLLARY 3.1. *For $\{z_k^N\}_{k=0}^{\rho_N}$ generated by an approximation scheme satisfying the hypotheses of Theorem 3.3, it follows that*

$$|\pi_1 z_k^N((\eta, \phi), u) - x(t_k^N; \eta, \phi, u)| \rightarrow 0$$

as $N \rightarrow \infty$ uniformly in k , $k = 0, 1, 2, \dots, \rho_N$, and uniformly in u , $u \in \mathcal{E}$, where x denotes the unique solution to the initial value problem (2.1), (2.2).

4. Construction of convergent approximation schemes. In this section we construct approximation schemes which are based upon the framework described in § 3 and which satisfy the hypotheses of Theorem 3.3. Since the schemes described below and the verification of the fact that they satisfy the hypotheses of Theorem 3.3 have appeared elsewhere [2], [4], [8], [25], the relevant results are outlined and the details omitted. A similar description and outline of these methods appears in [26]. However, so as to make our presentation complete and self-contained and, more importantly, since the formulation of the schemes which is required here is somewhat simpler than that which is required for the parameter estimation problem, we have included the following brief summary.

Each of our approximation schemes is composed of two interrelated components: the state discretization, as is characterized by the choice of Z_N , P_n , and \mathcal{A}_n , and the temporal discretization, which is determined by the rational functions $C(z)$ and $D(z)$. The interrelation which exists between the two components is a consequence of the conditions under which our fundamental convergence result, Theorem 3.3, applies. We begin with a description of two state approximations and then discuss and characterize families of rational functions which, when coupled with these state approximations, lead to convergent approximation schemes.

The averaging state approximation (AVE), the more primitive of the two state approximations to be discussed, is based upon finite difference approximations and is defined as follows. For each $N = 1, 2, \dots$ let $\chi_j^N, j = 1, 2, \dots, N$, denote the characteristic function on the interval $[-j(r/N), -(j-1)r/N]$ and let

$$Z_N = \left\{ (\eta, \phi) \in Z : \phi = \sum_{j=1}^N v_j \chi_j^N, v_j \in \mathbb{R}^n \right\}.$$

We note that $\dim Z_N = n(N+1)$, and that the orthogonal (with respect to the standard Z inner product) projections $P_N: Z \rightarrow Z_N$ are given by

$$P_N(\eta, \phi) = \left(\eta, \sum_{j=1}^N \phi_j^N \chi_j^N \right)$$

where $\phi_j^N = (N/r) \int_{-j(r/N)}^{-(j-1)r/N} \phi(\theta) d\theta$. It is not difficult to show that $P_N z \rightarrow z$ as $N \rightarrow \infty$ for each $z \in Z$. Let the operators $L_N: Z_N \rightarrow \mathbb{R}^n$ and $D_N: Z_N \rightarrow L_2^n(-r, 0)$ be given by

$$L_N \left(\eta, \sum_{j=1}^N v_j \chi_j^N \right) = A_0 \eta + \sum_{i=1}^n \sum_{j=1}^N A_i v_j \chi_j^N(-\tau_i) + \frac{r}{N} \sum_{j=1}^N A_j^N v_j$$

where $A_j^N = (N/r) \int_{-j(r/N)}^{-(j-1)r/N} A(\theta) d\theta, j = 1, 2, \dots, N$, and

$$D_N \left(\eta, \sum_{j=1}^N v_j \chi_j^N \right) = \sum_{j=1}^N \frac{N}{r} (v_{j-1} - v_j) \chi_j^N$$

where $v_0 = \eta$, respectively. Define $\mathcal{A}_N: Z_N \rightarrow Z_N$ by

$$\mathcal{A}_N(\eta, \phi) = (L_N(\eta, \phi), D_N(\eta, \phi)),$$

and for $t \geq 0$ let $S_N(t) = e^{\mathcal{A}_N t}$. A sequence of inner products on Z , $\langle \cdot, \cdot \rangle_N$, can be constructed for which there exists an $M > 0$, independent of N , such that

$$(4.1) \quad |(\eta, \phi)|_N \leq |(\eta, \phi)| \leq M |(\eta, \phi)|_N$$

for all $(\eta, \phi) \in Z_N$. Furthermore, there exists a $\beta > 0$ independent of N , for all N sufficiently large, for which the operators $\mathcal{A}_N - \beta I$ are maximal dissipative with respect to the $\langle \cdot, \cdot \rangle_N$ inner product on Z_N . It follows therefore that $\mathcal{A}_N \in G(M, \beta)$ and $|S_N(t)| \leq M e^{\beta t}$ for all N sufficiently large. It is in fact the case that the \mathcal{A}_N as defined above satisfy a somewhat stronger condition. It can be shown that there exists an $\alpha > 0$ for which

$$(4.2) \quad \left| I + \frac{r}{N} \mathcal{A}_N \right|_N \leq 1 + \alpha \frac{r}{N},$$

for all N sufficiently large. The significance of (4.2) will become apparent when we discuss the choice of the rational function component of the approximation scheme below.

If we let $D_1 = D(\mathcal{A}^2)$ and $D_2 = D(\mathcal{A})$, then $D_1 \subset D(\mathcal{A})$ is a dense subset of Z , and for all $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda > \beta$, $R(\lambda; \mathcal{A}) D_2 = D_1$. Moreover, it can be shown that for each $z \in D_1$

$$|\mathcal{A}_N P_N z - \mathcal{A} z| = O(N^{-1/2})$$

as $N \rightarrow \infty$.

We shall next describe a spline based state approximation. The discussions which follow will be restricted to constructions involving linear or first order spline functions. However, the results given below are easily generalized so as to be applicable to state

approximations employing higher order spline functions. For each $N = 1, 2, \dots$ and $\theta \in [-r, 0]$ let

$$\begin{aligned}\phi_0^N(\theta) &= \begin{cases} \frac{N}{r}(\theta - t_1^N), & t_1^N \leq \theta \leq 0, \\ 0 & \text{otherwise;} \end{cases} \\ \phi_j^N(\theta) &= \begin{cases} \frac{N}{r}(t_{j-1}^N - \theta), & t_j^N \leq \theta \leq t_{j-1}^N, \\ \frac{N}{r}(\theta - t_{j+1}^N), & t_{j+1}^N \leq \theta \leq t_j^N, \quad j = 1, 2, \dots, N-1, \\ 0 & \text{otherwise;} \end{cases} \\ \phi_N^N(\theta) &= \begin{cases} \frac{N}{r}(t_{N-1}^N - \theta), & -r \leq \theta \leq t_{N-1}^N, \\ 0 & \text{otherwise,} \end{cases}\end{aligned}$$

where $t_j^N = -j(r/N)$, $j = 0, 1, 2, \dots, N$, and define Z_N by

$$Z_N = \left\{ (\phi(0), \phi) \in Z : \phi = \sum_{j=0}^N v_j \phi_j^N, v_j \in \mathbb{R}^n \right\}.$$

It is immediately clear that $\dim Z_N = n(N+1)$, $Z_N \subset D(\mathcal{A})$, and Z_N consists of all those elements $(\eta, \phi) \in Z$ for which $\eta = \phi(0)$ and ϕ is a first order spline function with knots at $\{t_j^N\}_{j=0}^N$. Let $P_N: Z \rightarrow Z_N$ denote the orthogonal projection from Z onto Z_N computed with respect to the weighted inner product on Z , $\langle \cdot, \cdot \rangle_g$, defined in § 2. Finally we define the operators $\mathcal{A}_N: Z_N \rightarrow Z_N$ by

$$\mathcal{A}_N = P_N \mathcal{A}.$$

Using the fact that the P_N are orthogonal projections, it follows from (2.3) that for $z_N \in Z_N$

$$(4.3) \quad \langle \mathcal{A}_N z_N, z_N \rangle_g = \langle P_N \mathcal{A} z_N, z_N \rangle_g = \langle \mathcal{A} z_N, P_N z_N \rangle_g = \langle \mathcal{A} z_N, z_N \rangle_g \leq \omega \langle z_N, z_N \rangle_g,$$

and hence that $\mathcal{A}_N \in G(\sqrt{\nu}, \omega)$. Furthermore, using the properties of interpolatory splines, it is not difficult to show that $P_N z \rightarrow z$ as $N \rightarrow \infty$ for each $z \in Z$, and that

$$(4.4) \quad |\mathcal{A}_N P_N z - \mathcal{A} z| = O(N^{-1})$$

for each $z \in D_1 \equiv D(\mathcal{A}^3)$. If we choose $D_2 = D(\mathcal{A}^2)$, then $R(\lambda; \mathcal{A})D_2 = D_1$ and all of the hypotheses and conditions of Theorem 3.3 concerning the state approximation only hold for the linear spline scheme defined above. We note that for state approximations employing higher order spline functions, the order of convergence in (4.4) and therefore in the integration method itself (see Remark 3.1) can be increased.

For either the AVE or the spline based state approximations, a rational function $C(z)$ satisfying the conditions of Theorem 3.1 must be chosen for which the operators $\{C((r/N)\mathcal{A}_N)\}_{l=0}^{\rho_N}$ are uniformly bounded in N for all N sufficiently large. It is clear from condition (5a) that we are seeking rational function approximations to the exponential. While there are many families of approximating rational functions from which to choose, we have restricted our attention to the well-known Padé approximants

[28] which are given by $P_{jk}(z) = N_{jk}(z)/D_{jk}(z)$ where

$$(4.5) \quad N_{jk}(z) = \sum_{i=0}^k \frac{(j+k-i)!k!}{(j+k)!i!(k-i)!} z^i$$

and

$$(4.6) \quad D_{jk}(z) = \sum_{i=0}^j \frac{(j+k-i)!j!}{(j+k)!i!(j-i)!} (-z)^i.$$

It can be shown that

$$|P_{jk}(z) - e^z| = O(|z|^{j+k+1}) \quad \text{as } |z| \rightarrow 0,$$

and hence the Padé approximants satisfy condition (5b) since $\deg P_{jk}(z) = k - j \leq k + j + 1$. It is immediately clear from (4.5), (4.6) that $\{P_{0k}(z)\}_{k=0}^{\infty}$ are the Maclaurin polynomials for e^z and therefore satisfy condition (5c). Furthermore, Ehle [11], in his study of the use of the Padé approximants in the construction of A -stable integration schemes for stiff systems of ordinary differential equations, has shown that for $z \in \{z \in C: \operatorname{Re} z \leq 0\}$

$$(4.7) \quad |P_{jk}(z)| \leq 1, \quad j = k, k+1, k+2, \quad k = 0, 1, 2, \dots$$

Consequently, from the standpoint of the constraint that condition 5 of Theorem 3.1 be satisfied, $C(z)$ can be chosen from among the entries in the top row, the principal diagonal and the first two subdiagonals of the Padé table. However, the convergence of approximation schemes constructed using these rational functions and the AVE or spline based state approximations defined above is guaranteed by Theorem 3.3 only if the uniform boundedness of the operators $\{P_{jk}((r/N)\mathcal{A}_N)^l\}_{l=0}^{\infty}$ can be demonstrated.

Using the von Neumann theory of spectral sets [22] and a result due to Hersh and Kato [14], the following result can be obtained.

THEOREM 4.1. *Let T be a bounded linear operator on a Hilbert space H for which there exists a $\beta > 0$ such that $\langle Tx, x \rangle \leq \beta \langle x, x \rangle$ for all $x \in H$, and let $r(z)$ be a rational function satisfying condition (5) of Theorem 3.1. Then if $|r(z)| \leq 1$ for all $z \in \{z \in C: \operatorname{Re} z \leq 0\}$ we have*

$$|r(hT)| \leq 1 + \beta Kh$$

where K is a positive constant independent of h and T .

It follows immediately from the dissipative properties of the operators \mathcal{A}_N defined as a part of the AVE state approximation, (4.1), (4.7) and Theorem 4.1 that for $j = k, k+1, k+2, k = 1, 2, \dots$ and $l = 0, 1, 2, \dots, \rho_N$

$$\left| P_{jk} \left(\frac{r}{N} \mathcal{A}_N \right)^l \right| \leq M \left(1 + \beta K_{jk} \frac{r}{N} \right)^l \leq M e^{\beta K_{jk} \rho_N (r/N)} M e^{\beta K_{jk} T}.$$

Similarly, for the spline based state approximations, it follows from (4.3) that

$$\left| P_{jk} \left(\frac{r}{N} \mathcal{A}_N \right)^l \right| \leq \sqrt{\nu} e^{\omega K_{jk} T}.$$

In addition, for the AVE state approximation which satisfies (4.2), it can be shown independently of Theorem 4.1 that for $k = 1, 2, \dots$ and $l = 0, 1, 2, \dots, \rho_N$

$$\left| P_{0k} \left(\frac{r}{N} \mathcal{A}_N \right)^l \right| \leq M e^{\alpha T}.$$

Although, as far as the convergence of the approximation scheme is concerned, it would suffice to choose $D(z) \equiv 1$, and hence $D(\lambda(r/N)\mathcal{A}_N) = I$ (see Theorem 3.2), empirical evidence can be given, and an intuitive argument can be made for choosing $D(z)$ as a rational function approximation to the exponential. It is easily verified that any rational function approximation to the exponential which is a suitable choice for $C(z)$ is a suitable choice for $D(z)$ as well. In addition, for the spline based state approximations and $k = 1, 2, \dots$ it can be shown that $|P_{0k}(\lambda(r/N)\mathcal{A}_N)P_{Nz} - z| \rightarrow 0$ as $N \rightarrow \infty$ for each $z \in Z$. A more detailed description of the role played by the rational function $D(z)$ and its effect upon the overall performance of the approximation scheme can be found in [25].

The results in this section are summarized in the following theorem.

THEOREM 4.2. *For $\{Z_N, P_N, \mathcal{A}_N, C(z), D(z)\}$ an approximation scheme for the initial value problem (2.1), (2.2), the hypotheses and conditions of Theorem 3.3 are satisfied if*

- (1) Z_N, P_N, \mathcal{A}_N is an AVE state approximation and $C(z), D(z) \in \mathcal{D}_p \cup \mathcal{M}_p$ or
- (2) Z_N, P_N, \mathcal{A}_N is a spline based state approximation, $C(z) \in \mathcal{D}_p$ and $D(z) \in \mathcal{D}_p \cup \mathcal{M}_p$, where $\mathcal{D}_p = \{P_{jk}(z)\}$, $j = k, k+1, k+2, k = 1, 2, \dots$ and $\mathcal{M}_p = \{P_{0k}(z)\}$, $k = 1, 2, \dots$.

5. Application to optimal control problems. In this section we consider the application of the approximation results discussed above to the solution of optimal control problems in which the state is governed by a nonlinear hereditary system of the form (2.1). In particular let $\phi_1: R^n \rightarrow R^1$, $\phi_2: L_2^n(0, T) \rightarrow R^1$ be continuous and let $\phi_3: L_2^m(0, T) \rightarrow R^1$ be continuous and convex. Let U be a closed convex subset of $L_2^m(0, T)$ and define problem (P) as follows:

- (P) Minimize $\Phi(u) = \phi_1(x(T; \eta, \phi, u)) + \phi_2(x(\cdot; \eta, \phi, u)) + \phi_3(u)$
over all $u \in U$ where $x(\cdot; \eta, \phi, u)$ denotes the unique solution to (2.1), (2.2) corresponding to $u \in U$.

The approach we take is to consider a sequence of approximating optimal control problems $\{(P_N)\}$, in each of which the governing state equation is a finite dimensional discrete difference equation constructed in accordance with the approximation framework developed in § 3. Let $\{Z_N, P_N, \mathcal{A}_N, C(z), D(z)\}$ be an approximation scheme for (2.1), (2.2) which satisfies the hypotheses of Theorem 3.3 and for $z_0 = (\eta, \phi)$, $u \in L_2^m(0, T)$ and $k = 0, 1, 2, \dots, \rho_N$ let

$$z_k^N(z_0, u) = (x_k^N(z_0, u), y_k^N(z_0, u))$$

where $\{z_k^N(z_0, u)\}_{k=0}^{\rho_N}$ are given by (3.2) with $x_k^N(z_0, u) \in R^n$ and $y_k^N(z_0, u) \in L_2^n(0, T)$. Define $x^N \in L_2^n(0, T)$ by

$$x^N(\theta) = x^N(\theta; \eta, \phi, u) = \sum_{j=0}^{\rho_N} x_j^N(z_0, u) \chi_{[j(r/N), (j+1)r/N)}(\theta)$$

and for each $N = 1, 2, \dots$ let problem (P_N) be given by

- (P_N) Minimize $\Phi_N(u) = \phi_1(x_{\rho_N}^N(z_0, u)) + \phi_2(x^N(\cdot; \eta, \phi, u)) + \phi_3(u)$
over all $u \in U$.

Remark 5.1. While it is true that for each $N = 1, 2, \dots$ problem (P_N) is not fully discrete in that the minimization of Φ is being considered over a function space, it is in fact possible to define the problem in a form which is directly suitable for solution on the computer. Indeed, if we consider the minimization over the set $U_N \equiv$

$Q_N U \subset \times_0^{\rho_N-1} R^m$ where $Q_N: L_2^m(0, T) \rightarrow \times_0^{\rho_N-1} R^m$ is defined by

$$(Q_N u)_j = \frac{N}{r} \int_{j(r/N)}^{(j+1)r/N} u(\tau) d\tau, \quad j = 0, 1, 2, \dots, \rho^N - 1,$$

then by placing relatively minor restrictions on the choice of the set U , all of the convergence results for the solutions to the sequence of problems $\{(P_N)\}$ to be discussed below can be shown to hold for the fully discrete problems as well. In order to simplify the presentation, however, we shall restrict our attention to the approximating problems as given.

It is our ultimate goal to demonstrate that in some sense, solutions to problem (P_N) approximate solutions to problem (P) . However, before this can be accomplished, the existence of solutions to problems (P) and (P_N) must be considered. In order to insure the convexity of Φ and Φ_N with respect to u , it is necessary that we restrict f , the nonlinear part of the state equation, to being affine in the controls. Following Banks [2], henceforth we shall assume that $f: R^1 \times R^n \times L_2^n(-r, 0) \times R^m \rightarrow R^n$ is of the form

$$(5.1) \quad f(t, \eta, \phi, v) = f_1(t, \eta, \phi) + (f_2(t, \eta, \phi) + B(t))v$$

where B is continuous and $f_1: R^1 \times R^n \times L_2^n(-r, 0) \rightarrow R^n$ and $f_2: R^1 \times R^n \times L_2^n(-r, 0) \rightarrow R^{n \times m}$ satisfy the following hypotheses:

(1) The mappings $(t, \eta, \phi) \rightarrow f_i(t, \eta, \phi)$, $i = 1, 2$, are continuous on $R^1 \times R^n \times L_2^n(-r, 0)$.

(2) For any bounded subset \mathcal{D} of $R^n \times L_2^n(-r, 0)$ there exist $m_i = m_i(\mathcal{D})$, $m_i \in L_\infty^{\text{loc}}$, $i = 1, 2$, such that for $t \in R^1$ and $(\eta, \phi), (\xi, \psi) \in \mathcal{D}$ one has

$$|f_i(t, \eta, \phi) - f_i(t, \xi, \psi)| \leq m_i(t) \{|\eta - \xi| + |\phi - \psi|\}.$$

(3) For $i = 1, 2$, $f_i(t, 0, 0) = 0$ and there exist functions $\hat{m}_i \in L_\infty^{\text{loc}}$ such that for $t \in R^1$

$$|f_i(t, \eta, \phi)| \leq \hat{m}_i(t) \{|\eta| + |\phi|\}$$

for $(\eta, \phi) \in R^n \times L_2^n(-r, 0)$ with $|\eta| + |\phi|$ sufficiently large.

It is immediately clear that any function f of the form (5.1) satisfying (1)–(3) above will also satisfy hypotheses (H1)–(H4).

In addition, it is necessary that we make either one or the other of the following two assumptions:

(A1) The set U is bounded.

(A2) The mappings ϕ_i , $i = 1, 2, 3$, satisfy

- (i) $\phi_i \geq 0$, $i = 1, 2$,
- (ii) $\phi_3(u) \rightarrow \infty$ if $|u| \rightarrow \infty$.

We note that problem (P) is most commonly stated with $U = L_2^m(0, T)$ and Φ a quadratic of the form

$$(5.2) \quad \begin{aligned} \Phi(u) = & x(t; \eta, \phi, u)^T G x(T; \eta, \phi, u) \\ & + \int_0^T x(s; \eta, \phi, u)^T Q x(s; \eta, \phi, u) ds + \int_0^T u(s)^T R u(s) ds \end{aligned}$$

where G and Q are positive semidefinite $n \times n$ matrices and R is a positive definite $m \times m$ matrix. In this case, assumption (A2) holds.

LEMMA 5.1. For f of the form (5.1) satisfying hypotheses (1)–(3), $z_0 \in Z$ and $u_l \rightarrow u$ weakly in $L_2^m(0, T)$ we have

$$(5.3) \quad |z(t; z_0, u_l) - z(t; z_0, u)| \rightarrow 0$$

as $l \rightarrow \infty$ uniformly in t for $t \in [0, T]$, and for $N = 1, 2, \dots$ fixed we have

$$(5.4) \quad |z_k^N(z_0, u_l) - z_k^N(z_0, u)| \rightarrow 0$$

as $l \rightarrow \infty$ uniformly in k for $k = 0, 1, 2, \dots, \rho_N$ where $z(t; z_0, u)$ and $z_k^N(z_0, u)$ are given by (2.4) and (3.2), respectively.

The proof of (5.3) follows from [2, Thm. 3.2] while similar arguments and Lemma 3.1 can be used to verify (5.4).

COROLLARY 5.1. *Under the hypotheses of Lemma 5.1, if $\{u_N\}$ is a sequence in $L_2^m(0, T)$ for which $u_N \rightarrow u$ weakly, then*

$$|z_k^N(z_0, u_N) - z(t_k^N; z_0, u)| \rightarrow 0$$

as $N \rightarrow \infty$ uniformly in k , $k = 0, 1, 2, \dots, \rho_N$.

Proof. Since

$$(5.5) \quad \begin{aligned} & |z_k^N(z_0, u_N) - z(t_k^N; z_0, u)| \\ &= |z_k^N(z_0, u_N) - z(t_k^N; z_0, u_N)| + |z(t_k^N; z_0, u_N) - z(t_k^N; z_0, u)| \end{aligned}$$

and $u_N \rightarrow u$ weakly implies that $\{u_N\}$ lies in a bounded subset of $L_2^m(0, T)$, the first term on the right-hand side of (5.5) tends toward zero as $N \rightarrow \infty$ uniformly in k , $k = 0, 1, 2, \dots, \rho_N$, as a consequence of Theorem 3.3, while Lemma 5.1 insures that the second term tends toward zero in the stated manner as well.

THEOREM 5.1. *If either assumption (A1) or (A2) holds and if f is of the form (5.1) satisfying hypotheses (1)–(3), then problems (P) and (P_N) have solutions.*

Proof. Lemma 5.1, ϕ_i continuous, $i = 1, 2, 3$ and ϕ_3 convex imply that Φ and Φ_N are weakly semicontinuous from below. Therefore, if U is bounded, Φ and Φ_N will assume their infimum on U (see [19, Existence Thm, p. 90]), and the theorem is proven.

On the other hand, suppose assumption (A2) holds, and let $\{u_i\} \in U$ be such that

$$\Phi(u_i) \rightarrow \alpha = \inf \{\Phi(u) : u \in U\}.$$

Note that $\phi_i \geq 0$, $i = 1, 2, 3$, implies that $0 \leq \alpha < \infty$. Since U is closed and convex (and therefore weakly sequentially closed) and $\{u_i\}$ is bounded (assumption (A2)), $\{u_i\}$ must contain a weakly convergent subsequence $\{u_{i_j}\}$, $u_{i_j} \rightarrow \bar{u} \in U$, weakly. However, Φ weakly semicontinuous from below implies that

$$\alpha \leq \Phi(\bar{u}) \leq \liminf \Phi(u_{i_j}) = \alpha$$

and hence $\Phi(\bar{u}) = \alpha$, and \bar{u} is a solution to problem (P). A similar argument may be used to demonstrate the existence of a solution $\bar{u}_N \in U$ to problem (P_N) .

THEOREM 5.2. *Suppose that the hypotheses of Theorem 5.1 hold and for each $N = 1, 2, \dots$, \bar{u}_N denotes a solution to problem (P_N) . Then $\{\bar{u}_N\}$ contains a subsequence $\{\bar{u}_{N_k}\}$ for which $\bar{u}_{N_k} \rightarrow \bar{u} \in U$ weakly. Moreover, \bar{u} is a solution to problem (P) and $\Phi_{N_k}(\bar{u}_{N_k}) \rightarrow \Phi(\bar{u})$ as $k \rightarrow \infty$.*

Proof. Under either assumption (A1) or (A2), the sequence $\{\bar{u}_N\}$ is bounded. It therefore must contain a weakly convergent subsequence $\{\bar{u}_{N_k}\}$. If $\bar{u} \in U$ is such that $\bar{u}_{N_k} \rightarrow \bar{u}$ weakly as $k \rightarrow \infty$, then Corollary 3.1, Corollary 5.1 and the weak semicontinuity from below of ϕ_3 (it being continuous and convex) imply that

$$\begin{aligned} \Phi(\bar{u}) &= \phi_1(x(T; \eta, \phi, \bar{u})) + \phi_2(x(\cdot; \eta, \phi, \bar{u})) + \phi_3(\bar{u}) \\ &\leq \lim_{k \rightarrow \infty} \phi_1(x_{\rho_N}^N((\eta, \phi), \bar{u}_{N_k})) + \lim_{k \rightarrow \infty} \phi_2(x^N(\cdot; \eta, \phi, \bar{u}_{N_k})) + \liminf_{k \rightarrow \infty} \phi_3(\bar{u}_{N_k}) \\ &= \lim_{k \rightarrow \infty} \inf \Phi_{N_k}(\bar{u}_{N_k}) \leq \lim_{k \rightarrow \infty} \sup \Phi_{N_k}(\bar{u}_{N_k}) \\ &\leq \lim_{k \rightarrow \infty} \sup \Phi_{N_k}(u) = \lim_{k \rightarrow \infty} \Phi_{N_k}(u) = \Phi(u) \end{aligned}$$

for arbitrary $u \in U$, and hence that \bar{u} is a solution to problem (P). The fact that $\Phi_{N_k}(\bar{u}_{N_k}) \rightarrow \Phi(\bar{u})$ as $k \rightarrow \infty$ follows from

$$\Phi(\bar{u}) \leq \liminf_{k \rightarrow \infty} \Phi_{N_k}(\bar{u}_{N_k}) \leq \limsup_{k \rightarrow \infty} \Phi_{N_k}(\bar{u}_{N_k})$$

$$\leq \limsup_{k \rightarrow \infty} \Phi_{N_k}(\bar{u}) = \lim_{k \rightarrow \infty} \Phi_{N_k}(\bar{u}) = \Phi(\bar{u}).$$

Remark 5.2. Since it is difficult to determine the convexity properties of the functional Φ , it is not possible to say anything about the uniqueness of solutions to problem (P). However, if in fact problem (P) has a unique solution, then the sequence itself, $\{\bar{u}_N\}$ will converge to \bar{u} weakly as $N \rightarrow \infty$.

Remark 5.3. If Φ is of the form (5.2), then it is possible to show that $|\bar{u}_{N_k}| \rightarrow |\bar{u}|$ as well, and hence that $\bar{u}_{N_k} \rightarrow \bar{u}$ strongly as $k \rightarrow \infty$. Once again if problem (P) admits a unique solution \bar{u} , then $\bar{u}_N \rightarrow \bar{u}$ strongly as $N \rightarrow \infty$.

6. Numerical results. In this section we present numerical results obtained through the implementation of the approximation schemes described above. The schemes employed have been constructed using the AVE and spline based (SPL) state approximations together with the Padé rational function approximations to the exponential. In all of the examples below, however, we have chosen $C(z) = D(z) = P_{22}(z)$ and $\lambda = \frac{1}{2}$. The effect of varying the choice of the rational function components of the approximation scheme (from among those in the Padé table for which the hypotheses of Theorem 3.3 are satisfied) was studied extensively in [25].

We have included one example involving the integration of an initial value problem of the form (2.1), (2.2) only, and three other examples which involve the solution of an optimal control problem of the form given by problem (P) in §5. We have deliberately chosen to include examples which have been used by other authors to test other approximation schemes for the integration of FDE and the solution of FDE control problems so that our methods can be compared to theirs. The other places where each example has appeared have been noted.

All programming was done in FORTRAN and implemented on the Digital Equipment Corporation DEC system 10 computer at Bowdoin College. The optimization in each of the approximating problems (P_N) was carried out using the IMSL [15] routine ZXMIN, an iterative quasi-Newton algorithm for finding the minimum of a scalar valued function of several variables. The discretization of the admissible control space U in the approximating optimal control problems (P_N) was done in two different ways. One involved the use of the space $U_N = \times_{0^N}^{\rho_N} R^m$ as an approximation to $L_2^m(0, T)$ (see Remark 5.1). In this case the number of parameters over which the minimization takes place increases with the degree of approximation N . The second approach was to minimize over the space $\tilde{U} = \times_0^L R^m$ where L is a fixed constant independent of N . A cubic spline interpolation scheme was then used to obtain the values of the control which are required to evaluate (3.2). The approximate solutions resulting from the two methods were virtually indistinguishable. However, the number of iterations required to obtain the minimizing control increased like $O(N)$ for the first method, while the iteration count remained essentially constant for all values of N for the second method.

Since, with the exception of Example 6.2 which has a linear state equation, it is impossible to obtain exact solutions to the optimal control problems below, we have included approximate solutions which were obtained using methods independent from our own. These alternate approximate solutions, which can be used for comparison, were computed by Daniel [10] using a fourth order integration scheme for FDE

developed by Tavernini [27] to solve the mixed retarded/advanced two-point boundary value problem which results from the application of the necessary conditions for optimality (see [6]) to problem (P).

Example 6.1 (Banks [2, Example 4.1]). We consider the integration of the equation

$$\dot{x}(t) = -1.5x(t) - 1.25x(t-1) + x(t) \sin x(t)$$

on the interval $0 \leq t \leq 5$ with initial data

$$x(0) = 1, \quad x_0(s) = 10s + 1, \quad -1 \leq s \leq 0.$$

The approximate solutions generated by the AVE and SPL state approximations are given in Tables 6.1 and 6.2, respectively. The values in the last column of each of the tables were computed using the method of steps [12] together with a fourth order Runge-Kutta routine for ordinary differential equations, and may be used for comparison purposes.

TABLE 6.1

t	$x_4^{\text{AVE}}(t)$	$x_8^{\text{AVE}}(t)$	$x_{16}^{\text{AVE}}(t)$	$x_{32}^{\text{AVE}}(t)$	$x(t)$
0.0	1.0	1.0	1.0	1.0	1.0
.5	3.0954	3.1924	3.2531	3.2840	3.3142
1.0	2.1375	2.2051	2.2522	2.2841	2.3317
1.5	.9759	.7151	.5163	.3877	.2294
2.0	-.2258	-.6233	-.8116	-.9020	-.9909
2.5	-.5984	-.6920	-.7221	-.7331	-.7399
3.0	-.3491	-.2599	-.1715	-.1073	-.0245
3.5	-.0573	.1091	.2409	.3251	.4259
4.0	.1024	.2389	.3244	.3711	.4195
4.5	1.2229	.1598	.1532	.1370	.1081
5.0	.0634	.0150	-.0469	-.0919	-.1480

TABLE 6.2

t	$x_4^{\text{SPL}}(t)$	$x_8^{\text{SPL}}(t)$	$x_{16}^{\text{SPL}}(t)$	$x_{32}^{\text{SPL}}(t)$	$x(t)$
0.0	1.0038	1.0010	1.0003	1.0001	1.0
.5	3.5036	3.3623	3.3344	3.3236	3.3142
1.0	2.1694	2.2636	2.2992	2.3157	2.3317
1.5	.3642	.2834	.2538	.2405	.2294
2.0	-1.0308	-.9972	-.9929	-.9919	-.9909
2.5	-.7248	-.7332	-.7345	-.7367	-.7399
3.0	-.0612	-.0218	-.0188	-.0205	-.0245
3.5	.4055	.4166	.4230	.4251	.4259
4.0	.4180	.4145	.4157	.4173	.4195
4.5	.1572	.1187	.1099	.1081	.1081
5.0	-.1124	-.1391	-.1454	-.1473	-.1480

Example 6.2 (Banks, Burns and Cliff [5, Example C7], Rockey [23, Test Problem 5.6]). In this example we consider an optimal control problem whose state equation is a linear harmonic oscillator with delayed damping:

$$\text{Minimize} \quad \Phi(u) = 5y(2)^2 + \frac{1}{2} \int_0^2 u(s)^2 ds$$

over $u \in U = L_2^1(0, 2)$ subject to

$$(6.1) \quad \ddot{y}(t) + \dot{y}(t-1) + y(t) = u(t)$$

with initial conditions

$$(6.2) \quad y(0) = 10, \quad y_0(s) = 10, \quad -1 \leq s \leq 0,$$

$$(6.3) \quad \dot{y}(0) = 0, \quad \dot{y}_0(s) = 0, \quad -1 \leq s \leq 0.$$

For this problem, the true optimal control \bar{u} may be computed, and is given by

$$\bar{u}(t) = \begin{cases} \delta \sin(2-t) + \frac{\delta}{2}(1-t) \sin(t-1), & 0 \leq t \leq 1, \\ \delta \sin(2-t), & 1 \leq t \leq 2, \end{cases}$$

where $\delta \approx 2.5599$, with $\Phi(\bar{u}) = 3.3991$. This example may be put in the form of problem (P) by transforming (6.1), (6.2), (6.3) into an equivalent first order system, which is given by

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} x(t-1) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t),$$

$$x(0) = \begin{bmatrix} 10 \\ 0 \end{bmatrix}, \quad x_0(s) = \begin{bmatrix} 10 \\ 0 \end{bmatrix}, \quad -1 \leq s \leq 0,$$

where

$$x(t) = \begin{bmatrix} y(t) \\ \dot{y}(t) \end{bmatrix}.$$

The payoff functional Φ would now take the form

$$\Phi(u) = x^T(2) \begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix} x(2) + \int_0^2 u(s)^2 ds.$$

Tables 6.3 and 6.4 contain the resulting approximating optimal controls.

TABLE 6.3

t	$\bar{u}_4^{\text{AVE}}(t)$	$\bar{u}_8^{\text{AVE}}(t)$	$\bar{u}_{16}^{\text{AVE}}(t)$	$\bar{u}_{32}^{\text{AVE}}(t)$	$\bar{u}(t)$
0.0	1.2757	1.2797	1.2746	1.2336	1.2506
.25	1.4515	1.6358	1.7463	1.8024	1.8645
.50	1.7195	1.9506	2.0888	2.1706	2.2467
.75	1.8076	2.0427	2.1827	2.2642	2.3501
1.00	1.7070	1.9094	2.0263	2.0911	2.1541
1.25	1.4333	1.5844	1.6641	1.7075	1.7449
1.50	1.0255	1.1216	1.1718	1.2018	1.2273
1.75	.5324	.5794	.6043	.6164	.6333
2.0	.2708	.1473	.0776	.0337	0.0000
$\Phi_N(\bar{u}_N)$	2.1705	2.6781	3.0054	3.1931	3.3991

TABLE 6.4

t	$\bar{u}_4^{\text{SPL}}(t)$	$\bar{u}_8^{\text{SPL}}(t)$	$\bar{u}_{16}^{\text{SPL}}(t)$	$\bar{u}_{32}^{\text{SPL}}(t)$	$\bar{u}(t)$
0.0	1.6887	1.4468	1.3456	1.2776	1.2506
.25	1.9415	1.8856	1.8748	1.8686	1.8645
.50	2.3024	2.2635	2.2553	2.2501	2.2467
.75	2.3675	2.3570	2.3583	2.3521	2.3501
1.00	2.1634	2.1573	2.1482	2.1539	2.1541
1.25	1.7592	1.7489	1.7458	1.7443	1.7449
1.50	1.2238	1.2269	1.2261	1.2276	1.2273
1.75	.61496	.6301	.6285	.6339	.6333
2.0	.2999	.1548	.0798	.0392	0.000
$\Phi_N(\bar{u}_N)$	3.5664	3.4438	3.411	3.4021	3.3991

Example 6.3 (Banks [2, Example 4.4], Banks, Burns and Cliff [5, Example C11], Daniel [10, Example 4.5], Rockey [23, Test Problem 5.10]). In this example we consider an optimal control problem with a one-dimensional nonlinear state equation

$$\text{Minimize} \quad \Phi(u) = \frac{1}{2}x(2)^2 + \frac{1}{2} \int_0^2 x(s)^2 + u(s)^2 ds$$

over $u \in L_2^1(0, 2)$ subject to

$$\dot{x}(t) = x(t-1) + x(t) \sin x(t) + u(t)$$

with initial conditions given by

$$x(0) = 0, \quad x_0(s) = \begin{cases} -10s, & -\frac{1}{2} \leq s \leq 0, \\ 10(s+1), & -1 \leq s \leq -\frac{1}{2}. \end{cases}$$

The approximating minimizing controls for the AVE and SPL state approximations are given in Tables 6.5 and 6.6 respectively, while Tables 6.7 and 6.8 contain the corresponding optimal trajectories.

TABLE 6.5

t	$\bar{u}_4^{\text{AVE}}(t)$	$\bar{u}_8^{\text{AVE}}(t)$	$\bar{u}_{16}^{\text{AVE}}(t)$	$\bar{u}_{32}^{\text{AVE}}(t)$	$\bar{u}(t)$
0.0	-2.1967	-2.2417	-2.2681	-2.2817	-2.3028
.25	-2.0860	-2.1699	-2.2295	-2.2662	-2.3164
.50	-1.8082	-1.9655	-2.0971	-2.1893	-2.3189
.75	-1.4605	-1.5635	-1.6386	-1.6853	-1.7470
1.00	-1.1242	-1.1470	-1.1443	-1.1333	-1.1031
1.25	-.8467	-.8273	-.7972	-.7747	-.7483
1.50	-.6332	-.6072	-.5838	-.5708	-.5619
1.75	-.4665	-.4484	-.4376	-.4349	-.4440
2.0	-.3921	-.3477	-.3282	-.3223	-.3230
$\Phi_N(\bar{u}_N)$	1.9914	2.1673	2.3020	2.3953	2.5334

TABLE 6.6

t	$\bar{u}_4^{\text{SPL}}(t)$	$\bar{u}_8^{\text{SPL}}(t)$	$\bar{u}_{16}^{\text{SPL}}(t)$	$\bar{u}_{32}^{\text{SPL}}(t)$	$\bar{u}(t)$
0.0	-2.2389	-2.2372	-2.2596	-2.2741	-2.3028
.25	-2.3139	-2.3019	-2.3023	-2.3041	-2.3164
.50	-2.1999	-2.2596	-2.2908	-2.3022	-2.3189
.75	-1.6929	-1.7129	-1.7295	-1.7364	-1.7470
1.00	-1.1384	-1.1198	-1.1117	-1.1070	-1.1031
1.25	-.7830	-.7702	-.7610	-.7569	-.7483
1.50	-.6151	-.5874	-.5749	-.5682	-.5619
1.75	-.5032	-.4643	-.4531	-.4469	-.4440
2.0	-.4417	-.3676	-.3461	-.3351	-.3230
$\Phi_N(\bar{u}_N)$	2.5119	2.4996	2.5103	2.5133	2.5334

TABLE 6.7

t	$\bar{x}_4^{\text{AVE}}(t)$	$\bar{x}_8^{\text{AVE}}(t)$	$\bar{x}_{16}^{\text{AVE}}(t)$	$\bar{x}_{32}^{\text{AVE}}(t)$	$\bar{x}(t)$
0.0	0.0	0.0	0.0	0.0	0.0
.25	-.0087	-.1034	-.1672	-.2051	-.2473
.50	.1537	.1434	.1357	.1271	.1078
.75	.2757	.3540	.4329	.4931	.5663
1.00	.3282	.4199	.5009	.5562	.6186
1.25	.3368	.3901	.4182	.4259	.4127
1.50	.3314	.3393	.3233	.3006	.2474
1.75	.3337	.3139	.2886	.2687	.2272
2.00	.3486	.3264	.3165	.3159	.3053

TABLE 6.8

t	$\bar{x}_4^{\text{SPL}}(t)$	$\bar{x}_8^{\text{SPL}}(t)$	$\bar{x}_{16}^{\text{SPL}}(t)$	$\bar{x}_{32}^{\text{SPL}}(t)$	$\bar{x}(t)$
0.0	-.0034	-.0010	-.0003	-.0001	0.0
.25	-.2538	-.2415	-.2425	-.2440	-.2473
.50	.1721	.1259	.1155	.1136	.1078
.75	.6048	.5994	.5749	.5737	.5663
1.00	.6607	.6234	.6257	.6278	.6186
1.25	.4115	.4222	.4222	.4231	.4127
1.50	.2005	.2599	.2665	.2629	.2474
1.75	.2398	.2617	.2494	.2456	.2272
2.0	.4021	.3485	.3359	.3282	.3053

Example 6.4 (Daniel [10, Example 4.2]). In this example we consider an inertial control problem (see [9])

$$\text{Minimize} \quad \Phi(u) = \frac{1}{2}y(2)^2 + \frac{1}{2} \int_0^2 \dot{u}(s)^2 ds$$

over $u \in U = \{u \in H_1^1(0, 2) : u(0) = 0\}$ subject to

$$\dot{y}(t) = y(t-1) + \frac{1}{2}t^2 \sin y(t) + u(t)$$

with initial conditions

$$y(0) = 1, \quad y_0(s) = 1, \quad -1 \leq s \leq 0.$$

Although this example is not in the form of problem (P), it can be transformed into an equivalent optimal control problem to which the theory developed above applies. If we let

$$x(t) = \begin{bmatrix} y(t) \\ u(t) \end{bmatrix}, \quad v(t) = \dot{u}(t),$$

then the problem becomes

$$\text{Minimize} \quad \Phi(v) = x(2)^T \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{bmatrix} x(2) + \frac{1}{2} \int_0^2 v(s)^2 ds$$

over $v \in L_2^1(0, 2)$ subject to

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} x(t-1) + \begin{bmatrix} \frac{1}{2} t^2 \sin x_1(t) \\ v(t) \end{bmatrix}$$

with initial conditions

$$x(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad x_0(s) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad -1 \leq s \leq 0.$$

The approximating optimal controls $\bar{u}_N = (\bar{x}_N)_2$ are given in Tables 6.9 and 6.10, and the corresponding optimal trajectories $\bar{y}_N = (\bar{x}_N)_1$ are given in Tables 6.11 and 6.12.

TABLE 6.9

t	$\bar{u}_4^{\text{AVE}}(t)$	$\bar{u}_8^{\text{AVE}}(t)$	$\bar{u}_{16}^{\text{AVE}}(t)$	$\bar{u}_{32}^{\text{AVE}}(t)$	$\bar{u}(t)$
0.0	0.0	0.0	0.0	0.0	0.0
.25	-.6687	-.6836	-.6870	-.6880	-.6858
.50	-1.2308	-1.2329	-1.2332	-1.2336	-1.2291
.75	-1.6634	-1.6569	-1.6555	-1.6558	-1.6494
1.00	-1.9840	-1.9723	-1.9707	-1.9714	-1.9645
1.25	-2.2085	-2.1937	-2.1932	-2.1950	-2.1891
1.50	-2.3520	-2.3349	-2.3352	-2.3382	-2.3332
1.75	-2.4293	-2.4095	-2.4098	-2.4132	-2.4087
2.00	-2.4644	-2.4342	-2.4320	-2.4349	-2.4303
$\Phi_N(\bar{u}_N)$	2.5484	2.4804	2.4617	2.4570	2.4649

TABLE 6.10

t	$\bar{u}_4^{\text{SPL}}(t)$	$\bar{u}_8^{\text{SPL}}(t)$	$\bar{u}_{16}^{\text{SPL}}(t)$	$\bar{u}_{32}^{\text{SPL}}(t)$	$\bar{u}(t)$
0.0	0.0	0.0	0.0	0.0	0.0
.25	-.6491	-.6799	-.6860	-.6877	-.6858
.50	-1.2161	-1.2282	-1.2319	-1.2331	-1.2291
.75	-1.6506	-1.6524	-1.6539	-1.6553	-1.6494
1.00	-1.9727	-1.9692	-1.9699	-1.9716	-1.9645
1.25	-2.2070	-2.1957	-2.1950	-2.1966	-2.1891
1.50	-2.3639	-2.3436	-2.3404	-2.3411	-2.3332
1.75	-2.4483	-2.4207	-2.4157	-2.4167	-2.4087
2.00	-2.4816	-2.4447	-2.4382	-2.4388	-2.4303
$\Phi_N(\bar{u}_N)$	2.5826	2.4986	2.4719	2.4624	2.4649

TABLE 6.11

t	$\bar{y}_4^{\text{AVE}}(t)$	$\bar{y}_8^{\text{AVE}}(t)$	$\bar{y}_{16}^{\text{AVE}}(t)$	$\bar{y}_{32}^{\text{AVE}}(t)$	$\bar{y}(t)$
0.0	1.0000	1.0000	1.0000	1.0000	1.0000
.25	1.1698	1.1654	1.1638	1.1633	1.1636
.50	1.2019	1.1911	1.1881	1.1873	1.1889
.75	1.1454	1.1273	1.1213	1.1193	1.1228
1.00	1.0462	1.0208	1.0095	1.0034	1.0041
1.25	.9371	.9066	.8919	.8835	.8862
1.50	.8376	.8031	.7878	.7805	.7927
1.75	.7590	.7178	.7008	.6932	.7167
2.00	.7103	.6561	.6315	.6195	.6564

TABLE 6.12

t	$\bar{y}_4^{\text{SPL}}(t)$	$\bar{y}_8^{\text{SPL}}(t)$	$\bar{y}_{16}^{\text{SPL}}(t)$	$\bar{y}_{32}^{\text{SPL}}(t)$	$\bar{y}(t)$
0.0	1.0000	1.0000	1.0000	1.0000	1.0000
.25	1.1703	1.1662	1.1639	1.1633	1.1636
.50	1.2102	1.1932	1.1887	1.1875	1.1889
.75	1.1500	1.1279	1.1214	1.1195	1.1228
1.00	1.0418	1.0111	1.0017	.9984	1.0041
1.25	.9340	.8944	.8824	.8775	.8862
1.50	.8507	.8055	.7875	.7799	.7927
1.75	.7949	.7332	.7084	.6968	.7162
2.00	.7674	.6820	.6440	.6252	.6564

Based upon the examples presented here, and several others which we have looked at, the following observations can be made:

(1) The schemes which we have proposed represent feasible and relatively efficient approximation methods for solving certain classes of nonlinear hereditary control problems.

(2) Since the resulting approximating problems are governed by discrete difference equations, the programming required is relatively simple. Moreover, since no additional discretization is necessary when the schemes are implemented on the computer, no further stability analysis is required in order to guarantee convergence of the approximating solutions.

(3) The spline based schemes, although somewhat more difficult to program and costlier to run, outperform the averaging schemes. However, the difference appears to be more pronounced in the case of simple integration of initial value problems as opposed to the solution of optimal control problems.

(4) The accuracy of the approximating controls and trajectories is quite good even for relatively small values of N . This is especially true for the schemes employing the spline based state approximation.

(5) Our results are comparable to those obtained by Rockey [23] and to those obtained via the semidiscrete schemes developed by Banks [2], [5] and Daniel [10].

We have also applied our schemes to the design of an open loop controller for the mach number-guide vane angle control loop of the National Transonic Wind Tunnel Facility (NTF) at the NASA Langley Research Center in Hampton, Virginia (see [1], [10]). Although the operation of the NTF is best described by a complex system of nonlinear partial differential equations, the dynamics of the system near

steady state operating conditions can be modeled by a linear hereditary system in which either the guide vane angle actuator or the guide vane angle actuator rate acts as a control. If we assume that a disturbance has occurred at time $t = 0$, the problem is to choose the control so as to drive the system back to equilibrium as quickly as possible without exceeding the physical limitations of the components of the system. This leads to a linear quadratic optimal control problem in which the dynamics are governed by a linear FDE of the form (2.1) with $f(t, \eta, \phi, u) = Bu$. While an approximation to the closed loop solution to this problem (in the form of approximating feedback gains matrices) would be more desirable (and is accessible through the techniques discussed in [7], [13] and [20]), we have generated approximating open loop solutions using the schemes developed above. This permitted us to test our methods on systems of higher dimension ($n = 3$ and 4) with the optimization being carried out over an extended time interval ($T = 30$). Both the averaging and spline based state approximations were employed with values of N as large as 24. We compared our results to the open loop solutions to this problem which appear in [10] and to the open loop form of the closed loop solutions computed in [1] and [7]. Our schemes performed comparably, both qualitatively and quantitatively, and provided acceptable approximating solutions for all values of $N \geq 4$.

Acknowledgments. The author would like to thank Professor H. T. Banks for his valuable comments and Professors P. L. Daniel and J. Crowley for their assistance in the preparation of the numerical results presented in § 6.

REFERENCES

- [1] E. S. ARMSTRONG AND J. S. TRIPP, *An application of multivariable design techniques to the control of the National Transonic Facility*, NASA Technical Paper 1887, Langley Research Center, Hampton, VA, August 1981.
- [2] H. T. BANKS, *Approximation of nonlinear functional differential equation control systems*, J. Optim. Theory Appl., 29 (1979), pp. 383–408.
- [3] H. T. BANKS AND J. A. BURNS, *An abstract framework for approximate solutions to optimal control problems governed by hereditary systems*, in Proc. International Conference on Differential Equations (Univ. of Southern California, Sept. 1974), H. A. Antosiewicz, ed., Academic Press, New York, 1975, pp. 10–25.
- [4] ———, *Hereditary control problems: Numerical methods based on averaging approximations*, this Journal, 16 (1978), pp. 169–208.
- [5] H. T. BANKS, J. A. BURNS AND E. M. CLIFF, *A comparison of numerical methods for identification and optimization problems involving control systems with delays*, LCDS Technical Report 79-7, Providence, RI, 1979.
- [6] H. T. BANKS, J. A. BURNS, E. M. CLIFF AND P. R. THRIFT, *Numerical solutions of hereditary control problems via an approximation technique*, LCDS Technical Report 75-6, Brown Univ., Providence, RI, 1975.
- [7] H. T. BANKS, K. ITO AND I. G. ROSEN, *A spline based technique for computing Riccati operators and feedback controls in regulator problems for delay equations*, ICASE Report 82-31, Institute for Computer Applications in Science and Engineering, Hampton, VA, 1982; SIAM J. Sci. Stat. Comput., 5 (1984), to appear.
- [8] H. T. BANKS AND F. KAPPEL, *Spline approximations for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496–522.
- [9] P. L. DANIEL, *Spline-based approximation methods for the identification and control of nonlinear functional differential equations*, Ph.D. Thesis, Brown Univ., Providence, RI, June, 1981.
- [10] ———, *Spline approximations for nonlinear hereditary control systems*, ICASE Report 82-10, Institute for Computer Applications in Science and Engineering, Hampton, VA, 1982.
- [11] R. L. EHLE, *A-stable methods and Padé approximations to the exponential*, SIAM J. Math. Anal., 4 (1973), pp. 671–680.

- [12] L. E. EL'SGOL'TS, *Introduction to the Theory of Differential Equations with Deviating Arguments*, Holden-Day, San Francisco, 1966.
- [13] J. S. GIBSON, *Linear quadratic optimal control of hereditary differential systems: Infinite dimensional Riccati equations and numerical approximations*, this Journal, 21 (1983), pp. 95–139.
- [14] R. HERSH AND T. KATO, *High-accuracy stable difference schemes for well-posed initial-value problems*, SIAM J. Numer. Anal., 16 (1979), pp. 670–682.
- [15] International Mathematical and Statistical Libraries, Library 1, ed. 6, ZXMIN-1-3, IMSL Inc., Houston, 1977.
- [16] F. KAPPEL, *An approximation scheme for delay equations*, in Proceedings of the International Conference on Nonlinear Phenomena in the Math Sciences (Univ. of Texas, Arlington, TX, June 16–20, 1980), Academic Press, New York, to appear.
- [17] F. KAPPEL AND W. SCHAPPACHER, *Autonomous nonlinear functional differential equations and averaging approximations*, Nonlinear Analysis Theory, Methods and Appl., 2 (1978), pp. 391–422.
- [18] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
- [19] W. KRABS, *Optimization and Approximation*, John Wiley, New York, 1979.
- [20] K. KUNISCH, *Approximation schemes for the linear quadratic optimal control problem associated with delay equations*, this Journal, 20 (1982), pp. 506–540.
- [21] D. C. REBER, *A finite difference technique for solving optimization problems governed by linear functional differential equations*, J. Differential Equations, 32 (1979), pp. 192–232.
- [22] F. REISZ AND B. SZ.-NAGY, *Functional Analysis*, Ungar, New York, 1975.
- [23] S. A. ROCKEY, *Discrete methods of state approximation, parameter identification and optimal control for hereditary systems*, Ph.D. thesis, Brown Univ., Providence, RI, June 1982.
- [24] I. G. ROSEN, *A discrete approximation framework for hereditary systems*, Ph.D. thesis, Brown Univ., Providence, RI, June 1980.
- [25] ———, *A discrete approximation framework for hereditary systems*, J. Differential Equations, 40 (1981), pp. 377–449.
- [26] ———, *Discrete approximation methods for parameter identification in delay systems*, this Journal, 22 (1984), pp. 95–120.
- [27] L. TAVERNINI, *One-step methods for the numerical solution of Volterra functional differential equations*, SIAM J. Numer. Anal., 8 (1971), pp. 786–795.
- [28] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

A COMPLETE OPTIMALITY CONDITION IN THE INVERSE PROBLEM OF OPTIMAL CONTROL*

TAKAO FUJII† AND MASARU NARAZAKI‡

Abstract. A complete optimality condition in the standard inverse problem of optimal control for the multi-input case is established. The optimality of a feedback control law is characterized completely in terms of a new geometric condition as well as the well-known return difference condition. The new condition not only provides a better insight into the well-known sensitivity reduction property of optimal control for the multi-input case, but also indicates an essential difference between the solutions of the single- and multi-input inverse problem.

Key words. linear system, optimal control, sensitivity reduction, linear matrix inequality, geometric approach

1. Introduction. This paper treats the most familiar type of inverse optimal control problem for a linear time-invariant system

$$(1.1) \quad \dot{x} = Ax + Bu, \quad x(0) = x_0,$$

and a quadratic cost of the form:

$$(1.2) \quad J = \int_0^\infty (x^T Q x + u^T u) dt.$$

Here x is an n -vector of states and u is an r -vector of piecewise-continuous controls; A , B and Q are real matrices of appropriate dimensions with Q symmetric nonnegative definite ($Q \geq 0$). As is well known, this problem is stated as follows: Given a linear constant feedback control law

$$(1.3) \quad u = -Kx,$$

which is assumed to be a stabilizing control (i.e., such that the corresponding trajectory of (1.1) is asymptotically stable):

(I) Find necessary and sufficient conditions on the matrices A , B and K such that the control law (1.3) minimizes a cost of the form (1.2) for some $Q \geq 0$.

(II) Determine all such costs (i.e., all such Q).

This problem was first posed by Kalman [12], who gave a complete solution to Problem (I) for the single-input case: that is the so-called "return difference condition", which provides a beautiful link between classical and modern control theory. This result was generalized later to the multi-input case by Anderson [1], [2], who obtained both a necessary condition and a sufficient condition (but not a necessary and sufficient condition) for optimality, which are expressed as follows.

C.1. (*Necessary condition*). $\Phi(i\omega) \geq 0$ for all real ω .

C.2. (*Sufficient condition*). $\Phi(i\omega) > 0$ for all real ω . Equivalently, $\Phi(s)$ satisfies C.1 as well as the rank condition:

$$\text{rank } \Phi(i\omega) = r \quad \text{for all real } \omega,$$

* Received by the editors October 23, 1981, and in revised form March 1, 1983.

† Department of Control Engineering, Faculty of Engineering Science, Osaka University, 1-1 Machikaneyama, Toyonaka, Osaka 560, Japan.

‡ Mitsubishi Precision Co., Ltd., 345 Kamimachi, Kamakura, Japan.

where

$$(1.4) \quad \Phi(s) = T(-s)^T T(s) - I,$$

$$(1.5) \quad T(s) = I + K(sI - A)^{-1}B \quad (\text{return difference matrix}).$$

It should be noted here that Condition C.2 is, of course, important from the system sensitivity aspect [1], but theoretically this is rather strong as a sufficient condition for optimality particularly in the multi-input case. This situation does not change even if we replace C.2 by just a little weaker sufficient condition.

C.3. $\Phi(i\omega) \geq 0$ for all real ω ; normal rank $\Phi(s) = r$.

(See Remark 4.1 for sufficiency of this condition.) In fact, there exists an essential gap between Conditions C.1 and C.3 in the multi-input case ($r > 1$) as will be shown in § 2. It is thus theoretically interesting to reduce this gap.

This paper is concerned with Problem (I) and provides a first significant complete solution in terms of Condition C.1 and an additional condition of geometric type (§ 4). A similar solution has been obtained also in the first version of this paper [10] together with some basic solutions to Problem (II), but it is just a technical one which is hard to interpret physically. On the contrary, the solution obtained here has a nice physical interpretation related to the sensitivity reduction of the optimal closed-loop system, which provides some additional insight into the inverse problem for the multi-input case. From this result, we also obtain a practically important observation that the necessary Condition C.1 itself is a sufficient condition for optimality in a *generic* sense, in addition to some useful sufficient conditions for optimality including Condition C.2.

The inverse problems of other types for the linear time-invariant case have been studied by several authors [11], [16]. First, in the “inverse problem of stable regulator” [16], admissible controls are restricted to *stabilizing* controls, and under this restriction, C.1 is shown to be a solution to Problem (I). On the other hand, in [11] they remove the nonnegative definiteness requirement on the weighting matrix Q , in addition to the preceding restriction on admissible controls. It should be noted that such modifications of the problem make the inverse problem fairly easy to solve. It is thus important to note that no simplifying assumptions at all are made in our paper on both admissible controls and the weighting matrix Q . In summary, this paper studies the original inverse problem of optimal control posed by Kalman and provides a complete optimality condition without any assumptions.

2. Review of optimal control and its inverse problems. Throughout the paper the pair (A, B) is assumed to be controllable as usual. In view of $Q \geq 0$, it will be convenient to express the cost (1.2) in the form

$$(2.1) \quad J = \int_0^\infty (x^T C^T C x + u^T u) dt$$

where C is an $m \times n$ real matrix of rank m ($m = \text{rank } Q$) such that $Q = C^T C$. Following [1], [12] we often refer to the feedback control law (1.3) simply as a “control law” K , and we say that the control law K is *stable* if all the eigenvalues of $A - BK$ have negative real parts, denoted by $\text{Re } \lambda(A - BK) < 0$, and *optimal* or *optimal for some* C (or $C^T C$) if it minimizes the cost (2.1) for some C . Moreover, the adjectives *maximal* (*minimal*) qualifying real symmetric matrices mean l.u.b. (g.l.b.) with respect to their usual partial ordering by nonnegative definiteness.

In this section, we state some useful results of optimal control and its inverse problem, which are pertinent to our development. The first one is a less known result in optimal control problem.

FACT 2.1. *The optimal control for the cost (2.1) always exists and is given in the form of state feedback control law (1.3) with the following properties.*

P.1. *The optimal control law K is uniquely given by*

$$(2.2) \quad K = B^T \bar{P},$$

where \bar{P} is the minimal nonnegative definite solution of the Riccati equation

$$(2.3) \quad PA + A^T P - PBB^T P + C^T C = 0,$$

and it always exists.

P.2. *The optimal control law K is stable, or equivalently, the minimal solution $\bar{P} \geq 0$ of (2.3) satisfies*

$$(2.4) \quad \operatorname{Re} \lambda(A - BB^T \bar{P}) < 0,$$

if and only if (C, A) is detectable; moreover, in this case \bar{P} is the unique nonnegative definite solution of (2.3).

Proof. The first part is well known (see e.g. [2]). Property P.1, except uniqueness of K , is due to Mårtensson [15]; for the uniqueness proof, see [19, Appendix A] and use the observable canonical form (see [19, p. 625]) of the triple (A, B, C) .

The “if” and the latter part of P.2 follows directly from [14, Thm. 3]. The “only if” part also follows from this by combining the *maximality* of a real symmetric solution \bar{P} satisfying (2.4) [20, Lemma 3] with its minimality just defined. This completes the proof.

The second result is the one about the optimality of a control law K for the *given* cost (2.1), which is a straightforward generalization of [1] where (C, A) was assumed to be *observable*.

FACT 2.2. *Let K be a control law for the system (1.1), and the pair (C, A) be detectable. Then the following statements are equivalent.*

(a) *K is optimal for the cost (2.1).*

(b) *There exists a solution $P \geq 0$ of (2.3) such that*

$$(2.5) \quad K = B^T P.$$

(c) *K is stable, and moreover,*

$$(2.6) \quad \Phi(s) = B^T (-sI - A^T)^{-1} C^T C (sI - A)^{-1} B.$$

Proof. (a) \Rightarrow (b). Immediate from P.1 of Fact 2.1. (b) \Rightarrow (c). By P.2 of Fact 2.1 detectability of (C, A) implies $P = \bar{P}$ and (2.4), so that stability of K follows from (2.5); (2.6) follows from (2.3) and (2.5) as in [1, p. 15]. (c) \Rightarrow (a). See the proof of [1, Thm. 6] or [3, § 6].

It is observed from the implication (a) \Rightarrow (c) that the optimal control law K for the cost (2.1) satisfies Condition C.1 in general,¹ but can satisfy Condition C.3 only when $m \geq r$. Consequently there exists an essential gap between these two conditions in the multi-input case ($r > 1$). In the single-input case, however, it follows from (2.6) and controllability of (A, B) that C.3 fails only in the trivial case of $C = 0$, or equivalently, $K = 0$. This observation leads to the well-known result by Kalman for the single-input case [12, Thm. 6]: Condition C.3 is a necessary and sufficient condition for optimality of a stable control law K with (K, A) *observable*.

¹ Note that this is true even if (C, A) is not detectable (see the proof of Fact 2.2).

As the third result, we give a precise statement of Anderson's result on the inverse problem [1, Thm. 8] described in § 1.

FACT 2.3. *Let K be a stable control law for the system (1.1), and the pair (K, A) be observable. Then K is optimal for some C if Condition C.2 holds.*

Remark 2.1. It should be noted that some authors [8], [18] misquote this result or that of [16] as mentioned in § 1. They state that the necessary Condition C.1 is also a sufficient condition for optimality of a stable control law K in the class of *piecewise-continuous* controls. This statement is, however, false as the following counterexample shows.

Let A , B and K in (1.1) and (1.3) be given by

$$(2.7) \quad A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = I, \quad K = 2I.$$

It is then obvious that (A, B) is controllable, (K, A) is observable, and the control law K is stable; moreover, $\Phi(s)$ defined by (1.4) becomes

$$(2.8) \quad \Phi(s) = \begin{bmatrix} \frac{8}{(1+s)(1-s)} & 0 \\ 0 & 0 \end{bmatrix},$$

which implies Condition C.1. But the control law K is not optimal as we shall show in the following. Suppose to the contrary that K is optimal for some C . Then by P.1 of Fact 2.1, K is uniquely given by (2.2), and moreover, B is nonsingular. Consequently, from (2.2), \bar{P} is uniquely determined by $\bar{P} = (B^T)^{-1}K = 2I$, and thus, the associated weighting matrix $C^T C$ for which K is optimal is *uniquely* given by

$$(2.9) \quad C^T C = \bar{P}^2 - \bar{P}A - A^T \bar{P} = \begin{bmatrix} 8 & 0 \\ 0 & 0 \end{bmatrix},$$

so that C must be either

$$(2.10) \quad C = [\sqrt{8} \ 0] \quad \text{or} \quad C = [-\sqrt{8} \ 0].$$

It is then easy to see that (C, A) is not detectable. This contradicts the "only if" part of P.2 in Fact 2.1, since K is both stable and optimal; more directly, the optimality of K for C given by (2.10) contradicts the existence of another control law $K_1 = \text{diag}(2, 0)$ which gives a lower value of J than K for the same C and $x_0 = [0 \ 1]^T$. Consequently, K is not optimal for any C , although this is optimal for C given by (2.10) in the limited class of stabilizing controls (see [16]).

The final result of this section is an important and new observation about the optimality of a stable control law K for some C .

FACT 2.4. *Let K be a stable control law. Then the following statements are equivalent:*

- i) K is optimal for some C .
- ii) (a) *There exist matrices $P \geq 0$ and C satisfying (2.3) and (2.5).*
 (b) *The pair (C, A) is detectable.*
- iii) K is optimal for some C with (C, A) detectable.

Proof. The implications i) \Rightarrow ii) and ii) \Rightarrow iii) are ensured respectively by Facts 2.1 and 2.2, and finally iii) \Rightarrow i) is obvious.

Remark 2.2. It is observed from this result that the optimality of a *stable* control law K is related essentially to *detectability* of (C, A) for the associated weighting matrix $C^T C$. This is in striking contrast to the inverse problem of stable regulator

[16], where such detectability condition is not required at all, but in some case, there is an irrelevant optimality of K for those C with (C, A) *undetectable*, as in the case of the counterexample shown in Remark 2.1; the optimality of this type may be irrelevant from the practical viewpoint, since no penalties are attached to some *unstable* modes in this case. It is this detectability requirement of (C, A) that makes the inverse problem treated here considerably difficult to solve, as compared with the inverse problem of stable regulator. In later sections our efforts will be devoted mainly to find conditions on A, B and K such that (C, A) is detectable when condition iia) of Fact 2.4 holds.

3. Preliminaries. In this section we obtain some basic optimality characterization of the control law together with its related system properties.

3.1. Basic characterization of optimality. In connection with the optimality criterion (a), (b) of Fact 2.4, we introduce the following matrix for a real symmetric matrix P :

$$(3.1) \quad \Gamma(P) = \begin{bmatrix} PA + A^T P - K^T K & PB - K^T \\ B^T P - K & 0 \end{bmatrix}.$$

Then we observe that (2.3) and (2.5) are equivalent to

$$(3.2) \quad \Gamma(P) = - \begin{bmatrix} C^T \\ 0 \end{bmatrix} [C \quad 0],$$

and note that the matrix P satisfying (2.3) and (2.5) for some C is always a solution of the linear matrix inequality (LMI):

$$(3.3) \quad \Gamma(P) \leq 0.$$

Conversely, a solution P of (3.3) always satisfies (2.3) and (2.5) for some C . As we shall see later, LMI plays a key role in the inverse problem, which is in contrast to the regulator problem where the Riccati equation plays the same role.

First, we need the following result about LMI for basic characterization of optimality.

LEMMA 3.1. *Let K be stable and Condition C.1 be satisfied. Then the following are true.*

(a) *There exist the minimal solution $\bar{P} \geq 0$ of LMI, as well as the $\bar{m} \times n$ real matrix \bar{C} such that*

$$(3.4a) \quad \bar{m} = \text{rank } \Phi(s),$$

$$(3.4b) \quad \Gamma(\bar{P}) = - \begin{bmatrix} \bar{C}^T \\ 0 \end{bmatrix} [\bar{C} \quad 0],$$

$$(3.4c)^2 \quad \text{rank} \begin{bmatrix} A - sI & B \\ \bar{C} & 0 \end{bmatrix} = n + \bar{m}, \quad \text{Re } s > 0,$$

$$(3.4d) \quad \Phi(s) = B^T (-sI - A^T)^{-1} \bar{C}^T \bar{C} (sI - A)^{-1} B.$$

(b) *Let C be any matrix obtained by factoring $\Gamma(P)$ as in (3.2) for some solution P of (3.3). Then if (C, A) is detectable, so is (\bar{C}, A) .*

Proof. (a) Let $F = A - BK$ and P_0 be a unique real symmetric matrix such that

$$P_0 F + F^T P_0 = -K^T K.$$

² Equivalently, (\bar{C}, A, B) is right invertible and minimum phase.

Then $P_0 \geq 0$ by $\operatorname{Re} \lambda(F) < 0$, and C.1 is shown [1, p. 22] to be equivalent to

$$(3.5) \quad \Psi(i\omega) \geq 0 \quad \text{for all real } \omega,$$

where $\Psi(s)$ is defined by

$$\Psi(s) = B^T(-sI - F^T)^{-1}(K - B^T P_0)^T + (K - B^T P_0)(sI - F)^{-1}B$$

and is related to $\Phi(s)$ via

$$(3.6) \quad \Psi(s) = [T(-s)^T]^{-1}\Phi(s)T(s)^{-1}.$$

Moreover, it is easy to show that \bar{P} is the minimal solution to LMI iff $\bar{X} = \bar{P} - P_0$ is the minimal solution to a linear matrix inequality of the form,

$$(3.7) \quad \tilde{\Gamma}(X) = \begin{bmatrix} XF + F^T X & XB - (K - B^T P_0)^T \\ B^T X - (K - B^T P_0) & 0 \end{bmatrix} \leq 0,$$

and in addition, these two solutions satisfy

$$(3.8) \quad \tilde{\Gamma}(\bar{X}) = \Gamma(\bar{P}).$$

By the assumed stability of F and Anderson's results [3, Thms. 1 and 3] (see also [5], [6]), Condition (3.5) ensures the existence of the minimal solution $\bar{X} \geq 0$ to (3.7) (and hence that of $\bar{P} \geq P_0 \geq 0$) as well as the $\bar{m} \times n$ matrix \bar{C} such that

$$(3.9a) \quad \bar{m} = \operatorname{rank} \Psi(s),$$

$$(3.9b) \quad \tilde{\Gamma}(\bar{X}) = - \begin{bmatrix} \bar{C}^T \\ 0 \end{bmatrix} [\bar{C} \quad 0],$$

$$(3.9c) \quad \operatorname{rank} \bar{C}(sI - F)^{-1}B = \bar{m}, \quad \operatorname{Re} s > 0.$$

By (3.6) and (3.8), it follows (3.9a) \Leftrightarrow (3.4a) and (3.9b) \Leftrightarrow (3.4b); (3.9c) \Leftrightarrow (3.4c) follows immediately from the definition of F ; (3.4d) follows immediately by pre- and post-multiplying (3.4b) respectively by $[B^T(-sI - A^T)^{-1}I]$ and $[B^T(sI - A^T)^{-1}I]^T$.

(b) Let (\bar{C}, A) be undetectable, so that there exist a scalar λ and a vector $x \neq 0$ such that

$$(3.10) \quad Ax = \lambda x, \quad \bar{C}x = 0, \quad \operatorname{Re} \lambda \geq 0.$$

Post- and pre-multiplying (3.2) respectively by $[x^T \ 0]^T$ and $[x^* \ 0]$ (where x^* denotes the conjugate transpose of x), we have by (3.10)

$$(3.11) \quad 2(\operatorname{Re} \lambda)x^*Px - \|Kx\|^2 + \|Cx\|^2 = 0.$$

Similarly, from (3.4b) and (3.10), we have

$$(3.12) \quad 2(\operatorname{Re} \lambda)x^*\bar{P}x - \|Kx\|^2 = 0.$$

Subtracting (3.12) from (3.11) yields

$$(3.13) \quad 2(\operatorname{Re} \lambda)x^*(P - \bar{P})x + \|Cx\|^2 = 0.$$

Since $P \geq \bar{P}$ and $\operatorname{Re} \lambda \geq 0$ by assumption, this implies $Cx = 0$, so that from (3.10) the pair (C, A) is also undetectable.

Combining this lemma with Facts 2.2 and 2.4 yields a basic characterization of optimality, which is a basis for our later development.

PROPOSITION 3.1. *Let K be a stable control law for the system (1.1), and Condition C.1 be satisfied. Then K is optimal if and only if the pair (\bar{C}, A) is detectable, where \bar{C} is the matrix defined in Lemma 3.1 for matrices A, B and K . Moreover, K is always optimal for \bar{C} , if it is optimal for some C .*

Proof. By Fact 2.4, optimality of K for C implies existence of P satisfying (3.2) as well as detectability of (C, A) , which implies detectability of (\bar{C}, A) by Lemma 3.1. Conversely, detectability of (\bar{C}, A) implies optimality of K for \bar{C} by Fact 2.2, since by Lemma 3.1 there exist $P = \bar{P} \geq 0$ and $C = \bar{C}$ satisfying (3.2), namely (2.3) and (2.5).

3.2. System properties. In this section, we derive some important properties of the linear time-invariant system

$$(3.14) \quad \dot{x} = Ax + Bu, \quad y = Cx,$$

where x and u are the state and control vectors as defined in § 1, and y is the m -vector of outputs. These properties will be used in the next section for expressing the optimality condition of Proposition 3.1 in terms of A, B and K . Below we write $S = (C, A, B)$ for the system (3.14), and $x(t; \xi, u)$, $y(t; \xi, u)$ respectively for its state and output forced by $u(t)$ with the initial state $x(0) = \xi$.

In the sequel, we use the standard notation of linear algebra in connection with the system (3.14): R^n is the real n -dimensional vector space; $\text{Im } A$, $\text{Ker } A$ and $\sigma(A)$ are the image, the kernel and the spectrum of a map A , respectively, and $\lambda(A)$ is an arbitrary element of $\sigma(A)$ —i.e. eigenvalue of A ; $A|_{\mathcal{V}}$ is the restriction of A to a subspace \mathcal{V} . We write \mathcal{V}^* for the largest (A, B) -invariant subspace contained in $\text{Ker } C$, and \mathcal{R}^* for the largest (A, B) -controllability subspace contained in $\text{Ker } C$. It is assumed that the reader is familiar with the basic concepts and properties of (A, B) -invariant and controllability subspace as well as those of \mathcal{V}^* and \mathcal{R}^* [22]. We note that these subspaces are initially real, but we shall introduce their complexifications without comment. For example, for any complex λ , we consider $\text{Ker } (A - \lambda I)$ as a complex subspace of the complexification of R^n . Finally, let $a(s)$ and $b(s)$ be any polynomials in s with coefficients in the real or complex field, then $a|b$ means that $a(s)$ divides $b(s)$.

With these notations, we are now ready to state a useful result on a canonical representation for (C, A, B) .

LEMMA 3.2 [7]. *Let F^* be any matrix such that*

$$(3.15) \quad (A + BF^*)\mathcal{V}^* \subset \mathcal{V}^*,$$

and decompose R^n as

$$(3.16a) \quad R^n = \mathcal{R}^* \oplus \mathcal{S} \oplus \mathcal{W}$$

where \mathcal{S} and \mathcal{W} are any subspaces such that

$$(3.16b) \quad R^n = \mathcal{V}^* \oplus \mathcal{W}, \quad \mathcal{V}^* = \mathcal{R}^* \oplus \mathcal{S},$$

and choose the coordinate basis of R^n such that

$$(3.16c) \quad \mathcal{R}^* = \text{Im} \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix}_{\substack{n_1 \\ n_2 \\ n_3}}, \quad \mathcal{S} = \text{Im} \begin{bmatrix} 0 \\ I \\ 0 \end{bmatrix}_{\substack{n_1 \\ n_2 \\ n_3}}, \quad \mathcal{W} = \text{Im} \begin{bmatrix} 0 \\ 0 \\ I \end{bmatrix}_{\substack{n_1 \\ n_2 \\ n_3}}.$$

Then $A + BF^*$ has a representation of the form

$$(3.17) \quad A + BF^* = \begin{bmatrix} \overbrace{A_{11}}^{n_1} & \overbrace{A_{12}}^{n_2} & \overbrace{A_{13}}^{n_3} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix} \begin{matrix} \left. \vphantom{\begin{matrix} \overbrace{A_{11}}^{n_1} \\ \overbrace{A_{12}}^{n_2} \\ \overbrace{A_{13}}^{n_3} \end{matrix}} \right\} n_1 \\ \left. \vphantom{\begin{matrix} A_{23} \\ A_{33} \end{matrix}} \right\} n_2 \\ \left. \vphantom{A_{33}} \right\} n_3 \end{matrix}.$$

Moreover, the invariant polynomials of $sI - A_{22}$ (or A_{22}), together with a number equal to 1, constitute the invariant polynomials of the system matrix of (C, A, B) defined by

$$(3.18) \quad P(s) = \begin{bmatrix} A - sI & B \\ C & 0 \end{bmatrix}.$$

From this lemma we obtain the following result.

PROPOSITION 3.2. Let \mathcal{M} be any subspace such that

$$(3.19) \quad A\mathcal{M} \subset \mathcal{M} \subset \mathcal{V}^*,$$

and $a_i(s)$ ($1 \leq i \leq k$), $p_i(s)$ ($1 \leq i \leq l$) be respectively the invariant polynomials of $A_m = A|_{\mathcal{M}}$ and the system matrix $P(s)$ of (3.18), which are ordered so that $a_i|a_{i-1}$ and $p_i|p_{i-1}$. Then $\mathcal{M} \subset \mathcal{R}^*$ if $a_1(s)$ and $p_1(s)$ are coprime.

Proof. Let M be a matrix composed of a basis for \mathcal{M} . By (A, B) -invariance of \mathcal{V}^* and (3.19) there exists a matrix F^* such that

$$(3.20) \quad (A + BF^*)\mathcal{V}^* \subset \mathcal{V}^*, \quad F^*|_{\mathcal{M}} = 0,$$

(set $F^* = F - F^0$ for any F with $(A + BF)\mathcal{V}^* \subset \mathcal{V}^*$ and F^0 with $F^0[M \tilde{M}] = [FM \ 0]$, where \tilde{M} is a full column rank matrix with $\text{Im}[M \tilde{M}] = \mathcal{V}^*$). Now we apply Lemma 3.2 to the triple (C, A, B) and F^* so defined. First, we decompose R^n as (3.16) and partition M as

$$(3.21) \quad M = [M_1^T \ M_2^T \ M_3^T]^T$$

in accordance with this decomposition. Then, by (3.19) and (3.16) $M_3 = 0$; hence by (3.16c) we have

$$(3.22) \quad \mathcal{M} \subset \mathcal{R}^* \quad \text{iff} \quad M_2 = 0,$$

and also, by (3.20b) and (3.17), as well as the definition of A_m ,

$$(3.23) \quad \begin{bmatrix} M_1 \\ M_2 \\ 0 \end{bmatrix} A_m = AM = (A + BF^*) \begin{bmatrix} M_1 \\ M_2 \\ 0 \end{bmatrix} = \begin{bmatrix} A_{11}M_1 + A_{12}M_2 \\ A_{22}M_2 \\ 0 \end{bmatrix},$$

so that we have the key relation,

$$(3.24) \quad A_{22}M_2 = M_2A_m.$$

Also, we note that $p_i(s)$, $i = 1, \dots, l$ are the invariant polynomials of A_{22} by Lemma 3.2. By definition, $a_1(s)$ and $p_1(s)$ are the minimal polynomials of A_m and A_{22} respectively, so that by assumption $\sigma(A_m) \cap \sigma(A_{22}) = 0$ in (3.24). This implies $M_2 = 0$ as is well known, so the result follows by (3.22).

COROLLARY 3.1. *Let (C, A, B) be minimum phase, that is, its system matrix $P(s)$ of (3.18) satisfies*

$$(3.25) \quad \text{rank } P(s) = \text{normal rank } P(s), \quad \text{Re } s > 0.$$

Then

$$(3.26) \quad \text{Ker } (A - \lambda I) \cap \text{Ker } C \subset \mathcal{R}^*, \quad \text{Re } \lambda > 0.$$

Proof. This follows immediately from Proposition 3.2, since $\mathcal{M} = \text{Ker } (A - \lambda I) \cap \text{Ker } C$ satisfies (3.19); and moreover $a_1(s) = s - \lambda$ with $\text{Re } \lambda > 0$ by definition, while $p_1(s)$ has all zeros in $\text{Re } s \leq 0$ by (3.25).

The following is a useful characterization of \mathcal{R}^* , which is a slight modification of [17, Thm. 6.1] derived in a less transparent way than the one shown below.

PROPOSITION 3.3. *The subspace \mathcal{R}^* equals to the set of states reachable from the origin by such control inputs that the resulting outputs are identically zero, that is,*

$$(3.27) \quad \mathcal{R}^* = \{x \in R^n \mid x(T; 0, u) = x, y(t; 0, u) \equiv 0 \text{ for some finite } T \geq 0 \text{ and } u\}.$$

Proof. This is almost obvious from the definition of \mathcal{R}^* as well as the two equivalent definitions of an (A, B) -controllability subspace [21], [22], that is, a subspace $\mathcal{R} \subset R^n$ is an (A, B) -controllability subspace iff (C.1) \mathcal{R} is the controllable subspace of $(A + BF, BG)$ for some real matrices F and G , or equivalently, (C.2) every state $x \in \mathcal{R}$ can be reached at a finite time from the origin along a controlled trajectory that is wholly contained in \mathcal{R} . In fact, let \mathcal{R} be the subspace on the right side of (3.27). Then $\mathcal{R}^* \subset \mathcal{R}$ is obvious from Definition 1 and $\mathcal{R}^* \subset \text{Ker } C$, while $\mathcal{R} \subset \mathcal{R}^*$ is obvious from Definition 2 and the definition of \mathcal{R}^* , since \mathcal{R} satisfies Condition C.2 and $\mathcal{R} \subset \text{Ker } C$ by definition.

4. Main results. In this section, we first derive a complete set of optimality conditions of a control law K (i.e. a complete solution of Problem (I)) as well as some useful sufficient conditions for optimality of K (§ 4.1), and then give its nice physical interpretation (§ 4.2).

4.1. Optimality conditions. To begin with, it may be convenient to state the following result before deriving the complete optimality conditions.

PROPOSITION 4.1. *Let K be a stable control law for the system (1.1), and C.1 be satisfied. Then K is optimal if and only if*

$$(4.1) \quad \text{Ker } (A - sI) \cap \mathcal{X}^0 = 0, \quad \text{Re } s > 0.$$

Here \mathcal{X}^0 is the set of states of (1.1) reachable from the origin by such control inputs $u(t)$ that

$$(4.2) \quad \Phi(s)U(s) \equiv 0,$$

where $U(s)$ is the Laplace transform of $u(t)$.

Proof. First note that (3.4) holds by Lemma 3.1 and the assumption made above. By Proposition 3.1, it is enough to show that (4.1) is equivalent to detectability of (\bar{C}, A) , i.e.,

$$(4.3) \quad \mathcal{M}(s) \triangleq \text{Ker } (A - sI) \cap \text{Ker } \bar{C} = 0, \quad \text{Re } s \geq 0.$$

Note here that

$$(4.4) \quad \mathcal{M}(s) = 0, \quad \text{Re } s = 0,$$

since otherwise post- and pre-multiplying a *nonzero* $x \in \mathcal{M}(\lambda)$ ($\operatorname{Re} \lambda = 0$) and its conjugate transpose x^* respectively on both sides of (1.1) block in (3.4b) and then substituting $Ax = \lambda x$, $\bar{C}x = 0$ would yield $Kx = 0$, implying that $A - BK$ has an eigenvalue λ with $\operatorname{Re} \lambda = 0$, a contradiction to stability assumption of K .

Let $\bar{\mathcal{R}}^*$ be the largest (A, B) -controllability subspace contained in $\operatorname{Ker} \bar{C}$. Since (\bar{C}, A, B) is minimum phase by (3.4c), $\mathcal{M}(s) \subset \bar{\mathcal{R}}^* \subset \operatorname{Ker} \bar{C}$ by Corollary 3.1 and the definition of $\bar{\mathcal{R}}^*$, so that

$$(4.5) \quad \mathcal{M}(s) = \mathcal{M}(s) \cap \bar{\mathcal{R}}^* = \operatorname{Ker} (A - sI) \cap \bar{\mathcal{R}}^*.$$

Moreover, by (3.4c),

$$\text{normal rank } [\bar{C}(-sI - A)^{-1}B]^T = \bar{m},$$

so that by (3.4d)

$$\Phi(s)U(s) \equiv 0 \Leftrightarrow \bar{C}(sI - A)^{-1}BU(s) \equiv 0.$$

Noting this relation and applying Proposition 3.3 to (\bar{C}, A, B) yields $\mathcal{X}^0 = \bar{\mathcal{R}}^*$, and hence by (4.5)

$$\mathcal{M}(s) = \operatorname{Ker} (A - sI) \cap \mathcal{X}^0.$$

Combining this with (4.4) establishes (4.1) \Leftrightarrow (4.3) as claimed.

We have assumed C.1 in Proposition 4.1 for simplicity. Since this is a necessary condition for optimality of K as shown in § 2, it is possible to remove this assumption. As a result, we obtain the following result, which is a complete solution to Problem (I).

THEOREM 4.1. *Let K be a stable control law for the system (1.1). Then K is optimal if and only if Conditions C.1 and (4.1) are both satisfied. Moreover, K is always optimal, if it is, with respect to the cost (2.1) for some C such that (C, A) is detectable and (C, A, B) is right invertible and minimum phase.*

Proof. The former part is obvious from the preceding observation and Proposition 4.1. The latter part is due to Proposition 3.1 and (a) of Lemma 3.1 (set $C = \bar{C}$).

Remark 4.1. Let $\det \Phi(s) \neq 0$. This obviously implies $\mathcal{X}^0 = 0$ by definition, so that it follows from Theorem 4.1 that C.3 in § 1 ensures the optimality of a stable control law K . Hence, Anderson's result (Fact 2.3) is recovered easily.

Remark 4.2. Let $\Phi(s) \equiv 0$; for example,

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad K = [0 \quad 2].$$

Then the input $u(t)$ satisfying (4.2) is arbitrary and (A, B) is controllable by assumption, so that \mathcal{X}^0 is the whole state space. In this case, therefore, Theorem 4.1 reveals that a stable control law K is optimal iff A is stable. Furthermore, combining this fact with P.1 of Fact 2.1 and (a) of Lemma 3.1 yields another fact that $K = 0$ is the only stable optimal control law in this case. Hence, the foregoing stable control law is not optimal.

Remark 4.3. It is important to note that under C.1 and stability of K , Condition (4.1) fails only in a very rare case where

$$(4.6) \quad -\lambda \in \sigma(A - BK) \quad \text{for some } \lambda \in \sigma(A).$$

In fact, suppose (4.1) fails or, equivalently, there exists a nonzero $x \in \text{Ker}(A - \lambda I) \cap \text{Ker } \bar{C}$ for some $\lambda \in \sigma(A)$ with $\text{Re } \lambda > 0$ (see the proof of Proposition 4.1). Since, by (3.4b),

$$(4.7) \quad \bar{P}(A - \lambda I) + (A - BK + \lambda I)^T \bar{P} + \bar{C}^T \bar{C} = 0,$$

we obviously obtain $(A - BK + \lambda I)^T \bar{P}x = 0$, so that $\det(A - BK + \lambda I) = 0$. Otherwise $\bar{P}x = 0$, and hence, by (3.4b), $(A - BK)x = (A - BB^T \bar{P})x = Ax = \lambda x$, a contradiction to stability of K . Consequently, $\lambda \in \sigma(A)$ and $-\lambda \in \sigma(A - BK)$. Note also that this proof also shows $\text{Ker}(A - \lambda I) \cap \mathcal{X}^0 = 0$ for all $\lambda \in \sigma(A)$ with $\text{Re } \lambda > 0$ and $-\lambda \notin \sigma(A - BK)$ under the same condition. Therefore, in checking (4.1), it is enough to check it only for those $s \in \sigma(A)$ with $-s \in \sigma(A - BK)$.

From Remarks 4.1 and 4.3, we obtain the following corollary.

COROLLARY 4.1. *Let Condition C.1 hold. Then a stable control law K is optimal if either of the following conditions holds.*

- (a) *normal rank $\Phi(s) = r$.*
- (b) *$\lambda(A) + \lambda(A - BK) \neq 0$.*

Remark 4.4. As we have shown in § 2, (a) holds only for the limited class of optimal control laws K , while (b) holds *generically* for all control laws K . In this respect Condition (b) is definitely superior to Condition (a) as a sufficient condition for optimality; and what is more, this condition together with the necessity of Condition C.1 for optimality implies a practically important observation.

COROLLARY 4.2. *Condition C.1 is a necessary and sufficient condition for a stable control law K to be generically optimal.*

Now it may be helpful for better understanding of the results obtained above to show how these results can be used to check the optimality of a given control law K . Let us consider the example system and control law given by (2.7). First, it is apparent from (2.7) and (2.8) that Condition C.1 holds, but neither sufficient condition (a) nor (b) of Corollary 4.1 holds. It thus remains to check if Condition (4.1) holds or not. By (2.8) all control inputs $u(t)$ satisfying (4.2) are given in the form

$$u(t) = \begin{bmatrix} 0 \\ u_2(t) \end{bmatrix},$$

where $u_2(t)$ is an arbitrary piecewise-continuous function, so that by the particular form of A, B and the definition of \mathcal{X}^0

$$\mathcal{X}^0 = \text{Im} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \text{Ker}(A - I),$$

implying that Condition (4.1) does not hold. We thus conclude from Theorem 4.1 that K is not optimal in this case, which is the same conclusion as drawn in a different way in Remark 2.1.

Up to this point we have assumed the stability of a control law K . This assumption, though it is a fundamental one in the inverse problem, can also be removed in return for restricting the cost matrix. As a result, we obtain the following optimality condition of a *general* (not necessarily stable) control law K for some restricted type of C .

THEOREM 4.2. *Let K be a control law for the system (1.1). Then K is optimal for some C such that (C, A) is detectable if and only if K is stable and satisfies Conditions C.1 and (4.1).*

Proof. By Theorem 4.1, it suffices to show the stability of K in the “only if” part, which is a direct consequence of Fact 2.2.

Remark 4.5. Note that the detectability of (C, A) is known as the most general constraint on the cost matrix appearing in the standard optimal regulator problem, and the optimal control law is stable only for this type of cost matrix, as shown in Fact 2.1. We also note that Theorem 4.2 corresponds to the result in the inverse problem of stable regulator [16, Corollary 6.2], where the first two conditions in the theorem are shown to be the optimality conditions.

4.2. Interpretation of the optimality condition. In the previous section, we have obtained two types of conditions as a complete set of optimality conditions: one is the frequency domain condition C.1 (as often called the return difference condition), and the other is a new condition (4.1) of geometric type. The former condition has been well known as a condition on sensitivity reduction of the closed-loop system (1.1), (1.3) to system parameter variations as compared to the equivalent open-loop system. It may thus be interesting if we can interpret the latter one as some condition related to the sensitivity of the closed-loop system. This is indeed what we shall do in the sequel. For this purpose we need to take the comparison sensitivity approach developed by Cruz and Perkins [9] and applied to optimal control system by Kreindler [13], which is as follows (see [2, § 7.1, § 7.2], [13], [18]).

Consider two different configurations of the optimal control system, namely pure closed-loop and pure open-loop configurations (see Fig. 1). The basic idea of the approach is to compare the effects of parameter variations in the system matrices A ,

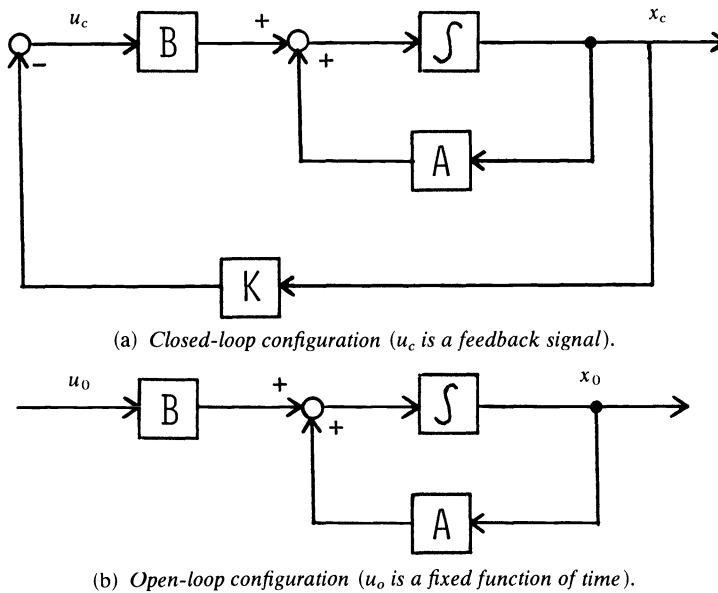


FIG. 1. Two configurations of the optimal control system.

B on the responses for these two configurations that produce identical responses for nominal parameter values (hence, they are called nominally equivalent). Let A and B depend on a parameter vector q , and subscripts o and c refer to the open- and closed-loop configurations, respectively. Then the equations for the first order trajectory changes $\delta x_o(t)$ and $\delta x_c(t)$ due to small variations δq of q are described by using the corresponding first-order changes δA , δB of A and B as follows:

$$(4.8) \quad \delta \dot{x}_o = A \delta x_o + \delta A \bar{x}_o + \delta B \bar{u}_o, \quad \delta x_o(0) = 0,$$

$$(4.9) \quad \delta \dot{x}_c = A \delta x_c + \delta A \bar{x}_c + \delta B \bar{u}_c - B K \delta x_c, \quad \delta x_c(0) = 0,$$

where the bars refer to their nominal values. Hence, by definition,

$$(4.10) \quad \bar{x}_o(t) \equiv \bar{x}_c(t), \quad \bar{u}_o(t) \equiv \bar{u}_c(t) = -K\bar{x}_c(t).$$

Note here that u_o does not depend on q , while u_c depends on q through the feedback (1.3), which is reflected in the last term on the right side of (4.9); the presence of this term—the only difference between (4.8) and (4.9)—leads to reduction of the effect of δq on x_c as compared to x_o . Now, a simple calculation shows [2, § 7.1], [9], [13],

$$(4.11) \quad \begin{aligned} \delta J &\triangleq \int_0^\infty \delta x_o^T K^T K \delta x_o dt - \int_0^\infty \delta x_c^T K^T K \delta x_c dt \\ &= \frac{1}{2\pi} \int_{-\infty}^\infty \delta X_c^T(-i\omega) K^T \Phi(i\omega) K \delta X_c(i\omega) d\omega, \end{aligned}$$

where $\delta X_c(s)$ is the Laplace transform of $\delta x_c(t)$. Obviously, by this relation Condition C.1 ensures a well-known property contained in [2], [13], that the optimal closed-loop configuration has lower sensitivity (in Kx) than the equivalent open-loop one in the sense of $\delta J \geq 0$.³ It should be noted, however, that this sense of sensitivity reduction includes the special case of no sensitivity reduction (i.e. $\delta J = 0$); namely, *there are some cases where no sensitivity reduction at all can be achieved by optimal feedback*. The new condition (4.1) will prove to be closely related to this fact.

Let $\delta x_d = \delta x_c - \delta x_o$, or equivalently, express δx_c as

$$(4.12) \quad \delta x_c = \delta x_o + \delta x_d,$$

and write the input by feedback in (4.9) as $\delta u_c(t) = -K\delta x_c(t)$ and $\delta U_c(s) = -K\delta X_c(s)$. Then, by (4.8) to (4.11),

$$(4.13) \quad \delta \dot{x}_d = A\delta x_d + B\delta u_c, \quad \delta x_d(0) = 0,$$

$$(4.14) \quad \Phi(s)\delta U_c(s) \equiv 0 \Leftrightarrow \delta J = 0.$$

Hence, we can characterize δx_d as that component of δx_c due to feedback which compensates for the other component change δx_o due directly to the variations of A and B , while \mathcal{X}^0 is characterized as the maximal (l.u.b.) set of δx_d reachable by those feedbacks that achieve *no sensitivity reduction*. In other words, there is indeed a sensitivity reduction at least for that δx_c with $\delta x_d \notin \mathcal{X}^0$. These characterizations of δx_d and \mathcal{X}^0 yield the following interpretation of (4.1) in connection with the sensitivity reduction of the optimal closed-loop configuration.

THEOREM 4.3. *In the optimal closed-loop configuration, there is always a strict sense of sensitivity reduction (i.e. $\delta J > 0$) for such trajectory change δx_c that its compensating component δx_d by feedback contains an unstable mode of A in the sense that*

$$(4.15) \quad \delta x_d(t) \in \text{Ker}(A - \lambda I) \quad \text{for all } t > 0 \text{ and some } \lambda \in \sigma^+(A),$$

where $\sigma^+(A)$ is the set of $\lambda \in \sigma(A)$ with $\text{Re } \lambda > 0$.

Remark 4.6. Let $\dim \text{Ker}(A - \lambda I) = 1$ for simplicity. Then it follows easily from the constraint (4.13) on δx_d that (4.15) restricts the form of δx_d as $\delta x_d(t) = e^{\lambda t} h(t) v$ where $\lambda \in \sigma^+(A)$, $v = \text{Ker}(A - \lambda I)$ and $h(t)$ is some continuous function with $h(0) = 0$. This obviously justifies the use of the above expression “contains an unstable mode of A ”.

³ Strictly speaking, the sensitivity measure δJ here is a bit weaker than that in [2], [13], but used here for simplifying the development.

Note that $\delta x_c(t)$ contains no unstable modes of A in general, since $F = A - BK$ is stable by assumption and $\delta X_c(s) = (sI - F)^{-1}(\delta A - \delta BK)(sI - F)^{-1}x(0)$ by (4.9), (4.10). Hence, by (4.12), whenever δx_a contains an unstable mode of A , δx_o must also contain the same mode so that these two same modes of A cancel in δx_c as a consequence of their summation. In other words, if some parameter variation excites an unstable mode of A in δx_o according to the constraint (4.8), then this mode is automatically cancelled in δx_c by such prompt reaction of feedback control that *excites the same mode in δx_a* according to the constraint (4.13); hence such a mode may be called a *hidden* mode of A . This is considered as the way of control in which the optimal feedback control law controls an unstable *hidden* mode of A excited in the response due to parameter variation. As is checked easily, this is actually a common feature of all (not necessarily optimal) *stable feedback controls*. Theorem 4.3 shows, however, that as far as an optimal one is concerned, its process of control for unstable hidden modes of A mentioned above leads inevitably to sensitivity reduction in the *strict* sense. In other words, *under the presence of parameter variations, the optimal feedback control always controls any unstable hidden mode of A in such a way that the sensitivity of the resulting closed-loop response is reduced strictly*. This is a new additional sensitivity reduction property of the optimal feedback control, which provides not only a better insight into the well-known sensitivity reduction property derived from the return difference condition C.1 alone, but also a better understanding of how *optimal* stable feedback controls differ from *other* ones from the sensitivity point of view.

Another interesting feature of the new sensitivity property as related to unstable modes of A alone is its significance peculiar essentially to the *multi-input* case. In the single-input case, there proves to be a sensitivity reduction for *all* modes of A as in the multi-input case of $\det \Phi(s) \neq 0$, and hence there is no essential distinction between stable and unstable modes of A from the sensitivity point of view. Consequently, the intrinsic value of this property lies only in the case of multi-input and *particularly* $\det \Phi(s) \equiv 0$, which obviously makes an essential difference between the solutions of the single- and multi-input inverse problem. The point is: *unstable modes of A play no particular role in sensitivity reduction (or optimality property), provided that there is a full degree of sensitivity reduction in the sense of normal rank $\Phi(s) = r$* .

In summary, Condition C.1 ensures sensitivity reduction of the optimal closed-loop response of *any* type (but in the broad sense of $\delta J \geq 0$)—a well-known property; Condition (4.1) ensures, on the assumption of C.1, the *strict* sense of sensitivity reduction ($\delta J > 0$) for that *particular* type of optimal closed-loop response containing an unstable hidden mode of A —a new property which certainly provides an additional and better insight into the well-known sensitivity reduction property of optimal control mentioned above.

5. Concluding remarks. A new geometric condition as well as the known return difference condition has been obtained as a complete set of optimality conditions in the inverse optimal control problem. The principal contribution of the new condition to the inverse problem is that it provides a better insight into the well-known sensitivity reduction property of optimal control, by revealing an additional sensitivity reduction property of optimal control related to unstable modes of A . Another interesting contribution of practical importance is that it provides a new characterization of the return difference condition as a necessary and sufficient condition for generic optimality of a feedback control. Finally, although we have treated only the case where the weighting matrix R for the control input is I (the unit matrix), other cases with $R \neq I$

can also be treated similarly if R is known, by replacing B with $BR^{-1/2}$. However, the inverse problem for the case with R being unknown is a different problem which remains to be solved in the future.

Acknowledgments. This work was carried out under the direction of Prof. N. Suda. The authors wish to thank Prof. N. Suda, and also Prof. H. Kimura and Dr. Y. Inouye for their helpful suggestions during the course of the research. Thanks are also due to Prof. B. D. O. Anderson for his generous invaluable comments on this work.

REFERENCES

- [1] B. D. O. ANDERSON, *The inverse problem of linear optimal control*, Rep. SEL-66-039, Stanford Univ., Stanford, CA, 1966.
- [2] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
- [3] B. D. O. ANDERSON, *Algebraic properties of minimal degree spectral factors*, Automatica, 9 (1973), pp. 491–500.
- [4] B. D. O. ANDERSON AND J. B. MOORE, *Network Analysis and Synthesis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [5] B. D. O. ANDERSON, K. L. HITZ AND N. D. DIEN, *Recursive algorithm for spectral factorization*, IEEE Trans. Circuits and Systems, CAS-21 (1974), pp. 742–750.
- [6] B. D. O. ANDERSON, *Corrections to "Algebraic properties of minimal degree spectral factors"*, Automatica, 11 (1975), pp. 321–322.
- [7] ———, *A note on transmission zeros of a transfer function matrix*, IEEE Trans. Automatic Control, AC-21 (1976), pp. 589–590.
- [8] J. CASTI, *The linear-quadratic control problem: some recent results and outstanding problems*, SIAM Rev., 22 (1980), pp. 459–485.
- [9] J. B. CRUZ AND W. R. PERKINS, *A new approach to the sensitivity problem in multivariable feedback design*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 216–223.
- [10] T. FUJII AND M. NARAZAKI, *A complete solution to the inverse problem of optimal control*, Proc. 1982 IEEE Conference on Decision and Control, IEEE, New York, 1982, pp. 289–294.
- [11] A. JAMESON AND E. KREINDLER, *Inverse problem of linear optimal control*, SIAM J. Control, 11 (1973), pp. 1–19.
- [12] R. E. KALMAN, *When is a linear control system optimal?* Trans. ASME J. Basic Engr. 86D (1964), pp. 51–60.
- [13] E. KREINDLER, *Closed-loop sensitivity reduction of linear optimal control systems*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 254–262.
- [14] V. KUČERA, *A contribution to matrix quadratic equations*, IEEE Trans. Automatic Control, AC-17 (1972), pp. 344–347.
- [15] K. MÄRTENSSON, *On the matrix Riccati equation*, Inform. Sci., 3 (1971), pp. 17–49.
- [16] B. P. MOLINARI, *The stable regulator problem and its inverse*, IEEE Trans. Automatic Control, AC-18 (1973), pp. 454–459.
- [17] A. S. MORSE AND W. M. WONHAM, *Decoupling and pole assignment by dynamic compensation*, SIAM J. Control, 8 (1970), pp. 446–465.
- [18] W. R. PERKINS AND J. B. CRUZ, *Feedback properties of linear regulators*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 659–664.
- [19] N. SUDA AND T. FUJII, *The optimality property of an optimal regulator incorporating an observer*, Int. J. Control, 33 (1981), pp. 617–647.
- [20] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, AC-16 (1971), pp. 621–634.
- [21] W. M. WONHAM, *Geometric state-space theory in linear multivariable control: A status report*, Automatica, 15 (1979), pp. 5–13.
- [22] ———, *Linear Multivariable Control, A Geometric Approach*, 2nd ed., Springer-Verlag, New York, 1979, pp. 594–600.

FEEDBACK CONTROL OF SECOND ORDER EVOLUTION EQUATIONS WITH DAMPING*

YOSHIYUKI SAKAWA†

Abstract. Feedback control is developed for a class of distributed systems described by second order evolution equations with slight damping. In order to increase the degree of stability of the system, a dynamic compensator is designed on the basis of a finite-dimensional model of the system, and a feedback control system is constructed by using sensor outputs. It is shown that the degree of stability of the whole system including the compensator is improved by "modal control" based on a finite-dimensional modal model. This proves the mathematical validity of the modal control of distributed systems.

Key words. feedback stabilization, second order evolution equation, finite-dimensional dynamic compensator, observer

1. Introduction. We consider feedback control of a class of distributed systems described by second order evolution equations with slight damping. The damping term in the equation reflects the dissipation of energy which is empirically observed in nature. We are given a finite number of control inputs to the system and a finite number of outputs from sensors. In order to increase the degree of stability of the system, we construct a finite-dimensional dynamic compensator using the sensor outputs. Our formulation includes the active control of vibrations in mechanically flexible systems, which has application to attitude control of flexible spacecraft and active tendon control of structures, for example [3], [19].

First, it will be proved that the solution of the evolution equation can be represented by an analytic semigroup, and spectral properties of the semigroup will be examined. Then, on the basis of a finite-dimensional modal model of the system, finite-dimensional observers are constructed as a dynamic compensator. It will be shown that the degree of stability of the whole system including the compensator is improved by the modal control, in spite of the undesirable effects of control and observation "spillovers" [2]. Thus, this paper gives mathematical validity to modal control on the basis of a finite-dimensional model of distributed systems.

2. Damped evolution equations. Let Ω be a bounded open domain in a finite-dimensional Euclidean space, and let $L^2(\Omega)$ denote the Hilbert space of all square-integrable functions with the inner product

$$(u_1, u_2) = \int_{\Omega} u_1(x) \bar{u}_2(x) dx,$$

where \bar{u} denotes the complex conjugate of u .

We consider control systems described by the second order (in t) damped evolution equation in $L^2(\Omega)$:

$$(2.1) \quad \frac{d^2 u(t)}{dt^2} + 2\alpha A \frac{du(t)}{dt} + Au(t) = Bf(t) = \sum_{k=1}^r b^k f^k(t), \quad t > 0,$$

where $u(t) \in L^2(\Omega)$, $b^k \in L^2(\Omega)$, $f^k(t)$ are scalar functions Hölder-continuous on $[0, \infty)$, α is a small positive constant, A is an unbounded operator and

$$B = [b^1, \dots, b^r], \quad f(t) = [f^1(t), \dots, f^r(t)]^T.$$

* Received by the editors April 21, 1982, and in final revised form February 15, 1983.

† Department of Control Engineering, Faculty of Engineering Science, Osaka University, Toyonaka, Osaka 560, Japan.

We assume that A is a *selfadjoint positive-definite* operator with domain $D(A)$ dense in $L^2(\Omega)$, and that A^{-1} exists and is *compact*.

The state of the system is measured by p averaging sensors, whose outputs are expressed as

$$(2.2) \quad y^k(t) = (c^k, u(t)), \quad k = 1, \dots, p,$$

where the c^k are sensor influence functions in $L^2(\Omega)$. We introduce the output vector function $y(t) = [y^1(t), \dots, y^p(t)]^T \in R^p$, which can be written as

$$(2.3) \quad y(t) = Cu(t),$$

where C is a bounded operator mapping $L^2(\Omega)$ into R^p defined by

$$Cu = [(c^1, u), \dots, (c^p, u)]^T.$$

By the *solution* of (2.1) we mean a function $u(t) \in L^2(\Omega)$ satisfying the following conditions [11]:

- (i) $u(t)$ and $\dot{u}(t) = du(t)/dt$ are continuous on $[0, T]$, where $T > 0$ is an arbitrary number.
- (ii) $u(t) \in D(A)$ for any $t \geq 0$, and $Au(t)$ is continuous on $[0, T]$. $\dot{u}(t) \in D(A)$ for any $t > 0$ and $A\dot{u}(t)$ is continuous on $(0, T]$.
- (iii) $u(t)$ is twice continuously differentiable and satisfies (2.1) on $(0, T]$.
- (iv) $u(t)$ and $\dot{u}(t)$ satisfy the initial condition

$$(2.4) \quad u(0) = u_0, \quad \dot{u}(0) = u_{t0},$$

where $u_0 \in D(A)$ and $u_{t0} \in L^2(\Omega)$.

From the assumption that A has a bounded inverse, it is easy to see that A is *closed*. Also, from the Hilbert–Schmidt theory for compact selfadjoint operators, it is well known that there exist eigenvalues λ_i and corresponding eigenfunctions $\phi_{ij}(x)$ of the operator A satisfying the following conditions [13], [17]:

(i)

$$(2.5) \quad 0 < \lambda_1 < \lambda_2 < \dots < \lambda_i < \dots, \quad \lim \lambda_i = \infty.$$

(ii)

$$(2.6) \quad A\phi_{ij} = \lambda_i \phi_{ij}, \quad j = 1, \dots, m_i, \quad i = 1, 2, \dots,$$

where $m_i < \infty$ for each i .

- (iii) The set $\{\phi_{ij}(\cdot)\}$ of the eigenfunctions forms a *complete orthonormal system* in $L^2(\Omega)$.

Since $u \in L^2(\Omega)$ has the unique expression

$$(2.7) \quad u(\cdot) = \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} (u, \phi_{ij}) \phi_{ij}(\cdot),$$

$D(A)$ consists of all elements $u \in L^2(\Omega)$ such that

$$(2.8) \quad \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} \lambda_i^2 |(u, \phi_{ij})|^2 < \infty.$$

Here m_i is the multiplicity of the eigenvalue λ_i .

We assume that

$$(2.9) \quad \alpha^2 \lambda_i^2 - \lambda_i \neq 0, \quad i = 1, 2, \dots$$

and that $\alpha > 0$ is so small that

$$(2.10) \quad \alpha \lambda_1 < \frac{1}{2\alpha}.$$

Now let us introduce a function

$$(2.11) \quad g(\lambda) = \sqrt{\alpha^2 \lambda^2 - \lambda}.$$

If $\alpha^2 \lambda^2 - \lambda < 0$, then

$$g(\lambda) = i\sqrt{\lambda - \alpha^2 \lambda^2}.$$

Since A is a selfadjoint operator, it is possible to define an operator $g(A)$ by [6]

$$(2.12) \quad D(g(A)) = \left\{ u \in L^2(\Omega) : \sum_{i=1}^{\infty} \sum_{j=1}^{m_j} |g(\lambda_i)(u, \phi_{ij})|^2 < \infty \right\},$$

$$(2.13) \quad g(A)u = \sum_{i=1}^{\infty} \sum_{j=1}^{m_j} g(\lambda_i)(u, \phi_{ij})\phi_{ij}.$$

Let n be a sufficiently large number. Then the inequality

$$\alpha^2 \lambda_i^2 \left[1 - \left(\frac{1}{\alpha^2 \lambda_n} \right) \right] \leq |g(\lambda_i)|^2 \leq \alpha^2 \lambda_i^2$$

holds for any $i \geq n$. Hence, in view of (2.8) and (2.12), we see that

$$(2.14) \quad D(g(A)) = D(A).$$

In the same way, we can define the inverse operator $g^{-1}(A)$ and its domain $D(g^{-1}(A))$ by replacing $g(\lambda_i)$ by $g^{-1}(\lambda_i)$ in (2.12) and (2.13), respectively. It is easily seen that $g^{-1}(A)$ is a bounded operator, because the sequence $\{|g^{-1}(\lambda_i)|^2\}$ converges to zero, and that

$$g^{-1}(A)g(A) = I \quad \text{in } D(A), \quad g(A)g^{-1}(A) = I \quad \text{in } L^2(\Omega).$$

Suppose a solution $u(t)$ of (2.1) exists. Let us introduce a new function $v(t)$, $t \geq 0$, by

$$(2.15) \quad v(t) = g^{-1}(A)(\dot{u}(t) + \alpha Au(t)).$$

Since $v(t) \in D(g(A)) = D(A)$, we obtain

$$(2.16) \quad \dot{u}(t) = -\alpha Au(t) + g(A)v(t), \quad t \geq 0.$$

Also, since A is closed, we see that

$$\frac{d(Au(t))}{dt} = A\dot{u}(t).$$

Differentiating (2.15) with respect to t and using the relation (2.1) yields

$$(2.17) \quad \dot{v}(t) = g^{-1}(A)[- \alpha A\dot{u}(t) - Au(t) + Bf(t)], \quad t > 0.$$

Furthermore, substituting (2.16) into (2.17) gives

$$\dot{v}(t) = g^{-1}(A)[(\alpha^2 A^2 - A)u(t) - \alpha Ag(A)v(t) + Bf(t)].$$

Since $g^2(A) = \alpha^2 A^2 - A$, we finally obtain

$$(2.18) \quad \dot{v}(t) = g(A)u(t) - \alpha Av(t) + g^{-1}(A)Bf(t), \quad t > 0.$$

Now let us introduce the new functions

$$(2.19) \quad \xi(t) = u(t) + v(t), \quad \eta(t) = u(t) - v(t).$$

We see from (2.16) and (2.18) that these functions satisfy the equations

$$(2.20) \quad \dot{\xi}(t) = A^+ \xi(t) + g^{-1}(A) B f(t), \quad \dot{\eta}(t) = A^- \eta(t) - g^{-1}(A) B f(t), \quad t > 0,$$

where

$$(2.21) \quad A^+ = -\alpha A + g(A), \quad A^- = -\alpha A - g(A).$$

Equation (2.20) can be rewritten as

$$(2.22) \quad \dot{\zeta}(t) = \mathcal{A} \zeta(t) + \mathcal{B} f(t), \quad t > 0,$$

where

$$(2.23) \quad \zeta(t) = \begin{bmatrix} \xi(t) \\ \eta(t) \end{bmatrix}, \quad \mathcal{A} = \begin{bmatrix} A^+ & 0 \\ 0 & A^- \end{bmatrix}, \quad \mathcal{B} = \begin{bmatrix} g^{-1}(A) B \\ -g^{-1}(A) B \end{bmatrix}.$$

3. Spectral property of \mathcal{A} and existence of solutions. We consider spectral property of the operators $A^\pm = -\alpha A \pm g(A)$. Let $u \in D(A)$. Then

$$(3.1) \quad A^\pm u = \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} (-\alpha \lambda_i \pm g(\lambda_i))(u, \phi_{ij}) \phi_{ij}.$$

Setting

$$(3.2) \quad \mu_i^+ = -\alpha \lambda_i + g(\lambda_i), \quad \mu_i^- = -\alpha \lambda_i - g(\lambda_i),$$

from (3.1) we obtain

$$(3.3) \quad A^\pm \phi_{ij} = \mu_i^\pm \phi_{ij}, \quad j = 1, \dots, m_i, \quad i = 1, 2, \dots.$$

Thus we see that μ_i^+ and μ_i^- are the eigenvalues of A^+ and A^- , respectively, and that ϕ_{ij} are the corresponding eigenfunctions of both A^+ and A^- .

From assumption (2.10), we see that

$$\alpha^2 \lambda_1^2 - \lambda_1 < -\frac{\lambda_1}{2} < 0.$$

Therefore there is an integer $\nu \geq 1$ such that

$$(3.4) \quad \alpha^2 \lambda_\nu^2 - \lambda_\nu < 0, \quad \alpha^2 \lambda_{\nu+1}^2 - \lambda_{\nu+1} > 0.$$

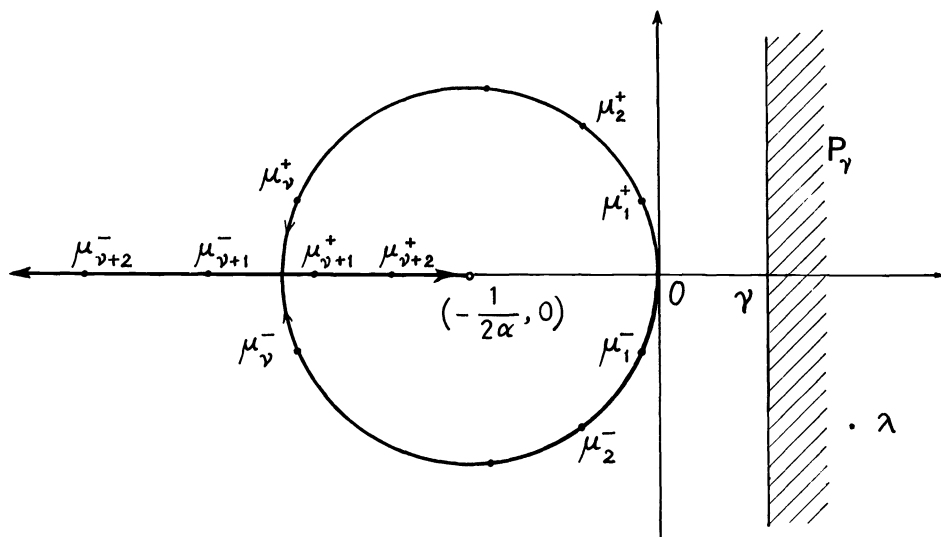
If $i \leq \nu$, μ_i^+ and μ_i^- are complex conjugates of each other. Since $\lambda_i \rightarrow \infty$ ($i \rightarrow \infty$),

$$(3.5) \quad \lim_{i \rightarrow \infty} \mu_i^+ = -\frac{1}{2\alpha}, \quad \lim_{i \rightarrow \infty} \mu_i^- = -\infty.$$

Also, it is easily seen that, if $i \leq \nu$, μ_i^+ and μ_i^- are located in the complex plane on the circle of radius $(1/2\alpha)$ with center at $-(1/2\alpha) + \sqrt{-1} \cdot 0$, and that μ_i^+ and μ_i^- are on the negative real axis, if $i > \nu$. Thus μ_i^+ and μ_i^- are distributed as shown in Fig. 1. It is clear that A^+ is a bounded operator.

LEMMA 3.1. *The operators A^+ and A^- generate analytic semigroups. Thus the operator \mathcal{A} also generates an analytic semigroup $e^{\mathcal{A}t}$, and the solution of (2.22) satisfying the initial condition*

$$(3.6) \quad \lim_{t \rightarrow +0} \zeta(t) = \zeta_0 = [\xi_0, \eta_0]^T,$$

FIG. 1. Distribution of eigenvalues of \mathcal{A} .

where $\xi_0, \eta_0 \in L^2(\Omega)$, can be uniquely expressed as

$$(3.7) \quad \zeta(t) = e^{\mathcal{A}t} \zeta_0 + \int_0^t e^{\mathcal{A}(t-s)} \mathcal{B}f(s) ds.$$

Proof. Since A^+ is a bounded operator, it is clear that A^+ generates an analytic semigroup. A necessary and sufficient condition for A^- to generate an analytic semigroup is that for each $\gamma > 0$ the half-plane $P_\gamma = \{\lambda: \operatorname{Re} \lambda > \gamma\}$ is contained in the resolvent set of A^- , and there exists a constant C such that the relation

$$(3.8) \quad \|(\lambda - A^-)^{-1}\| \leq C(1 + |\lambda - \gamma|)^{-1}$$

holds for any $\lambda \in P_\gamma$ [16, Remark 3.3.2]. In view of Fig. 1, since the spectrum of A^- is contained in the left half-plane $\{\lambda: \operatorname{Re} \lambda < 0\}$, it is clear that P_γ is contained in the resolvent set of A^- .

To prove (3.8), assume that $u \in D(A)$ and set $v = (\lambda - A^-)u$. Since $u \in L^2(\Omega)$ can be expressed as in (2.7), we see that

$$(3.9) \quad v = \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} (\lambda - \mu_i^-)(u, \phi_{ij}) \phi_{ij}.$$

From (3.9) we obtain

$$\|v\| \geq \inf_i |\lambda - \mu_i^-| \|u\|.$$

Thus, we see that

$$(3.10) \quad \|(\lambda - A^-)^{-1}\| = \sup \left(\frac{\|u\|}{\|v\|} \right) \leq \frac{1}{\inf_i |\lambda - \mu_i^-|}.$$

Therefore, what we have to prove is that

$$(3.11) \quad C \inf_i |\lambda - \mu_i^-| \geq (1 + |\lambda - \gamma|).$$

In view of Fig. 1 and (3.5), $\inf_i |\lambda - \mu_i^-|$ is attained on the circle or at the point $\mu_{\nu+1}^-$. Let $\inf_i |\lambda - \mu_i^-| = |\lambda - \mu_m^-|$. In view of Fig. 1, there is a constant β such that $|\mu_m^- - \gamma| \leq \beta$. Consequently,

$$(3.12) \quad |\lambda - \gamma| \leq |\lambda - \mu_m^-| + |\mu_m^- - \gamma| \leq |\lambda - \mu_m^-| + \beta.$$

It is clear that

$$(3.13) \quad |\lambda - \mu_m^-| \geq \gamma.$$

Thus, if

$$(3.14) \quad C \geq 1 + \frac{1+\beta}{\gamma},$$

then we see that

$$(3.15) \quad (C-1)|\lambda - \mu_m^-| \geq 1 + \beta.$$

Using (3.12) and (3.15), we obtain (3.11) for any $\lambda \in P_\gamma$. Therefore, if we choose a constant C satisfying (3.14), then (3.8) holds. Thus, it has been proved that the operator \mathcal{A} generates an analytic semigroup.

Since the function $f(t)$ is assumed to be Hölder continuous, there is a unique solution of (2.22) satisfying the initial condition (3.6) which is expressed as in (3.7) [13], [16]. \square

On the basis of Lemma 3.1, we obtain

THEOREM 3.1. *Given any $u_0 \in D(A)$ and $u_{t0} \in L^2(\Omega)$, there exists a unique solution of (2.1) satisfying the initial condition (2.4).*

Proof. Following (2.15), set

$$(3.16) \quad v_0 = g^{-1}(A)(u_{t0} + \alpha A u_0).$$

It is clear that $v_0 \in D(A)$. Let $\xi(t)$ and $\eta(t)$, $t \geq 0$, be the solution of (2.20) satisfying the initial conditions

$$\xi(0) = u_0 + v_0 \in D(A), \quad \eta(0) = u_0 - v_0 \in D(A),$$

respectively. From (2.19), $u(t)$ and $v(t)$, $t \geq 0$, can be defined by

$$(3.17) \quad u(t) = \frac{\xi(t) + \eta(t)}{2}, \quad v(t) = \frac{\xi(t) - \eta(t)}{2}.$$

It is clear that the functions $u(t)$ and $v(t)$ thus defined satisfy (2.16) and (2.18), respectively, and that (2.15) is obtained from (2.16).

Since the functions $u(t)$ and $v(t)$, $t > 0$, are analytic, differentiating (2.16), substituting (2.18) into it, and using (2.15) gives (2.1). Thus, we see that the function $u(t) = [\xi(t) + \eta(t)]/2$ is the unique solution of (2.1) satisfying the initial conditions

$$u(0) = u_0, \quad \dot{u}(0) = u_{t0},$$

where $u_0 \in D(A)$, $u_{t0} \in L^2(\Omega)$. \square

4. A finite-dimensional system, controllability and observability. Equation (2.1) is clearly stable, because it has a dissipative term $2\alpha A \dot{u}(t)$. However, since the damping coefficient α is usually very small, it is desirable to increase the degree of stability by feedback control.

Let us introduce the orthogonal projection operators P_n and Q_n in $L^2(\Omega)$ by

$$(4.1) \quad P_n u = \sum_{i=1}^n \sum_{j=1}^{m_i} (u, \phi_{ij}) \phi_{ij}, \quad Q_n u = (1 - P_n)u = \sum_{i=n+1}^{\infty} \sum_{j=1}^{m_i} (u, \phi_{ij}) \phi_{ij}.$$

Set

$$\tilde{P}_n = \begin{bmatrix} P_n & 0 \\ 0 & P_n \end{bmatrix}, \quad \tilde{Q}_n = \begin{bmatrix} Q_n & 0 \\ 0 & Q_n \end{bmatrix}.$$

Applying these operators to (2.22) gives

$$(4.2) \quad \tilde{P}_n \dot{\zeta}(t) = \mathcal{A} \tilde{P}_n \zeta(t) + \tilde{P}_n \mathcal{B} f(t),$$

$$(4.3) \quad \tilde{Q}_n \dot{\zeta}(t) = \mathcal{A} \tilde{Q}_n \zeta(t) + \tilde{Q}_n \mathcal{B} f(t).$$

The solution of the finite-dimensional equation (4.2) with the initial condition $\tilde{P}_n \zeta(0) = \tilde{P}_n \zeta_0 = \tilde{P}_n [\xi_0, \eta_0]^T$ can be expressed as

$$(4.4) \quad \tilde{P}_n \zeta(t) = \sum_{i=1}^n \sum_{j=1}^{m_i} \begin{bmatrix} \xi_{ij}(t) \\ \eta_{ij}(t) \end{bmatrix} \phi_{ij},$$

where $\xi_0, \eta_0 \in L^2(\Omega)$, and $\xi_{ij}(t)$ and $\eta_{ij}(t)$ are solutions of

$$(4.5) \quad \begin{aligned} \dot{\xi}_{ij}(t) &= \mu_i^+ \xi_{ij} + g^{-1}(\lambda_i) b_{ij} f(t), \\ \dot{\eta}_{ij}(t) &= \mu_i^- \eta_{ij} - g^{-1}(\lambda_i) b_{ij} f(t), \end{aligned}$$

satisfying the initial conditions $\xi_{ij}(0) = (\xi_0, \phi_{ij})$, $\eta_{ij}(0) = (\eta_0, \phi_{ij})$, respectively. In (4.5), b_{ij} is a row vector defined by

$$b_{ij} = [b_{ij}^1, \dots, b_{ij}^r],$$

where b_{ij}^k is given by $b_{ij}^k = (b^k, \phi_{ij})$.

Now suppose $i \leq \nu$. Since μ_i^+ and μ_i^- and $\xi_{ij}(0)$ and $\eta_{ij}(0)$ are pairs of complex conjugates, $\xi_{ij}(t)$ and $\eta_{ij}(t)$ are also complex conjugates. Let

$$\xi_{ij}(t) = \alpha_{ij}(t) + i\beta_{ij}(t).$$

Then (4.5) can be written in real form as

$$(4.6) \quad \begin{aligned} \dot{\alpha}_{ij}(t) &= -\alpha \lambda_i \alpha_{ij}(t) - \omega_i \beta_{ij}(t), & \dot{\beta}_{ij}(t) &= \omega_i \alpha_{ij}(t) - \alpha \lambda_i \beta_{ij}(t) - \frac{b_{ij}}{\omega_i} f(t), \end{aligned}$$

where

$$\omega_i = \sqrt{\lambda_i - \alpha^2 \lambda_i^2}.$$

Equation (4.6) is equivalent to

$$(4.7) \quad \dot{\gamma}_{ij}(t) = M_i \gamma_{ij}(t) - \frac{1}{\omega_i} \begin{bmatrix} 0 \\ b_{ij} \end{bmatrix} f(t),$$

where

$$\gamma_{ij}(t) = \begin{bmatrix} \alpha_{ij}(t) \\ \beta_{ij}(t) \end{bmatrix}, \quad M_i = \begin{bmatrix} -\alpha \lambda_i & -\omega_i \\ \omega_i & -\alpha \lambda_i \end{bmatrix}.$$

By introducing the following vectors and matrices

$$\gamma_i(t) = \begin{bmatrix} \gamma_{i1}(t) \\ \vdots \\ \gamma_{im_i}(t) \end{bmatrix}, \quad \hat{M}_i = \text{block diag} [M_i, \dots, M_i],$$

$$\hat{B}_i = \begin{bmatrix} 0 \\ b_{i1} \\ \vdots \\ 0 \\ b_{im_i} \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 \\ b_{i1}^1 & \cdots & b_{i1}^r \\ \vdots & & \vdots \\ 0 & \cdots & 0 \\ b_{im_i}^1 & \cdots & b_{im_i}^r \end{bmatrix},$$

(4.7) is rewritten as

$$(4.8) \quad \dot{\gamma}_i(t) = \hat{M}_i \gamma_i(t) - \frac{1}{\omega_i} \hat{B}_i f(t).$$

Next, suppose $i \geq \nu + 1$. Since μ_i^\pm are real numbers in this case, by introducing the following vectors and matrix

$$\xi_i(t) = \begin{bmatrix} \xi_{i1}(t) \\ \vdots \\ \xi_{im_i}(t) \end{bmatrix}, \quad \eta_i(t) = \begin{bmatrix} \eta_{i1}(t) \\ \vdots \\ \eta_{im_i}(t) \end{bmatrix},$$

$$\tilde{B}_i = \begin{bmatrix} b_{i1} \\ \vdots \\ b_{im_i} \end{bmatrix} = \begin{bmatrix} b_{i1}^1 & \cdots & b_{i1}^r \\ \vdots & & \vdots \\ b_{im_i}^1 & \cdots & b_{im_i}^r \end{bmatrix},$$

(4.5) can be written as

$$(4.9) \quad \begin{aligned} \dot{\xi}_i(t) &= \mu_i^+ \xi_i(t) + g^{-1}(\lambda_i) \tilde{B}_i f(t), \\ \dot{\eta}_i(t) &= \mu_i^- \eta_i(t) - g^{-1}(\lambda_i) \tilde{B}_i f(t). \end{aligned}$$

Let σ be any number such that

$$(4.10) \quad -\left(\frac{1}{2\alpha}\right) < -\sigma < -\alpha\lambda_1 < 0,$$

and let l be an integer such that $l \leq \nu$, and such that

$$(4.11) \quad \operatorname{Re} \mu_{l+1}^\pm < -\sigma \leq \operatorname{Re} \mu_l^\pm.$$

Let $n > l$, and let $L = m_1 + \cdots + m_l$, $N = m_1 + \cdots + m_n$. It is clear that $N > L$. Furthermore, let us define a $2L$ -dimensional vector $x_1(t)$, a $2L \times r$ matrix B_1 and a $2L \times 2L$ block diagonal matrix A_1 by

$$(4.12) \quad x_1(t) = \begin{bmatrix} \gamma_1(t) \\ \vdots \\ \gamma_l(t) \end{bmatrix}, \quad B_1 = - \begin{bmatrix} \hat{B}_1/\omega_1 \\ \vdots \\ \hat{B}_l/\omega_l \end{bmatrix},$$

$$A_1 = \text{block diag} [\hat{M}_1, \dots, \hat{M}_l].$$

Then from (4.8) we obtain

$$(4.13) \quad \dot{x}_1(t) = A_1 x_1(t) + B_1 f(t).$$

Similarly, let us define a $2(N-L)$ -dimensional vector $x_2(t)$, a $2(N-L) \times r$ matrix B_2 and a $2(N-L) \times 2(N-L)$ block diagonal matrix A_2 by

$$(4.14) \quad x_2(t) = \begin{bmatrix} \gamma_{l+1}(t) \\ \vdots \\ \gamma_\nu(t) \\ \xi_{\nu+1}(t) \\ \eta_{\nu+1}(t) \\ \vdots \\ \xi_n(t) \\ \eta_n(t) \end{bmatrix}, \quad B_2 = \begin{bmatrix} -\hat{B}_{l+1}/\omega_{l+1} \\ \vdots \\ -\hat{B}_\nu/\omega_\nu \\ g^{-1}(\lambda_{\nu+1})\tilde{B}_{\nu+1} \\ -g^{-1}(\lambda_{\nu+1})\tilde{B}_{\nu+1} \\ \vdots \\ g^{-1}(\lambda_n)\tilde{B}_n \\ -g^{-1}(\lambda_n)\tilde{B}_n \end{bmatrix},$$

$$(4.15) \quad A_2 = \text{block diag} [\hat{M}_{l+1}, \dots, \hat{M}_\nu, \mu_{\nu+1}^+ I_{m_{\nu+1}}, \mu_{\nu+1}^- I_{m_{\nu+1}}, \dots, \mu_n^+ I_{m_n}, \mu_n^- I_{m_n}],$$

where I_m denotes an $m \times m$ unit matrix. Then from (4.8) and (4.9) we obtain

$$(4.16) \quad \dot{x}_2(t) = A_2 x_2(t) + B_2 f(t).$$

Since $u(t) = [\xi(t) + \eta(t)]/2$, $P_n u(t)$ can be expressed as

$$P_n u(t) = \sum_{i=1}^n \sum_{j=1}^{m_i} \frac{1}{2} [\xi_{ij}(t) + \eta_{ij}(t)] \phi_{ij}.$$

In view of Fig. 1 and (4.5), the damping characteristic of zero-input response of (2.1) will be determined by the first finite number of eigenvalues μ_i^+ and μ_i^- of A^+ and A^- , respectively. In the case where the damping coefficient α is very small, the damping characteristic of (2.1) will be poor. Thus, our problem is to provide enough damping to the system by the feedback control. To be concrete, our problem is to place the closed-loop eigenvalues of (4.13) in some desirable way by using pole assignment theory [18]. Since the state vector $x_1(t)$ is not directly measurable, observers using the sensor outputs will be constructed to estimate the state variables.

Using the relation

$$u(t) = \frac{P_n(\xi(t) + \eta(t))}{2} + \frac{Q_n(\xi(t) + \eta(t))}{2},$$

the sensor output functions (2.2) are expressed as

$$y^k(t) = \sum_{i=1}^n \sum_{j=1}^{m_i} \frac{c_{ij}^k(\xi(t) + \eta(t))}{2} + \frac{(Q_n c^k, Q_n(\xi(t) + \eta(t)))}{2}, \quad k = 1, \dots, p,$$

where $c_{ij}^k = (c^k, \phi_{ij})$. By defining the matrices

$$\hat{C}_i = \begin{bmatrix} c_{i1}^1 & 0 & \dots & c_{im_i}^1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ c_{i1}^p & 0 & \dots & c_{im_i}^p & 0 \end{bmatrix}, \quad \tilde{C}_i = \begin{bmatrix} c_{i1}^1 & \dots & c_{im_i}^1 \\ \vdots & & \vdots \\ c_{i1}^p & \dots & c_{im_i}^p \end{bmatrix},$$

$$(4.17) \quad C_1 = [\hat{C}_1, \hat{C}_2, \dots, \hat{C}_l],$$

$$(4.18) \quad C_2 = [\hat{C}_{l+1}, \dots, \hat{C}_\nu, \frac{1}{2}\tilde{C}_{\nu+1}, \frac{1}{2}\tilde{C}_{\nu+1}, \dots, \frac{1}{2}\tilde{C}_n, \frac{1}{2}\tilde{C}_n],$$

the output vector function can be expressed as

$$(4.19) \quad y(t) = C_1 x_1(t) + C_2 x_2(t) + \frac{S_n Q_n(\xi(t) + \eta(t))}{2},$$

where $S_n Q_n(\xi(t) + \eta(t))/2$ is called the *observation spillover* [2], and the operator S_n mapping $Q_n L^2(\Omega)$ into R^p is defined by

$$(4.20) \quad S_n u = \begin{bmatrix} (Q_n c^1, u) \\ \vdots \\ (Q_n c^p, u) \end{bmatrix}, \quad u \in Q_n L^2(\Omega).$$

LEMMA 4.1. *The linear system (C_1, A_1, B_1) is controllable and observable if and only if*

$$(4.21) \quad \text{rank } \tilde{B}_i = \text{rank } \tilde{C}_i = m_i, \quad i = 1, \dots, l.$$

Proof. The linear system (4.13) is controllable if and only if the rows of $\exp(-A_1 t) B_1$ are linearly independent on $[0, \infty)$ [4, p. 177]. We see that

$$(4.22) \quad -\exp(-A_1 t) B_1 = \begin{bmatrix} \exp(-\hat{M}_1 t) \hat{B}_1 / \omega_1 \\ \vdots \\ \exp(-\hat{M}_l t) \hat{B}_l / \omega_l \end{bmatrix}.$$

Since

$$\exp(-M_i t) = e^{\alpha \lambda_i t} \begin{bmatrix} \cos \omega_i t & \sin \omega_i t \\ -\sin \omega_i t & \cos \omega_i t \end{bmatrix},$$

we obtain

$$(4.23) \quad \exp(-\hat{M}_i t) \hat{B}_i = e^{\alpha \lambda_i t} \begin{bmatrix} b_{i1} \sin \omega_i t \\ b_{i1} \cos \omega_i t \\ \vdots \\ b_{im_i} \sin \omega_i t \\ b_{im_i} \cos \omega_i t \end{bmatrix}.$$

Therefore, (A_1, B_1) is controllable if and only if the rows b_{i1}, \dots, b_{im_i} are linearly independent for each i . In other words, (A_1, B_1) is controllable if and only if the first equality of (4.21) holds for $i = 1, \dots, l$.

In the same way, (C_1, A_1) is observable if and only if the columns of $C_1 \exp(A_1 t)$ are linearly independent on $[0, \infty)$ [4, p. 188]. We see that

$$(4.24) \quad C_1 \exp(A_1 t) = [\hat{C}_1 \exp(\hat{M}_1 t), \dots, \hat{C}_l \exp(\hat{M}_l t)],$$

$$(4.25) \quad \hat{C}_i \exp(\hat{M}_i t) = e^{-\alpha \lambda_i t} [c_{i1} \cos \omega_i t, -c_{i1} \sin \omega_i t, \dots, c_{im_i} \cos \omega_i t, -c_{im_i} \sin \omega_i t],$$

where

$$c_{ij} = [c_{ij}^1, \dots, c_{ij}^p]^T.$$

Clearly, all columns of $C_1 \exp(A_1 t)$ are linearly independent on $[0, \infty)$ if and only if the columns c_{i1}, \dots, c_{im_i} are linearly independent for each i . Therefore, (C_1, A_1) is observable if and only if the second equality of (4.21) holds for $i = 1, \dots, l$.

Remark 4.1. In order that the rank conditions (4.21) hold, the number r of control inputs and the number p of sensors should satisfy

$$r, p \geq \max \{m_1, \dots, m_l\}.$$

Remark 4.2. It can be easily seen that the infinite-dimensional system (2.1) is controllable if and only if

$$\text{rank } \tilde{B}_i = m_i, \quad i = 1, 2, \dots.$$

Similarly, (2.1) is observable by the observation (2.3) if and only if

$$\text{rank } \tilde{C}_i = m_i, \quad i = 1, 2, \dots$$

These controllability and observability conditions coincide with the controllability and observability conditions for the first order diffusion system

$$\frac{du(t)}{dt} + Au(t) = Bf(t), \quad y(t) = Cu(t),$$

respectively, where the operators B and C are as defined above [14], [15].

5. Feedback control using observers. Let us construct two kinds of finite-dimensional observers defined by

$$(5.1) \quad \dot{z}_1(t) = (A_1 - G_1 C_1)z_1(t) + G_1[y(t) - C_2 z_2(t)] + B_1 f(t),$$

$$(5.2) \quad \dot{z}_2(t) = A_2 z_2(t) + B_2 f(t),$$

where $z_1(t)$ is a $2L$ -dimensional vector estimating $x_1(t)$, $z_2(t)$ is a $2(N-L)$ -dimensional vector estimating $x_2(t)$, and G_1 is a $2L \times p$ matrix to be determined.

It is clear from (4.16) and (5.2) that

$$(5.3) \quad x_2(t) - z_2(t) = e^{A_2 t}(x_{20} - z_{20}),$$

where $x_{20} = x_2(0)$, and $z_{20} = z_2(0)$. Let us define a $4L$ -dimensional vector $q_1(t)$ by

$$(5.4) \quad q_1(t) = \begin{bmatrix} x_1(t) \\ z_1(t) \end{bmatrix}.$$

Let the control input vector function $f(t)$ be given by

$$(5.5) \quad f(t) = F_1 z_1(t),$$

where F_1 is an $r \times 2L$ matrix to be determined. Substituting (5.5) into (4.3), (4.13) and (5.1), and using (4.19) and (5.3) yields

$$(5.6) \quad \frac{d}{dt} \begin{bmatrix} q_1(t) \\ \tilde{Q}_n \zeta(t) \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} q_1(t) \\ \tilde{Q}_n \zeta(t) \end{bmatrix} + \begin{bmatrix} \psi(t) \\ 0 \end{bmatrix},$$

where

$$(5.7) \quad \begin{aligned} A_{11} &= \begin{bmatrix} A_1 & B_1 F_1 \\ G_1 C_1 & A_1 - G_1 C_1 + B_1 F_1 \end{bmatrix}, \\ A_{12} &= \begin{bmatrix} 0 \\ G_1 S_n(I, I)/2 \end{bmatrix}, \\ A_{21} &= [0, \tilde{Q}_n B F_1], \quad A_{22} = \mathcal{A} \tilde{Q}_n, \end{aligned}$$

$$(5.8) \quad \psi(t) = \begin{bmatrix} 0 \\ G_1 C_2 e^{A_2 t}(x_{20} - z_{20}) \end{bmatrix}.$$

We see that

$$(5.9) \quad \begin{aligned} & \begin{bmatrix} I_L & 0 \\ -I_L & I_L \end{bmatrix} \begin{bmatrix} A_1 & B_1 F_1 \\ G_1 C_1 & A_1 - G_1 C_1 + B_1 F_1 \end{bmatrix} \begin{bmatrix} I_L & 0 \\ -I_L & I_L \end{bmatrix}^{-1} \\ &= \begin{bmatrix} A_1 + B_1 F_1 & B_1 F_1 \\ 0 & A_1 - G_1 C_1 \end{bmatrix}. \end{aligned}$$

Suppose the rank conditions (4.21) hold. Then the finite-dimensional system (C_1, A_1, B_1) is controllable and observable. Consequently, there exist matrices F_1 and G_1 such that all the eigenvalues of the matrices $A_1 + B_1 F_1$ and $A_1 - G_1 C_1$ take arbitrarily the preassigned values $\{-\mu_1, -\mu_2, \dots, -\mu_{4L}\}$ [12], [18]. Here, the real numbers $\mu_i > 0$ are selected in such a way that

$$(5.10) \quad -\mu_{4L} < \dots < -\mu_2 < -\mu_1 \leq \operatorname{Re} \mu_{l+1}^\pm.$$

Combining (4.11), and (5.10) yields

$$(5.11) \quad -\mu_1 \leq \operatorname{Re} \mu_{l+1}^\pm < -\sigma.$$

Let us construct the matrices F_1 and G_1 as stated above. In view of (5.7) and (5.9), we see that the matrix A_{11} is similar to the diagonal matrix $\operatorname{diag}[-\mu_1, -\mu_2, \dots, -\mu_{4L}]$, and the matrix $e^{A_{11}t}$ is also similar to the diagonal matrix $\operatorname{diag}[e^{-\mu_1 t}, e^{-\mu_2 t}, \dots, e^{-\mu_{4L} t}]$. In other words, there is a nonsingular matrix T_1 such that [1], [8]

$$(5.12) \quad T_1 e^{A_{11}t} T_1^{-1} = \operatorname{diag}[e^{-\mu_1 t}, \dots, e^{-\mu_{4L} t}].$$

From (5.12) we obtain

$$(5.13) \quad \|e^{A_{11}t}\| \leq M_1 e^{-\mu_1 t}, \quad t \geq 0,$$

where M_1 is the so-called condition number of the nonsingular matrix T_1 defined by

$$(5.14) \quad M_1 = \|T_1\| \|T_1^{-1}\| \geq 1.$$

Suppose that the integer n is selected in such a way that

$$(5.15) \quad \operatorname{Re} \mu_{n+1}^\pm \leq -\left(\frac{1}{2\alpha}\right).$$

Let $\tilde{Q}_n \zeta = [Q_n \xi, Q_n \eta]^T \in Q_n L^2(\Omega) \times Q_n L^2(\Omega)$. Since

$$A_{22} = \mathcal{A} \tilde{Q}_n = \begin{bmatrix} A^+ Q_n & 0 \\ 0 & A^- Q_n \end{bmatrix},$$

for $t \geq 0$ we obtain

$$(5.16) \quad e^{A_{22}t} \tilde{Q}_n \zeta = \sum_{i=n+1}^{\infty} \sum_{j=1}^{m_i} \begin{bmatrix} e^{\mu_i^+ t} & 0 \\ 0 & e^{\mu_i^- t} \end{bmatrix} \begin{bmatrix} (Q_n \xi, \phi_{ij}) \\ (Q_n \eta, \phi_{ij}) \end{bmatrix} \phi_{ij}.$$

From this equation, we see that

$$\|e^{A_{22}t}\| \leq \max \{e^{\operatorname{Re} \mu_i^+ t}, e^{\operatorname{Re} \mu_i^- t}, i \geq n+1\}.$$

In view of (3.5), (5.15) and Fig. 1, it follows that

$$(5.17) \quad \|e^{A_{22}t}\| \leq e^{-(1/2\alpha)t}, \quad t \geq 0.$$

Let us define the operators

$$(5.18) \quad \tilde{A} = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} 0 & A_{12} \\ A_{21} & 0 \end{bmatrix},$$

where \tilde{A} is unbounded, whereas \tilde{B} is bounded. Then

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \tilde{A} + \tilde{B}.$$

Let us introduce the vectors

$$(5.19) \quad w(t) = \begin{bmatrix} q_1(t) \\ \tilde{Q}_n \zeta(t) \end{bmatrix}, \quad \tilde{\psi}(t) = \begin{bmatrix} \psi(t) \\ 0 \end{bmatrix},$$

with the norm $\|w(t)\| = [\|q_1(t)\|^2 + \|\tilde{Q}_n \zeta(t)\|^2]^{1/2}$. Then (5.6) can be rewritten as

$$(5.20) \quad \dot{w}(t) = (\tilde{A} + \tilde{B})w(t) + \tilde{\psi}(t).$$

It is clear that

$$e^{\tilde{A}t} = \begin{bmatrix} e^{A_{11}t} & 0 \\ 0 & e^{A_{22}t} \end{bmatrix},$$

and that

$$\|e^{\tilde{A}t}\| \leq \max(\|e^{A_{11}t}\|, \|e^{A_{22}t}\|).$$

Since $M_1 \geq 1$ and $-\mu_1 \leq \operatorname{Re} \mu_{l+1}^\pm$, we see from (5.13) and (5.17) that

$$(5.21) \quad \|e^{\tilde{A}t}\| \leq M_1 e^{-\gamma t}, \quad t \geq 0,$$

where

$$(5.22) \quad -\gamma = \max \left\{ -\left(\frac{1}{2\alpha}\right), \operatorname{Re} \mu_{l+1}^\pm \right\}.$$

From (5.18) we see that

$$\|\tilde{B}\| \leq \max(\|A_{12}\|, \|A_{21}\|) \leq \max\left(\|\tilde{Q}_n \mathcal{B} F_1\|, \|G_1\| \frac{\|S_n\|}{\sqrt{2}}\right).$$

Also from (2.23) we obtain

$$\|\tilde{Q}_n \mathcal{B}\| \leq \sqrt{2} \sup_{i \geq n+1} |g^{-1}(\lambda_i)| \left(\sum_{k=1}^r \|Q_n b^k\|^2 \right)^{1/2}.$$

From (4.20) we obtain

$$\|S_n\| \leq \left(\sum_{k=1}^p \|Q_n c^k\|^2 \right)^{1/2}.$$

Therefore, we see that

$$(5.23) \quad \|\tilde{B}\| \leq \max \left\{ \sqrt{2} \|F_1\| \sup_i |g^{-1}(\lambda_i)| \left(\sum_{k=1}^r \|Q_n b^k\|^2 \right)^{1/2}, \|G_1\| \frac{(\sum_{k=1}^p \|Q_n c^k\|^2)^{1/2}}{\sqrt{2}} \right\}.$$

Now, applying the perturbation theory of semigroups [10, p. 495], [16, p. 71], we obtain

$$(5.24) \quad \|e^{(\tilde{A} + \tilde{B})t}\| \leq M_1 e^{-\beta_1 t}, \quad t \geq 0,$$

where

$$-\beta_1 = -\gamma + M_1 \|\tilde{B}\|.$$

Since $b^k \in L^2(\Omega)$ ($k = 1, \dots, r$), $c^k \in L^2(\Omega)$ ($k = 1, \dots, p$), and since M_1 is independent of n , for any small $\varepsilon > 0$ there is an integer n ($> l$) such that

$$(5.25) \quad M_1 \|\tilde{B}\| < \varepsilon.$$

By choosing n properly, we see from (5.11) that

$$(5.26) \quad -\mu_1 \leq -\gamma \leq -\beta_1 < -\sigma.$$

The solution of (5.20) is clearly written as

$$(5.27) \quad w(t) = e^{(\tilde{A} + \tilde{B})t} w_0 + \int_0^t e^{(\tilde{A} + \tilde{B})(t-s)} \tilde{\psi}(s) ds.$$

By using (5.24), $\|w(t)\|$ can be estimated as

$$(5.28) \quad \|w(t)\| \leq M_1 e^{-\beta_1 t} \left[\|w_0\| + \int_0^t e^{\beta_1 s} \|\tilde{\psi}(s)\| ds \right].$$

Let us estimate $\|\tilde{\psi}(s)\|$. From (4.15) we see that

$$(5.29) \quad \|e^{A_2 t}\| \leq e^{-\gamma t} \leq e^{-\beta_2 t}, \quad t \geq 0,$$

where $-\beta_2$ is any number such that

$$(5.30) \quad -\beta_1 < -\beta_2 < -\sigma.$$

Using (5.8) and (5.29), we obtain

$$(5.31) \quad \|\tilde{\psi}(t)\| \leq \|G_1\| \|C_2\| e^{-\beta_2 t} \|x_{20} - z_{20}\|.$$

Substituting (5.31) into (5.28) yields

$$(5.32) \quad \|w(t)\| \leq M_2 e^{-\beta_2 t}, \quad t \geq 0,$$

where

$$M_2 = M_1 [\|w_0\| + \|G_1\| \|C_2\| (\beta_1 - \beta_2)^{-1} \|x_{20} - z_{20}\|].$$

Let us introduce a $4(N-L)$ -dimensional vector $q_2(t)$ defined by

$$q_2(t) = \begin{bmatrix} x_2(t) \\ z_2(t) \end{bmatrix}.$$

From (4.16) and (5.2) we obtain

$$(5.33) \quad \dot{q}_2(t) = \hat{A} q_2(t) + \hat{B} z_1(t),$$

where

$$\hat{A} = \begin{bmatrix} A_2 & 0 \\ 0 & A_2 \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} B_2 F_1 \\ B_2 F_1 \end{bmatrix}.$$

Integrating (5.33) gives

$$(5.34) \quad q_2(t) = e^{\hat{A}t} q_{20} + \int_0^t e^{\hat{A}(t-s)} \hat{B} z_1(s) ds,$$

where $q_{20} = q_2(0)$. Using the estimates

$$(5.35) \quad \begin{aligned} \|e^{\hat{A}t}\| &\leq \|e^{A_2 t}\| \leq e^{-\beta_2 t} \leq e^{-\sigma t}, \quad t \geq 0, \\ \|\hat{B}\| &\leq \sqrt{2} \|F_1\| \|B_2\|, \\ \|z_1(t)\| &\leq \|w(t)\| \leq M_2 e^{-\beta_2 t}, \quad t \geq 0, \end{aligned}$$

we obtain

$$(5.36) \quad \|q_2(t)\| \leq e^{-\sigma t} [\|q_{20}\| + \sqrt{2} M_2 \|F_1\| \|B_2\| (\beta_2 - \sigma)^{-1}].$$

Putting (5.32) and (5.36) together, we finally obtain

$$(5.37) \quad \|w(t)\| \leq M_1[\|w_0\| + \sqrt{2}\|G_1\| \|C_2\|(\beta_1 - \beta_2)^{-1}\|q_{20}\|] e^{-\sigma t},$$

$$(5.38) \quad \|q_2(t)\| \leq M_1[\sqrt{2}\|F_1\| \|B_2\|(\beta_2 - \sigma)^{-1}\|w_0\| + \{1 + 2\|F_1\| \|G_1\| \|B_2\| \|C_2\|(\beta_2 - \sigma)^{-1}(\beta_1 - \beta_2)^{-1}\}\|q_{20}\|] e^{-\sigma t}$$

for $t \geq 0$. Define an infinite-dimensional vector $\tilde{w}(t)$ by

$$\tilde{w}(t) = \begin{bmatrix} q_1(t) \\ q_2(t) \\ \tilde{Q}_n \xi(t) \end{bmatrix} \in R^{4N} \times Q_n L^2(\Omega) \times Q_n L^2(\Omega).$$

It is obvious that $\tilde{w}(t)$ represents the state of the distributed system as well as the state of the dynamic controller. From (5.37) and (5.38) we see that

$$(5.39) \quad \|\tilde{w}(t)\| \leq K e^{-\sigma t} \|\tilde{w}(0)\|, \quad t \geq 0,$$

where K is a constant dependent on l, n , etc.

Thus we can summarize what we have discussed so far as follows:

THEOREM 5.1. *Given an arbitrary damping constant σ such that $0 < \alpha\lambda_1 < \sigma < (1/2\alpha)$, suppose that the rank conditions (4.21) hold, where l is an integer satisfying (4.11). Then a finite-dimensional feedback dynamic controller described by (5.1), (5.2) and (5.5) can be constructed in such a way that the state $\tilde{w}(t)$ of the overall system satisfies (5.39), where K is a constant dependent on l, n , etc.*

Remark 5.1. If we know the eigenvalues λ_i and the eigenfunctions $\phi_{ij}(\cdot)$ of the operator A , the control influence functions $b^k(\cdot) \in L^2(\Omega)$, $k = 1, \dots, r$, and the sensor influence functions $c^k(\cdot) \in L^2(\Omega)$, $k = 1, \dots, p$, then by specifying the integers l and n , we can calculate the matrices A_1, B_1, C_1 and A_2, B_2, C_2 . Therefore, it is possible to construct the observers O_1 and O_2 governed by (5.1) and (5.2), respectively. The structure of the dynamic compensator for the distributed parameter system (D.P.S.) is as shown in Fig. 2, where the matrices G_1 and F_1 are selected in such a way that the matrices $A_1 + B_1 F_1$ and $A_1 - G_1 C_1$ have the preassigned eigenvalues.

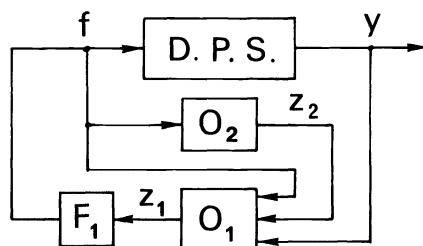


FIG. 2. Block diagram of dynamic compensator.

Remark 5.2. An algorithm for determining the matrix F_1 such that the matrix $A_1 + B_1 F_1$ has the preassigned eigenvalues $\{-\mu_1, \dots, -\mu_{2L}\}$, where $\mu_i \neq \mu_j$ ($i \neq j$) and each μ_i does not coincide with the eigenvalues of A_1 , is as follows.

1. Seek a set of r -dimensional vectors $\{\xi_i\}$ such that the vectors

$$(5.40) \quad v_i = -(\mu_i I + A_1)^{-1} B_1 \xi_i, \quad i = 1, \dots, 2L,$$

are linearly independent. It has been proved that such a set of vectors $\{\xi_i\}$ exists [20].

2. The matrix F_1 is given by

$$(5.41) \quad F_1 = [\xi_1, \dots, \xi_{2L}][v_1, \dots, v_{2L}]^{-1}.$$

It is clear that the vectors v_i are eigenvectors of $A_1 + B_1 F_1$ corresponding to the eigenvalues $-\mu_i$, because from (5.40) and (5.41) we obtain

$$(\mu_i I + A_1 + B_1 F_1)v_i = 0.$$

The matrix G_1 such that the matrix $A_1 - G_1 C_1$ has the preassigned eigenvalues can be determined by using the same algorithm.

Remark 5.3. The solutions $u(t)$ and $v(t)$ of (2.16) and (2.18) are expressed as

$$(5.42) \quad \begin{aligned} u(t) &= \sum_{i=1}^{\nu} \sum_{j=1}^{m_i} \alpha_{ij}(t) \phi_{ij} + \frac{1}{2} \sum_{i=\nu+1}^{\infty} \sum_{j=1}^{m_i} (\xi_{ij}(t) + \eta_{ij}(t)) \phi_{ij}, \\ v(t) &= \sqrt{-1} \sum_{i=1}^{\nu} \sum_{j=1}^{m_i} \beta_{ij}(t) \phi_{ij} + \frac{1}{2} \sum_{i=\nu+1}^{\infty} \sum_{j=1}^{m_i} (\xi_{ij}(t) - \eta_{ij}(t)) \phi_{ij}, \end{aligned}$$

where $\alpha_{ij}(t)$ and $\beta_{ij}(t)$ are solutions of (4.6), and $\xi_{ij}(t)$ and $\eta_{ij}(t)$ are solutions of (4.5). From (5.42) we see that

$$(5.43) \quad \|u(t)\|^2 + \|v(t)\|^2 = \sum_{i=1}^{\nu} \sum_{j=1}^{m_i} [\alpha_{ij}^2(t) + \beta_{ij}^2(t)] + \sum_{i=\nu+1}^{\infty} \sum_{j=1}^{m_i} \frac{1}{2} [\xi_{ij}^2(t) + \eta_{ij}^2(t)].$$

Therefore,

$$(5.44) \quad \begin{aligned} (\|u(t)\|^2 + \|v(t)\|^2)^{1/2} &\leq (\|x_1(t)\|^2 + \|x_2(t)\|^2 + \|\tilde{Q}_n \zeta(t)\|^2)^{1/2} \\ &\leq \|\tilde{w}(t)\| \leq K e^{-\sigma t} \|\tilde{w}(0)\|. \end{aligned}$$

Remark 5.4. The bounded operator \tilde{B} defined by (5.18) results from the control and observation spillovers [2]. If

$$\begin{aligned} b^k(\cdot) &\in P_n L^2(\Omega), & k &= 1, \dots, r, \\ c^k(\cdot) &\in P_n L^2(\Omega), & k &= 1, \dots, p, \end{aligned}$$

for some integer n , then $\tilde{B} = 0$.

Remark 5.5. If $l = n$, (5.25) does not hold. Because in this case the constant M_1 depends on $l = n$ and $M_1 \rightarrow \infty$ ($l \rightarrow \infty$), in general. Thus, boundedness of $M_1 \|\tilde{B}\|$ with respect to n is not clear. The key point of this paper lies in the introduction of two kinds of observers. In the first observer equation (5.1), since the output $y(t)$ contains the observation spillover, its effect has been reduced by subtracting $C_2 z_2(t)$ from $y(t)$, where $z_2(t)$ is the output of the second observer (5.2).

Remark 5.6. In this paper, an identity observer has been used for estimating the state of the first L modes of the distributed system. It is also possible to construct a feedback dynamic controller by use of a reduced order observer [12].

Remark 5.7. From (4.10) we see that the upper limit of the decay rate σ by the feedback control is $(1/2\alpha)$. Therefore, the smaller the damping constant α of the original distributed system is, the larger is the decay rate σ which can be obtained by the feedback.

6. Examples.

6.1. Active tendon control of structures. We consider a slender structure that can be modeled as a uniform cantilever beam. The equation of motion of slender and

flexible structures with internal viscous damping of the Voigt type can be written as [19]

$$(6.1) \quad \frac{\partial^2 u(t, x)}{\partial t^2} + 2\alpha \frac{\partial^5 u(t, x)}{\partial t \partial x^4} + \frac{\partial^4 u(t, x)}{\partial x^4} = b(x)f(t),$$

where $0 < x < L$. The structure is clamped at $x = 0$, and it is free at the other end $x = L$, where the bending moment and the shearing force vanish. Therefore, the boundary conditions are given by

$$(6.2) \quad \begin{aligned} u(t, 0) &= \frac{\partial u(t, 0)}{\partial x} = 0, \\ \frac{\partial^2 u(t, L)}{\partial x^2} + 2\alpha \frac{\partial^3 u(t, L)}{\partial x^2 \partial t} &= 0, \\ \frac{\partial^3 u(t, L)}{\partial x^3} + 2\alpha \frac{\partial^4 u(t, L)}{\partial x^3 \partial t} &= 0. \end{aligned}$$

The output from a sensor is given by

$$(6.3) \quad y(t) = \int_0^L c(x)u(t, x) dx.$$

Now we define an operator A in $L^2(0, L)$ by

$$(6.4) \quad D(A) = \{u: u \in H^4(0, L), u(0) = u'(0) = 0, u''(L) = u'''(L) = 0\},$$

$$(6.5) \quad Au = \frac{\partial^4 u}{\partial x^4},$$

where $H^4(0, L)$ denotes the fourth order *Sobolev space* on $(0, L)$, and a prime denotes the derivative with respect to x . All the derivatives in (6.4) and (6.5) are taken in the sense of distribution on $(0, L)$. Thus (6.1) together with (6.2) can be expressed as (2.1). If $u, v \in D(A)$, then we see that

$$(Au, v) = (u'', v'') = (u, Av).$$

Therefore, A is a symmetric operator.

Let us consider the eigenvalue problem

$$(6.6) \quad (A - \lambda I)u = 0, \quad u \in D(A).$$

By the elementary computation [7], we see that (6.6) has a nonzero solution if and only if $\lambda = (\beta/L)^4$, where β is a positive solution of

$$(6.7) \quad \cosh \beta \cos \beta + 1 = 0.$$

Let β_i be the solutions of (6.7) such that $0 < \beta_1 < \beta_2 < \dots$. Then the eigenvalues of A are given by

$$(6.8) \quad \lambda_i = \left(\frac{\beta_i}{L}\right)^4, \quad i = 1, 2, \dots.$$

It is easy to see that the corresponding normalized eigenfunctions are given by

$$(6.9) \quad \phi_i(x) = \left[\cosh\left(\frac{\beta_i x}{L}\right) - \cos\left(\frac{\beta_i x}{L}\right) - \gamma_i \left(\sinh\left(\frac{\beta_i x}{L}\right) - \sin\left(\frac{\beta_i x}{L}\right) \right) \right] / \sqrt{L},$$

where

$$\gamma_i = \frac{\cosh \beta_i + \cos \beta_i}{\sinh \beta_i + \sin \beta_i}.$$

Since all the eigenvalues of A are positive, A is clearly a selfadjoint positive-definite operator with the domain $D(A)$ dense in $L^2(0, L)$. Also, because of the existence of the Green's function [6, p. 1330], A^{-1} is compact. If

$$(6.10) \quad b_i = (b, \phi_i) \neq 0, \quad c_i = (c, \phi_i) \neq 0, \quad i = 1, 2, \dots, l,$$

then the result of Theorem 5.1 can be applied to the system (6.1).

In this example, $m_i = 1$ for all i . The solution of (6.1) is expressed as

$$(6.11) \quad u(t, x) = \sum_{i=1}^{\nu} \alpha_i(t) \phi_i(x) + \sum_{i=\nu+1}^{\infty} \frac{1}{2} [\xi_i(t) + \eta_i(t)] \phi_i(x).$$

In (6.11), $\alpha_i(t)$ are the solutions of

$$(6.12) \quad \frac{d}{dt} \begin{bmatrix} \alpha_i(t) \\ \beta_i(t) \end{bmatrix} = M_i \begin{bmatrix} \alpha_i(t) \\ \beta_i(t) \end{bmatrix} - \begin{bmatrix} 0 \\ b_i/\omega_i \end{bmatrix} f(t),$$

where

$$\omega_i = \sqrt{\lambda_i - \alpha^2 \lambda_i^2} \quad \text{and} \quad M_i = \begin{bmatrix} -\alpha \lambda_i & -\omega_i \\ \omega_i & -\alpha \lambda_i \end{bmatrix},$$

and $\xi_i(t), \eta_i(t)$ are the solutions of

$$(6.13) \quad \begin{aligned} \dot{\xi}_i &= \mu_i^+ \xi_i + g^{-1}(\lambda_i) b_i f(t), \\ \dot{\eta}_i &= \mu_i^- \eta_i - g^{-1}(\lambda_i) b_i f(t), \end{aligned}$$

where $g(\lambda) = \sqrt{\alpha^2 \lambda^2 - \lambda}$ and $\mu_i^\pm = -\alpha \lambda_i \pm g(\lambda_i)$. The matrices are given as follows:

$$A_1 = \text{block diag} [M_1, \dots, M_l], \quad B_1 = - \begin{bmatrix} 0, \frac{b_1}{\omega_1}, \dots, 0, \frac{b_l}{\omega_l} \end{bmatrix}^T, \quad C_1 = [c_1, 0, \dots, c_l, 0],$$

$$A_2 = \text{block diag} [M_{l+1}, \dots, M_\nu, \mu_{\nu+1}^+, \mu_{\nu+1}^-, \dots, \mu_n^+, \mu_n^-].$$

$$B_2 = \begin{bmatrix} 0, -\frac{b_{l+1}}{\omega_{l+1}}, \dots, 0, -\frac{b_\nu}{\omega_\nu}, g^{-1}(\lambda_{\nu+1}) b_{\nu+1}, \\ -g^{-1}(\lambda_{\nu+1}) b_{\nu+1}, \dots, g^{-1}(\lambda_n) b_n, -g^{-1}(\lambda_n) b_n \end{bmatrix}^T,$$

$$C_2 = [c_{l+1}, 0, \dots, c_\nu, 0, \frac{1}{2} c_{\nu+1}, \frac{1}{2} c_{\nu+1}, \dots, \frac{1}{2} c_n, \frac{1}{2} c_n].$$

By using these matrices, the observers O_1 and O_2 governed by (5.1) and (5.2) can be constructed, respectively.

6.2. Strongly damped wave equation. Our next example is the strongly damped wave equation

$$(6.14) \quad u_{tt} - 2\alpha \Delta u_t - \Delta u = \sum_{k=1}^r b^k(x) f^k(t),$$

where Δ denotes the Laplacian on a bounded domain Ω . The boundary condition is assumed to be of the Dirichlet type

$$(6.15) \quad u(t, x) = 0, \quad x \in \Gamma,$$

where Γ is a sufficiently smooth boundary of Ω . In this example, the operator A is defined by

$$(6.16) \quad D(A) = \{u : u \in H_0^2(\Omega)\},$$

$$(6.17) \quad Au = -\Delta u.$$

Here $H_0^2(\Omega)$ is the closure of $C_0^\infty(\Omega)$ in the Sobolev space $H^2(\Omega)$, $C_0^\infty(\Omega)$ being the space of infinitely differentiable functions with compact support on Ω . It is well-known that A is a selfadjoint positive-definite operator with the dense domain in $L^2(\Omega)$ and that A^{-1} is compact [13, p. 152]. Thus, (6.14) together with (6.15) can be expressed as (2.1).

Acknowledgment. The author wishes to thank Dr. N. Fujii for his valuable discussions.

REFERENCES

- [1] F. AYRES, JR., *Matrices*, Schaum's Outline Series, McGraw-Hill, New York, 1962.
- [2] M. J. BALAS, *Modal control of certain flexible dynamic systems*, this Journal, 16 (1978), pp. 450–462.
- [3] ———, *Feedback control of flexible systems*, IEEE Trans. Automat. Contr., AC-23 (1978), pp. 673–679.
- [4] C. T. CHEN, *Introduction to Linear System Theory*, Holt, Rinehart and Winston, New York, 1970.
- [5] R. F. CURTAIN, *Finite-dimensional compensator design for parabolic distributed systems with point sensors and boundary input*, IEEE Trans. Automat. Contr., AC-27 (1982), pp. 98–104.
- [6] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part II*, Interscience, New York, 1963.
- [7] S. I. GAIDUK, *A problem on the transverse oscillations of a viscoelastic rod*, Differential Equations (English translation from Russian), 3 (1967), pp. 793–801.
- [8] F. R. GANTMACHER, *The Theory of Matrices, Vol. I*, Chelsea, New York, 1960.
- [9] J. S. GIBSON, *An analysis of optimal modal regulation: convergence and stability*, this Journal, 19 (1981), pp. 686–707.
- [10] T. KATO, *Perturbation Theory for Linear Operators*, Springer, Berlin, 1966.
- [11] S. G. KREIN, *Linear Differential Equations in Banach Space*, American Mathematical Society, Providence, 1971.
- [12] D. G. LUENBERGER, *An introduction to observers*, IEEE Trans. Automat. Contr., AC-16 (1971), pp. 596–602.
- [13] S. MIZOHATA, *The Theory of Partial Differential Equations*, Cambridge Univ. Press, Cambridge, 1973.
- [14] Y. SAKAWA, *Controllability for partial differential equations of parabolic type*, this Journal, 12 (1974), pp. 389–400.
- [15] ———, *Observability and related problems for partial differential equations of parabolic type*, this Journal, 13 (1975), pp. 14–27.
- [16] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.
- [17] A. E. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York, 1958.
- [18] W. H. WONHAM, *On pole assignment in multiple-input controllable linear systems*, IEEE Trans. Automat. Contr., AC-12 (1967), pp. 660–665.
- [19] J. N. YANG AND F. GIANNPOULOS, *Active tendon control of structures*, J. Engineering Mechanics Division, ASCE, 104 (1978), pp. 551–568.
- [20] H. KIMURA, *Pole assignment by gain output feedback*, IEEE Trans. Automat. Contr., AC-20 (1975), pp. 509–516.

OUTPUT FEEDBACK AND GENERIC STABILIZABILITY*

C. I. BYRNES[†] AND B. D. O. ANDERSON[‡]

Abstract. We consider questions of pole placement and stabilization for generic linear systems with prescribed state, input and output dimensions, where the controller must be implemented by linear memoryless output feedback. We present a criterion, in terms of a special pole placement property, for generic stabilizability and apply this to describe constraints on the dimensions which are consistent with generic stabilizability. We also discuss the rationality and solvability by radicals of stabilizing or pole positioning gains, and we describe how decision algebra can theoretically handle existence questions for generic systems.

Key words. multivariable control, output feedback, stabilizability of multivariable systems, decision algebra, Galois theory, solvability of pole placement equations by radicals

1. Introduction. In this paper, we are concerned with questions of pole assignability and stabilizability for real linear input-output systems

$$(1.1) \quad \frac{dx}{dt} = Fx + Gu, \quad y = Hx,$$

or

$$(1.1)' \quad x(t+1) = Fx(t) + Gu(t), \quad y(t) = Hx(t)$$

where we allow constant gains $u = Ky$ as feedback. The equations of pole assignability are real polynomials, and it is natural to attempt to solve these equations by eliminating the unknown variable K . Similar remarks apply to the equations of stabilizability which include, however, algebraic inequalities arising for example from the Routh-Hurwitz criteria. In what follows, we shall use various results from classical algebraic geometry, including elimination theory and the Schubert calculus of enumerative geometry, which apply to the equations of pole placement.

Put geometrically, elimination theory consists in the study of a projection

$$(1.2) \quad p_1 : X \times Y \rightarrow K$$

restricted to an algebraic, or semialgebraic, set $Z \subset X \times Y$, where X and Y can be taken to real or complex vector spaces, e.g. $X = \mathbb{R}^N$, $Y = \mathbb{R}$. The main problem in elimination theory consists in finding a description of the set

$$p_1(Z) = \{x : \exists y \text{ such that } (x, y) \in Z\}$$

in terms of Z . A basic example is given by

$$(1.3) \quad Z = \{(x, y) : x = y^2\},$$

which is algebraic but for which $p_1(Z)$ is only semialgebraic if we take real coefficients.¹

* Received by the editors September 25, 1981, and in revised form January 14, 1983. This research was partially supported by: the National Aeronautics and Space Administration, grant NSG-2265, the National Science Foundation grant ENG-79-09459 and the Australian Research Grants Committee.

[†] Department of Mathematics and Division of Applied Sciences, Harvard University, Cambridge, Massachusetts 02138. Written while this author was a Gast Professor an der Universität Bremen.

[‡] Department of Electrical and Computer Engineering, University of Newcastle, New South Wales, 2308, Australia. Current address: Department of Systems Engineering, Institute of Advanced Studies, Australian National University, Canberra, A.C.T. 2600, Australia.

¹ Semialgebraic sets are defined in (3.4), § 3.

In relation to the pole assignability question for a prescribed F, G, H , we can identify the entries of K with the space Y and the coefficients of the closed loop characteristic polynomial, call them p_1, \dots, p_n , with X . Then

$$Z = \left\{ (p_1, \dots, p_n, K) : \det(sI - F - GKH) = s^n + \sum_{i=1}^n p_i s^{n-i} \right\},$$

and pole assignability of a generic closed-loop polynomial holds if and only if $p_1(Z)$ coincides with all of R^n save a proper subvariety.

Among the results we obtain using classical algebraic geometry is: the condition $mp \leq n$ is necessary for the stabilizability of the generic (F, G, H) . This condition is well known to be necessary for pole assignability of the generic (F, G, H) , and our result raises the question as to whether or not, in terms of the values m, n, p , these two questions might not be equivalent. As unlikely as this may be, at the time we write there is no counterexample (although there is evidence in this direction for $m = 2, n = 9$, and $p = 6$, see [5]). We also show that if a stabilizing gain exists, then such a gain can be found by a rational procedure. On the other hand, we show that if $mp = n$, a rational procedure for finding a gain K which assigns a given characteristic polynomial (assuming such a K exists) does not exist unless $\min(m, p) = 1$, in which case a linear formula can be found. Moreover, using square roots as well as rational operations only helps if $\min(m, p) = \max(m, p) = 2$. This is of course in contrast with pole assignment by state feedback, and answers in the negative a question raised in [1].

We also argue that one can in principle determine by rational calculations whether, given m, n, p , generic F, G, H are pole assignable, generically pole assignable, or stabilizable. We say "in principle" since the number of calculations required is enormous.

We use several tools to prove the results. One of the theorems, due to Tarski-Seidenberg, asserts that if Z is semialgebraic, then $p_1(Z)$ is semialgebraic. This theorem can be used iteratively to reduce the question of the existence of a solution $x \in \mathbb{R}^n$ to a set of semialgebraic equations to the question of existence of a solution to another set of semialgebraic equations in, for example, the unknown $x_1 \in \mathbb{R}$. Such existence can be decided by a rational procedure in the coefficients of the resulting semialgebraic equations. The Tarski-Seidenberg theorem is extremely qualitative, and "worst-case" analysis [7] shows that such a decision procedure takes at least 2^{k^n} steps, where $k > 0$ is a constant and n is the length of the input formula.

We also use a classical form of elimination theory, over \mathbb{C} : if $Z \subset \mathbb{C}^N \times \mathbb{C}^M$ is defined by equations which are homogeneous in y , then $p_1(Z) \subset \mathbb{C}^N$ is definable by polynomial equations. In particular, $p_1(Z)$ is closed.

A topological form of this elimination theorem also holds over \mathbb{R} and is crucial in showing that (for $mp \leq n$) the image of the pole-placement map is Euclidean closed in \mathbb{R}^n for the generic system [4]. Our proof of Theorem 1 relies on this result.

We must also use rather explicit elimination arguments which have appeared in the literature. Among these are the works by Willems-Hesselink [23] and, more recently, Morse-Wolovich-Anderson [19] which treat the case $m = p = 2$, and $m = 2, p = 3$. These authors, after considerable calculation, obtain a single explicit equation in a single unknown, and it is possible to obtain some quantitative and qualitative results from the form of the equations. Finally, we use the results of Brockett-Byrnes [3] who determined the degree of this equation, for general m, p , using methods of the Schubert calculus. This calculus was developed in the 19th century in order to deduce the degree of the final equation one would obtain in certain problems of enumerative geometry, without going through the elimination theory first. It is a

fortunate fact that the return difference equation corresponds to a classical equation of enumerative geometry, enabling one to determine this degree as a function of m and p .

2. Statements of the main results. Let us suppose that (F, G, H) is a triple of matrices which correspond to either a discrete or a continuous time system having m inputs, n states and p outputs. We consider the questions, for m, n, p fixed:

Question 1. Is it true that for all (F, G, H) , except perhaps those contained in a proper algebraic set, one can arbitrarily assign the (closed-loop) eigenvalues of $F + GKH$ by suitable choice of output feedback K ?

Question 2. Is it true that one can stabilize all (F, G, H) , except perhaps those contained in a proper algebraic set, by some output feedback K ?

Concerning Question 1, it is known [13], [23] that $mp \geq n$ is a necessary condition on the parameters m, n, p . In § 3 we derive a stabilizability criterion as a limiting form of the equivalence of generic stabilizability for continuous and for discrete time systems. This can be thought of as an equivalence between generic stabilizability and the generic existence to an output feedback deadbeat control problem for nondegenerate systems (in the sense of [3], [4]):

THEOREM 1. *If $mp \leq n$, the following statements are equivalent:*

- (i) *m, n, p are such that the generic (F, G, H) is stabilizable;*
- (ii) *m, n, p are such that for any nondegenerate (F, G, H) there exists a gain K such that the closed loop polynomial is s^n .*

This result holds for $mp > n$ as well, with “nondegenerate” replaced by the weaker term “generic”. Since we do not need the general result here, we shall only prove it in the case $mp \leq n$. From Theorem 1 we obtain

THEOREM 2. *$mp \geq n$ is necessary for generic stability.*

This result of course implies that $mp \geq n$ is necessary for Question 1 as well, but also raises the question as to whether the answers to Questions 1 and 2 might not agree, as functions of the parameters m, n , and p . On the one hand, if $\max(m, p) \geq n$ then generically either G or H is of rank n so that one is in the state feedback situation where the answer to Question 1, and therefore to Question 2, is well known to be in the affirmative under the generic hypothesis of reachability. On the other extreme, Theorem 2 shows that for $mp < n$ the answer to both questions is in the negative, so that explicit calculations for $mp \sim n$ are therefore quite interesting. However, aside from a few special cases, our knowledge is incomplete.

Example 1. ($m = p = 2$). If $n = 4$, it has been shown by Willems–Hesselink [23] that pole placement does not hold for an open subset of (F, G, H) . In [3] it is shown that pole placement does not hold unless the transfer function $T(s) = H(sI - F)^{-1}G$ has rank 1. In particular, pole placement does not hold for (F, G, H) in an open, dense set. In [19], necessary and sufficient conditions for generic pole placement for a particular system of this dimension are derived.

Thus, by Kimura’s theorem [16] and the Willems–Hesselink counterexample, the answer to Question 1 is yes if, and only if, $n \leq 3$. In [23] it is asserted that a modification by P. Molander of the techniques in [23] shows that the answer to Question 2 is in the negative if $n = 4$. Thus, the answer to both Questions 1 and 2, if $m = p = 2$, is yes if, and only if, $n \leq 3$. Since this result is unpublished, in § 4 we present a verification of Molander’s conclusion as a corollary to our generic stabilizability criterion. This of course gives another proof of the Willems–Hesselink theorem.

THEOREM 3 (Molander). *There is a nonempty open set of (nondegenerate) 2×2 systems of degree 4 which are not stabilizable by constant output gain feedback.*

Example 2. ($m = 2$, $p = 2^k - 1$). It is known in this case that the answer to Question 1, and therefore to Question 2, is in the affirmative [3] provided $mp \geq n$. By Theorem 2, the answer to both questions, for these values of m, p , is therefore yes if, and only if, $mp \geq n$.

Example 3. ($m = 2$, $p = 4$). At present, one is able to deduce from the results proved in [3] and more refined topological methods that the answer to Question 1, and therefore to Question 2, is in the affirmative whenever $n \leq 7$. Theorem 2 then asserts that the only case which remains to be analyzed is $n = 8$, where it has been conjectured [6] that the answer to Question 1 is in the negative.

We should mention, however, that there are cases (e.g. $m = 2$, $p = 6$, $n = 9$) where generic stabilizability is known to hold, but where Question 1 remains unanswered [5].

Until now, we have only discussed the existence of solutions to the problem of pole positioning and stabilization. Equally important is the consideration of what kind of algorithm might exist for finding a gain K which places the poles, or stabilizes the system, provided such a gain exists. In §§ 5 and 6 we analyze each of these questions and prove

THEOREM 4. *Suppose there exists a gain which stabilizes the system (F, G, H) . Then, one can find such a K by an algorithm which is rational in the coefficients of (F, G, H) .*

In [1] the question was raised as to whether rational formulae exist for a gain K which places the closed loop characteristic polynomial at $p(s) = s^n + p_1 s^{n-1} + \dots + p_n$. That is, provided such a gain K exists, can one find K as a rational function of $(F, G, H, p_1, \dots, p_n)$? This holds for the case of state feedback and, in particular, where $\min(m, p) = 1$ and $\max(m, p) \geq n$. In this case, a linear formula for K follows from consideration of the phase-variable canonical form. However, as the equation obtained by Willems–Hesselink (see also [3], [19]) shows for the case $m = p = 2$, $n = 4$, there exist precisely 2 gains (possibly a complex conjugate pair) counted with multiplicity which place a given real monic polynomial

$$s^4 + p_1 s^3 + \dots + p_4.$$

Moreover, the coefficients of such a 2×2 gain K are given by the solution formula for a quadratic equation. Thus, in general, a rational formula does not exist. If $mp = n$, we can give a more precise answer to the question raised in [1]:

THEOREM 5. *If $mp = n$, the following statements are equivalent for the generic (F, G, H) and monic polynomial $p(s)$:*

- (a) *There exists a rational formula, in the coefficients of $p(s)$ and entries of (F, G, H) , for some K which places the closed loop polynomial at $p(s)$;*
- (b) *There exists a linear formula, in the coefficients of $p(s)$ and entries of (F, G, H) , for such a K ;*
- (c) $\min(m, p) = 1$ and $\max(m, p) = n$.

THEOREM 6. *If $mp = n$, the following statements are equivalent for the generic (F, G, H) and monic polynomial $p(s)$:*

- (a) *There exists a formula, involving rational expressions and square roots, for some K which places the closed loop polynomial at $p(s)$;*
- (b) *Either $\min(m, p) = 1$ or $\min(m, p) = \max(m, p) = 2$.*

Indeed, if $mp = n$ we conjecture that the only cases for which there exists formulae for K involving rational operations and radicals are

- (i) $\min(m, p) = 1$ and $\max(m, p) = n$; or
- (ii) $\min(m, p) = \max(m, p) = 2$.

This conjecture appears natural in the light of our techniques (§ 6), which are an application of Galois theory and of the methods used in [3] enabling one to express the number $d_{m,p}$ of (perhaps complex) gains K which place the poles of a given generic (= nondegenerate) system at a given monic polynomial if $mp = n$. In fact

$$d_{m,p} = \frac{1! \cdots (p-1)!(mp)!}{m! \cdots (m+p-1)!}.$$

This agrees with the Willems–Hesselink calculation [20] that $d_{2,2} = 2$, and with the recent calculation made by Morse–Wolovich–Anderson [19] that² $d_{2,3} = d_{3,2} = 5$.

Our methods for proving Theorem 4 rely quite heavily on the Tarski–Seidenberg theorem (Prop. 3.2). In the course of the proof we need several other results from “decision algebra”. With these results in hand, it only requires modest additional effort to show that the question raised in this paper, i.e. whether or not Questions 1 and 2 are equivalent for any fixed m, n, p triple, can in fact be answered by decision algebra. This is shown in the Appendix.

The actual application of a decision-algebra-based checking procedure is of course extremely impractical to implement, but we should emphasize that, at present, this is the only method which is even in principle capable of answering this equivalence question for arbitrary m, n, p . For this reason, we feel it is worthwhile to give a proof of this statement.

3. Proof of Theorems 1 and 2. We shall begin by proving that, for m, n, p fixed, stabilizability for the generic $(F, G, H) \in \mathbb{R}^{n^2+n(m+p)}$ is equivalent to the property that $(s - \rho)^n$, $\rho \in \mathbb{R}$, may be assigned as the closed loop characteristic polynomial for the generic $(F, G, H) \in \mathbb{R}^{n^2+n(m+p)}$. It is intuitively clear that Question 2 should not distinguish between continuous time and discrete time stabilizability. This follows from the first lemma where $\varepsilon = 1$ and $\rho = 0$:

LEMMA 3.1. *The following statements are equivalent:*

- (i) m, n, p are such that for all (F, G, H) —except perhaps those contained in a proper algebraic set—there exists a stabilizing gain K ;
- (ii) m, n, p are such that for all (F, G, H) —except perhaps those contained in a proper algebraic set—for all real ρ and all $\varepsilon > 0$, there exists a gain K such that the eigenvalues of $F + GKH$ are contained in an ε -disc centered about ρ .

Proof. We first note that to say (1.1) is stabilizable is to say the system

$$(3.1) \quad \dot{x} = Fx + Gu, \quad y = Hx + Ju,$$

with J arbitrary but fixed, is stabilizable. For, if K is a stabilizing gain for (1.1), and $I + JK$ is nonsingular, then the gain $u = \tilde{K}y$, where $K(I + JK)^{-1}$ stabilizes (3.1). $I + JK$ is singular; we may choose \tilde{K} sufficiently close to K so that \tilde{K} is a well-defined stabilizing gain for (3.1).

Now consider the conformal transformation

$$\phi(z) = \left(\frac{z - \rho}{\varepsilon + 1} \right) \left(\frac{z - \rho}{\varepsilon - 1} \right)^{-1}$$

and define the rational matrix valued function

$$(3.2) \quad V(z) = W(\phi(z)) = \tilde{H}(zI - \tilde{F})^{-1}\tilde{G} + \tilde{J}$$

² Based on our techniques and those in [12], the authors of [6] have confirmed our conjecture in the case $m = 2$, $p = 3$ by showing that the Galois group of the output feedback problem is the full symmetric group, S_5 .

where $W(z)$ is the open loop transfer function,

$$(3.3) \quad W(z) = H(zI - F)^{-1}G + J.$$

Now let \bar{K} be a gain such that the closed-loop poles of

$$W(z)(I + KW(z))^{-1}$$

are at z_1, \dots, z_n . Then, generically, the poles $\phi(z_1), \dots, \phi(z_n)$ of $V(z)(I + KV(z))^{-1}$ will be finite. Since

$$\operatorname{Re}[z] < 0 \quad \text{if and only if} \quad |\phi(z) - \rho| < \varepsilon,$$

\bar{K} stabilizes $W(z)$ with respect to $\operatorname{Re}[z] < 0$ if, and only if, it stabilizes $V(z)$ with respect to the ε -disc centered about ρ . We claim that, consequently, a generic (F, G, H, J) is stabilizable with respect to $\operatorname{Re}[z] < 0$ if, and only if, a generic $(\tilde{F}, \tilde{G}, \tilde{H}, \tilde{J})$ is stabilizable with respect to the ε -disc $B(\rho; \varepsilon)$. Assuming the claim, by our first observation the “direct part” J may be omitted, and the lemma is proved.³

To verify the claim, we first develop $W(z)$ in a Laurent series

$$W(z) = J + \sum_{i=1}^{\infty} L_i z^{-i}$$

and form the $n \times n, p \times m$ block Hankel matrix

$$h_w = [L_{i+j-1}].$$

Then $W(z)$ determines, and is determined by, a point in the set

$$\mathcal{H}_{m,p}^n = \{(J, L_1, \dots, L_{2n}) : \operatorname{rank} h_w = n\}.$$

$\mathcal{H}_{m,p}^n$ is, by definition, an open subset of an algebraic set of matrices. Moreover, $\mathcal{H}_{m,p}^n$ is the image of the rational map

$$\Pi: \mathcal{M} \subset \mathbb{R}^{n^2 + n(m+p) + mp} \rightarrow \mathcal{H}_{m,p}^n$$

defined on the open dense set \mathcal{M} of minimal systems by

$$\Pi(F, G, H, J) = (J, L_1, \dots, L_{2n})$$

where of course

$$H(sI - F)^{-1}G + J = J + \sum_{i=1}^{\infty} L_i z^{-i}.$$

Therefore, $\mathcal{H}_{m,p}^n$ is irreducible, as the image of an irreducible algebraic set [21]. In this language, we have:

(i) ϕ induces, via (3.2), a rational map

$$\Phi: \mathcal{H}_{m,p}^n \rightarrow \mathcal{H}_{m,p}^n$$

with singularities on the algebraic set $\{h_w: W \text{ has a pole at } 1\}$, since $V(z) = \Phi(W(z)) = W(\phi(z))$ is proper if, and only if, $W(1)$ is finite.

(ii) Image $\Phi = \mathcal{H}_{m,p}^n - \{h_v: V \text{ has a pole at } \varepsilon + \rho\}$ for similar reasons to those in (i).

Furthermore, since stability of minimal systems is an input-output property, if \mathcal{D} is a self-conjugate subset of \mathbb{C} , then

(iii) The set

$$U = \{\sigma = (F, G, H, J) : \sigma \text{ is stabilizable with respect to } \mathcal{D}\}$$

³ Argument along these lines has been developed independently by J. C. Willems.

is open and dense in \mathcal{M} , if and only if,

$$\Pi(U) \subset \mathcal{H}_{m,p}^n$$

is open and dense in $\mathcal{H}_{m,p}^n$.

The claim then follows from (i), (ii) and (iii). Q.E.D.

Remark. A similar, perhaps well-known, result is that for fixed m, n, p stabilizability is generic if, and only if, for generic (F, G, H) there exists a gain K such that the closed-loop spectrum lies in $\operatorname{Re}[s] < \sigma$ or $\operatorname{Re}[s] > \sigma$, with $\sigma \in \mathbb{R}$ arbitrary.

The next proof relies on the following result which is stated in the notation of (1.2). For f, g polynomials, set:

$$(3.4) \quad \begin{aligned} U\{f_i\} &= \{x \in \mathbb{R}^n : f_i(x) > 0, \forall i\}, \\ V\{g_i\} &= \{x \in \mathbb{R}^n : g_i(x) \geq 0, \forall i\}. \end{aligned}$$

A subset $Z \subset \mathbb{R}^n$ is called *semialgebraic* if it is a finite union of finite intersections of sets of the form (3.4). For example, the algebraic set

$$Z = \{x \in \mathbb{R}^n : g(x) = 0\}$$

is semialgebraic. A subset of the form $U\{f_i\}$ is called a *basic open semialgebraic set*, and those of the form $V\{g_i\}$ are called *basic closed semialgebraic sets*.

PROPOSITION 3.2. *If $Z \subset X \times Y$ is a semialgebraic set, then $p_1(Z) \subset X$ is a semialgebraic set. Thus, the existence of Y such that*

$$p_1(x_0, y) = x_0$$

can be checked by a finite number of rational operations in x_0 .

This theorem is of course a version of the Tarski–Seidenberg theorem. It is worth noting that a recent improvement on this result has been made [8], [9], viz. if it is known that $p_1(Z)$ is Euclidean closed (or open), then $p_1(Z)$ is a finite union of basic closed (or open) semialgebraic sets. Of course, $p_1(Z)$ is not necessarily closed, even if Z is closed.

LEMMA 3.3. *If $mp \leq n$, then the following statements are equivalent:*

- (i) *m, n, p are such that the generic (F, G, H) is stabilizable;*
- (ii) *m, n, p are such that for all real ρ and for the generic (F, G, H) , there exists a gain K such that the closed loop characteristic polynomial is $(s - \rho)^n$.*

Proof. Statement (ii) obviously implies (i). For the converse, consider the function, for $\sigma = (F, G, H)$,

$$(3.5) \quad \chi_\sigma : \mathbb{R}^{mp} \rightarrow \mathbb{R}^n, \quad \text{defined via } \chi_\sigma(K) = (p_1, \dots, p_n),$$

where

$$s^n + p_1 s^{n-1} + \dots + p_n = \det(sI - F - GKH).$$

If statement (i) holds, then for each r there exists an open dense subset $U_r \subset \mathbb{R}^{n^2} \times \mathbb{R}^{nm} \times \mathbb{R}^{np} = \mathbb{R}^N$ such that for $(F, G, H) \in U_r$

$$(p_1, \dots, p_n) \in \text{image}(\chi_\sigma)$$

where the roots of $s^n + p_1 s^{n-1} + \dots + p_n$ lie in an $1/r$ -disc centered about ρ . By the Baire category theorem,

$$U = \bigcap_{r=1}^{\infty} U_r$$

is a dense subset of \mathbb{R}^N such that for $(F, G, H) \in U$,

$$(\bar{p}_1, \dots, \bar{p}_n) \in \overline{\text{image}(\chi_\sigma)}$$

where

$$s_n + \bar{p}_1 s^{n-1} + \dots + \bar{p}_n = (s - \rho)^n.$$

Now, according to [4, § 4, Thm.] provided $mp \leq n$ there exists an open dense subset $W \subset \mathbb{R}^N$ —the set of nondegenerate systems—such that $\text{image}(\chi_\sigma)$ is Euclidean closed for $(F, G, H) \in W$. Thus, if

$$(F, G, H) \in U = U' \cap W,$$

then

$$(\bar{p}_1, \dots, \bar{p}_n) \in \text{image}(\chi_\sigma).$$

Now, any real gain K may be regarded as a point in \mathbb{R}^{mp} , and we may consider the real algebraic set

$$(3.6) \quad V^\rho = \{(F, G, H, K) : \det(sI - F - GKH) = (s - \rho)^n\} \subset \mathbb{R}^N \times \mathbb{R}^{mp}.$$

By the Tarski–Seidenberg theorem (Prop. 3.2),

$$p_1(V^\rho) \subset \mathbb{R}^N,$$

the projection onto the first factor, is a semialgebraic set in \mathbb{R}^N ; i.e., $p_1(V^\rho)$ is defined by a finite set of equations and inequations as in (3.4). Since

$$U \subset p_1(V^\rho) = \mathbb{R}^N$$

is dense, it follows that $p_1(V^\rho)$ may be defined by algebraic conditions (perhaps disjunctive)

$$f_1(F, G, H) > 0, \dots, f_r(F, G, H) > 0,$$

from which it follows that $p_1(V^\rho)$ is open and dense. Since $(F, G, H) \in p_1(V^\rho)$ if, and only if, there exists a K such that the closed-loop characteristic polynomial is $(s - \rho)^n$, the lemma is proved. Q.E.D.

For the more precise assertion in part (ii) of Theorem 1, we need the following:

LEMMA 3.4. For any $p = (p_1, \dots, p_n) \in \mathbb{R}^n$, the subset

$$V_p = \{\sigma = (F, G, H) \in W : \chi_\sigma(K) = p \text{ for some } K\}$$

is closed in W .

Remark. The corresponding assertion for (F, G, H) minimal can be false. This is quite analogous to the fact that the set

$$\{x \in \mathbb{R} : \exists y \in \mathbb{R} \text{ such that } xy = 1\}$$

is not closed in \mathbb{R} , while the set

$$\{x \in \mathbb{R} - \{0\} : \exists y \in \mathbb{R} \text{ such that } xy = 1\}$$

is closed in the open dense subset $W = \mathbb{R} - \{0\} \subset \mathbb{R}$.

Proof. As in [4], we may think of $K \in \mathbb{R}^{mp}$ as a point in $\text{Grass}(p, m + p)$ —the set of p -planes in \mathbb{R}^{m+p} —via the assignment

$$K \mapsto \text{graph}(K) = \{(y, Ky)\} \subset \mathbb{R}^p \oplus \mathbb{R}^m.$$

It is known (see e.g. [4] and references cited therein) that $\text{Grass}(p, m+p)$ may be regarded as a compact manifold of dimension mp . Moreover,

$$\text{Grass}(p, m+p) = \mathbb{R}^{mp} \cup \sigma(\infty)$$

where $\sigma(\infty)$ is the closed subset defined by

$$\sigma(\infty) = \{\Pi \in \text{Grass}(p, m+p) : \dim(\Pi \cap \mathbb{R}^m) \geq 1\}.$$

That is, $\Pi \in \sigma(\infty)$ if, and only if, Π is not complementary to U . Thus, $\Pi \notin \sigma(\infty)$ if, and only if,

$$\Pi = \text{graph}(K), \quad \text{for some linear } K: \mathbb{R}^p \rightarrow \mathbb{R}^m.$$

On the other hand, one may regard the monic polynomial

$$p(s) = s^n + p_1 s^{n-1} + \cdots + p_n$$

as a point $(p_1, \dots, p_n) \in \mathbb{R}^n$ and therefore [4] as a point, via the homogeneous coordinates

$$[p_1, \dots, p_n, 1] \in \mathbb{RP}^n,$$

in real projective n -space. Of course, $\mathbb{RP}^n = \text{Grass}(1, n+1)$ by definition. According to [4, Remarks, p. 103], for nondegenerate σ the map χ_σ extends continuously to a map

$$\chi_\sigma: \text{Grass}(p, m+p) \rightarrow \mathbb{RP}^n,$$

satisfying

$$(3.7a) \quad \chi_\sigma(\Pi) = [p_1, \dots, p_n, 0]$$

if, and only if,

$$(3.7b) \quad \Pi \in \sigma(\infty).$$

Matters being so, consider the continuous function

$$\chi: W \times \text{Grass}(p, m+p) \rightarrow \mathbb{RP}^n$$

defined via

$$\chi(F, G, H, \Pi) = \chi(\sigma, \Pi) = \chi_\sigma(\Pi).$$

Therefore, if $\bar{p} = [1, 0, \dots, 0, 1]$ corresponds to $\bar{p}(s) = s^n$,

$$Z = \chi^{-1}(\bar{p}) \subset W \times \text{Grass}(p, m+p)$$

is a closed subset. Since $\text{Grass}(p, m+p)$ is compact,

$$p_1(Z) \subset W$$

is closed and, by virtue of (3.6),

$$p_1(Z) = \{\sigma = (F, G, H) : \chi_\sigma(K) = \bar{p}, \text{ for some } K\} = V_{\bar{p}}. \quad \text{Q.E.D.}$$

On the other hand, $U \cap W \subset V_{\bar{p}}$ is dense in W by the Baire category theorem, and therefore

$$V_{\bar{p}} = W,$$

from which (ii), and Theorem 1, follow. Q.E.D.

We now turn to a proof of Theorem 2. Clearly, it suffices to consider the case $mp \leq n$; thus, the preceding lemmata and Theorem 1 are applicable.

Consider, then, the algebraic set of nilpotent $n \times n$ real matrices

$$\mathcal{N} = \{N: N^k = 0, \text{ for some } k\}$$

and the algebraic set $V = V^0$ obtained by setting $\rho = 0$ in (3.6). We define the polynomial mapping

$$(3.8) \quad \Phi: \mathcal{N} \times \mathbb{R}^{nm} \times \mathbb{R}^{np} \times \mathbb{R}^{mp} \rightarrow \mathbb{R}^{n^2} \times \mathbb{R}^{nm} \times \mathbb{R}^{np}$$

via

$$\Phi(N, G, H, K) = (N - GKH, G, H).$$

From Theorem 1, we have:

LEMMA 3.5. *If $mp \leq n$ and if the generic system is stabilizable, then the image of Φ contains an open, dense set.*

Denote by $\mathcal{N}_{\mathbb{C}}$ the algebraic set of nilpotent $n \times n$ complex matrices. It is known (see e.g. [17], [20]) that $\mathcal{N}_{\mathbb{C}}$ is an irreducible algebraic set. Therefore there exists an open dense subset U of $\mathcal{N}_{\mathbb{C}}$ which is itself a complex manifold and therefore has a dimension. Indeed [17], [20],

$$\dim_{\mathbb{C}}(U) = n^2 - n.$$

The points of U are called simple, and one of the thorny points in real algebraic geometry [18] is that in general an irreducible real algebraic set $V_{\mathbb{R}}$ may contain none of the simple points of $V_{\mathbb{C}}$. This, for example, is the reason for the failure of the Hilbert Nullstellensatz over \mathbb{R} , and the best-known example of this phenomenon is

$$W_{\mathbb{R}} = \{(x, y): x^2 + y^2 = 0\}.$$

If $V_{\mathbb{R}}$ contains a simple point of $V_{\mathbb{C}}$, then for example $\dim_{\mathbb{R}}(V_{\mathbb{R}})$ is defined as above and

$$(3.9) \quad \dim_{\mathbb{R}}(V_{\mathbb{R}}) = \dim_{\mathbb{C}}(V_{\mathbb{C}}).$$

It is an elementary computation to check that the real matrix

$$N = \begin{bmatrix} 0 & 1 & & & \\ & 0 & \cdot & & 0 \\ & & \cdot & \cdot & \\ 0 & & & \cdot & \cdot \\ & & & \cdot & 1 \\ & & & & 0 \end{bmatrix}$$

is a simple point of $\mathcal{N}_{\mathbb{C}}$. Thus, $\dim_{\mathbb{R}} N$ satisfies (3.9). We will now give a self-contained proof of

LEMMA 3.6. $\dim_{\mathbb{C}}(\mathcal{N}) = n^2 - n$.

Proof. Since the matrix N consists of a single Jordan block, the dimension of the centralizer

$$Z(N) = \{T \in GL(n, \mathbb{C}): TN = NT\}$$

is n , according to the Frobenius dimension formula ([15, Vol. II, Thm. 19, p. 111]). Now consider the orbit of N under $GL(n, \mathbb{C})$:

$$\mathcal{O}(N) = \{TNT^{-1}: T \in GL(n, \mathbb{C})\} \approx \frac{GL(n, \mathbb{C})}{Z(N)}.$$

In particular,

$$\dim_{\mathbb{C}} \mathcal{O}(N) = \dim GL(n, \mathbb{C}) - \dim Z(N) = n^2 - n.$$

We claim $\overline{\mathcal{O}(N)} = \mathcal{N}$, from which follows:

- (i) \mathcal{N} is irreducible, since $\mathcal{O}(N)$ is irreducible; and
- (ii) $\dim_{\mathbb{C}} \mathcal{O}(N) = \dim_{\mathbb{C}} (\mathcal{N})$, by the closed orbit lemma [14] and (3.9).

Following [20], note that if N_i is any nilpotent Jordan canonical form, then clearly there is a 1-parameter diagonal subgroup $T_\lambda \in GL(n, \mathbb{C})$ such that

$$\lim_{\lambda \rightarrow \infty} T_\lambda N T_\lambda^{-1} = N_i.$$

Therefore, $\overline{\mathcal{O}(N)} = \mathcal{N}$. Q.E.D.

Now suppose that m , n and p are such that the generic system is stabilizable, and $mp \leq n$. By Lemma 3.5 and [21, Thm. 7, p. 60] one has

$$(3.10) \quad \dim \mathcal{N}_n + n(m+p) + mp \geq n^2 + n(m+p).$$

In the light of Lemma 3.6 and (3.10),

$$n^2 - n + mp \geq n^2,$$

yielding

$$mp \geq n.$$

In conclusion, if $mp \leq n$ then $mp = n$ is necessary for generic stabilizability, whence Theorem 2. Q.E.D.

4. Proof of Theorem 3. In the proof of Theorem 1 (cf. Lemma 3.2), we made use of certain facts concerning $p \times m$ systems of degree n which also allow us to show, together with Theorem 1, that for $n = 4$, $m = p = 2$, generic stabilizability is not possible. Specifically:

- (i) if $mp \leq n$, then the class W of nondegenerate systems is open and dense in $\mathbb{R}^{n^2} \times \mathbb{R}^{mn} \times \mathbb{R}^{np}$; and
- (ii) for any monic polynomial $p(s)$ of degree n , the set

$$V_p = \{(F, G, H) \in W : \det(sI - F - GKH) = p(s), \text{ for some } K\}$$

is closed in W (Lemma 3.4).

In light of Theorem 1, if $\bar{p}(s) = s^n$, then generic stabilizability implies that $V_{\bar{p}}$ is dense and closed in W , hence coincides with W . Therefore, to find one nondegenerate system for which $\bar{p}(s)$ is not assignable as a closed-loop polynomial is to prove that stabilizability is not generic.

We shall now give a “frequency domain” criterion [3] (which can be taken as a definition, compare [4]) for nondegeneracy. If $T(s)$ is the transfer function

$$(4.1) \quad T(s) = H(sI - F)^{-1}G$$

of the system, denote by $t_i(s)$ the i th column of the $(p+m) \times m$ matrix

$$\mathcal{T}(s) = \begin{bmatrix} T(s) \\ I \end{bmatrix}.$$

If $\phi(y, u)$ is a complex linear functional on $\mathbb{C}^p \oplus \mathbb{C}^m$, then we can form the scalar rational function

$$\phi(t_i(s)) \quad \text{for } i = 1, \dots, m.$$

Now suppose $\Phi = \{\phi_1, \dots, \phi_p\}$ is any linearly independent set of linear functionals on $(m+p)$ -space, and form the determinant

$$(4.2) \quad \Phi(s) = \det [\phi_i(t_j(s))].$$

(F, G, H) is said to be nondegenerate provided

$$(4.3) \quad \Phi(s) \neq 0 \quad \text{in } s$$

for any choice of Φ .

Remarks 1. If (F, G, H) is scalar, then (F, G, H) is nondegenerate since (4.2)–(4.3) reduces, for $\phi(u, y) = au + by$, to $ag(s) + b \neq 0$ in s .

2. The zeros of the set $\Phi = \{\phi_1, \dots, \phi_m\}$ defines a p -plane in (u, y) -space which is the graph either of a linear function $u = Ky$, i.e. a finite constant gain, or of a linear relation between u and y , i.e. an infinite constant gain. The zeros of (4.2) are then, modulo pole-zero cancellation, the closed-loop poles at this gain, and (4.3) just asks that these zeros be finite in number, i.e. that the root-locus map χ be defined and continuous at this gain.

Example 4. Suppose $m = p = 2$, $n = 4$ and consider

$$G = \begin{bmatrix} I \\ 0 \end{bmatrix}, \quad H = [I \quad 0]$$

and

$$F = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 \end{bmatrix}.$$

We claim (F, G, H) is nondegenerate; to this end we compute (clearing denominators)

$$\mathcal{T}(s) = \begin{bmatrix} s^3 - 1 & -s \\ s & s^3 \\ s^4 + s - 1 & 0 \\ 0 & s^4 + s - 1 \end{bmatrix}$$

and consider 2 linear functionals

$$\phi_1(y, u) = a_1 y_1 + a_2 y_2 + a_3 u_1 + a_4 u_2, \quad \phi_2(u, y) = b_1 y_1 + b_2 y_2 + b_3 u_1 + b_4 u_2.$$

Thus,

$$(4.4) \quad \Phi(s) = \det [\phi_i(t_j(s))] = \det \begin{bmatrix} \alpha_{11}(s) & \alpha_{12}(s) \\ \alpha_{21}(s) & \alpha_{22}(s) \end{bmatrix}$$

where

$$(4.5) \quad \begin{aligned} \alpha_{11}(s) &= a_3 s^4 + a_1 s^3 + (a_2 + a_3)s - a_3 - a_1, \\ \alpha_{12}(s) &= a_4 s^4 + a_2 s^3 + (a_4 - a_1)s + a_4, \\ \alpha_{21}(s) &= b_3 s^4 + b_1 s^3 + (b_2 - b_3)s + b_3 - b_1, \\ \alpha_{22}(s) &= b_4 s^4 + b_2 s^3 + (b_4 - b_1)s + b_4. \end{aligned}$$

Now, (4.4) vanishes just in case there exists c_s —a priori depending on s —such that

$$(4.6) \quad c_s \alpha_{11}(s) = \alpha_{21}(s), \quad c_s \alpha_{12}(s) = \alpha_{22}(s)$$

for all but finitely many $s \in \mathbb{C}$. Comparing coefficients shows that c_s is constant for all but finitely many, and hence all, s and therefore an inspection of (4.5)–(4.6) shows that

$$c\phi_1 = \phi_2,$$

contradicting linear independence of the functionals ϕ_i .

Recall that in the proof of Lemma 3.3, the fact that image (χ) is closed for all nondegenerate (F, G, H) was used rather crucially. If

$$K = \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix}.$$

it is readily verified that

$$(4.7) \quad \begin{aligned} -\det(sI - F - GKH) &= s^4 + (-x_1 - x_4)s^3 + (x_1x_4 - x_2x_3)s^2 \\ &\quad + (1 - x_2 + x_3)s + (1 + x_1). \end{aligned}$$

By the quadratic formula, it is easily verified that image (χ) is a closed semialgebraic set. Furthermore, if (4.7) is to be s^4 , we require

$$(4.7)' \quad x_1 + x_4 = x_1x_4 - x_2x_3 = 1 - x_2 + x_3 = 1 + x_1 = 0,$$

whence

$$x_2x_3 = -1, \quad x_2 = 1 + x_3,$$

whence

$$(4.8) \quad x_3^2 + x_3 + 1 = 0.$$

This equation (4.8) cannot be satisfied by any real x_3 , i.e. there is no real gain producing closed-loop poles at $s = 0$. Since (F, G, H) is nondegenerate, our previous remarks imply Theorem 3, thereby verifying Molander's conclusion.

5. Proof of Theorem 4. In addition to the Tarski–Seidenberg theorem (Prop. 3.2), we shall also need a somewhat different result from decision algebra, which deals with the question of describing the set of x_0 for which $p_1^{-1}(x_0) \cap Z = \{x_0\} \times Y$, i.e. for which $(x_0, y) \in Z$ for all y . In the course of deriving this result we also will state the Tarski–Seidenberg theorem in what is perhaps a more familiar form [15, Vol. III], [22].

Notational conventions are as follows: x, y, z denote collections of indeterminates, with each of x, y, z considered to be shorthand for a number of indeterminates x_1, \dots, x_n etc. Particular real values taken by these quantities will be denoted by $\hat{x}, \hat{y}, \hat{z}$; p, q, r, s , perhaps with subscripts will denote polynomials in x, y, z with real coefficients. We shall regard $p(x, y) = 0$ or $q(x, y) \geq 0$ as examples of equations or inequations, (i.e. descriptions of problems for which solutions are sought, should they exist), and we shall regard $p(\hat{x}, \hat{y}) = 0$ or $q(\hat{x}, \hat{y}) \geq 0$ as examples of equalities or inequalities (i.e. statements of fact that can be verified by arithmetic, and which show that \hat{x}, \hat{y} are solutions of $p(x, y) = 0$ or $q(x, y) \geq 0$).

We shall reserve script letters \mathcal{S}, \mathcal{T} , etc. to denote collections of a finite number of equations and inequations or inequalities and inequalities of the following type. $\mathcal{S}(x)$ is an abbreviation for:

$$\begin{aligned} &\text{either } \{p_{i1}(x) = 0 \text{ and } q_{j1}(x) > 0 \quad \text{and } r_{k1}(x) \neq 0 \text{ and } s_{l1}(x) \geq 0\} \\ &\text{or } \quad \{p_{i2}(x) = 0 \text{ and } q_{j2}(x) > 0 \quad \text{and } r_{k2}(x) \neq 0 \text{ and } s_{l2}(x) \geq 0\} \\ &\text{or } \quad \vdots \\ &\text{or } \quad \{p_{it}(x) = 0 \text{ and } q_{jt}(x) > 0 \quad \text{and } r_{kt}(x) \neq 0 \text{ and } s_{lt}(x) \geq 0\}, \end{aligned}$$

where it is understood that $p_{i\alpha}(x) = 0$ is shorthand for $p_{1\alpha}(x) = 0$ and $P_{2\alpha}(x) = 0$ and \dots and $p_{i\alpha}(x) = 0$, and similarly for $q_{j\alpha}$ etc. Naturally, $\mathcal{S}(\hat{x})$ is an abbreviation for the associated set of equalities and inequalities. We can talk of the problem of solving $\mathcal{S}(x)$ and of $\mathcal{S}(\hat{x})$ holding, or of \hat{x} being a solution of $\mathcal{S}(x)$.

The above type of $\mathcal{S}(x)$ is more or less standard in decision algebra. However, we shall sometimes use a simple modification. Each $s_{\alpha\beta} \geq 0$ is a disjunction: $s_{\alpha\beta} > 0$ or $s_{\alpha\beta} = 0$. This means that any $\mathcal{S}(x)$ and thus any $\mathcal{S}(\hat{x})$ can be rewritten to exclude inequations or inequalities of the \geq type.

LEMMA 5.1. *The statement $\mathcal{S}(\hat{x})$ does not hold is equivalent to a statement $\tilde{\mathcal{S}}(\hat{x})$ holds where $\tilde{\mathcal{S}}(x)$, termed the negator of \mathcal{S} , is itself a collection of equations and inequations of the standard form.*

Proof. “ $\mathcal{S}(\hat{x})$ holds” is a disjunction (“or” statement) of conjunctions (“and” statements) of formulas of the type $p(\hat{x}) = 0$, $q(\hat{x}) > 0$, $r(\hat{x}) \neq 0$ and $s(\hat{x}) \geq 0$. Hence “ $\mathcal{S}(\hat{x})$ does not hold” is a conjunction of disjunctions of negations of these formulas, i.e. of $p(\hat{x}) \neq 0$, $-q(\hat{x}) \geq 0$, $r(\hat{x}) = 0$ and $-s(\hat{x}) > 0$. Any conjunction of disjunctions can be rearranged as a disjunction of conjunctions, and in this way, $\tilde{\mathcal{S}}(x)$ is defined.

Obviously, $\tilde{\tilde{\mathcal{S}}} = \mathcal{S}$.

Next, we recall the main result of decision algebra, the Tarski–Seidenberg theorem. We break it into two parts.

PROPOSITION 5.2. (A) *Consider an equation/inequation set $\mathcal{S}(x, y)$. Then one can determine by a finite number of rational calculations a second such set $\mathcal{T}(y)$ such that $\mathcal{T}(\hat{y})$ holds if and only if there exists at least one \hat{x} such that $\mathcal{S}(\hat{x}, \hat{y})$ holds.*

(B) *The solvability of any equation/inequation set $\mathcal{T}(y)$ is determinable by a finite number of rational calculations.*

We remark that the set $\mathcal{T}(y)$ in part A may be empty: this would imply that there are no pairs \hat{x}, \hat{y} for which $\mathcal{S}(\hat{x}, \hat{y})$ holds.

PROPOSITION 5.3. *Consider an equation/inequation set $\mathcal{S}(x, y)$. Then the set of values \hat{y} of y such that for all \hat{x} , $\mathcal{S}(\hat{x}, \hat{y})$ holds, is definable by an equation/inequation set $\mathcal{T}(y)$.*

Proof. Let $\tilde{\mathcal{S}}(x, y)$ be the negator of $\mathcal{S}(x, y)$, existing by Lemma 5.1. By Proposition 5.2A, we can find $\tilde{\mathcal{T}}(y)$ such that $\tilde{\mathcal{T}}(\hat{y})$ holds if and only if there exists at least one \hat{x} such that $\tilde{\mathcal{S}}(\hat{x}, \hat{y})$ holds. Let \mathcal{T} be the negator of $\tilde{\mathcal{T}}$. Then $\mathcal{T}(\hat{y})$ holds if and only if there exists no \hat{x} such that $\tilde{\mathcal{S}}(\hat{x}, \hat{y})$ holds, i.e. if and only if for all \hat{x} , $\mathcal{S}(\hat{x}, \hat{y})$ holds.

The following algorithm, in conjunction with Propositions 5.2 and 5.3, gives a proof of Theorem 4. We find it convenient to break this into two parts.

Part I. Find a cube containing a stabilizing gain K .

I.1. Choose $N > 0$, and consider the semialgebraic set

$$Z \subset \mathbb{R}^{mp-1} \times \mathbb{R}$$

defined by $Z = Z_1 \cap Z_2$, with

$$Z_1 = \{K : \det(sI - F - GKH) \text{ is Hurwitz}\},$$

$$Z_2 = \{K : \sum (k_{ij})^2 < N\}.$$

From the Routh–Hurwitz criterion, it follows that Z_1 is a (basic open) semi-algebraic set and it is then clear that Z is semialgebraic. Using Proposition 5.2 inductively, we can decide by rational operations whether there exists a gain $K \in Z$. If $Z \neq \emptyset$, go to Step II.1. Otherwise, go to Step I.2.

I.2. Replace N by $2N$ and go to Step I.1. Since a stabilizing gain exists by hypothesis, we will eventually move to Part II.

Part II. Find a cube contained in the set of stabilizing gains K .

II.1. We suppose there is a stabilizing gain K in the cube $\|K\| < N$. Using Proposition 5.3 inductively, we can decide by rational operations whether all such K are stabilizing. If so, choose any K such that $\|K\| < N$. If not, go to Step II.2.

II.2. Divide the cube into 2^{mp} cubes with sides of length N . Return to Step I.1 with this list of cubes.

This algorithm will stop at some stage, since the set of stabilizing gains is open and therefore contains a cube of sufficiently small size. Q.E.D.

Example 5. One might ask whether one can bound the number of steps in this program simply in terms of m, n, p . The answer is no, as we now illustrate. Consider the open loop system with transfer function

$$w(s) = \frac{1}{s^3 + as^2 + bs}$$

where $a, b > 0$. For negative feedback with gain k , the closed loop characteristic polynomial is $s^3 + as^2 + bs + k$ and therefore is Hurwitz if, and only if, $k \in (0, ab)$. It follows that the size of a cube (here, an interval) contained in the open set of stabilizing gains can be made arbitrarily small by suitable choice of ab . In turn, the number of steps in Part II of the algorithm can be made arbitrarily large, though for fixed (a, b) it is of course finite.

6. Proof of Theorems 5 and 6. Since we have already demonstrated the existence of linear formulae for the appropriate values of m, n, p , it is enough to show that these are the only values for which such formulae can exist. Moreover, it suffices to prove this last assertion over $\mathbb{C} = R(\sqrt{-1})$. Consider the closed loop characteristic coefficient map χ , defined in (3.4), extended to gains with complex coefficients

$$(6.1) \quad \chi_{\mathbb{C}}: \mathbb{C}^{mp} \rightarrow \mathbb{C}^n$$

where (F, G, H) is understood to be a generic, but fixed, system with $n = mp$. We first analyze the question as to whether there exists a formula for $(k_{ij}) \in \chi^{-1}(p)$ which is rational in the coordinates of $p = (p_l) \in \mathbb{C}^n$. Thus, we consider the field K_1 of all rational expressions (or functions) in the p_l , and the field K_2 of all rational functions in the (k_{ij}) :

$$(6.2) \quad K_1 = \mathbb{C}(p_l), \quad K_2 = \mathbb{C}(k_{ij}).$$

Since $\chi_{\mathbb{C}}$ is polynomial, if $f \in K_1$ then $f \circ \chi_{\mathbb{C}} \in K_2$. For generic (F, G, H) , image $\chi_{\mathbb{C}}$ contains an open set [13] so that

$$(6.3) \quad f \circ \chi_{\mathbb{C}} = 0 \Rightarrow f = 0.$$

By virtue of (6.3), we can think of K_1 as a subfield of K_2 , i.e.

$$(6.4) \quad K_1 \approx \chi_{\mathbb{C}}^* K_1 \subset K_2$$

where $\chi_{\mathbb{C}}^* f = f \circ \chi_{\mathbb{C}}$, and an easy dimension argument shows that (6.4) is a finite field extension. That is, K_2 , as a vector space over the field of scalars K_1 , is finite dimensional. For example, to say rational formulae for $(k_{ij}) \in \chi_{\mathbb{C}}^{-1}(p_l)$ exist is to say the dimension of this vector space

$$(6.5) \quad \delta = [K_2 : K_1] = \dim_{K_1}(K_2)$$

is equal to 1, i.e. $K_1 = K_2$. We shall now give a formula for δ , in terms of m, p . In [4] it was shown that $\chi_{\mathbb{C}}$ is proper and it follows from the proof in [4] that

$$R_1 \approx \chi_{\mathbb{C}}^* R_1 \subset R_2$$

is an integral ring extension, where

$$R_1 = \mathbb{C}[p_l], \quad R_2 = \mathbb{C}[k_{ij}].$$

In this case (since the field \mathbb{C} has characteristic zero), δ is given by the number d of solutions, counted with multiplicity, to the equation [18, pp. 116–117]

$$\chi_{\mathbb{C}}(K) = p.$$

On the other hand, d has been computed using methods of the Schubert calculus in [3] to be

$$(6.6) \quad d = \frac{1! \cdots (p-1)!(mp)!}{m! \cdots (m+p-1)!}.$$

Thus, Theorem 5 follows from the following elementary observation:

LEMMA 6.1. In (6.6), $d = 1 \Leftrightarrow \min(m, p) = 1$.

As for Theorem 6, from the explicit form of the solution to the pole-placement equations, derived via elimination methods by Willems–Hesselink [23], it is clear that (over \mathbb{R} or \mathbb{C}) quadratic formulae and rational expressions are sufficient to express K as a function of (p_1, \dots, p_n) for generic (F, G, H) when $m = p = 2$, and $n = 4$. We shall now prove that, except for the linear cases $\min(m, p) = 1$, this is the only case when formulae—involving square roots and rational operations—for K in terms of (p_1, \dots, p_n) exist.

To this end, we consider a Galois extension

$$(6.7) \quad K_1 \subset K,$$

that is, a minimal normal extension of $K_1 = \mathbb{C}(p)$ which contains all of the roots to the equation

$$(6.8) \quad \chi_{\mathbb{C}}(K) = (p).$$

If a solution expressible by square roots and rational operations alone exists, then

$$\delta' = [K : K_1]$$

is a power of 2 [2]. On the other hand, by Artin's theorem of the primitive element [2], we may regard $K_2 \subset K$ and therefore

$$\delta = [K_2 : K_1] \quad \text{divides} \quad [K : K_1],$$

from which it follows that

$$\delta = d_{m,p} = 2^r, \quad \text{for some } r.$$

Theorem 6 therefore follows from the following result:

LEMMA 6.2. If $\min(m, p) \geq 2$ and $m + p \geq 5$, then $d_{m,p}$ is divisible by an odd prime.

Remark. The proof we present here is based on an application of the strong form of Bertrand's postulate [11, p. 373] shown to us by W. H. Gustafson.

Proof. By the strong form of Bertrand's postulate, there is a prime q satisfying

$$(6.9) \quad m + p - 1 < q < 2(m + p) - 4,$$

under the hypothesis $m + p \geq 5$. Clearly, q does not divide the denominator of $d_{m,p}$. On the other hand, if $\min(m, p) \geq 2$, then

$$mp > q,$$

so that q divides the numerator of $d_{m,p}$. Hence, $q | d_{m,p}$. Q.E.D.

Appendix. "In principle" answers to Questions 1 and 2 by decision algebra. In [1], indications of the applicability of decision algebra to problems of systems theory were given. In particular, it was shown that one can determine, at least in principle, by rational operations whether a given system (F, G, H) can be stabilized. We shall extend these results to show that one can answer Questions 1 and 2 by rational operations using decision theoretic techniques, but we emphasize that such results are very qualitative. In fact, a "worst-case" analysis [7] shows that any decision procedure takes at least $2^k n$ steps, where $k > 0$ is a constant and n is the length of the input formula.

However, in the absence of any other technique which allows one, for example, even in principle to distinguish between Questions 1 and 2, we thought it worthwhile to point out that this is a question which can be answered by the Tarski-Seidenberg theory. An interesting special case is whether or not we can place poles for generic 2×2 systems with McMillan degree 8. One does know that there exist 14 complex solutions to the pole-placement equations, but at present one does not know whether any of these are real.

The new ingredient here is the consideration of the generic system (F, G, H) rather than a particular choice of system (F_0, G_0, H_0) , and we shall need to present some further results from decision algebra. The notation is as in § 5.

LEMMA A.1. *Consider an equation/inequation set $\mathcal{S}(x, y, z)$. Then there exists a set $\mathcal{T}(y)$ such that $\mathcal{T}(\hat{y})$ holds if and only if for all \hat{z} , there exists \hat{x} depending on \hat{y}, \hat{z} with $\mathcal{S}(\hat{x}, \hat{y}, \hat{z})$ holding.*

Proof. By Proposition 5.1', there exists $\mathcal{R}(y, z)$ such that $\mathcal{R}(\hat{y}, \hat{z})$ holds if and only if $\mathcal{S}(x, \hat{y}, \hat{z})$ is solvable, i.e. if and only if there exists at least one \hat{x} , depending on \hat{y} and \hat{z} , such that $\mathcal{S}(\hat{x}, \hat{y}, \hat{z})$ holds. By Proposition 5.3, there exists $\mathcal{T}(y)$ such that $\mathcal{T}(\hat{y})$ holds if and only if $\mathcal{R}(\hat{y}, \hat{z})$ holds for all \hat{z} . Then clearly, $\mathcal{T}(\hat{y})$ holds if and only if, for all \hat{z} , there exists \hat{x} such that $\mathcal{S}(\hat{x}, \hat{y}, \hat{z})$ holds.

In Proposition 5.3 and Lemma A.1, the set $\mathcal{T}(y)$ may be empty. The following lemma replaces the "all \hat{x} " in Proposition 5.3 by "almost all", and in this sense may enable one to get a practical result when the $\mathcal{T}(y)$ of this proposition is empty.

LEMMA A.2. *Consider an equation/inequation set $\mathcal{S}(x, y)$. Then there exists an equation/inequation set $\mathcal{T}(y)$ such that $\mathcal{T}(\hat{y})$ holds if and only if $\mathcal{S}(\hat{x}, \hat{y})$ holds for all \hat{x} save a set contained in a proper variety depending on \hat{y} .*

Proof. Given a polynomial $p(x, y)$, it is clear that there exists a possibly empty $\mathcal{P}(y)$ such that $\mathcal{P}(\hat{y})$ holds if and only if $p(x, \hat{y})$ is the zero polynomial, i.e. $p(\hat{x}, \hat{y}) = 0$ for all \hat{x} . Further, if $p(\hat{x}, \hat{y}) = 0$ for all \hat{x} save those lying in a proper variety, $p(\hat{x}, \hat{y}) = 0$ for all \hat{x} .

Given a polynomial $r(x, y)$, it is clear that there exists $\mathcal{R}(y)$ such that $\mathcal{R}(\hat{y})$ holds if and only if $r(x, \hat{y}) \neq 0$ is solved by all x save those on a proper variety depending on \hat{y} .

Given a polynomial $s(x, y)$, it is clear that there exists $\bar{\mathcal{S}}(y)$ such that $\bar{\mathcal{S}}(y)$ holds if and only if $s(\hat{x}, \hat{y}) < 0$ for some \hat{x} . Hence $\mathcal{S}(\hat{y})$ holds if and only if $s(\hat{x}, \hat{y}) \geq 0$ for all \hat{x} . Further, if $s(\hat{x}, \hat{y}) \geq 0$ for all \hat{x} save those in a proper variety, $s(\hat{x}, \hat{y}) \geq 0$ for all \hat{x} .

Given a polynomial $q(x, y)$, it is clear that there exists $\mathcal{Q}_1(\hat{y})$ such that $q(\hat{x}, \hat{y}) \geq 0$ for all \hat{x} and $\mathcal{Q}_2(\hat{y})$ such that $q(\hat{x}, \hat{y}) \neq 0$ for all \hat{x} save those in a proper variety. Let $\mathcal{Q}(y)$ denote the conjunction of $\mathcal{Q}_1(y)$ and $\mathcal{Q}_2(y)$. Then $\mathcal{Q}(\hat{y})$ holds if and only if $q(\hat{x}, \hat{y}) > 0$ for all \hat{x} save those in a proper variety depending on \hat{y} .

Suppose now that $\mathcal{S}(x, y)$ is the disjunction of equation/inequation sets $\mathcal{S}_i(x, y)$ where each $\mathcal{S}_i(x, y)$ is a conjunction of

$$p_{\alpha i}(x, y) = 0, \quad q_{\beta i}(x, y) > 0, \quad r_{\gamma i}(x, y) \neq 0, \quad s_{\delta i}(x, y) \geq 0.$$

By the discussion above, it is clear that there exists $\mathcal{T}_i(y)$ such that $\mathcal{T}_i(\hat{y})$ holds if and only if $\mathcal{S}_i(\hat{x}, \hat{y})$ holds for all \hat{x} save those in a proper variety depending on \hat{y} . $\mathcal{T}(y)$ is obtained as the disjunction of the $\mathcal{T}_i(y)$.

Now consider the system (1.1), subject to output feedback $u = Ky$. The coefficients of the closed-loop characteristic polynomial, as a function of K , give rise to the polynomial mapping (3.4)

$$\chi: \mathbb{R}^{mp} \rightarrow \mathbb{R}^n,$$

and we wrote $\chi_{(F,G,H)}$ to emphasize the dependence on the open loop system (1.1). Then, Question 1 asks whether $\chi_{(F,G,H)}$ is surjective for the generic (F, G, H) and we claim that this question can be answered within the scope of decision algebra. To this end, let $X = \mathbb{R}^{mp}$, $Y = \mathbb{R}^{n^2+nm+np}$ and $Z = \mathbb{R}^n$, so that $(K, (F, G, H), (p_l)) \in X \times Y \times Z$, and consider the algebraic subset $W \subset X \times Y \times Z$ defined by the equations

$$(A.1) \quad \mathcal{S}(x, y, z): \chi_{(F,G,H)}(K) = (p_l).$$

By Lemma A.1, there exists an equation/inequation set \mathcal{T} in $y = \{F, G, H\}$ such that $\mathcal{T}(\hat{y})$ holds if and only if for all \hat{z} , i.e. for all p_l , there exists \hat{x} , i.e. a value of K , such that $\mathcal{S}(\hat{x}, \hat{y}, \hat{z})$ holds, i.e. such that (A.1) holds.

Let $\bar{\mathcal{T}}(y)$ denote the negator of \mathcal{T} , and write $\bar{\mathcal{T}}_i(y)$ as a disjunction of conjunctions $\bar{\mathcal{T}}_i$. As observed in § 4, we can assume without loss of generality that each $\bar{\mathcal{T}}_i$ contains equations $p_{\alpha i}(y) = 0$, and inequations $q_{\beta i}(y) > 0$ and $r_{\gamma i}(y) \neq 0$, *without* inequations of the type $s_{\delta i}(y) \geq 0$. We can determine (see Prop. 5.2B) whether any $\bar{\mathcal{T}}_i$ defines an empty set of solutions; if so, we discard it.

Now with $\bar{\mathcal{T}}_i$ of the form just noted, and with each possessing a solution, we can readily answer Question 1.

If $\bar{\mathcal{T}}_i(\hat{F}, \hat{G}, \hat{H})$ holds for any i , pole positionability for all α_i via choice of K is not possible, and conversely. It follows that if each $\bar{\mathcal{T}}_i$ includes one or more equalities, then the set of $\hat{F}, \hat{G}, \hat{H}$ for which pole positionability is not possible lies within a proper variety, and that for almost all $\hat{F}, \hat{G}, \hat{H}$, pole positionability for all p_l can be achieved.

On the other hand, if $\bar{\mathcal{T}}_i$ contains no equalities then it is clear that there exists a neighborhood of any one solution of $\bar{\mathcal{T}}_i(y)$ which consists entirely of solutions. (The fact that $\bar{\mathcal{T}}_i$ contains no inequations of the type $s_{\delta i}(y) \geq 0$ is crucial.) In this case, it cannot be true that for almost all $\hat{F}, \hat{G}, \hat{H}$, pole positionability can be achieved for all p_l .

This analysis of the $\bar{\mathcal{T}}_i$ answers Question 1.

Now one can also ask whether image (χ) is almost all of \mathbb{R}^n , for almost all (F, G, H) . Let us identify K with x and F, G, H and the p_l with y . Equations (A.1) yield a collection $\mathcal{S}(x, y)$ of polynomial equations. By Proposition 5.2A, there exists $\mathcal{T}(y) = \mathcal{T}(F, G, H, p_l)$ such that $\mathcal{T}(\hat{y})$ holds if and only if $\mathcal{S}(x, \hat{y})$ is solvable. Using arguments like those above, it is easy to check whether or not the set of \hat{y} for which $\mathcal{T}(\hat{y})$ is true is contained in a proper variety. If it is, then and only then will it be true that for almost all F, G, H , the map χ is almost onto \mathbb{R}^n .

We shall now turn to an analysis of Question 2.

If the closed-loop characteristic polynomial has all roots in the half plane $\text{Re}[s] < 0$, certain polynomial inequalities in the p_l obtainable from the Hurwitz determinants (see [4]) must hold, and conversely. Accordingly, we have

$$(A.2) \quad \begin{aligned} p_i(F, G, H, K) &= p_i, & i &= 1, \dots, n, \\ q_j(p_i) &> 0, & j &= 1, \dots, n. \end{aligned}$$

Identify K and p with x and F, G, H , with y . Regard (A.2) as an equation/inequation set $\mathcal{S}(x, y)$. By Tarski–Seidenberg–A, there exists $\mathcal{T}(y)$ such that $\mathcal{T}(\hat{y}) = \mathcal{T}(\hat{F}, \hat{G}, \hat{H})$ holds if and only if (A.2) can be satisfied by some K, p_i . If the set of \hat{y} such that $\mathcal{T}(\hat{y})$ holds is contained in a proper variety, then and only then Question 2 has an affirmative answer. The discussion of Question 1 described how one could check whether the set of \hat{y} , such that $\mathcal{T}(\hat{y})$ holds, is or is not contained in a proper variety.

Acknowledgments. We would like to thank the referees for several helpful suggestions, especially leading to a simplification of the proof of Theorem 2. We also thank Professor W. H. Gustafson for showing us Bertrand’s postulate on which the proof of Lemma 6.2 reposes.

REFERENCES

- [1] B. D. O. ANDERSON, N. K. BOSE AND E. I. JURY, *Output feedback stabilization and related problems—Solution via decision algebra methods*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 53–66.
- [2] E. ARTIN, *Galois Theory*, Univ. Notre Dame Press, Notre Dame, IN, 1971.
- [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, *Algebraic and geometric aspects of the analysis of feedback systems*, in Geometric Methods in Control Theory, C. I. Byrnes and C. F. Martin, eds., D. Reidel, Dordrecht, Holland, 1980.
- [5] ———, *High gain feedback and the stabilizability of multivariable systems*, in Proc. Vth International Conference on Analysis and Optimization of Systems, Springer-Verlag, Berlin, 1982.
- [6] C. I. BYRNES AND P. K. STEVENS, *Global properties of the root-locus map*, in Feedback Control of Linear and Nonlinear Systems, Lecture Notes in Control and Inf. Sciences, 39, Springer-Verlag, Berlin, 1982.
- [7] G. E. COLLINS, *Quantifier elimination for real closed fields by cylindrical algebraic decomposition*, in Proc. of 2nd Conference on Automata Theory and Formal Languages, Lecture Notes in Computer Sci., 33, Springer-Verlag, Berlin, 1975, pp. 134–183.
- [8] M. COSTE AND M. F. COSTE, *Topologies for real algebraic geometry*, in Topos Theoretic Methods in Geometry, A. Kock, ed., Various Pub. Series, 30, Aarhus Univ., 1979.
- [9] C. DELZELL, Ph.D. Dissertation, Stanford Univ., Stanford, CA, 1980.
- [10] F. GANTMACHER, *Theory of Matrices*, Vol. II, Chelsea, New York, 1962.
- [11] G. H. HARDY AND E. M. WRIGHT, *An Introduction to the Theory of Numbers* 5th ed., Oxford Univ. Press, Cambridge, 1979.
- [12] J. HARRIS, *Galois groups of enumerative problems*, Duke Math. J., 46 (1979), pp. 685–724.
- [13] R. HERMANN AND C. F. MARTIN, *Application of algebraic geometry to system theory—Part I*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 19–25.
- [14] J. E. HUMPHREYS, *Linear Algebraic Groups*, Springer-Verlag, New York, 1975.
- [15] N. JACOBSON, *Basic Algebra*, Vols. II and III, W. H. Freeman, San Francisco, 1974.
- [16] H. KIMURA, *Pole assignability by gain output feedback*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 509–516.
- [17] B. KOSTANT, *Lie group representations on polynomial rings*, Amer. J. Math., 85 (1963), pp. 327–404.
- [18] J. MILNOR, *Singular Points on a Complex Hypersurface*, Annals of Math. Studies, 61, Princeton Univ. Press, Princeton, NJ, 1968.
- [19] A. S. MORSE, W. A. WOLOVICH AND B. D. O. ANDERSON, *Generic pole assignment: Preliminary results*, in Proc. 20th IEEE Conference on Decision and Control, San Diego, 1981.
- [20] D. MUMFORD AND K. SUOMINEN, *Introduction to the theory of moduli*, in Proc. of 5th Summer School in Mathematics, Oslo, 1970.
- [21] I. R. SHAFAREVITCH, *Basic Algebraic Geometry*, Springer-Verlag, New York, 1974.
- [22] A. TARSKI, *A Decision Method for Elementary Algebra and Geometry*, Rand Corp., 1948.
- [23] J. C. WILLEMS AND W. H. HESSELINK, *Generic properties of the pole-placement problem*, in Proc. 7th IFAC Congress, 1978, pp. 1725–1729.

CALCULUS RULES ON THE APPROXIMATE SECOND-ORDER DIRECTIONAL DERIVATIVE OF A CONVEX FUNCTION*

J.-B. HIRIART-URRUTY†

Abstract. Given a real-valued convex function f , the approximate second-order directional derivative $f''_\varepsilon(x_0; d, d)$ of f at x_0 in the direction d is an object which is defined whenever the parameter ε is chosen strictly positive. The aim of this work is to derive expressions of the approximate second-order directional derivative of a function f which has been constructed from other functions f_i whose properties are better known; we address ourselves to the problem of calculating f''_ε , having the $(f_i)''_\eta$, $\eta > 0$, at our disposal. Calculus rules are given for the main functional operations preserving convexity: composition with an affine mapping, sum of functions, image of a function under a linear mapping, maximum of functions.

Key words. convex functions, approximate first-order directional derivative, approximate second-order directional derivative

Introduction. Given a convex function $f: \mathbb{R}^n \rightarrow \mathbb{R}$, x and d in \mathbb{R}^n , the *approximate first-order directional derivative* $f'_\varepsilon(x; d)$ of f at x in the direction d is defined for all $\varepsilon \geq 0$ by:

$$(0.1) \quad f'_\varepsilon(x; d) = \inf_{\lambda > 0} \{ [f(x + \lambda d) - f(x) + \varepsilon] \lambda^{-1} \}.$$

When $\varepsilon = 0$, the infimum of the difference quotient $[f(x + \lambda d) - f(x)] \lambda^{-1}$ over \mathbb{R}_+^* is also its limit when $\lambda \rightarrow 0^+$ and the resulting $f'_0(x; d)$, which is denoted simply by $f'(x; d)$, is what is called the first-order directional derivative of f at x in the d direction. As a general rule, the function $x \rightarrow f'(x; d)$ is not expected to be differentiable at all points x_0 of \mathbb{R}^n . However, as explained in our earlier work [5], it is not a tough condition on the behavior of f on the half-line $x_0 + \mathbb{R}_+ d$ to assume that $[f'(x_0 + \alpha d; d) - f'(x_0; d)] \alpha^{-1}$ has a limit when $\alpha \rightarrow 0^+$. This limit, denoted by $f''(x_0; d, d)$, is called the *second-order directional derivative of f at x_0 in the direction d* . Nevertheless, it remains that $f''(x_0; d, d)$ cannot be defined for all x_0 and d . With regard to this problem, to have “perturbed by ε ” the definition of $f'(x; d)$ has wrought good effects. Among them, we retain that, given $\varepsilon > 0$, the function $x \rightarrow f'_\varepsilon(x; d)$ admits a directional derivative at all x_0 and in all directions δ . For reasons which will be sketched later, we only consider the “diagonal case”, i.e., when $\delta = d$. The difference quotient $[f'_\varepsilon(x_0 + \alpha d; d) - f'_\varepsilon(x_0; d)] \alpha^{-1}$ does have a limit when $\alpha \rightarrow 0^+$, a limit which we denote by $f''_\varepsilon(x_0; d, d)$. That is precisely what is called the *approximate second-order directional derivative of f at x_0 in the direction d* . $f''_\varepsilon(x_0; d, d)$ indeed plays the role of an approximation of $f''(x_0; d, d)$ when the latter exists; we proved earlier (cf. [5, § III]) that:

$$(0.2) \quad \lim_{\varepsilon \rightarrow 0^+} f''_\varepsilon(x_0; d, d) = f''(x_0; d, d)$$

whenever the second-order directional derivative $f''(x_0; d, d)$ is defined. The charm of f''_ε lies in the fact that it is defined for all $\varepsilon > 0$ and therefore could serve as a substitute for f'' in devising second-order minimization methods. Minimization procedures are called second-order ones if the definition of x_{n+1} from x_n requires us to calculate a direction d_n by means of the functions $d \rightarrow f'_{\varepsilon_n}(x_n; d)$ and $d \rightarrow f''_{\varepsilon_n}(x_n; d, d)$. Considerations from the point of view of algorithms suggest that the directional derivative of $f'_\varepsilon(\cdot; d)$ at x in the direction $\delta = d$ is of main interest, which is the reason

* Received by the editors October 5, 1982, and in final form July 13, 1983.

† Université Paul Sabatier (Toulouse III), 118, route de Narbonne, 31062 Toulouse Cédex, France.

why we focused our attention on calculus rules on $f''_\varepsilon(x_0; d, d)$ rather than $f''_\varepsilon(x_0; d, \delta)$. The main features of $f''_\varepsilon(x; d, d)$ as a function of the parameters x , ε or d are portrayed in the author's survey paper [6].

Even having at our disposal the chain rules on the approximate first-order directional derivatives [3], the direct definition of $f''_\varepsilon(x_0; d, d)$ as the directional derivative of $f'_\varepsilon(\cdot; d)$ at x_0 in the d direction is not easily amenable to calculus. Instead, we shall use the expression for $f''_\varepsilon(x_0; d, d)$ such as worked out by Lemaréchal and Nurminskii initially [11] and generalized by Auslender [1]. We have that:

$$(0.3) \quad f''_\varepsilon(x_0; d, d) = \min \{ \mu [f'_\varepsilon(x_0; d) - f'(x_0; d)] \mid \mu \in M^f(\varepsilon) \},$$

where $M^f(\varepsilon)$ stands for the set of $\mu \geq 0$ minimizing the function $\mu \rightarrow \mu[f(x_0 + d/\mu) - f(x_0) + \varepsilon]$ on \mathbb{R}_+ . It is clear that calculus rules on f'_ε are a necessary ingredient in deriving calculus rules on f''_ε . Nevertheless, the key-ingredient remains the set $M^f(\varepsilon)$, and for all the functional operations we shall consider, our first task will be to express $M^f(\varepsilon)$ in terms of the $M^{f_i}(\eta)$, where the f_i are the functions from which f has been constructed. For that purpose, duality results, especially the one relating $f'_\varepsilon(x_0; d)$ and $M^f(\varepsilon)$ (cf. § 1), will be of constant use.

The paper is divided as follows. A first section contains all the necessary background on $f'_\varepsilon(x_0; d)$, $M^f(\varepsilon)$ and $f''_\varepsilon(x_0; d, d)$. Each of the further sections is devoted to a particular functional operation preserving convexity, namely: composition with an affine mapping, sum of functions, image of a function under a linear mapping, infimal convolution of functions and maximum of functions. Typically, a section devoted to a given functional operation contains:

- (i) a recall of the corresponding chain rule on f'_ε ; this "first-order" calculus rule allows us to settle some parameters which are necessary for the calculation of f''_ε ;
- (ii) a calculus rule on the intervals $M^f(\varepsilon)$, which is the key-result in the section;
- (iii) a general calculus rule on f''_ε , usually followed by a sharper result in the case where the f_i involved in the construction of f are strictly convex;
- (iv) a look at the limiting case $\varepsilon = 0$, when this case offers some interest.

All the operations considered except one "preserve smoothness" in a certain sense. By that we mean that if the involved f_i are C^2 and convex, the new function constructed from them is again C^2 and convex (at least under suitable assumptions). So, in such cases, one can reasonably expect to express f''_ε *exactly* in terms of $(f_i)''_{\eta_i}$ for appropriate values of parameters η_i . As for the last operation which consists of taking the maximum of functions ($f = \max_i f_i$), it is typically a *nonsmooth* operation. As a result, there is no hope to be able to express *exactly* f''_ε in terms of the $(f_i)''_{\eta_i}$, the difficulty being due to the fairly complicated expression of $M^f(\varepsilon)$. The last section is devoted to this functional operation which is by far the most difficult to handle, but also one of the most useful for applications.

1. Preliminary results. All the necessary background on convex analysis can be found in Rockafellar's book [12] which we refer to for basic definitions. Let f be a real-valued convex function on \mathbb{R}^n . Given a nonnegative ε , we denote by $f'_\varepsilon(x_0; d)$ the *approximate directional derivative* (or ε -*directional derivative*) of f at x_0 in the direction d , that is,

$$(1.1) \quad f'_\varepsilon(x_0; d) = \inf_{\lambda > 0} \{ [f(x_0 + \lambda d) - f(x_0) + \varepsilon] \lambda^{-1} \}.$$

As $\varepsilon \downarrow 0$, $f'_\varepsilon(x_0; d)$ decreases to $f'(x_0; d) = \lim_{\lambda \rightarrow 0^+} \{ [f(x_0 + \lambda d) - f(x_0)] \lambda^{-1} \}$. The *recession function* (or asymptotic function) f_∞ of f is the positively homogeneous closed

proper convex function defined by:

$$(1.2) \quad \forall d \in \mathbb{R}^n \quad f_\infty(d) = \sup_{\lambda > 0} \{[f(x_0 + \lambda d) - f(x_0)]\lambda^{-1}\}.$$

For fixed x_0 and d , we denote by $r_{x_0, d}^f$ (or simply r^f if the context clearly indicates what the x_0 and d are) the function defined on \mathbb{R} as follows:

$$r_{x_0, d}^f(\mu) = \begin{cases} \mu \left[f\left(x_0 + \frac{d}{\mu}\right) - f(x_0) \right] & \text{if } \mu > 0, \\ f_\infty(d) & \text{if } \mu = 0, \\ +\infty & \text{if } \mu < 0. \end{cases}$$

r^f is a closed proper convex function [12, p. 35] and, by definition, $f'_\varepsilon(x_0; d) = \inf \{r^f(\mu) + \varepsilon\mu \mid \mu \in \mathbb{R}\}$.

The conjugate function of r^f can be expressed in terms of $f'_\varepsilon(x_0; d)$. For that, let $t_{x_0, d}^f$ (or t^f) be defined on \mathbb{R} as

$$t_{x_0, d}^f(\varepsilon) = \begin{cases} -f'_\varepsilon(x_0; d) & \text{if } \varepsilon \geq 0, \\ +\infty & \text{if } \varepsilon < 0. \end{cases}$$

It comes from the definition of $f'_\varepsilon(x_0; d)$ that t^f is a *decreasing convex function*. Moreover, the functions t^f and r^f are related as follows (cf. [5, § II]):

$$(1.3) \quad t^f(\varepsilon) = (r^f)^*(-\varepsilon) \quad \text{for all } \varepsilon \in \mathbb{R},$$

$$(1.4) \quad r^f(\mu) = \sup_{\varepsilon > 0} [f'_\varepsilon(x_0; d) - \varepsilon\mu] \quad \text{for all } \mu \geq 0.$$

Also we note that $f'_{\varepsilon}(x_0; d)$ increases to $f_\infty(d) = \sup_{\varepsilon > 0} f'_\varepsilon(x_0; d)$ as $\varepsilon \uparrow +\infty$.

If f and g are two convex functions which coincide at x_0 (i.e., $f(x_0) = g(x_0)$), we derive from (1.1) and (1.4) the following comparison result:

$f(x_0 + \lambda d) \leq g(x_0 + \lambda d)$ for all $\lambda > 0$ if and only if $f'_\varepsilon(x_0; d) \leq g'_\varepsilon(x_0; d)$ for all $\varepsilon > 0$.

The concept of *approximate second-order directional derivative* f''_ε of f is defined by means of the ε -directional derivative of f . Given x_0 , $\varepsilon > 0$ and d , we set

$$f''_\varepsilon(x_0; d, d) = \lim_{\alpha \rightarrow 0^+} \{[f'_\varepsilon(x_0 + \alpha d; d) - f'_\varepsilon(x_0; d)]\alpha^{-1}\}.$$

This limit does exist and can be expressed in terms of $f'_\varepsilon(x_0; d)$, $f'(x_0; d)$ and a set $M_{x_0, d}^f(\varepsilon)$ defined as:

$$(1.5) \quad M_{x_0, d}^f(\varepsilon) = \{\mu \geq 0 \mid r^f(\mu) + \varepsilon\mu = f'_\varepsilon(x_0; d)\}.$$

Since $r^f(\mu) + \varepsilon\mu$ goes to $+\infty$ as $\mu \rightarrow +\infty$ (because $\varepsilon > 0$), the set $M_{x_0, d}^f(\varepsilon)$ (or simply $M^f(\varepsilon)$) is a nonempty compact interval of \mathbb{R}_+ . $\mu_{x_0, d}^f(\varepsilon)$ (or $\mu^f(\varepsilon)$) denotes the unique element of $M_{x_0, d}^f(\varepsilon)$ whenever it is single-valued. We now can recall the formulation of $f''_\varepsilon(x_0; d, d)$ [11], [1].

THEOREM 1.1. For all $\varepsilon > 0$,

$$(1.6) \quad f''_\varepsilon(x_0; d, d) = \bar{\mu}_{x_0, d}^f(\varepsilon)[f'_\varepsilon(x_0; d) - f'(x_0; d)],$$

where $\bar{\mu}_{x_0, d}^f(\varepsilon)$ stands for $\min \{\mu \mid \mu \in M_{x_0, d}^f(\varepsilon)\}$.

It is convenient to define $M^f(0)$ by extending the definition given in (1.5) to $\varepsilon = 0$. For that, let $a_f^*(x_0, d)$ (or simply a_f^*) denote the supremum of all $a \geq 0$ for which

$$f(x_0 + \lambda d) = f(x_0) + \lambda f'(x_0; d) \quad \text{for all } \lambda \in [0, a].$$

If $0 < a_f^* < +\infty$, that means that f restricted to the segment $x_0 + [0, a_f^*]d$ is an affine function. Having $a_f^* = 0$ means that

$$[f(x_0 + \lambda d) - f(x_0)]\lambda^{-1} > f'(x_0; d) \quad \text{for all } \lambda > 0,$$

while $a_f^* = +\infty$ corresponds precisely to the case where f is affine on the half-line $x_0 + \mathbb{R}_+ d$. $M^f(0)$ can be described as the segment $[1/a_f^*, +\infty[$ ($1/+\infty = 0$ and $1/0 = +\infty$ by convention).

$M^f(0)$ is nonempty if and only if $a_f^* > 0$. In such a case, the formula (1.6) written for $\varepsilon = 0$ yields that $f''_0(x_0; d, d) = 0$. This is consistent with the fact that $[f'(x_0 + \alpha d; d) - f'(x_0; d)]\alpha^{-1}$ is null for $\alpha > 0$ small enough. More generally, under mild assumptions on the behavior of f on $x_0 + \mathbb{R}_+ d$, the limit of $f''_\varepsilon(x_0; d, d)$ when $\varepsilon \rightarrow 0^+$ does exist and coincides with what is expected, namely the *second-order (directional) derivative* of f at x_0 in the d direction. We recall that f is said to have a second-order derivative at x_0 in the direction d if

$$(1.7) \quad \lim_{\alpha \rightarrow 0^+} \{[f'(x_0 + \alpha d; d) - f'(x_0; d)]\alpha^{-1}\}$$

$$(1.8) \quad = \lim_{\alpha \rightarrow 0^+} \left\{ 2 \left[\frac{f(x_0 + \alpha d) - f(x_0)}{\alpha} - f'(x_0; d) \right] \alpha^{-1} \right\}$$

exists in \mathbb{R}_+ . The common limit, denoted by $f''(x_0; d, d)$, is called the second-order derivative of f at x_0 in the d direction. For calculus rules on f'' , one can indifferently use Dini's formulation (i.e., (1.7)) or that of de la Vallée-Poussin (i.e., (1.8)). The following result on the behavior of $f''_\varepsilon(x_0; d, d)$ when $\varepsilon \rightarrow 0^+$ is from [5, § III].

THEOREM 1.2. *Suppose f has a second-order derivative at x_0 in the direction d . Then $f''_\varepsilon(x_0; d, d)$ converges to $f''(x_0; d, d)$ as $\varepsilon \rightarrow 0^+$.*

The following characterization of $M^f(\varepsilon)$ turns out to be a key-duality result for calculus rules on M^f .

THEOREM 1.3 [5]. *For all $\varepsilon \geq 0$, the subdifferential of t^f at ε is exactly $-M^f(\varepsilon)$.*

As verified in [1], a sufficient condition for $M^f(\varepsilon)$ to be single-valued for all $\varepsilon > 0$ is that f be strictly convex on the half-line $x_0 + \mathbb{R}_+ d$. This assumption also implies that a_f^* is null. f is affine on $x_0 + \mathbb{R}_+ d$ if and only if $M^f(\varepsilon) = \{0\}$ for all $\varepsilon > 0$. If $a_f^* < +\infty$, there then exists $\bar{\varepsilon} > 0$ satisfying:

$$\exists \lambda > 0 \quad [f(x_0 + \lambda d) - f(x_0) + \bar{\varepsilon}]\lambda^{-1} < f_\infty(d).$$

Define $\varepsilon_f^*(x_0, d)$ (or simply ε_f^*) as the supremum of all the $\bar{\varepsilon} > 0$ for which the above holds. Clearly, ε_f^* takes into account the behavior of $f(x_0 + \lambda d)$ when $\lambda \rightarrow +\infty$, and $\varepsilon_f^* = +\infty$ if and only if f is *coercive* in the d direction (i.e., $f_\infty(d) = +\infty$). It is convenient to extend the definition of ε_f^* to cover the case where f is affine on $x_0 + \mathbb{R}_+ d$ by posing $\varepsilon_f^* = 0$. We proved earlier [4], [5] that for $\varepsilon \in \mathbb{R}_+^*$:

$$(i) \quad M^f(\varepsilon) = \{0\} \Leftrightarrow \varepsilon > \varepsilon_f^*;$$

$$(ii) \quad 0 \notin M^f(\varepsilon) \Leftrightarrow f'_\varepsilon(x_0; d) < f_\infty(d) \Leftrightarrow f''_\varepsilon(x_0; d, d) > 0 \Leftrightarrow \varepsilon < \varepsilon_f^*.$$

Finally, observe that the definition itself of $M^f_{x_0, d}(\varepsilon)$ and the continuity of f make the multifunctions $\varepsilon \rightrightarrows M^f_{x_0, d}(\varepsilon)$ and $d \rightrightarrows M^f_{x_0, d}(\varepsilon)$ upper-semicontinuous.

2. Composition with an affine mapping. Given a convex function $f: \mathbb{R}^m \rightarrow \mathbb{R}$ and an affine mapping $A: \mathbb{R}^n \rightarrow \mathbb{R}^m$, let $f \circ A$ denote the convex function defined on \mathbb{R}^n by

$$\forall x \in \mathbb{R}^n \quad (f \circ A)(x) = f(Ax).$$

For all $\varepsilon \geq 0$ we have that

$$\begin{aligned}(f \circ A)'_{\varepsilon}(x; d) &= \inf_{\lambda > 0} \{[(f \circ A)(x + \lambda d) - (f \circ A)(x) + \varepsilon] \lambda^{-1}\} \\ &= \inf_{\lambda > 0} \{[f(Ax + \lambda A_0 d) - f(Ax) + \varepsilon] \lambda^{-1}\},\end{aligned}$$

where A_0 is the linear mapping associated with A . Thus

$$(2.1) \quad (f \circ A)'_{\varepsilon}(x; d) = f'_{\varepsilon}(Ax; A_0 d) \quad \text{for all } \varepsilon \geq 0.$$

The following results from the definition itself of $(f \circ A)''_{\varepsilon}(x_0; d, d)$.

THEOREM 2.1. *For all $\varepsilon > 0$*

$$(2.2) \quad (f \circ A)''_{\varepsilon}(x_0; d, d) = f''_{\varepsilon}(Ax_0; A_0 d, A_0 d).$$

As for the limiting case $\varepsilon = 0$, the expression of $(f \circ A)''(x_0; d, d)$ readily comes from its definition and is of the same kind as (2.2).

PROPOSITION 2.2. *Assume f has a second-order derivative at Ax_0 in the direction $A_0 d$. Then $f \circ A$ has a second-order derivative at x_0 in the direction d and*

$$(2.3) \quad (f \circ A)''(x_0; d, d) = f''(Ax_0; A_0 d, A_0 d).$$

The next example illustrates the utilization of the calculus rule (2.2).

Example. Given x_0 and d , let φ be defined on the real line by

$$\varphi(\lambda) = f(x_0 + \lambda d) \quad \text{for all } \lambda \in \mathbb{R}.$$

For a fixed $\lambda_0 \in \mathbb{R}$ we denote by $\varphi''_{\varepsilon}(\lambda_0)$ what should strictly be $\varphi''_{\varepsilon}(\lambda_0; 1, 1)$. We then have

$$(2.4) \quad \varphi''_{\varepsilon}(\lambda_0) = f''_{\varepsilon}(x_0 + \lambda_0 d; d, d).$$

3. Sum of functions. Given two convex functions f and g defined on \mathbb{R}^n , we consider their sum $f + g$. We know from [3, Theorem 2.1] that for all $\varepsilon \geq 0$:

$$(3.1) \quad (f + g)'_{\varepsilon}(x_0; d) = \max_{\substack{\varepsilon_1 \geq 0, \varepsilon_2 \geq 0 \\ \varepsilon_1 + \varepsilon_2 = \varepsilon}} \{f'_{\varepsilon_1}(x_0; d) + g'_{\varepsilon_2}(x_0; d)\}.$$

So, given $\varepsilon \geq 0$, there are nonnegative $\bar{\varepsilon}_1$ and $\bar{\varepsilon}_2$ adding up to ε for which

$$(3.2) \quad (f + g)'_{\varepsilon}(x_0; d) = f'_{\bar{\varepsilon}_1}(x_0; d) + g'_{\bar{\varepsilon}_2}(x_0; d).$$

Such a pair $(\bar{\varepsilon}_1, \bar{\varepsilon}_2)$ is called an *admissible pair*.

The expressions of $(f + g)'_{\varepsilon}(x_0; d)$ and $(f + g)''_{\varepsilon}(x_0; d, d)$ look pretty much alike. Recall that $\bar{\mu}^h(\varepsilon) = \min \{\mu | \mu \in M^h(\varepsilon)\}$.

THEOREM 3.1. *Given $\varepsilon > 0$, let $(\bar{\varepsilon}_1, \bar{\varepsilon}_2)$ be an admissible pair. Then*

$$(3.3) \quad \bar{\mu}^{f+g}(\varepsilon) = \max [\bar{\mu}^f(\bar{\varepsilon}_1), \bar{\mu}^g(\bar{\varepsilon}_2)]$$

so that

$$(3.4) \quad \begin{aligned}(f + g)''_{\varepsilon}(x_0; d, d) &= \max [\bar{\mu}^f(\bar{\varepsilon}_1), \bar{\mu}^g(\bar{\varepsilon}_2)] \\ &\quad \cdot \{[f'_{\bar{\varepsilon}_1}(x_0; d) - f'(x_0; d)] + [g'_{\bar{\varepsilon}_2}(x_0; d) - g'(x_0; d)]\}.\end{aligned}$$

In particular

$$(3.5) \quad (f + g)''_{\varepsilon}(x_0; d, d) \geq f''_{\bar{\varepsilon}_1}(x_0; d, d) + g''_{\bar{\varepsilon}_2}(x_0; d, d).$$

It may happen that $(f+g)'_\varepsilon(x_0; d) = f'(x_0; d) + g'_\varepsilon(x_0; d)$, that is to say $(0, \varepsilon)$ is an admissible pair. What this result implicitly says is that $a_f^* > 0$ in such a case and

$$(3.6) \quad (f+g)''_\varepsilon(x_0; d, d) = \max [1/a_f^*, \bar{\mu}^g(\varepsilon)] \cdot [g'_\varepsilon(x_0; d) - g'(x_0; d)].$$

In particular, if f is affine on $x_0 + \mathbb{R}_+ d$ (i.e., $a_f^* = +\infty$), we get that

$$(3.7) \quad (f+g)''_\varepsilon(x_0; d, d) = g''_\varepsilon(x_0; d, d) \quad \text{for all } \varepsilon > 0.$$

Only the relation (3.3) in Theorem 3.1 needs to be proved. It will be derived from the next key-result which relates $M^{f+g}(\varepsilon)$ to $M^f(\varepsilon_1)$ and $M^g(\varepsilon_2)$.

THEOREM 3.2. *Given $\varepsilon > 0$, let ε_1 and ε_2 be nonnegative real numbers adding up to ε . Then $(\varepsilon_1, \varepsilon_2)$ is an admissible pair if and only if $M^f(\varepsilon_1) \cap M^g(\varepsilon_2)$ is nonempty. In such a case,*

$$(3.8) \quad M^{f+g}(\varepsilon) = M^f(\varepsilon_1) \cap M^g(\varepsilon_2).$$

Proof. Consider convex functions t^{f+g} , t^f and t^g such as defined in § 1. The formula (3.1) can be rephrased as follows:

$$t^{f+g}(\varepsilon) = \inf_{\varepsilon_1 + \varepsilon_2 = \varepsilon} \{t^f(\varepsilon_1) + t^g(\varepsilon_2)\} \quad \text{for all } \varepsilon.$$

In other words, t^{f+g} is the *infimal convolution* $t^f \nabla t^g$ of t^f and t^g [12, p. 34]. This infimal convolution is exact at $\varepsilon \geq 0$ if there are nonnegative ε_1 and ε_2 adding up to ε for which

$$t^{f+g}(\varepsilon) = t^f(\varepsilon_1) + t^g(\varepsilon_2).$$

They are precisely the ε_1 and ε_2 for which

$$(f+g)'_\varepsilon(x_0; d) = f'_{\varepsilon_1}(x_0; d) + g'_{\varepsilon_2}(x_0; d),$$

i.e., the admissible pairs $(\varepsilon_1, \varepsilon_2)$.

Now, since the subdifferential of t^h at $\varepsilon \geq 0$ is $-M^h(\varepsilon)$ (see Theorem 1.3), it remains to apply the results on the subdifferential of an infimal convolution [10, p. 368] to get that: $\partial t^f(\varepsilon_1) \cap \partial t^g(\varepsilon_2)$ is nonempty if and only if the infimal convolution of t^f and t^g is exact at $\varepsilon = \varepsilon_1 + \varepsilon_2$ and, in such a case,

$$(3.9) \quad \partial(t^f \nabla t^g)(\varepsilon) = \partial t^f(\varepsilon_1) \cap \partial t^g(\varepsilon_2).$$

Hence the announced result is proved. \square

Remark 1. The calculus rule (3.9) is also valid for $\varepsilon = 0$ [10, p. 367] so that:

$$M^{f+g}(0) = M^f(0) \cap M^g(0).$$

Since $M^h(0)$ is $[1/a_h^*, +\infty[$ (see § 1), we therefore have:

$$(3.10) \quad a_{f+g}^* = \min(a_f^*, a_g^*) \quad \bar{\mathbb{R}}_+.$$

Along the same lines, one easily checks that

$$(3.11) \quad \varepsilon_{f+g}^* = \varepsilon_f^* + \varepsilon_g^* \quad \text{in } \bar{\mathbb{R}}_+.$$

Remark 2. If $(0, \varepsilon)$ is an admissible pair and if M^g is single-valued at ε , we necessarily have that $\mu^g(\varepsilon) \geq 1/a_f^*$. Thus the relationship (3.6) is made more precise in such a case with

$$(3.12) \quad (f+g)''_\varepsilon(x_0; d, d) = g''_\varepsilon(x_0; d, d).$$

It is clear that if one can find an admissible pair $(\bar{\varepsilon}_1, \bar{\varepsilon}_2)$ for which

$$M^f(\bar{\varepsilon}_1) = \{\mu^f(\bar{\varepsilon}_1)\} \quad \text{and} \quad M^g(\bar{\varepsilon}_2) = \{\mu^g(\bar{\varepsilon}_2)\},$$

equality then holds in (3.5). A sufficient condition for that is to assume that f and g are strictly convex on $x_0 + \mathbb{R}_+d$.

COROLLARY 3.3. *Suppose f and g are strictly convex on $x_0 + \mathbb{R}_+d$. Then, for every $\varepsilon > 0$, the $\bar{\varepsilon}_1$ and $\bar{\varepsilon}_2$ of any admissible pair are positive and:*

$$(3.13) \quad \begin{aligned} \mu^{f+g}(\varepsilon) &= \mu^f(\bar{\varepsilon}_1) = \mu^g(\bar{\varepsilon}_2), \\ (f+g)''_\varepsilon(x_0; d, d) &= f''_{\bar{\varepsilon}_1}(x_0; d, d) + g''_{\bar{\varepsilon}_2}(x_0; d, d). \end{aligned}$$

Proof. Since f and g are assumed strictly convex on $x_0 + \mathbb{R}_+d$, $M^f(0) = M^g(0) = \phi$ and both M^f and M^g are single-valued at all $\varepsilon > 0$. So, given $\varepsilon > 0$, the $\bar{\varepsilon}_1$ and $\bar{\varepsilon}_2$ of any admissible pair are positive. The rest follows from (3.8). \square

Example. Let f and g be defined on \mathbb{R} by

$$\forall x \in \mathbb{R} \quad f(x) = \max(-2x, -x-2, x-8), \quad g(x) = x^2.$$

We consider the case where $x_0 = 0$ and $d = 1$. The calculation of $(f+g)''_\varepsilon(x_0; d, d)$ will be illustrated for various values of ε . We observe that

$$\begin{aligned} M^f(0) &= [\tfrac{1}{2}, +\infty[, \\ M^f(\varepsilon) &= \begin{cases} \{\frac{1}{2}\} & \text{if } 0 < \varepsilon < 2, \\ \{\frac{1}{3}, \frac{1}{2}\} & \text{if } \varepsilon = 2, \\ \{\frac{1}{3}\} & \text{if } 2 < \varepsilon < 8, \\ [0, \frac{1}{3}] & \text{if } \varepsilon = 8, \\ \{0\} & \text{if } \varepsilon > 8; \end{cases} \\ M^g(0) &= \phi, \quad M^g(\varepsilon) = \{\varepsilon^{-1/2}\} \quad \text{if } \varepsilon > 0. \end{aligned}$$

Given $\varepsilon > 0$, the key-point is to find admissible pairs $(\bar{\varepsilon}_1, \bar{\varepsilon}_2)$. That will be done by using the characterization given in Theorem 3.2. First let $\varepsilon = 1$. The only admissible pair is $(0, 1)$ so that

$$M^{f+g}(1) = [\tfrac{1}{2}, +\infty[\cap\{1\} = \{1\}$$

and

$$(f+g)''_1(x_0; d, d) = g''_1(x_0; d, d) = 2.$$

This is an illustration of what has been said in Remark 2. Now let $\varepsilon = 5$. The only admissible pair is $(1, 4)$ so that

$$M^{f+g}(5) = M^f(1) = M^g(4) = \{\tfrac{1}{2}\}$$

and

$$(f+g)''_5(x_0; d, d) = f''_1(x_0; d, d) + g''_4(x_0; d, d) = \tfrac{1}{4} + 2 = \tfrac{9}{4}.$$

Finally let $\varepsilon = 10$. One easily checks that $(2, 8)$ is the unique admissible pair $(\varepsilon_1, \varepsilon_2)$. This is an example where

$$\bar{\mu}^{f+g}(\varepsilon) = \mu^g(\varepsilon_2) > \bar{\mu}^f(\varepsilon_1)$$

and

$$(f+g)''_\varepsilon(x_0; d, d) > f''_{\varepsilon_1}(x_0; d, d) + g''_{\varepsilon_2}(x_0; d, d).$$

Formula (3.4) yields the exact value of $(f+g)''_{10}(x_0; d, d)$, namely

$$(f+g)''_{10}(x_0; d, d) = \frac{1}{\sqrt{8}}\{[-1+2] + [2\sqrt{8}]\} = \frac{1}{\sqrt{8}} + 2.$$

g was a quadratic function in the above displayed example, so that g''_ε actually did not depend on ε . The formula giving $(f+g)''_\varepsilon$ can be somewhat simplified when g

turns out to be quadratic. To see this, consider a quadratic function g defined on \mathbb{R}^n as

$$\forall x \in \mathbb{R}^n \quad g(x) = \frac{1}{2} \langle Ax, x \rangle + \langle b, x \rangle + c,$$

where A is a symmetric positive semidefinite $n \times n$ matrix, b a vector in \mathbb{R}^n and c a real number. An easy calculation shows that

$$M^g(\varepsilon) = \left\{ \left(\frac{\langle Ad, d \rangle}{2\varepsilon} \right)^{1/2} \right\}$$

and

$$g''_\varepsilon(x_0; d, d) = \langle Ad, d \rangle \quad \text{for all } \varepsilon > 0.$$

If $Ad = 0$, g is affine on $x_0 + \mathbb{R}_+ d$ and, as already pointed out (cf. (3.7)), we have that

$$(f+g)''_\varepsilon(x_0; d, d) = f''_\varepsilon(x_0; d, d) \quad \text{for all } \varepsilon > 0.$$

When Ad is nonnull, the search for admissible pairs $(\varepsilon - \bar{\varepsilon}, \bar{\varepsilon})$ reduces to solving the equation

$$(3.14) \quad \left(\frac{\langle Ad, d \rangle}{2\bar{\varepsilon}} \right)^{1/2} \in M^f(\varepsilon - \bar{\varepsilon}), \quad \bar{\varepsilon} \in]0, \varepsilon].$$

COROLLARY 3.4. *Given a direction d such that $Ad \neq 0$ and $\varepsilon > 0$, there is one and only one $\bar{\varepsilon}$ solving the equation (3.14) and*

$$(3.15) \quad (f+g)''_\varepsilon(x_0; d, d) = \left(\frac{\langle Ad, d \rangle}{2\bar{\varepsilon}} \right)^{1/2} \{f'_{\varepsilon-\bar{\varepsilon}}(x_0; d) - f'(x_0; d) + (2\bar{\varepsilon}\langle Ad, d \rangle)^{1/2}\}.$$

If, moreover, M^f is single-valued at $\varepsilon - \bar{\varepsilon}$, we have that

$$(3.16) \quad (f+g)''_\varepsilon(x_0; d, d) = f''_{\varepsilon-\bar{\varepsilon}}(x_0; d, d) + \langle Ad, d \rangle.$$

Proof. Since $\langle Ad, d \rangle > 0$, $M^g(0)$ is empty so that $\bar{\varepsilon} > 0$ for any admissible pair $(\varepsilon - \bar{\varepsilon}, \bar{\varepsilon})$. Such admissible pairs do exist and are necessarily of the form $(\varepsilon - \bar{\varepsilon}, \bar{\varepsilon})$, with $\bar{\varepsilon}$ solving the equation (3.14) (cf. Theorem 3.2). It remains to be shown that there is only one $\bar{\varepsilon}$ solving (3.14). Assuming that $\bar{\varepsilon}$ and $\tilde{\varepsilon}$ solve (3.14), the monotonicity of the multifunction $\zeta \rightrightarrows M^f(\zeta)$ (cf. Theorem 1.3) gives

$$\left[\left(\frac{\langle Ad, d \rangle}{2\bar{\varepsilon}} \right)^{1/2} - \left(\frac{\langle Ad, d \rangle}{2\tilde{\varepsilon}} \right)^{1/2} \right] (\bar{\varepsilon} - \tilde{\varepsilon}) \geq 0.$$

Hence $\bar{\varepsilon} = \tilde{\varepsilon}$.

To get the expression (3.15) of $(f+g)''_\varepsilon(x_0; d, d)$, it suffices to apply (3.4) knowing that

$$g'_\varepsilon(x_0; d) = \langle \nabla g(x_0), d \rangle + (2\varepsilon \langle Ad, d \rangle)^{1/2}. \quad \square$$

Concerning the limiting case $\varepsilon = 0$, we have the following easy to prove result.

PROPOSITION 3.5. *Assume f and g have a second-order derivative at x_0 in the direction d . Then $f+g$ has a second-order derivative in the same direction with*

$$(3.17) \quad (f+g)''(x_0; d, d) = f''(x_0; d, d) + g''(x_0; d, d).$$

4. Image of a function under a linear mapping. Given a convex function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ and a surjective linear mapping A from \mathbb{R}^n to \mathbb{R}^m , we denote by Af the function

defined on \mathbb{R}^m by:

$$\forall y \in \mathbb{R}^m \quad (Af)(y) = \inf \{f(x) \mid Ax = y\}.$$

Such a function is sometimes called *image of f under A* [12, p. 38]. Af is clearly convex on \mathbb{R}^m and assumptions will be made on f to ensure that

$$(4.1) \quad X(y) = \{x \in \mathbb{R}^n \mid Ax = y \text{ and } f(x) = (Af)(y)\}$$

is nonempty for all $y \in \mathbb{R}^m$.

Our aim is to have an expression of $(Af)''_\varepsilon$ in terms of f''_ε which is valid for all $\varepsilon > 0$. To do that, the following is assumed throughout on f and A :

$$(H) \quad f_\infty(d) > 0 \quad \text{for all nonnull } d \text{ satisfying } Ad = 0.$$

This assumption is not necessary for deriving the next results. We however retain it for the sake of simplicity.¹

PROPOSITION 4.1. *Af is an everywhere finite convex function and*

$$(4.2) \quad (Af)_\infty(\delta) = \min \{f_\infty(d) \mid Ad = \delta\}.$$

Moreover, for each y , the set $X(y)$ is nonempty and compact.

Proof. The only thing to show is that $X(y)$ is compact; the rest follows from [12, Theorem 9].

Let $\theta: \mathbb{R}^n \times \mathbb{R}^m \rightarrow (-\infty, +\infty]$ be defined by

$$\theta(x, y) = f(x) + \psi_{\mathcal{A}}(x, y),$$

where \mathcal{A} denotes the graph of the mapping A . Clearly Af can be defined as:

$$\forall y \in \mathbb{R}^m \quad (Af)(y) = \inf_{x \in \mathbb{R}^n} \theta(x, y).$$

For each y , the partial function $\theta_y: x \rightarrow \theta_y(x) = \theta(x, y)$ is closed, proper and convex, and it results from calculus rules on the recession function [12, p. 66–67] that

$$(\theta_y)_\infty(d) = \theta_\infty(d, 0) = f_\infty(d) + \psi_{\mathcal{A}}(d, 0).$$

Thus $(\theta_y)_\infty(d) > 0$ for all nonnull d so that there is no nonnull direction d in which θ_y recedes [12, p. 69–70]. Hence the infimum of θ_y over \mathbb{R}^n is attained on a nonempty compact set. \square

Given $y_0 \in \mathbb{R}^m$ and $x_0 \in X(y_0)$, it comes from the ε -subgradient inequality that

$$(4.3) \quad \partial_\varepsilon(Af)(y_0) = \{y^* \in \mathbb{R}^m \mid A^*y^* \in \partial_\varepsilon f(x_0)\},$$

where A^* denotes the adjoint linear mapping.

Therefore, to get $(Af)'_\varepsilon(y_0; \delta)$, one has to express the support function of the inverse image of $\partial_\varepsilon f(x_0)$ under A^* .

THEOREM 4.2. *Given any $x_0 \in X(y_0)$, the following holds true for all $\varepsilon > 0$:*

$$(4.4) \quad (Af)'_\varepsilon(y_0; \delta) = \min_{Ad = \delta} f'_\varepsilon(x_0; d).$$

Proof. According to [12, Corollary 16.3.1], the formula (4.4) holds true if the image of A^* meets the relative interior of $\partial_\varepsilon f(x_0)$, i.e.,

$$(4.5) \quad \text{Im } A^* \cap \text{ri}(\partial_\varepsilon f(x_0)) \neq \emptyset.$$

¹ Likewise, A is assumed surjective to secure that $(Af)(y) < +\infty$ for all y . If not, the y_0 considered in the next results have to be taken in the interior of $\text{dom}(Af)$.

It comes from dual operations (cf. [12, Corollary 16.2.1]) that the condition above is equivalent to the following one:

$$(4.6) \quad f'_\varepsilon(x_0; d) > 0 \text{ or } f'_\varepsilon(x_0; -d) \leq 0 \quad \text{for all } d \in \ker A.$$

Since $x_0 \in X(y_0)$, we note that

$$f'(x_0; d) \geq 0 \quad \text{for all } d \in \ker A.$$

We therefore have

$$(4.7) \quad f_\infty(d) \geq f'(x_0; d) \geq 0 \quad \text{for all } d \in \ker A.$$

Consider first a nonnull direction $d \in \ker A$ for which $f_\infty(d) > f'(x_0; d)$. We know that

$$f_\infty(d) \geq f'_\varepsilon(x_0; d) > f'(x_0; d) \quad \text{for all } \varepsilon > 0.$$

Consequently, $f'_\varepsilon(x_0; d) > 0$ for $\varepsilon > 0$.

Now let d be a nonnull direction $d \in \ker A$ satisfying $f_\infty(d) = f'(x_0; d)$. In such a case we have that

$$f'(x_0; d) = f'_\varepsilon(x_0; d) = f_\infty(d) \quad \text{for all } \varepsilon > 0.$$

But, since $f_\infty(d) > 0$ (assumption (H)), $f'_\varepsilon(x_0; d) > 0$ for all $\varepsilon > 0$. Hence (4.6) is satisfied for all $\varepsilon > 0$ and the announced result is proved. \square

Given $\varepsilon > 0$, $x_0 \in X(y_0)$ and $\delta \in \mathbb{R}^m$, we denote by $D_\varepsilon(\delta)$ the (nonempty) set of directions $d \in \mathbb{R}^n$ satisfying

$$Ad = \delta, (Af)'_\varepsilon(y_0; \delta) = f'_\varepsilon(x_0; d).$$

THEOREM 4.3. $D_\varepsilon(\delta)$ is a compact convex set and

$$(4.8) \quad (Af)''_\varepsilon(y_0; \delta, \delta) \geq f''_\varepsilon(x_0; d, d)$$

for those $d \in D_\varepsilon(\delta)$ for which

$$(4.9) \quad \bar{\mu}_{y_0, \delta}^{Af}(\varepsilon) = \bar{\mu}_{x_0, d}^f(\varepsilon).$$

Before proving this result, we begin with the comparison of $M_{y_0, \delta}^{Af}(\varepsilon)$ to $M_{x_0, d}^f(\varepsilon)$ for $d \in D_\varepsilon(\delta)$.

THEOREM 4.4. We have that

$$(4.10) \quad M_{y_0, \delta}^{Af}(\varepsilon) = \bigcup_{d \in D_\varepsilon(\delta)} M_{x_0, d}^f(\varepsilon).$$

Proof. We firstly show that $D_\varepsilon(\delta)$ is a compact convex set. To see that, we mimic the proof of Proposition 4.1. Let $\theta: \mathbb{R}^n \times \mathbb{R}^m \rightarrow (-\infty, +\infty]$ be defined by:

$$\theta(d, \delta) = f'_\varepsilon(x_0; d) + \psi_{\mathcal{A}}(d, \delta).$$

We have that

$$(Af)'_\varepsilon(y_0; \delta) = \inf_{d \in \mathbb{R}^n} \theta(d, \delta).$$

Given δ , the function $\theta_\delta: d \rightarrow \theta(d, \delta)$ is a closed proper function and

$$(\theta_\delta)_\infty(d) = f'_\varepsilon(x_0; d) + \psi_{\mathcal{A}}(d, 0).$$

As we have already seen (cf. proof of Theorem 4.2), $f'_\varepsilon(x_0; d) > 0$ for all nonnull d in the kernel of A . Thus $(\theta_\delta)_\infty(d) > 0$ for all $d \neq 0$ so that there is no nonnull direction of recession for θ_δ . Hence the infimum of θ_δ over \mathbb{R}^n is attained on a (nonempty convex) compact set. Due to the definition itself of $D_\varepsilon(\delta)$, as also the continuity of the

function $\varepsilon \rightarrow f'_\varepsilon(x_0; d)$ on \mathbb{R}_+^* , it is easy to check that the multifunction $\varepsilon \rightrightarrows D_\varepsilon(\delta)$ is upper-semicontinuous on \mathbb{R}_+^* . From that (or from reasoning similar to that displayed above), we deduce that $\bigcup_{\varepsilon \in E} D_\varepsilon(\delta)$ is compact whenever E is a compact subset of \mathbb{R}_+^* .

Now fix $\varepsilon > 0$ and let $E = [\varepsilon, \bar{\varepsilon}] \subset \mathbb{R}_+^*$ be a compact neighborhood of ε . Posing $D(\delta) = \bigcup_{\varepsilon \in E} D_\varepsilon(\delta)$, we clearly have that

$$(4.11) \quad (Af)'_\varepsilon(y_0; \delta) = \min_{d \in D(\delta)} f'_\varepsilon(x_0; d) \quad \text{for all } \varepsilon \in E.$$

In other words,

$$t_{y_0, \delta}^{Af} = \max_{d \in D(\delta)} t_{x_0, d}^f \quad \text{in a neighborhood of } \varepsilon.$$

It then remains to apply the Rockafellar–Valadier theorem [10, p. 355] to get that

$$(4.12) \quad M_{y_0, \delta}^{Af}(\varepsilon) = \overline{\text{co}} \left\{ \bigcup_{d \in D_\varepsilon(\delta)} M_{x_0, d}^f(\varepsilon) \right\}.$$

But the multifunction $d \rightrightarrows M_{x_0, d}^f(\varepsilon)$ is upper-semicontinuous and $D_\varepsilon(\delta)$ is a compact convex set; thus the image of $D_\varepsilon(\delta)$ by this multifunction is a compact interval of \mathbb{R}_+^* . Hence the “ $\overline{\text{co}}$ ” is unnecessary in (4.12) and the desired equality is proved. \square

Proof of Theorem 4.3. For any $d \in D_\varepsilon(\delta)$ we have that

$$(4.13) \quad Ad = \delta \quad \text{and} \quad (Af)'_\varepsilon(y_0; \delta) = f'_\varepsilon(x_0; d).$$

If $Ad = \delta$, one easily verifies that

$$(Af)'_\varepsilon(y_0; \delta) \leq f'_\varepsilon(x_0; d).$$

Thus

$$(4.14) \quad (Af)'_\varepsilon(y_0; \delta) - (Af)'_\varepsilon(y_0; \delta) \geq f'_\varepsilon(x_0; d) - f'_\varepsilon(x_0; d) \quad \text{for all } d \in D_\varepsilon(\delta).$$

According to (4.10), there exists $\bar{d} \in D_\varepsilon(\delta)$ such that

$$\bar{\mu}_{y_0, \delta}^{Af}(\varepsilon) = \bar{\mu}_{x_0, \bar{d}}^f(\varepsilon).$$

The inequality (4.14) holds true for such \bar{d} so that

$$(Af)''_\varepsilon(y_0; \delta, \delta) = f''_\varepsilon(x_0; \bar{d}, \bar{d}).$$

Hence the theorem is proved. \square

COROLLARY 4.5. Assume f is strictly convex. Then Af is strictly convex and the multifunction X is single-valued so that

$$(4.15) \quad (Af)(y_0) = f(x(y_0)),$$

where $x(y_0)$ denotes the unique element of $X(y_0)$.

Moreover, there exists $\varepsilon^* > 0$ such that $D_\varepsilon(\delta)$ is reduced to one element $d_\varepsilon(\delta)$ whenever $\varepsilon \in]0, \varepsilon^*[$. Consequently, for ε small enough, we have:

$$(4.16) \quad (Af)'_\varepsilon(y_0; d) = f'_\varepsilon(x(y_0); d_\varepsilon(\delta)),$$

$$(4.17) \quad (Af)''_\varepsilon(y_0; \delta, \delta) \geq f''_\varepsilon(x(y_0); d_\varepsilon(\delta), d_\varepsilon(\delta)).$$

Proof. By considering the definition itself of Af , it is easy to see that Af is strictly convex and X single-valued. Since Af is strictly convex, $M_{y_0, \delta}^{Af}(\varepsilon)$ contains only one element $\mu_{y_0, \delta}^{Af}(\varepsilon)$ and it results from (4.10) that

$$(4.18) \quad \mu_{y_0, \delta}^{Af}(\varepsilon) = \mu_{x(y_0), d}^f(\varepsilon), \quad \text{for all } d \text{ in } D_\varepsilon(\delta).$$

This will allow us to show that D_ε is single-valued. Consider d_0 and d_1 in $D_\varepsilon(\delta)$, $d_0 \neq d_1$, and suppose that ε is small enough, namely $0 < \varepsilon < \varepsilon_{Af}^*(y_0, \delta)$. We have that

$$\begin{aligned}\mu_0 &= \mu_{y_0, \delta}^{Af}(\varepsilon) > 0, \\ Ad_0 &= Ad_1 = \delta, \\ (Af)'_\varepsilon(y_0; \delta) &= \mu_0 \left[f\left(x(y_0) + \frac{d_0}{\mu_0}\right) - f(x(y_0)) + \varepsilon \right] \\ &= \mu_0 \left[f\left(x(y_0) + \frac{d_1}{\mu_0}\right) - f(x(y_0)) + \varepsilon \right].\end{aligned}$$

Let $d_\alpha = \alpha d_1 + (1 - \alpha)d_0$ for some $\alpha \in]0, 1[$. The relations above and the strict convexity of f give

$$Ad_\alpha = \delta,$$

and

$$(Af)'_\varepsilon(y_0; \delta) > \mu_0 \left[f\left(x(y_0) + \frac{d_\alpha}{\mu_0}\right) - f(x(y_0)) + \varepsilon \right].$$

This contradicts the fact that $d_\alpha \in D_\varepsilon(\delta)$. Thus $D_\varepsilon(\delta)$ contains only one element $d_\varepsilon(\delta)$ and

$$\mu_{y_0, \delta}^{Af}(\varepsilon) = \mu_{x(y_0), d_\varepsilon(\delta)}^f(\varepsilon).$$

Thus (4.16) and (4.17) follow. \square

Example. Let $f: \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$ be a convex function and denote by φ the *marginal function* defined on \mathbb{R}^p by

$$\forall \xi_1 \in \mathbb{R}^p \quad \varphi(\xi_1) = \inf_{\xi_2 \in \mathbb{R}^q} f(\xi_1, \xi_2).$$

If $A: \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}^p$ is defined by $A(\xi_1, \xi_2) = \xi_1$, it is clear that φ is nothing else than Af . The assumption (H) is written here as

$$(H) \quad f_\infty(0, d_2) > 0 \quad \text{for all nonnull } d_2.$$

We then have

$$(4.19) \quad \varphi'_\varepsilon(\xi_1; d_1) = \min_{d_2 \in \mathbb{R}^q} f'_\varepsilon(\xi_1, \bar{\xi}_2; d_1, d_2),$$

$$(4.20) \quad \varphi''_\varepsilon(\xi_1; d_1, d_1) \geq \min_{d_2 \in D_\varepsilon(d_1)} f''_\varepsilon(\xi_1, \bar{\xi}_2; d_1, d_2; d_1, d_2),$$

where $f(\xi_1, \bar{\xi}_2) = \varphi(\xi_1)$ and $D_\varepsilon(d_1)$ denotes the set of d_2 for which $\varphi'_\varepsilon(\xi_1; d_1) = f'_\varepsilon(\xi_1, \bar{\xi}_2; d_1, d_2)$. The case of marginal functions is studied in full detail in [7] where assumptions weaker than (H) are considered.

Remark 1. The limiting case $\varepsilon = 0$ offers less interest in the present situation. Firstly, observe that even the assumption (H) does not ensure that

$$(4.21) \quad (Af)'(y_0; \delta) = \min_{Ad = \delta} f'(x_0; d).$$

A sufficient condition for that is:

$$f'(x_0; d) > 0 \text{ or } f'(x_0; -d) \leq 0 \quad \text{for all } d \in \ker A.$$

Assuming that (4.21) holds true, denote by $D_0(\delta)$ the set of d for which $Ad = \delta$ and $(Af)'(y_0; \delta) = f'(x_0; d)$. The comparison result between the second-order directional

derivatives of Af and f is as follows: if f has a second-order derivative at x_0 in the direction d for all $d \in D_0(\delta)$ and if Af has a second-order derivative at y_0 in the direction δ , then

$$(4.22) \quad (Af)''(y_0; \delta, \delta) \leq \inf_{d \in D_0(\delta)} f''(x_0; d, d).$$

This can easily be proved by considering the de la Vallée–Poussin formulation of second-order directional derivatives. As far as we can prove it, equality in (4.22) requires stronger assumptions (like twice differentiability) be made on f .

Remark 2. As already noticed, it comes from Theorem 4.3 that

$$(4.23) \quad (Af)''_\varepsilon(y_0; \delta, \delta) \geq \min_{d \in D_\varepsilon(\delta)} f''_\varepsilon(x_0; d, d).$$

Equality holds if, for example, f is differentiable at $x_0 \in X(y_0)$. In such a case, Af is differentiable at y_0 with $\nabla(Af)(y_0)$ as the unique element $y^* \in \mathbb{R}^m$ satisfying $A^*y^* = \nabla f(x_0)$. Thus, $D_0(\delta) = \{d | Ad = \delta\}$ and a look at the proof of Theorem 4.3 shows that

$$(Af)''_\varepsilon(y_0; \delta, \delta) = \min_{d \in D_\varepsilon(\delta)} f''_\varepsilon(x_0, d, d).$$

Moreover, the minimum is attained for those $d \in D_\varepsilon(\delta)$ for which $\bar{\mu}_{y_0, \delta}^{Af}(\varepsilon) = \bar{\mu}_{x_0, d}^f(\varepsilon)$.

5. Infimal convolution of functions. Given two convex functions f and g defined on \mathbb{R}^n , we recall that the infimal convolution $f \nabla g$ of f and g is defined in the following way:

$$\forall y \in \mathbb{R}^n \quad (f \nabla g)(y) = \inf_{\substack{x_1, x_2 \in \mathbb{R}^n \\ x_1 + x_2 = y}} \{f(x_1) + g(x_2)\}.$$

Of course, $f \nabla g$ can be viewed as an image of a function under a linear mapping so that results of the previous section apply. Nevertheless, due to its importance in convex analysis and optimization, the infimal convolution deserves that we linger on it.

As we will see by writing $f \nabla g$ in the form Ah , the assumption which corresponds to (H) in the previous section is:

$$(H) \quad f_\infty(d) + g_\infty(-d) > 0 \quad \text{for all nonnull } d.$$

We assume (H) throughout this section.

Recall that the infimal convolution of f and g is said to be exact at $y_0 = x_1 + x_2$ if one has

$$(5.1) \quad (f \nabla g)(y_0) = f(x_1) + g(x_2).$$

THEOREM 5.1. *Let $\varepsilon > 0$ and let (x_1, x_2) be a pair such that the infimal convolution of f and g is exact at $y_0 = x_1 + x_2$. Then*

$$(5.2) \quad \begin{aligned} (f \nabla g)'_\varepsilon(y_0; \delta) &= \min_{d_1 + d_2 = \delta} \max_{\substack{\varepsilon_1 \geq 0, \varepsilon_2 \geq 0 \\ \varepsilon_1 + \varepsilon_2 = \varepsilon}} \{f'_{\varepsilon_1}(x_1; d_1) + g'_{\varepsilon_2}(x_2; d_2)\} \\ &= \max_{\substack{\varepsilon_1 \geq 0, \varepsilon_2 \geq 0 \\ \varepsilon_1 + \varepsilon_2 = \varepsilon}} \inf_{d_1 + d_2 = \delta} \{f'_{\varepsilon_1}(x_1; d_1) + g'_{\varepsilon_2}(x_2; d_2)\}. \end{aligned}$$

Proof. Let $h: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ and $A: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ be defined as follows:

$$\begin{aligned} \forall (x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n \quad h(x_1, x_2) &= f(x_1) + g(x_2), \\ A(x_1, x_2) &= x_1 + x_2. \end{aligned}$$

It is clear that $f\nabla g = Ah$. We therefore obtain the first expression in (5.2) by applying Theorem 4.2 and (3.1).

For the second expression, we recall that

$$\partial_\varepsilon(f\nabla g)(y_0) = \bigcup_{\substack{\varepsilon_1 \geq 0, \varepsilon_2 \geq 0 \\ \varepsilon_1 + \varepsilon_2 = \varepsilon}} \{\partial_{\varepsilon_1} f(x_1) \cap \partial_{\varepsilon_2} g(x_2)\} \quad [3, \text{Theorem 3.1}].$$

Therefore, it results from calculus rules on support functions [12, p. 146–150] that

$$\begin{aligned} (f\nabla g)'_\varepsilon(y_0; \delta) &= \max_{\substack{\varepsilon_1 \geq 0, \varepsilon_2 \geq 0 \\ \varepsilon_1 + \varepsilon_2 = \varepsilon}} \psi_{\partial_{\varepsilon_1} f(x_1) \cap \partial_{\varepsilon_2} g(x_2)}^*(\delta), \\ \psi_{\partial_{\varepsilon_1} f(x_1) \cap \partial_{\varepsilon_2} g(x_2)}^*(\delta) &\leq \inf_{d_1 + d_2 = \delta} \{f'_{\varepsilon_1}(x_1; d_1) + g'_{\varepsilon_2}(x_2; d_2)\}. \end{aligned}$$

Thus

$$(f\nabla g)'_\varepsilon(y_0; \delta) \leq \max_{\substack{\varepsilon_1 \geq 0, \varepsilon_2 \geq 0 \\ \varepsilon_1 + \varepsilon_2 = \varepsilon}} \inf_{d_1 + d_2 = \delta} \{f'_{\varepsilon_1}(x_1; d_1) + g'_{\varepsilon_2}(x_2; d_2)\},$$

and the desired equality follows from the first expression in (5.2). \square

Remark 1. Calculating $(f\nabla g)'_\varepsilon(y_0; \delta)$ actually consists of performing two convolutions successively. Firstly we have that

$$\max_{\substack{\varepsilon_1 \geq 0, \varepsilon_2 \geq 0 \\ \varepsilon_1 + \varepsilon_2 = \varepsilon}} \{f'_{\varepsilon_1}(x_1; d_1) + g'_{\varepsilon_2}(x_2; d_2)\} = -(t_{x_1, d_1}^f \nabla t_{x_2, d_2}^g).$$

Performing this partial convolution (with respect to the variable ε) yields

$$(5.3) \quad \hat{f}_1(d_1) + \hat{g}_2(d_2) = f'_{\hat{\varepsilon}_1}(x_1; d_1) + g'_{\hat{\varepsilon}_2}(x_2; d_2).$$

Performing now the infimal convolution of \hat{f}_1 and \hat{g}_2 yields $(f\nabla g)'_\varepsilon(y_0; \delta)$. In other words, the relations (5.2) say that, in order to get $(f\nabla g)'_\varepsilon(y_0; \cdot)$, one is led to performing a convolution of $f'_\varepsilon(x_1; \cdot)$ and $g'_\varepsilon(x_2; \cdot)$ with respect to one variable followed by a convolution with respect to the other variable.

Remark 2. The function $K: (\varepsilon_1, \varepsilon_2, d_1, d_2) \rightarrow f'_{\varepsilon_1}(x_1; d_1) + g'_{\varepsilon_2}(x_2; d_2)$ is concave as a function of $(\varepsilon_1, \varepsilon_2)$ and convex as a function of (d_1, d_2) so that $(f\nabla g)'_\varepsilon(y_0; \delta)$ can be viewed as the saddle-value of K (with respect to maximizing over $C = \{(\varepsilon_1, \varepsilon_2) | \varepsilon_1 \geq 0, \varepsilon_2 \geq 0, \varepsilon_1 + \varepsilon_2 = \varepsilon\}$ and minimizing over $D = \mathbb{R}^n \times \mathbb{R}^n$). Therefore, saddle-points $(\bar{\varepsilon}_1, \bar{\varepsilon}_2, \bar{d}_1, \bar{d}_2)$ of K are such that

$$(5.4) \quad (f\nabla g)'_\varepsilon(y_0; \delta) = f'_{\bar{\varepsilon}_1}(x_1; \bar{d}_1) + g'_{\bar{\varepsilon}_2}(x_2; \bar{d}_2).$$

In accordance with our earlier notations, let $D_\varepsilon(\delta)$ denote the set of (d_1, d_2) satisfying:

$$\begin{aligned} (5.5) \quad d_1 + d_2 &= \delta, \\ (f\nabla g)'_\varepsilon(y_0; \delta) &= \max_{\substack{\varepsilon_1 \geq 0, \varepsilon_2 \geq 0 \\ \varepsilon_1 + \varepsilon_2 = \varepsilon}} \{f'_{\varepsilon_1}(x_1; d_1) + g'_{\varepsilon_2}(x_2; d_2)\} \\ &= h'_\varepsilon(x_1, x_2; d_1, d_2), \end{aligned}$$

where h is the function introduced in the proof of Theorem 5.1. According to Theorem 4.4 and Theorem 4.3, we have that:

$$(5.6) \quad M_{y_0, \delta}^{\nabla g}(\varepsilon) = \bigcup_{(d_1, d_2) \in D_\varepsilon(\delta)} M_{(x_1, x_2), (d_1, d_2)}^h(\varepsilon),$$

$$(5.7) \quad (f\nabla g)''_\varepsilon(y_0; \delta, \delta) \geq \min_{(d_1, d_2) \in D_\varepsilon(\delta)} h''_\varepsilon(x_1, x_2; (d_1, d_2), (d_1, d_2)).$$

To conclude, it remains now to apply the calculus rules of § 3. Recall that, given $(d_1, d_2) \in D_\varepsilon(\delta)$, $(\varepsilon_1, \varepsilon_2)$ is dubbed an admissible pair when

$$h'_\varepsilon(x_1, x_2; d_1, d_2) = f'_{\varepsilon_1}(x_1; d_1) + g'_{\varepsilon_2}(x_2; d_2).$$

It then results from Theorem 3.2 and Theorem 3.1 that

$$(5.8) \quad \begin{aligned} M^h_{(x_1, x_2), (d_1, d_2)}(\varepsilon) &= M^f_{x_1, d_1}(\varepsilon_1) \cap M^g_{x_2, d_2}(\varepsilon_2), \\ h''_\varepsilon(x_1, x_2; (d_1, d_2), (d_1, d_2)) &= \max[\bar{\mu}^f_{x_1, d_1}(\varepsilon_1), \bar{\mu}^g_{x_2, d_2}(\varepsilon_2)], \\ \{[f'_{\varepsilon_1}(x_1; d_1) - f'(x_1; d_1)] + [g'_{\varepsilon_2}(x_2; d_2) - g'(x_2; d_2)]\}. \end{aligned}$$

Hence we have an estimate of $(f\nabla g)''_\varepsilon(y_0; \delta, \delta)$ via the formula (5.7).

This somewhat complicated estimate of $(f\nabla g)''_\varepsilon(y_0; \delta, \delta)$ is made simpler when both f and g are strictly convex. Performing the infimal convolution of f and g yields a strictly convex function in such a case and, for all $y_0 \in \mathbb{R}^n$, there exists a unique x_0 for which

$$(f\nabla g)(y_0) = f(x_0) + g(y_0 - x_0).$$

Concerning $(f\nabla g)'_\varepsilon(y_0; \delta)$ and $(f\nabla g)''_\varepsilon(y_0; \delta, \delta)$, we derive the following from Corollary 4.5.

COROLLARY 5.2. *There is $\varepsilon^* > 0$ such that, for all $\varepsilon \in]0, \varepsilon^*[$, there exists a unique element $d_\varepsilon(\delta)$, an element $\varepsilon_1(\delta) \in [0, \varepsilon]$, such that;*

$$(5.9) \quad (f\nabla g)'_\varepsilon(y_0; \delta) = f'_{\varepsilon_1(\delta)}(x_0; d_\varepsilon(\delta)) + g'_{\varepsilon - \varepsilon_1(\delta)}(y_0 - x_0; \delta - d_\varepsilon(\delta)),$$

$$(5.10) \quad (f\nabla g)''_\varepsilon(y_0; \delta, \delta) \cong f''_{\varepsilon_1(\delta)}(x_0; d_\varepsilon(\delta), d_\varepsilon(\delta)) + g''_{\varepsilon - \varepsilon_1(\delta)}(y_0 - x_0; \delta - d_\varepsilon(\delta), \delta - d_\varepsilon(\delta)).$$

Example. A way of regularizing f consists in performing $f_r = f \nabla (r/2)\|\cdot\|^2$ for any $r > 0$, that is to say

$$(5.11) \quad \forall x \in \mathbb{R}^n \quad f_r(x) = \inf_u \left\{ f(u) + \frac{r}{2} \|x - u\|^2 \right\}.$$

The minimum in the definition of $f_r(x)$ above is achieved at an unique point which we denote by x_r . The following results on f_r are well known:

- (i) $f_r(x) \leq f(x)$ and $\lim_{r \rightarrow +\infty} f_r(x) = f(x)$ for all x ;
- (ii) x minimizes f on $\mathbb{R}^n \Leftrightarrow x = x_r \Leftrightarrow f_r(x) = f(x)$;
- (iii) f_r is differentiable and its gradient mapping is Lipschitz throughout \mathbb{R}^n .

The results of Theorem 5.1 allow us to obtain the exact expression of $(f_r)'_\varepsilon(x; \delta)$. From it are derived the following inequalities:

$$(5.12) \quad (f_r)'_\varepsilon(x; \cdot) \leq f'_\varepsilon(x_r; \cdot) \nabla \{r(x - x_r, \cdot) + \sqrt{2r\varepsilon}\|\cdot\|\},$$

$$(5.13) \quad (f_r)'_\varepsilon(x; \cdot) \geq f'_{\varepsilon_1}(x_r; \cdot) \nabla \{r(x - x_r, \cdot) + \sqrt{2r\varepsilon_2}\|\cdot\|\}$$

for all nonnegative ε_1 and ε_2 adding up to ε .

If we consider the particular case where f_r coincides with f at x , we then have that:

$$(5.14) \quad f'_\varepsilon(x; \cdot) \geq f'_\varepsilon(x_r; \cdot) \nabla \sqrt{2r\varepsilon}\|\cdot\| \geq (f_r)'_\varepsilon(x; \cdot),$$

$$(5.15) \quad (f_r)'_\varepsilon(x; \cdot) \geq f'_{\varepsilon_1}(x_r; \cdot) \nabla \sqrt{2r(\varepsilon - \varepsilon_1)}\|\cdot\| \quad \text{for all } \varepsilon_1 \in [0, \varepsilon].$$

To illustrate these formulae, take $f(x) = |x|$ on \mathbb{R} and the point $x = 0$. Since $f'_\varepsilon(0; \cdot) = |\cdot|$ for all $\varepsilon \geq 0$, the inequalities (5.14) and (5.15) imply that

$$(f_r)'_\varepsilon(0; \cdot) = |\cdot| \nabla \sqrt{2r\varepsilon}|\cdot|.$$

Hence $(f_r)'_{\varepsilon}(0; \cdot) = \min(1, \sqrt{2r\varepsilon})|\cdot|$ without further calculus. f_r and x_r can be calculated explicitly for this particular f , so that the inequalities (5.12) and (5.13) are easily checked.

As for the calculus of $(f_r)''_{\varepsilon}$, we suppose for the sake of simplicity that f is strictly convex. Then, for ε small enough, there exists a unique element $d_{\varepsilon}(\delta)$ and a unique element $\varepsilon_1(\delta)$ such that:

$$(5.16) \quad (f_r)'_{\varepsilon}(x; \delta) = f'_{\varepsilon_1(\delta)}(x_r; d_{\varepsilon}(\delta)) + r\langle x - x_r, \delta - d_{\varepsilon}(\delta) \rangle + \sqrt{2r(\varepsilon - \varepsilon_1(\delta))} \|\delta - d_{\varepsilon}(\delta)\|,$$

$$(5.17) \quad (f_r)''_{\varepsilon}(x; \delta, \delta) \cong f''_{\varepsilon_1(\delta)}(x_r; d_{\varepsilon}(\delta), d_{\varepsilon}(\delta)) + r\|\delta - d_{\varepsilon}(\delta)\|^2.$$

6. Maximum of functions. Let $\{f_i | i = 1, \dots, m\}$ be a collection of convex functions on \mathbb{R}^n and set $f = \max_{i=1, \dots, m} f_i$. It is known that the directional derivative of f at x_0 is the maximum of directional derivatives at x_0 of those f_i which satisfy $f_i(x_0) = f(x_0)$. The situation is different for the approximate directional derivative; due to its nonlocal nature, the formula giving $f'_{\varepsilon}(x_0; d)$ may require us to know $(f_i)'_{\varepsilon_i}(x_0; d)$ for all $i = 1, \dots, m$. More precisely, the following expansion holds for all $\varepsilon \geq 0$ (see [3], [8], [9]):

$$(6.1) \quad f'_{\varepsilon}(x_0; d) = \max \left\{ \sum_{i=1}^m (\alpha_i f_i)'_{\varepsilon_i}(x_0; d) \right\},$$

where the maximum is taken over the $(\alpha_1, \dots, \alpha_m)$ and $(\varepsilon_1, \dots, \varepsilon_m)$ satisfying:

$$(6.2) \quad \begin{aligned} \alpha_i &\geq 0 \quad \text{for all } i, & \sum_{i=1}^m \alpha_i &= 1, \\ \varepsilon_i &\geq 0 \quad \text{for all } i, & \sum_{i=1}^m \varepsilon_i + f(x_0) - \sum_{i=1}^m \alpha_i f_i(x_0) &= \varepsilon. \end{aligned}$$

Note that

$$(\alpha_i f_i)'_{\varepsilon_i}(x_0; d) = \begin{cases} \alpha_i (f_i)'_{\varepsilon_i/\alpha_i}(x_0; d) & \text{when } \alpha_i > 0, \\ 0 & \text{if } \alpha_i = 0. \end{cases}$$

Given x_0, d and $\varepsilon \geq 0$, we deduce from the above that there exist $(\bar{\alpha}_1, \dots, \bar{\alpha}_m)$ and $(\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_m)$ satisfying the conditions (6.2) and for which

$$f'_{\varepsilon}(x_0; d) = \sum_{i=1}^m (\bar{\alpha}_i f_i)'_{\bar{\varepsilon}_i}(x_0; d) = \sum_{\bar{\alpha}_i > 0} \bar{\alpha}_i (f_i)'_{\bar{\varepsilon}_i/\bar{\alpha}_i}(x_0; d).$$

$(\bar{\alpha}, \dots, \bar{\alpha}_m)$ and $(\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_m)$ form what we call an *admissible pair of vectors*. Only the $i \in \{1, \dots, m\}$ for which $\bar{\alpha}_i > 0$ are relevant for the calculation of $f'_{\varepsilon}(x_0; d)$. The other i are however of importance for determining $M^f(\varepsilon)$. As a general rule, it is not possible to express $f''_{\varepsilon}(x_0; d, d)$ in terms of the $(f_i)''_{\eta_i}$, $i = 1, \dots, m$. Imagine for example that f is the polyhedral function defined as the maximum of the m affine functions

$$f_i: x \rightarrow f_i(x) = \langle a_i, x \rangle + b_i, \quad i = 1, \dots, m.$$

As already seen, $(f_i)''_{\eta_i}(x_0; d, d)$ is null for all $\eta_i > 0$ while

$$M^{f_i}(\eta) = \begin{cases} \{0\} & \text{if } \eta > 0, \\ \mathbb{R}_+ & \text{if } \eta = 0. \end{cases}$$

So, the expression of $f''_{\varepsilon}(x_0; d, d)$ cannot be directly derived from those of $(f_i)''_{\eta_i}(x_0; d, d)$. The same example shows that expressing $M^f(\varepsilon)$ in terms of the $M^{f_i}(\eta)$ is not straightforward. That will precisely be our first task in this section.

For everything that follows, the key-functions are the functions $\theta_i: \mathbb{R}^2 \rightarrow (-\infty, +\infty]$ defined for all $i = 1, \dots, m$ as:

$$(6.3) \quad \theta_i(\varepsilon, \alpha) = \begin{cases} -(\alpha f_i)'_\varepsilon(x_0; d) & \text{if } \varepsilon \geq 0 \text{ and } \alpha \geq 0, \\ +\infty & \text{if not.} \end{cases}$$

For example, $\theta_i(\varepsilon, 0) = 0$ for all $\varepsilon \geq 0$ and $\theta_i(0, \alpha) = -\alpha (f_i)'(x_0; d)$ for all $\alpha \geq 0$. Actually, θ_i is of a very special structure, as we show now.

PROPOSITION 6.1. θ_i is the support function of the set

$$(6.4) \quad G_i = -\text{epi}(r^{f_i}).$$

Proof. We define $\sigma_i: \mathbb{R}^2 \rightarrow (-\infty, +\infty]$ as

$$(6.5) \quad \forall (\varepsilon, \alpha) \in \mathbb{R}^2 \quad \sigma_i(\varepsilon, \alpha) = \sup_{\mu > 0} [-\varepsilon\mu - \alpha r^{f_i}(\mu)].$$

By definition, we have that

$$\theta_i(\varepsilon, \alpha) = \sigma_i(\varepsilon, \alpha) \quad \text{for all } (\varepsilon, \alpha) \in \mathbb{R}_+^2.$$

Therefore,

$$(6.6) \quad \theta_i = \sigma_i + \psi_{\mathbb{R}_+^2} = \sigma_i + \psi_{\mathbb{R}_+^2}^*.$$

Thus, θ_i is a positively homogeneous closed proper convex function with \mathbb{R}_+^2 as domain. If G'_i denotes the set the support function of which is σ_i (i.e., $G'_i = \partial\sigma_i(0, 0)$), we infer from (6.6) that θ_i is the support function of

$$(6.7) \quad G_i = G'_i + \mathbb{R}_-^2.$$

According to its definition (see (6.5)), σ_i is the support function of

$$G'_i = \overline{\text{co}} \left\{ \bigcup_{\mu > 0} (-\mu, -r^{f_i}(\mu)) \right\}.$$

In other words, $-G'_i$ is the closed convex hull of the graph of r^{f_i} . Now, by the properties of the function r^{f_i} (see § 1), we have that

$$(6.8) \quad -G'_i + \mathbb{R}_+^2 = \text{epi } r^{f_i}.$$

Note incidentally that the addition of \mathbb{R}_+^2 to $-G'_i$ is unnecessary when $f_\infty(d) = +\infty$. In such a case, $\theta_i = \sigma_i$ and $G_i = G'_i$. If $f_\infty(d) < +\infty$, adding \mathbb{R}_+^2 to $-G'_i$ affects $-G'_i$ and yields $\text{epi } r^{f_i}$ precisely. Hence the announced result is proved. \square

G_i is the subdifferential of θ_i at $(0, 0)$. To get $\partial\theta_i(\varepsilon, \alpha)$ for the other points (ε, α) of \mathbb{R}_+^2 , we observe that

$$\partial\theta_i(\varepsilon, \alpha) = \{(\varepsilon^*, \alpha^*) \in G_i \mid \varepsilon\varepsilon^* + \alpha\alpha^* = \psi_{G_i}^*(\varepsilon, \alpha)\}.$$

The simplest way to determine $\partial\theta_i(\varepsilon, \alpha)$ is to solve the equation $\varepsilon\varepsilon^* + \alpha\alpha^* = \psi_{G_i}^*(\varepsilon, \alpha)$ graphically; see Fig. 1.

We obtain that

$$\partial\theta_i(0, \alpha) = -M^{f_i}(0) \times \{-f'(x_0; d)\} \quad \text{for } \alpha > 0,$$

$$\partial\theta_i(\varepsilon, \alpha) = \left\{ \left(-\mu, \frac{\varepsilon}{\alpha}\mu - f'_{\varepsilon/\alpha}(x_0; d) \right) \mid \mu \in M^{f_i}\left(\frac{\varepsilon}{\alpha}\right) \right\} \quad \text{if } \varepsilon > 0 \text{ and } \alpha > 0,$$

$$\partial\theta_i(\varepsilon, 0) = \{0\} \times]-\infty, -f_\infty(d)] \quad \text{if } \varepsilon > 0.$$

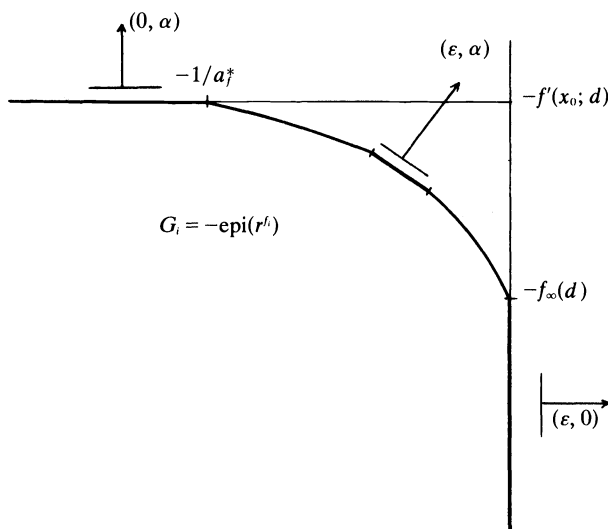


FIG. 1

Knowing an admissible pair of vectors $(\bar{\alpha}_1, \dots, \bar{\alpha}_n)$ and $(\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_m)$ enables us to calculate $f'_\varepsilon(x_0; d)$ since

$$(6.9) \quad -f'_\varepsilon(x_0; d) = \sum_{i=1}^m \theta_i(\bar{\varepsilon}_i, \bar{\alpha}_i).$$

Knowing $\partial\theta_i(\bar{\varepsilon}_i, \bar{\alpha}_i)$ for all $i = 1, \dots, m$ will now allow us to determine $M^f(\varepsilon)$.

THEOREM 6.2. *Given $\varepsilon \geq 0$, let $(\bar{\alpha}_1, \dots, \bar{\alpha}_m)$ and $(\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_m)$ form an admissible pair of vectors. Then $\mu \in M^f(\varepsilon)$ if and only if there exists α^* such that*

$$(6.10) \quad (-\mu, -\mu[f(x_0) - f_i(x_0)] + \alpha^*) \in \partial\theta_i(\bar{\varepsilon}_i, \bar{\alpha}_i) \quad \text{for all } i = 1, \dots, m.$$

Proof. Let $\Phi: \mathbb{R}^2 \rightarrow (-\infty, +\infty]$ be defined as follows:

$$\Phi(\varepsilon, \alpha) = \min \left\{ \sum_{i=1}^m \theta_i(\varepsilon_i, \alpha_i) \mid \sum_{i=1}^m \alpha_i = \alpha, \sum_{i=1}^m [\varepsilon_i + \alpha_i(f(x_0) - f_i(x_0))] = \varepsilon \right\}.$$

Clearly, Φ can be written as the *image of H under A* , where $H: \mathbb{R}^{2m} \rightarrow (-\infty, +\infty]$ and $A: \mathbb{R}^{2m} \rightarrow \mathbb{R}^2$ are defined as

$$H(\varepsilon_1, \dots, \varepsilon_m, \alpha_1, \dots, \alpha_m) = \sum_{i=1}^m \theta_i(\varepsilon_i, \alpha_i),$$

$$A(\varepsilon_1, \dots, \varepsilon_m, \alpha_1, \dots, \alpha_m) = \left(\sum_{i=1}^m [\varepsilon_i + \alpha_i(f(x_0) - f_i(x_0))], \sum_{i=1}^m \alpha_i \right).$$

Now, since $(\bar{\alpha}_1, \dots, \bar{\alpha}_m)$ and $(\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_m)$ form an admissible pair of vectors, we have that

$$\Phi(\varepsilon, 1) = \sum_{i=1}^m \theta_i(\bar{\alpha}_i, \bar{\varepsilon}_i).$$

Consequently, it results from calculus rules on subdifferentials that

$$(6.11) \quad \partial\varphi(\varepsilon, 1) = \{(\varepsilon^*, \alpha^*) \mid A^*(\varepsilon^*, \alpha^*) \in \partial H(\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_m, \bar{\alpha}_1, \dots, \bar{\alpha}_m)\}.$$

Due to the special structure of A^* and H , we deduce from the relation above that $(\varepsilon^*, \alpha^*) \in \partial\Phi(\varepsilon, 1)$ if and only if

$$(\varepsilon^*, \varepsilon^*[f(x_0) - f_i(x_0)] + \alpha^*) \in \partial\theta_i(\bar{\varepsilon}_i, \bar{\alpha}_i) \quad \text{for all } i = 1, \dots, m.$$

Φ has been defined precisely in such a way that $t^f(\varepsilon) = \Phi(\varepsilon, 1)$. Therefore

$$\partial t^f(\varepsilon) = \{\varepsilon^* | \exists \alpha^* \text{ such that } (\varepsilon^*, \alpha^*) \in \partial\Phi(\varepsilon, 1)\}$$

and the announced result is proved since $-M^f(\varepsilon) = \partial t^f(\varepsilon)$. \square

Let $\bar{\alpha}_i$ and $\bar{\varepsilon}_i$ be the i th components of an admissible pair of vectors. According to the different formulations of $\partial\theta_i(\bar{\varepsilon}_i, \bar{\alpha}_i)$ for $i = 1, \dots, m$, the i th condition on μ in (6.10) becomes:

$$(6.12) \quad r^{f_i}(\mu) \leq \mu[f(x_0) - f_i(x_0)] - \alpha^* \quad \text{if } \bar{\varepsilon}_i = \bar{\alpha}_i = 0,$$

$$(6.13) \quad \mu = 0 \text{ and } \alpha^* \leq -(f_i)_\infty(d) \quad \text{if } \bar{\varepsilon}_i > 0 \text{ and } \bar{\alpha}_i = 0,$$

$$(6.14) \quad \mu[f(x_0) - f_i(x_0)] - \alpha^* = f'_i(x_0; d) \text{ and } \mu \in M^{f_i}(0) \quad \text{if } \bar{\varepsilon}_i = 0 \text{ and } \bar{\alpha}_i > 0,$$

there exists $\mu_i \in M^{f_i}(\bar{\varepsilon}_i/\bar{\alpha}_i)$ such that $\mu = \mu_i$ and

$$(6.15) \quad \mu_i[f(x_0) - f_i(x_0)] - \alpha^* = (f_i)'_{\bar{\varepsilon}_i/\bar{\alpha}_i}(x_0; d) - \frac{\bar{\varepsilon}_i}{\bar{\alpha}_i} \mu_i,$$

if both $\bar{\varepsilon}_i$ and $\bar{\alpha}_i$ are > 0 .

These various conditions call for some comments:

(a) $\bar{\varepsilon}_i > 0$ and $\bar{\alpha}_i = 0$ cannot occur if f_i is coercive in the direction d ; if it occurs, $M^f(\varepsilon)$ is fully determined by this i th condition alone since it implies $M^f(0) = \{0\}$.

(b) $\bar{\varepsilon}_i = 0$ and $\bar{\alpha}_i > 0$ cannot happen if $a_{f_i}^* = 0$; if it happens, this i th condition provides the estimate $M^f(\varepsilon) \subset M^{f_i}(0)$.

(c) If both $\bar{\varepsilon}_i$ and $\bar{\alpha}_i$ are > 0 , the corresponding condition implies that $M^f(\varepsilon) \subset M^{f_i}(\bar{\varepsilon}_i/\bar{\alpha}_i)$.

It might happen that all the $\bar{\varepsilon}_i$ are null while there necessarily exists i such that $\bar{\alpha}_i > 0$. It thus results from (b) and (c) above that

$$(6.16) \quad M^f(\varepsilon) \subset \bigcap_{\bar{\alpha}_i > 0} M^{f_i}\left(\frac{\bar{\varepsilon}_i}{\bar{\alpha}_i}\right).$$

To illustrate the foregoing, consider again the case of a polyhedral function f defined as the maximum of the m affine functions $f_i(x) = \langle a_i, x \rangle + b_i$. We have:

$$(6.17) \quad f'_\varepsilon(x_0; d) = \max \left\{ \sum_{i=1}^m \alpha_i \langle a_i, d \rangle \mid \alpha_i \geq 0 \text{ for } i, \sum_{i=1}^m \alpha_i = 1 \text{ and } \sum_{i=1}^m \alpha_i (f(x_0) - f_i(x_0)) \leq \varepsilon \right\}.$$

If $\varepsilon > \max_i (f(x_0) - f_i(x_0))$, the last constraint is irrelevant and solving the maximization problem above yields $(\bar{\alpha}_1, \dots, \bar{\alpha}_m) = e_i$, where e_i is the i th unit vector of \mathbb{R}^m and i an index that

$$\langle a_i, d \rangle = \max_j \langle a_j, d \rangle = f'_\varepsilon(x_0; d).$$

$(\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_m)$ can be chosen so that $\bar{\varepsilon}_i > 0$ for all i . It then results from (6.13) or (6.15) that $M^f(\varepsilon) = \{0\}$.

If $\varepsilon \leq \max_i (f(x_0) - f_i(x_0))$, $(\bar{\alpha}_1, \dots, \bar{\alpha}_m)$ is obtained by solving the linear programming problem (6.17) while the $\bar{\varepsilon}_i$ are chosen so that they verify:

$$(6.18) \quad \begin{aligned} \bar{\varepsilon}_i &\geq 0 \quad \text{for all } i = 1, \dots, m, \\ \sum_{i=1}^m \bar{\varepsilon}_i + \sum_{i=1}^m \bar{\alpha}_i (f(x_0) - f_i(x_0)) &= \varepsilon. \end{aligned}$$

Usually the conditions above yield that $\bar{\varepsilon}_i = 0$ for all i . If so, $\mu \in M^f(\varepsilon)$ if and only if there is α^* such that (μ, α^*) belongs to the polyhedron Π of \mathbb{R}^2 defined as follows:

$$(6.19) \quad \begin{aligned} \alpha^* &\leq \mu[f(x_0) - f_i(x_0)] - \langle a_i, d \rangle \quad \text{for those } i \text{ for which } \bar{\alpha}_i = 0, \\ \mu &\geq 0 \text{ and } \alpha^* = \mu[f(x_0) - f_i(x_0)] - \langle a_i, d \rangle \quad \text{for those } i \text{ for which } \bar{\alpha}_i > 0. \end{aligned}$$

It might be easier to solve (6.17) in its dual form. Indeed,

$$f'_\varepsilon(x_0; d) = f'(x_0; d) + \rho(\varepsilon),$$

where $\rho(\varepsilon)$ is the optimal value of the dual program:

$$(6.17)^\circ \quad \begin{aligned} &\text{Minimize } -(\beta_1 + \beta_2 \varepsilon) \text{ subject to } \beta_2 \leq 0 \text{ and} \\ &\beta_1 + \beta_2 (f(x_0) - f_i(x_0)) \leq -\langle a_i, d \rangle + f'(x_0; d) \quad \text{for all } i = 1, \dots, m. \end{aligned}$$

In particular, if $\varepsilon < \tilde{\varepsilon} = \min \{f(x_0) - f_i(x_0) \mid f(x_0) > f_i(x_0)\}$, the program (6.17) $^\circ$ can easily be solved graphically and a solution is²:

$$\begin{aligned} \bar{\beta}_1 &= 0, \\ \bar{\beta}_2 &= \min \left\{ \frac{f'(x_0; d) - \langle a_i, d \rangle}{f(x_0) - f_i(x_0)} \mid f(x_0) > f_i(x_0) \text{ and } \langle a_i, d \rangle > f'(x_0; d) \right\}. \end{aligned}$$

Consequently, we have for all $\varepsilon \in]0, \tilde{\varepsilon}[$:

$$(6.20) \quad \begin{aligned} f'_\varepsilon(x_0; d) &= f'(x_0; d) - \bar{\beta}_2 \varepsilon, \\ M^f(\varepsilon) &= \{-\bar{\beta}_2\} \quad \text{and} \quad f''_\varepsilon(x_0; d, d) = (\bar{\beta}_2)^2 \varepsilon. \end{aligned}$$

Take for instance f_1, f_2 and f_3 defined on \mathbb{R} as:

$$f_1(x) = -2x, \quad f_2(x) = x, \quad f_3(x) = 2x - 1.$$

We consider $x_0 = 0$ and $d = 1$. If $\varepsilon > 1$ or < 1 , we know from the above that $M^f(\varepsilon)$ and $f''_\varepsilon(x_0; d, d)$ are fully determined without further calculus. If $\varepsilon = 1$, $M^f(\varepsilon)$ has to be calculated following the scheme we have drawn. Solving the linear program (6.17) yields $(\bar{\alpha}_1, \bar{\alpha}_2, \bar{\alpha}_3) = (0, 0, 1)$ while $(\bar{\varepsilon}_1, \bar{\varepsilon}_2, \bar{\varepsilon}_3)$ is necessarily the null vector. The polyhedron Π determined by the linear equalities or inequalities in (6.20) is then

$$\Pi = \{(\mu, \alpha^*) \in \mathbb{R}^2 \mid \mu \geq 0, \alpha^* \leq -1, \alpha^* = \mu - 2\}.$$

Thus $M^f(1)$ is $[0, 1]$.

To know whether or not $\bar{\varepsilon}_i$ and $\bar{\alpha}_i$ are strictly positive is of main importance since it determines which inequalities in (6.12)–(6.15) have to be used in calculating $M^f(\varepsilon)$. Some indications a priori can be given in that respect.

COROLLARY 6.3. *Suppose that $a_{f_i}^* = 0$ for all $i = 1, \dots, m$ and consider $\varepsilon \leq \varepsilon_f^*$. Then, for any admissible pair of vectors $(\bar{\alpha}_1, \dots, \bar{\alpha}_m)$ and $(\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_m)$, we have that either both $\bar{\alpha}_i$ and $\bar{\varepsilon}_i$ are null or both $\bar{\alpha}_i$ and $\bar{\varepsilon}_i$ are strictly positive.*

² Of course it is assumed there exists i such that $f(x_0) > f_i(x_0)$ and $\langle a_i, d \rangle > f'(x_0; d)$. Otherwise, it is clear that $M^f(\varepsilon) = \{0\}$ and $f''_\varepsilon(x_0; d, d) = 0$.

This result is a direct consequence of the remarks following conditions (6.12)–(6.15). To have assumed that $\varepsilon \leq \varepsilon_f^*$ secures that $M^f(\varepsilon)$ is not reduced to $\{0\}$.

COROLLARY 6.4. *Suppose that f_i is strictly convex on $x_0 + \mathbb{R}_+d$ for all $i = 1, \dots, m$ and let $\varepsilon \leq \varepsilon_f^*$. Then, for any admissible pair of vectors $(\bar{\alpha}_1, \dots, \bar{\alpha}_m)$ and $(\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_m)$, we have that:*

$$(6.21) \quad \mu^f(\varepsilon) = \mu^{f_i}(\bar{\varepsilon}_i/\bar{\alpha}_i) \quad \text{for all } i \text{ such that } \bar{\varepsilon}_i \cdot \bar{\alpha}_i > 0.$$

Proof. It is clear that f itself is strictly convex on $x_0 + \mathbb{R}_+d$. According to the previous corollary, there are only two possibilities on the $\bar{\varepsilon}_i$ and $\bar{\alpha}_i$: either both are null or both are strictly positive. In the latter case, the equality (6.21) derives from the inclusion (6.16). \square

The equality (6.21) combined with the expression of $f'_\varepsilon(x_0; d)$ asserts a remarkable “equilibrium property” whose geometrical interpretation is meaningful. Under the assumptions of Corollary 6.4, we have that:

$$\begin{aligned} f'_\varepsilon(x_0; d) &= \sum_{\bar{\varepsilon}_i \cdot \bar{\alpha}_i > 0} \bar{\alpha}_i (f_i)'_{\bar{\varepsilon}_i/\bar{\alpha}_i}(x_0; d), \\ \sum_{\bar{\varepsilon}_i \cdot \bar{\alpha}_i > 0} \bar{\alpha}_i \left[\frac{\bar{\varepsilon}_i}{\bar{\alpha}_i} + (f(x_0) - f_i(x_0)) \right] &= \varepsilon, \\ \mu^f(\varepsilon) &= \mu^{f_i} \left(\frac{\bar{\varepsilon}_i}{\bar{\alpha}_i} \right) \quad \text{if } \bar{\varepsilon}_i \cdot \bar{\alpha}_i > 0. \end{aligned}$$

So, $f'_\varepsilon(x_0; d)$ is a convex combination of the $(f_i)'_{\bar{\varepsilon}_i/\bar{\alpha}_i}(x_0; d)$ where the “dummy indices” i (i.e., those for which $\bar{\alpha}_i = \bar{\varepsilon}_i = 0$) are omitted. ε itself is a convex combination of the $\bar{\tau}_i = \bar{\varepsilon}_i/\bar{\alpha}_i + f(x_0) - f_i(x_0)$; $\bar{\tau}$ is exactly the quantity by which one has to move down from the point $(x_0, f(x_0))$ to calculate $(f_i)'_{\bar{\varepsilon}_i/\bar{\alpha}_i}(x_0; d)$. Finally, the equality $\mu^f(\varepsilon) = \mu^{f_i}(\bar{\varepsilon}_i/\bar{\alpha}_i)$ is a nice relationship between the “points of contact” of the epigraphs of f and f_i . Figure 2 is an illustration of that situation.

Before going into the problem of expressing $f''_\varepsilon(x_0; d, d)$, let us fix some notation. We set

$$\begin{aligned} I(x_0) &= \{i = 1, \dots, m \mid f(x_0) = f_i(x_0)\}, \\ I(x_0; d) &= \{i \in I(x_0) \mid f'(x_0; d) = f'_i(x_0; d)\}. \end{aligned}$$

Moreover, given $\bar{\alpha} = (\bar{\alpha}_1, \dots, \bar{\alpha}_m)$ from an admissible pair of vectors, we define the *support* of $\bar{\alpha}$ to be the set $\mathcal{S}(\bar{\alpha})$ of indices i such that $\bar{\alpha}_i > 0$. $\mathcal{S}(\bar{\alpha})$ depends on $\bar{\alpha}$ and thereby on ε .

Due to the constraint

$$\sum_{i=1}^m \bar{\varepsilon}_i + \sum_{i \notin I(x_0)} \bar{\alpha}_i (f(x_0) - f_i(x_0)) = \varepsilon,$$

it is easy to see that $\mathcal{S}(\bar{\alpha}) \cap I(x_0)$ is nonempty whenever $\varepsilon < \tilde{\varepsilon} = \min \{f(x_0) - f_i(x_0) \mid i \notin I(x_0)\}$. In fact,

$$(6.22) \quad \sum_{i \notin I(x_0)} \bar{\alpha}_i \leq \left(\varepsilon - \sum_{i=1}^m \bar{\varepsilon}_i \right) / \tilde{\varepsilon} \leq \varepsilon / \tilde{\varepsilon},$$

so that the smaller the ε , the smaller are the $\bar{\alpha}_i$ for $i \notin I(x_0)$.

The mathematical program (P_ε) described in (6.1)–(6.2) and whose optimal value is $f'_\varepsilon(x_0; d)$ is an “approximation” of the program (P_0) whose optimal value is $f'(x_0; d)$, i.e.

$$\begin{aligned} f'(x_0; d) &= \max_{i \in I(x_0)} f'_i(x_0; d) = f'_i(x_0; d) \quad \text{for any } i \in I(x_0; d) \\ (P_0) \quad &= \max \left\{ \sum_{i \in I(x_0)} \alpha_i f'_i(x_0; d) \mid \alpha_i \geq 0 \text{ for all } i, \sum_{i \in I(x_0)} \alpha_i = 1 \right\}. \end{aligned}$$

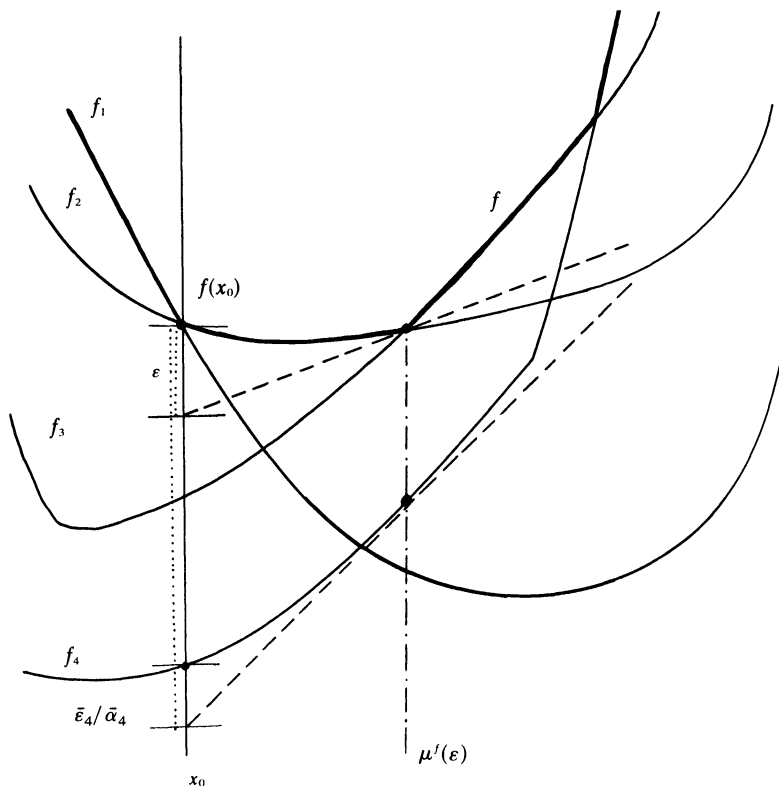


FIG. 2

Any admissible pair of vectors for (P_0) is of the form $(\bar{\alpha}, 0)$ where $\mathcal{S}(\bar{\alpha}) \subset I(x_0; d)$. It is not hard to verify that (P_ϵ) “converges” to (P_0) in the sense that the limit points of the solutions of (P_ϵ) [resp. the limit of the optimal value of (P_ϵ)] are solutions of (P_0) [resp. is the optimal value of (P_0)]. Consequently, $\mathcal{S}(\bar{\alpha}) \cap I(x_0; d)$ is nonempty for ϵ small enough and any $\bar{\alpha}$ from an admissible pair of vectors for (P_ϵ) .

THEOREM 6.5. *Given an admissible pair of vectors $(\bar{\alpha}_1, \dots, \bar{\alpha}_m)$ and $(\bar{\epsilon}_1, \dots, \bar{\epsilon}_m)$, we have that*

$$\begin{aligned} f''_\epsilon(x_0; d, d) &\geq \sum_{i \in \mathcal{S}(\bar{\alpha}) \cap I(x_0; d)} \bar{\alpha}_i (f_i)''_{\bar{\epsilon}_i/\bar{\alpha}_i}(x_0; d, d) \\ &+ \sum_{i \in \mathcal{S}(\bar{\alpha}) \cap I(x_0; d)} \bar{\alpha}_i \bar{\mu}^{f_i} \left(\frac{\bar{\epsilon}_i}{\bar{\alpha}_i} \right) [(f_i)'_{\bar{\epsilon}_i/\bar{\alpha}_i}(x_0; d) - f'(x_0; d)]. \end{aligned} \quad (6.23)$$

Equality holds if for example $\epsilon \leq \epsilon_f^*$ and the f_i are strictly convex on $x_0 + \mathbb{R}_+ d$.

Proof. The general formulation of $f''_\epsilon(x_0; d, d)$ is as follows:

$$f''_\epsilon(x_0; d, d) = \sum_{i \in \mathcal{S}(\bar{\alpha})} \bar{\alpha}_i \bar{\mu}^f(\epsilon) [(f_i)'_{\bar{\epsilon}_i/\bar{\alpha}_i}(x_0; d) - f'(x_0; d)]. \quad (6.24)$$

From the inclusion (6.16) we infer that

$$\bar{\mu}^f(\epsilon) \geq \max_{i \in \mathcal{S}(\bar{\alpha})} \bar{\mu}^{f_i} \left(\frac{\bar{\epsilon}_i}{\bar{\alpha}_i} \right). \quad (6.25)$$

Hence the inequality (6.22) is derived.

The case where equality holds follows from Corollary 6.4. \square

In an alternate formulation involving all the $(f_i)''_{\bar{\alpha}_i/\bar{\alpha}_i}(x_0; d, d)$ for $i \in \mathcal{S}(\bar{\alpha})$, the inequality (6.22) can be rewritten as follows:

$$(6.26) \quad \begin{aligned} f''_\varepsilon(x_0; d, d) \geq & \sum_{i \in \mathcal{S}(\bar{\alpha})} \bar{\alpha}_i (f_i)''_{\bar{\alpha}_i/\bar{\alpha}_i}(x_0; d, d) \\ & + \sum_{i \in \mathcal{S}(\bar{\alpha}) \setminus I(x_0; d)} \bar{\alpha}_i \bar{\mu}^{f_i} \left(\frac{\bar{\varepsilon}_i}{\bar{\alpha}_i} \right) [f'_i(x_0; d) - f'(x_0; d)]. \end{aligned}$$

Example. Let f be defined as the maximum of m quadratic functions

$$f_i: x \rightarrow f_i(x) = \frac{1}{2} \langle A_i x, x \rangle + \langle b_i, x \rangle + c_i, \quad i = 1, \dots, m,$$

where the A_i are supposed symmetric positive definite.

Given x_0, d and ε , the calculation of $f'_\varepsilon(x_0; d)$ requires us to solve the following maximization problem:

$$\text{Maximize } \left\{ \sum_{i=1}^m \alpha_i \langle A_i x_0 + b_i, d \rangle + \sum_{i=1}^m (2\varepsilon_i \alpha_i \langle A_i d, d \rangle)^{1/2} \right\}$$

$$\text{subject to: } \alpha_i \geq 0 \quad \text{for all } i, \quad \sum_{i=1}^m \alpha_i = 1,$$

$$\varepsilon_i \geq 0 \quad \text{for all } i, \quad \sum_{i=1}^m \varepsilon_i + \sum_{i=1}^m \alpha_i (f(x_0) - f_i(x_0)) = \varepsilon.$$

Due to the strict concavity of the objective function on the constraint set, the problem above has only one solution (one admissible pair of vectors) formed by $(\bar{\alpha}_1, \dots, \bar{\alpha}_m)$ and $(\bar{\varepsilon}_1, \dots, \bar{\varepsilon}_m)$. The set $\{1, \dots, m\}$ of indices can be divided into two: the set $\mathcal{S}(\bar{\alpha})$ of i for which both $\bar{\alpha}_i$ and $\bar{\varepsilon}_i$ are >0 , and the set of dummy indices i for which both $\bar{\alpha}_i$ and $\bar{\varepsilon}_i$ are null. We then have that

$$(6.27) \quad \begin{aligned} f''_\varepsilon(x_0; d, d) = & \sum_{i \in \mathcal{S}(\bar{\alpha})} \bar{\alpha}_i \langle A_i d, d \rangle \\ & + \sum_{i \in \mathcal{S}(\bar{\alpha}) \setminus I(x_0; d)} \bar{\alpha}_i \left(\frac{\bar{\alpha}_i \langle A_i d, d \rangle}{2\bar{\varepsilon}_i} \right)^{1/2} [\langle A_i x_0 + b_i, d \rangle - f'(x_0; d)] \end{aligned}$$

where $f'(x_0; d) = \max \{ \langle A_i x_0 + b_i, d \rangle \mid f_i(x_0) = f(x_0) \}$.

The next result concerns the limiting case $\varepsilon = 0$ and generalizes what is known when all the f_i are C^2 [2, Chap. III].

PROPOSITION 6.6. *Suppose that all the $f_i, i \in I(x_0; d)$, have a second-order derivative at x_0 in the direction d . Then f has a second-order derivative in the same direction with*

$$(6.28) \quad f''(x_0; d, d) = \max_{i \in I(x_0; d)} f''_i(x_0; d, d).$$

Proof. Since $f_i(x_0) = f(x_0)$ and $f'_i(x_0; d) = f'(x_0; d)$ whenever $i \in I(x_0; d)$, we have that

$$(6.29) \quad \liminf_{\alpha \rightarrow 0^+} \left\{ 2 \left[\frac{f(x_0 + \alpha d) - f(x_0)}{\alpha} - f'(x_0; d) \right] \alpha^{-1} \right\} \geq \max_{i \in I(x_0; d)} f''_i(x_0; d, d).$$

Due to the continuity of the f_i , there is $\alpha_0 > 0$ such that $I(x_0 + \alpha d) \subset I(x_0)$ for all $\alpha \in [0, \alpha_0]$. Now, due to the convexity of the f_i , the functions $\alpha \rightarrow f'_i(x_0 + \alpha d; d)$ and $\alpha \rightarrow f''_i(x_0 + \alpha d; d)$ are increasing and upper-semicontinuous. Consequently, one easily verifies there exists $\alpha_1 \in]0, \alpha_0]$ such that $I(x_0 + \alpha d; d) \subset I(x_0; d)$ for all $\alpha \in [0, \alpha_1]$.

Hence

$$(6.30) \quad \limsup_{\alpha \rightarrow 0^+} \{[f'(x_0 + \alpha d; d) - f'(x_0; d)]\alpha^{-1}\} \leq \max_{i \in I(x_0; d)} f_i''(x_0; d, d).$$

Combining the inequalities (6.29) and (6.30) yields the desired result. \square

REFERENCES

- [1] A. AUSLENDER, *Differential properties of the support function of the ε -subdifferential of a convex function*, Math. Programming, 24 (1982), pp. 257–268.
- [2] V. F. DEM'YANOV AND V. N. MALOZEMOV, *Introduction to Minimax*, John Wiley, New York, 1974.
- [3] J.-B. HIRIART-URRUTY, *ε -subdifferential calculus*, in Convex Analysis and Optimization, Research Notes in Mathematics Series 57, Pitman, London, 1982, pp. 43–92.
- [4] ———, *Approximating a second-order directional derivative for nonsmooth convex functions*, this Journal, 26 (1982), pp. 783–807.
- [5] ———, *Limiting behaviour of the approximate first-order and second-order directional derivatives for a convex function*, Nonlinear Analysis: Theory, Method & Applications, 1982, pp. 1309–1326.
- [6] ———, *The approximate first-order and second-order directional derivatives for a convex function*, in Proc. Conference on Mathematical Theories of Optimization, S. Margherita Ligure, 30 November–4 December 1981, to appear.
- [7] ———, *The approximate first-order and second-order directional derivatives of a marginal function in convex optimization*, J. Optim. Theory Appl. (special issue, A. Fiacco ed.), to appear.
- [8] S. S. KUTATELADZE, *Convex ε -programming*, Soviet, Math. Dokl., 20 (1979), pp. 391–393.
- [9] ———, *ε -subdifferentials and ε -optimality*, Siberian Math. J., 21 (1981), pp. 404–411.
- [10] P.-J. LAURENT, *Approximation et optimisation*, Hermann, Paris 1972.
- [11] C. LEMARECHAL AND E. A. NURMINSKII, *Sur la différentiabilité de la fonction d'appui du sous-différentiel approché*, Note aux Comptes Rendus Acad. Sci. Paris, 290 (1980), pp. 855–858.
- [12] R.-T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.

CONTROLLABILITY PROPERTIES OF LINEAR AND SEMILINEAR ABSTRACT CONTROL SYSTEMS*

HONG XING ZHOU†

Abstract. We consider control systems described by a semilinear abstract equation $y' + Ay = F(y, u) + Bu$ in Hilbert space. General conditions for exact reachability and approximate controllability are given which are related with two families of associated quadratic optimal control problems. The minimum norm optimal control and the construction of the reachable set of the corresponding linear system $y' + Ay = Bu$ are characterized respectively. Exact reachability for a class of semilinear control systems is obtained under some assumptions on the range of a nonlinear operator $\mathcal{F}_{(t_0, T)}$ (see definition in § 4) generated by the nonlinear function F .

Key words. linear and semilinear control systems, exact reachability

1. Introduction. We consider in this paper the semilinear abstract control system

$$(1.1) \quad \frac{dy(t)}{dt} + Ay(t) = F(y(t), u(t)) + Bu(t), \quad 0 < t < T,$$

where the state $y(t)$, $0 \leq t \leq T$, takes values in a Hilbert space X and satisfies the initial value condition

$$(1.2) \quad y(0) = \eta \in X,$$

and the control $u(t)$, $0 \leq t \leq T$, is in another Hilbert space U . Assume the operator $-A$ generates a strongly continuous semigroup $S(t)$ on X for $t \geq 0$. For a given $T > 0$ denote

$$(1.3) \quad M_T = \sup_{0 \leq t \leq T} \|S(t)\|_{\mathcal{L}[X \rightarrow X]}.$$

In (1.1) B is a bounded linear operator from U into X satisfying the following assumption:

$$(1.4) \quad \|u\|_U \leq K_b \|Bu\| \quad \text{for each } u \in U,$$

where K_b is some positive constant independent on $u \in U$ ($\|\cdot\|$ denotes the norm in X , i.e. $\|\cdot\| = \|\cdot\|_X$). Assumption (1.4) is very general (see § 4 in detail).

In [8], [9] some special cases with $F = F(y)$ of approximate controllability of a semilinear control system described by (1.1) are considered. If $X = L^2(0, 1)$, $U = R^1$ and (1.1) is a one-dimensional heat equation with $F = F(y)$ and $Bu = b(x)u(t)$, then it is proved [8] that the semilinear control system (1.1) is approximately controllable under a uniform boundedness assumption on $F(y)$. In [9] the author gives a sufficient condition of approximate controllability for (1.1) which can be applied to both infinite dimensional semilinear control systems or finite dimensional semilinear control systems described by an equation of the form (1.1). In general, approximate controllability for semilinear systems is related with both the linear operator B and the nonlinear function F and also with the relationship between B and F .

We study here exact reachability of semilinear abstract control systems described by (1.1). Some necessary preliminaries on abstract differential equations will be given later in this section. In § 2, some general conclusions on approximate controllability

* Received by the editors December 10, 1982, and in revised form April 19, 1983.

† Department of Mathematics, Shandong University, Jinan, Shandong Province, The People's Republic of China.

and exact reachability are introduced which are connected with two families of associated quadratic optimal control problems. In § 3, an abstract linear control system

$$(1.5) \quad \frac{dy(t)}{dt} + Ay(t) = Bu, \quad t > 0,$$

is independently considered which is called the linear system corresponding to the semilinear control system (1.1). The minimum norm optimal control and the construction of the reachable set $K_{(0,T)}$ of the linear system are characterized respectively. We discuss in § 4 exact reachability for a class of semilinear control systems by using an inclusion relation between the reachable set $K_{(0,T)}$ of the linear control system corresponding to (1.1) and the reachable set $K_{(0,T)}(F)$ of (1.1) under some assumption on the ranges of the linear operator B and the nonlinear function F .

Before discussing reachability problem and controllability problem for the semilinear control system (1.1), we need some preliminary knowledge on existence, uniqueness and continuous dependence of the solution of the abstract semilinear differential equation (1.1).

Assume that the nonlinear function F in (1.1) is defined on $X \times U$ and is uniformly Lipschitz on y and u :

$$(1.6) \quad \|F(y_1, u_1) - F(y_2, u_2)\| \leq K_1 \|y_1 - y_2\| + K_2 \|u_1 - u_2\|_U$$

for y_1, y_2 in X and u_1, u_2 in U ,

where K_1 and K_2 are positive constants independent on y_1, y_2 and u_1, u_2 .

As in [6], [9], under adequate assumptions on A, B and F it is well known that:

LEMMA 1.1. *Let t_0 be any given positive number in $[0, T)$. Then for each $u(\cdot)$ in $L^2(t_0, T; U)$ and η_0 in X the semilinear abstract differential equation (1.1) has a unique (mild) solution $y(t) = y(t; u) = y(t; u; t_0, \eta_0)$, $t_0 \leq t \leq T$ with initial value η_0 at time t_0 satisfying*

$$(1.7) \quad y(t; u) = S(t - t_0)\eta_0 + \int_{t_0}^t S(t - s)[F(y(s; u), u(s)) + Bu(s)] ds, \quad t_0 \leq t \leq T.$$

The unique solution $y(\cdot)$ in $C([t_0, T]; X)$ satisfies the following estimates:

$$(1.8) \quad \|y(t)\| \leq M_1[\|\eta_0\| + (t - t_0)\|F(0, 0)\| + \sqrt{t - t_0}(\|B\| + K_2)\|u(\cdot)\|_{L^2(t_0, T; U)}], \quad t_0 \leq t \leq T,$$

$$(1.9) \quad \|y(\cdot)\|_{L^2(t_0, T; X)} \leq \sqrt{2}M_1\|\eta_0\|(T - t_0)^{1/2} + \sqrt{\frac{2}{3}}M_1\|F(0, 0)\|(T - t_0)^{3/2} + M_1(T - t_0)(\|B\| + K_2)\|u(\cdot)\|_{L^2(t_0, T; U)},$$

where $M_1 = M_T e^{M_T K_1 T}$. Let $u_1(\cdot)$ and $u_2(\cdot)$ be in $L^2(t_0, T; U)$. Then

$$(1.10) \quad \|y(t; u_1) - y(t; u_2)\| \leq M_1 \sqrt{t - t_0}(\|B\| + K_2)\|u_1(\cdot) - u_2(\cdot)\|_{L^2(t_0, T; U)}, \quad t_0 \leq t \leq T$$

and

$$(1.11) \quad \|y(\cdot; u_1) - y(\cdot; u_2)\|_{L^2(t_0, T; X)} \leq M_1(T - t_0)(\|B\| + K_2)\|u_1(\cdot) - u_2(\cdot)\|_{L^2(t_0, T; U)}.$$

LEMMA 1.2. *The solution mappings $y(\cdot; \cdot)$ and $y(T; \cdot)$ are continuous from $L^2(t_0, T; U)$ into $C([t_0, T]; X)$ and X respectively.*

In order to establish weak compactness of the solution mapping $y(T; u)$ in control u a supplementary assumption on weak continuity of the function F is needed in the future.

Supplementary Assumption (F). The nonlinear function F in (1.1) is weakly continuous from $L^2(t_0, T; X) \times L^2(t_0, T; U)$ into $L^2(t_0, T; X)$. That is, if $y_n(\cdot) \in L^2(t_0, T; X)$ and $u_n(\cdot) \in L^2(t_0, T; U)$, $n = 1, 2, \dots$, and

$$w\text{-}\lim_{n \rightarrow \infty} y_n(\cdot) = \bar{y}(\cdot) \quad \text{in } L^2(t_0, T; X),$$

$$w\text{-}\lim_{n \rightarrow \infty} u_n(\cdot) = \bar{u}(\cdot) \quad \text{in } L^2(t_0, T; U),$$

where $\bar{y}(\cdot)$ and $\bar{u}(\cdot)$ are some elements in $L^2(t_0, T; X)$ and in $L^2(t_0, T; U)$ respectively, then there exist some subsequences $\{y_m(\cdot); m = 1, 2, \dots\} \subset \{y_n(\cdot); n = 1, 2, \dots\}$ and $\{u_m(\cdot); m = 1, 2, \dots\} \subset \{u_n(\cdot); n = 1, 2, \dots\}$ such that

$$w\text{-}\lim_{m \rightarrow \infty} F(y_m(\cdot), u_m(\cdot)) = F(\bar{y}(\cdot), \bar{u}(\cdot)) \quad \text{in } L^2(t_0, T; X).$$

LEMMA 1.3. The solution mappings $y(t; \cdot)$ for each $t \in (t_0, T]$ and $y(\cdot; \cdot)$ are weakly compact from $L^2(t_0, T; U)$ into X and $L^2(t_0, T; X)$ respectively, under Supplementary Assumption (F).

Remark 1.4. If we assume that the operator $-A$ in (1.1) is the infinitesimal generator of a compact semigroup $S(\cdot)$, then the mapping $y(t; \cdot)$, $t_0 < t \leq T$, is compact from $L^2(t_0, T; U)$ into X and the mapping $y(\cdot; \cdot)$ is compact from $L^2(t_0, T; U)$ into $C([t_0, T]; X)$, without Supplementary Assumption (F).

2. General conclusions. In this section, we are going to give some general conclusions on both approximate controllability and exact reachability. Firstly some definitions are introduced.

DEFINITION. Assume $\eta_0 = 0$ in Lemma 1.1. The nonempty subset $K_{(t_0, T)}(F)$ in X consisting of all terminal states of (1.1) at time T is called the *reachable set* at T of the semilinear system (1.1) starting at t_0 :

$$(2.1) \quad K_{(t_0, T)}(F) = \{y(T) = y(T; u; t_0, 0) \text{ for some } u(\cdot) \in L^2(t_0, T; U)\}.$$

DEFINITION. The control system (1.1) is called *approximately controllable* on $[t_0, T]$ if

$$(2.2) \quad \overline{K_{(t_0, T)}(F)} = X.$$

DEFINITION. For each $h \in X$ define

$$(2.3) \quad V_{(t_0, T)}[h] = \{u(\cdot) | u(\cdot) \in L^2(t_0, T; U) \text{ with } y(T; u; t_0, 0) = h\}.$$

If $V_{(t_0, T)}[h] \neq \emptyset$ (empty set in $L^2(t_0, T; U)$), then the semilinear control system (1.1) is called *h-exactly reachable* from the origin on $[t_0, T]$.

For the sake of brevity we usually call $K_{(t_0, T)}(F)$ the reachable set and omit $[t_0, T]$ when mentioning approximate controllability or *h-exact reachability*.

While discussing approximate controllability and exact reachability for the semilinear abstract control system (1.1), we consider two families of associated quadratic optimal control problems

$$(2.4) \quad (\text{Inf}) \quad J_\varepsilon(u; h) = \|y(T; u) - h\|^2 + \varepsilon \|u(\cdot)\|_{L^2(t_0, T; U)}^2 \quad \text{for } \varepsilon > 0$$

and

$$(2.5) \quad (\text{Inf}) \quad I_\varepsilon(u; h) = \frac{1}{\varepsilon} \|y(T; u) - h\|^2 + \|u(\cdot)\|_{L^2(t_0, T; U)}^2 \quad \text{for } \varepsilon > 0,$$

where $y(T; u) = y(T; u; t_0, 0)$ is the terminal state of the nonlinear system (1.1) at time T . It is easy to verify by Lemma 1.3 that for any given $h \in X$ and $\varepsilon > 0$ there exists some control $u_\varepsilon(\cdot) \in L^2(t_0, T; U)$ such that

$$(2.6) \quad J_\varepsilon(u_\varepsilon; h) = \inf_{u(\cdot) \in L^2(t_0, T; U)} J_\varepsilon(u; h)$$

and

$$(2.7) \quad I_\varepsilon(u_\varepsilon; h) = \inf_{u(\cdot) \in L^2(t_0, T; U)} I_\varepsilon(u; h).$$

The control $u_\varepsilon(\cdot)$ is called *minimization element* of the nonlinear functions $J_\varepsilon(u; h)$ and $I_\varepsilon(u; h)$.

The following two theorems establish general relations between controllability properties and the families (2.4) and (2.5) of associated quadratic optimal control problems.

THEOREM 2.1. (1). *Assume $h \in X$. Then h is in $\overline{K_{(t_0, T)}(F)}$ if and only if*

$$(2.8) \quad \lim_{\varepsilon \rightarrow 0} J_\varepsilon(u_\varepsilon; h) = 0.$$

(2). *The semilinear abstract control system (1.1) is approximately controllable if and only if (2.8) holds for every $h \in X$.*

Proof. (1). Let h be an arbitrary element in $\overline{K_{(t_0, T)}(F)}$. Then for any given integer $N > 0$ there exists some control $v_N(\cdot) \in L^2(t_0, T; U)$ such that

$$\|y(T; v_N) - h\| < \frac{1}{N}, \quad N = 1, 2, \dots$$

Thus

$$\lim_{\varepsilon \rightarrow 0} J_\varepsilon(u_\varepsilon; h) \leq \lim_{\varepsilon \rightarrow 0} J_\varepsilon(v_N; h) \leq \lim_{\varepsilon \rightarrow 0} \left(\frac{1}{N^2} + \varepsilon \|v_N(\cdot)\|_{L^2(t_0, T; U)}^2 \right) = \frac{1}{N^2}.$$

Taking $N \rightarrow \infty$ in above we obtain (2.8). Conversely, if (2.8) holds for some $h \in X$, then

$$\lim_{\varepsilon \rightarrow 0} \|y(T; u_\varepsilon) - h\|^2 \leq \lim_{\varepsilon \rightarrow 0} J_\varepsilon(u_\varepsilon; h) = 0,$$

and, equivalently, $h \in \overline{K_{(t_0, T)}(F)}$. The statement (2) follows directly from (1). \square

THEOREM 2.2. (1). *The abstract control system (1.1) is h -exactly reachable if and only if*

$$(2.9) \quad I_\varepsilon(u_\varepsilon; h) \text{ is uniformly bounded for } 0 < \varepsilon < \infty,$$

or equivalently

$$(2.10) \quad h \in \overline{K_{(t_0, T)}(F)} \text{ and } \|u_\varepsilon(\cdot)\|_{L^2(t_0, T; U)} \text{ is uniformly bounded for } 0 < \varepsilon < \infty.$$

(2). *If (1.1) is h -exactly reachable then there exists some control $u^*(\cdot) \in V_{(t_0, T)}[h]$ with minimum norm, i.e.*

$$(2.11) \quad \|u^*(\cdot)\|_{L^2(t_0, T; U)} = \min_{v \in V_{(t_0, T)}[h]} \|v(\cdot)\|_{L^2(t_0, T; U)}$$

and

$$(2.12) \quad \lim_{\varepsilon \rightarrow 0} I_\varepsilon(u_\varepsilon; h) = \overline{\lim_{\varepsilon \rightarrow 0}} \|u_\varepsilon(\cdot)\|_{L^2(t_0, T; U)}^2 = \|u^*(\cdot)\|_{L^2(t_0, T; U)}^2.$$

Proof. (1). Suppose the abstract control system (1.1) is h -exactly reachable and $v(\cdot)$ is an arbitrary control in $V_{(t_0, T)}[h]$. Then for any $\varepsilon > 0$

$$(2.13) \quad I_\varepsilon(u_\varepsilon; h) \leq I_\varepsilon(v; h) = \|v(\cdot)\|_{L^2(t_0, T; U)}^2.$$

On the other hand, if (2.9) holds for some $h \in X$, then

$$\lim_{\varepsilon \rightarrow 0} J_\varepsilon(u_\varepsilon; h) = \lim_{\varepsilon \rightarrow 0} \varepsilon I_\varepsilon(u_\varepsilon; h) = 0.$$

Moreover, there exists some constant $M(h)$ independent of $\varepsilon > 0$ such that

$$\|u_\varepsilon(\cdot)\|_{L^2(t_0, T; U)}^2 \leq I_\varepsilon(u_\varepsilon; h) \leq M(h).$$

Thus there exists some monotone sequence $\{\varepsilon_n; n = 1, 2, \dots\}$ with $\varepsilon_n \rightarrow 0$ as $n \rightarrow \infty$ such that

$$(2.14) \quad w\text{-}\lim_{n \rightarrow \infty} u_{\varepsilon_n}(\cdot) = u^*(\cdot) \quad \text{in } L^2(t_0, T; U),$$

where $u^*(\cdot)$ is some element in $L^2(t_0, T; U)$. By Lemma 1.3 and the lower semicontinuity of the norm, we have

$$\|y(T; u^*) - h\|^2 \leq \liminf_{n \rightarrow \infty} \|y(T; u_{\varepsilon_n}) - h\|^2 \leq \lim_{n \rightarrow \infty} J_{\varepsilon_n}(u_{\varepsilon_n}; h) = 0.$$

Thus, $u^*(\cdot) \in V_{(t_0, T)}[h]$ and the system (1.1) is h -exactly reachable.

The equivalent relation with (2.10) is obvious from the above.

(2). Let the system (1.1) be h -exactly reachable and $u^*(\cdot) \in V_{(t_0, T)}[h]$ be obtained by (2.14). Then

$$(2.15) \quad \begin{aligned} \|u^*(\cdot)\|_{L^2(t_0, T; U)}^2 &\leq \liminf_{n \rightarrow \infty} \|u_{\varepsilon_n}(\cdot)\|_{L^2(t_0, T; U)}^2 \\ &\leq \overline{\lim}_{\varepsilon \rightarrow 0} \|u_\varepsilon(\cdot)\|_{L^2(t_0, T; U)}^2 \leq \overline{\lim}_{\varepsilon \rightarrow 0} I_\varepsilon(u_\varepsilon; h) \leq \lim_{\varepsilon \rightarrow 0} I_\varepsilon(u^*; h) \\ &= \|u^*(\cdot)\|_{L^2(t_0, T; U)}^2. \end{aligned}$$

It implies

$$\lim_{n \rightarrow \infty} \|u_{\varepsilon_n}(\cdot)\|_{L^2(t_0, T; U)} = \|u^*(\cdot)\|_{L^2(t_0, T; U)}.$$

Since $L^2(t_0, T; U)$ is a Hilbert space we have

$$(2.16) \quad \lim_{n \rightarrow \infty} u_{\varepsilon_n}(\cdot) = u^*(\cdot) \quad \text{in } L^2(t_0, T; U).$$

By using (2.13) we have

$$(2.17) \quad \begin{aligned} \|u^*(\cdot)\|_{L^2(t_0, T; U)}^2 &= \lim_{n \rightarrow \infty} \|u_{\varepsilon_n}(\cdot)\|_{L^2(t_0, T; U)}^2 \\ &\leq \lim_{n \rightarrow \infty} I_{\varepsilon_n}(u_{\varepsilon_n}; h) \leq \lim_{n \rightarrow \infty} I_{\varepsilon_n}(v; h) = \|v(\cdot)\|_{L^2(t_0, T; U)}^2, \end{aligned}$$

where $v(\cdot)$ is an arbitrary control in $V_{(t_0, T)}[h]$. Hence (2.11) holds.

Since $u^*(\cdot) \in V_{(t_0, T)}[h]$ we have from (2.17)

$$(2.18) \quad \lim_{n \rightarrow \infty} I_{\varepsilon_n}(u_{\varepsilon_n}; h) = \|u^*(\cdot)\|_{L^2(t_0, T; U)}^2.$$

It is easy to see that $I_\varepsilon(u_\varepsilon; h)$ is monotonically decreasing in $\varepsilon > 0$. In fact, for $0 < \varepsilon_1 < \varepsilon_2 < \infty$, one has

$$\begin{aligned} I_{\varepsilon_2}(u_{\varepsilon_2}; h) &\leq I_{\varepsilon_2}(u_{\varepsilon_1}; h) = \frac{1}{\varepsilon_2} \|y(T; u_{\varepsilon_1}) - h\|^2 + \|u_{\varepsilon_1}(\cdot)\|_{L^2(t_0, T; U)}^2 \\ &\leq I_{\varepsilon_1}(u_{\varepsilon_1}; h). \end{aligned}$$

Thus (2.12) follows from (2.15) and monotonicity of $I_\varepsilon(u_\varepsilon; h)$ in ε immediately. \square

3. Linear systems. We now are going to discuss construction of the reachable set $K_{(0, T)}$ of the linear control system (1.5) by using Theorems 2.1 and 2.2. Before the discussion, some notations are introduced.

Define a bounded linear operator $(\mathcal{S}_T B)$ mapping $L^2(0, T; U)$ into X as

$$(3.1) \quad (\mathcal{S}_T B)u = \int_0^T S(T-t)Bu(t) dt.$$

Consequently, the dual operator $(\mathcal{S}_T B)^*$ is bounded from X into $L^2(0, T; U)$. The range of the dual operator $(\mathcal{S}_T B)^*$ is a subspace in $L^2(0, T; U)$ which is denoted by

$$(3.2) \quad \mathcal{U} = \{v(\cdot) | v(\cdot) = (\mathcal{S}_T B)^*(\cdot)\varphi \text{ with some } \varphi \in X\}.$$

The closure $\bar{\mathcal{U}}$ of \mathcal{U} in $L^2(0, T; U)$, in general, is a proper subspace, i.e. $\bar{\mathcal{U}} \neq L^2(0, T; U)$ [1].

Since the system (1.5) is linear it is known that, for each $h \in K_{(0, T)}$, there exists a unique optimal control $u^*(\cdot)$ with minimum norm in $V_{(0, T)}[h]$ which is called *minimum norm optimal control* corresponding to h . If a system (1.5) is finite dimensional: $0 < \dim U \leq \dim X < +\infty$, then the minimum norm optimal control $u^*(\cdot)$ has an explicit form [5]

$$(3.3) \quad u^*(\cdot) = (\mathcal{S}_T B)^*(\cdot)\varphi,$$

$$(3.4) \quad \varphi = [(\mathcal{S}_T B)(\mathcal{S}_T B)^*]^{-1}h.$$

If considering a time-optimal control problem, then the time-optimal control has the same expression (3.3) with some φ [2]. Differing from the finite dimensional case and time-optimal control problem, the minimum norm optimal control $u^*(\cdot)$ corresponding to $h \in K_{(0, T)}$ no longer has the same formula (3.3). In this case we claim the next theorem.

THEOREM 3.1. Assume $h \in K_{(0, T)}$. Then the corresponding minimum norm optimal control

$$(3.5) \quad u^*(\cdot) \in \bar{\mathcal{U}}.$$

Moreover, the reachable set $K_{(0, T)}$ may be characterized by

$$(3.6) \quad K_{(0, T)} = (\mathcal{S}_T B)\bar{\mathcal{U}}.$$

Proof. Suppose h is an arbitrary element in $K_{(0, T)}$, and $u^*(\cdot)$ is the unique minimum norm optimal control corresponding to h . By Theorem 2.2, there exists a sequence of positive numbers $\{\varepsilon_n; n = 1, 2, \dots\}$ which goes to 0 as $n \rightarrow \infty$ such that

$$(3.7) \quad \lim_{n \rightarrow \infty} u_{\varepsilon_n}(\cdot) = u^*(\cdot) \text{ in } L^2(0, T; U).$$

In the linear case the minimization element $u_\varepsilon(\cdot)$ for each $\varepsilon > 0$ has the following explicit formula

$$(3.8) \quad u_\varepsilon(\cdot) = (\varepsilon + \hat{G})^{-1}(\mathcal{S}_T B)^*h,$$

where \hat{G} is a bounded linear operator mapping $L^2(0, T; U)$ into $L^2(0, T; U)$, defined as

$$(3.9) \quad \hat{G} = (\mathcal{S}_T B)^* (\mathcal{S}_T B).$$

Obviously, the range of the operator \hat{G}

$$\mathcal{R}(\hat{G}) = (\mathcal{S}_T B)^* K_{(0,T)} \subset \mathcal{U}.$$

Hence the minimization element $u_\varepsilon(\cdot)$ for any $\varepsilon > 0$ may be rewritten as

$$(3.10) \quad \begin{aligned} u_\varepsilon &= (\varepsilon + \hat{G})^{-1} (\mathcal{S}_T B)^* h = (\varepsilon + \hat{G})^{-1} (\mathcal{S}_T B)^* (\mathcal{S}_T B) u^* \\ &= (\varepsilon + \hat{G})^{-1} \hat{G} u^* = \hat{G} (\varepsilon + \hat{G})^{-1} u^*, \end{aligned}$$

i.e. $u_\varepsilon \in \mathcal{U}$ for any $\varepsilon > 0$. Therefore, (3.5) follows from (3.7) and (3.10).

The characterization (3.6) of the reachable set $K_{(0,T)}$ is a direct result of (3.5). \square

Theorem 3.1 shows that even though the formula (3.3) for some minimum norm optimal control $u^*(\cdot)$ does not hold but may be approximated by a control sequence $\{u_{\varepsilon_n}(\cdot); n = 1, 2, \dots\}$, each control in it still has the formula (3.3). The characterization (3.6) also makes it known that as a definition the reachable set $K_{(0,T)}$ is $(\mathcal{S}_T B)L^2(0, T; U)$, but we may use a subspace \mathcal{U} (usually a proper subspace) in $L^2(0, T; U)$ instead of the whole space $L^2(0, T; U)$.

In the rest a subset of the reachable set $K_{(0,T)}$ is discussed; for each element belonging to it the corresponding minimum norm optimal control has the formula (3.3).

Suppose φ is an arbitrarily given element in X . Then $(\mathcal{S}_T B)^* \varphi \in \mathcal{U}$ and $(\mathcal{S}_T B)(\mathcal{S}_T B)^* \varphi \in X$. Define

$$(3.11) \quad G = (\mathcal{S}_T B)(\mathcal{S}_T B)^*.$$

Then it is a bounded linear operator from X into X and the range $\mathcal{R}(G) = GX = (\mathcal{S}_T B)\mathcal{U}$ of the operator G is a subset of the reachable set $K_{(0,T)}$. It is known that an abstract linear control system (1.5) is approximately controllable if and only if $\mathcal{N}(G) = \{0\}$, where $\mathcal{N}(G)$ is the null space of the operator G . In this case G^{-1} is an (unbounded) one-to-one operator from $\mathcal{R}(G)$ into X . If a system does not have approximate controllability, then $\mathcal{N}(G) \neq 0$. If it happens then $h = G(\varphi + z)$, for any $z \in \mathcal{N}(G)$, as $h = G\varphi$ and $G^{-1}h$ is not uniquely defined for any $h \in \mathcal{R}(G)$. But the element $\arg \min_{G\varphi=h} \|\varphi\|$ is uniquely defined for any $h \in \mathcal{R}(G)$ since the set $\{\varphi | G\varphi = h\}$ is a closed linear manifold in X . Thus we may define a one-to-one mapping from $\mathcal{R}(G)$ into X :

$$(3.12) \quad G^{-1}h = \arg \min_{G\varphi=h} \|\varphi\|.$$

THEOREM 3.2. *Let $h \in \mathcal{R}(G)$, $u^*(\cdot)$ be the corresponding minimum norm optimal control and $u_\varepsilon(\cdot)$ ($\varepsilon > 0$) be the corresponding minimization element. Then*

$$(3.13) \quad u^*(\cdot) \in \mathcal{U}$$

and

$$(3.14) \quad \lim_{\varepsilon \rightarrow 0} u_\varepsilon(\cdot) = u^*(\cdot) \quad \text{in } L^2(0, T; U).$$

Proof. Assume that $h \in \mathcal{R}(G)$, $\varphi = G^{-1}h$ is an element in X defined by (3.12). Thus, $h = G\varphi = (\mathcal{S}_T B)(\mathcal{S}_T B)^* \varphi$ and $(\mathcal{S}_T B)^* \varphi \in V_{(0,T)}[h]$. We claim

$$(3.15) \quad u^*(\cdot) = (\mathcal{S}_T B)^*(\cdot)\varphi.$$

Suppose $u(\cdot)$ is an arbitrary control in $V_{(0,T)}[h]$ and its projection on the closed subspace $\bar{\mathcal{U}}$ is denoted by $u_p(\cdot)$. For each ψ in X one has

$$\begin{aligned} (u_p(\cdot), (\mathcal{S}_T B)^*(\cdot)\psi)_{L^2(0,T;U)} &= (u(\cdot), (\mathcal{S}_T B)^*(\cdot)\psi)_{L^2(0,T;U)} \\ &= (h, \psi)_X = (G\varphi, \psi)_X \\ &= ((\mathcal{S}_T B)^*(\cdot)\varphi, (\mathcal{S}_T B)^*(\cdot)\psi)_{L^2(0,T;U)}, \end{aligned}$$

or

$$(u_p(\cdot) - (\mathcal{S}_T B)^*(\cdot)\varphi, (\mathcal{S}_T B)^*(\cdot)\psi)_{L^2(0,T;U)} = 0.$$

Since \mathcal{U} is dense in $\bar{\mathcal{U}}$ and $[u_p(\cdot) - (\mathcal{S}_T B)^*(\cdot)\varphi] \in \mathcal{U}$, we have

$$u_p(\cdot) = (\mathcal{S}_T B)^*(\cdot)\varphi \quad \text{for each } u(\cdot) \in V_{(0,T)}[h].$$

Therefore

$$\|(\mathcal{S}_T B)^*(\cdot)\varphi\|_{L^2(0,T;U)} = \|u_p(\cdot)\|_{L^2(0,T;U)} \leq \|u(\cdot)\|_{L^2(0,T;U)} \quad \text{for each } u(\cdot) \in V_{(0,T)}[h].$$

By uniqueness of the minimum norm optimal control of the linear system (1.5), (3.15) holds and $u^*(\cdot) \in \mathcal{U}$.

Since $u_\varepsilon = u^* - \varepsilon(\varepsilon + \hat{G})^{-1}u^*$ (see (3.10)), (3.14) is equivalent to

$$(3.16) \quad \lim_{\varepsilon \rightarrow 0} \varepsilon(\varepsilon + \hat{G})^{-1}u^* = 0 \quad \text{in } L_2(0, T; U).$$

If we consider another family with parameter $\varepsilon > 0$ of associated quadratic optimal control problems,

$$J_\varepsilon(v; \varphi) = \|(\mathcal{S}_T B)v - \varphi\|^2 + \varepsilon \|v(\cdot)\|_{L^2(0,T;U)}^2,$$

then $J_\varepsilon(v; \varphi)$ takes its minimum at $v = v_\varepsilon$ defined by

$$v_\varepsilon = (\varepsilon + \hat{G})^{-1}(\mathcal{S}_T B)^*\varphi = (\varepsilon + \hat{G})^{-1}u^*.$$

If we can show

$$(3.17) \quad \lim_{\varepsilon \rightarrow 0} \varepsilon \|v_\varepsilon(\cdot)\|_{L^2(0,T;U)}^2 = 0,$$

then

$$\begin{aligned} \lim_{\varepsilon \rightarrow 0} \varepsilon \|v_\varepsilon(\cdot)\|_{L^2(0,T;U)} &\leq \lim_{\varepsilon \rightarrow 0} \max \{\varepsilon; \varepsilon \|v_\varepsilon(\cdot)\|_{L^2(0,T;U)}\} \\ &\leq \lim_{\varepsilon \rightarrow 0} \max \{\varepsilon; \varepsilon \|v_\varepsilon(\cdot)\|_{L^2(0,T;U)}^2\} = 0. \end{aligned}$$

The last equation just is (3.16). The rest is to show (3.17) holds for $\varphi \in X$. (Notice, if $\varphi \in \bar{K}_{(0,T)}$ then (3.17) holds. Here we may show (3.17) holds for any given $\varphi \in X$.)

Suppose φ is arbitrarily given in X and

$$\varphi = \bar{\varphi} + \varphi^\perp,$$

where $\bar{\varphi} \in \bar{K}_{(0,T)}$ and $\varphi^\perp \in K_{(0,T)}^\perp$ —the orthogonal complement of $K_{(0,T)}$. Since $X = \bar{K}_{(0,T)} \oplus K_{(0,T)}^\perp$, one has

$$\|\varphi\|^2 = \|\bar{\varphi}\|^2 + \|\varphi^\perp\|^2 \quad \text{for any } \varphi \in X.$$

Denote

$$\bar{v}_\varepsilon = (\varepsilon + \hat{G})^{-1}(\mathcal{S}_T B)^*\bar{\varphi} \quad \text{and} \quad v_\varepsilon^\perp = (\varepsilon + \hat{G})^{-1}(\mathcal{S}_T B)^*\varphi^\perp.$$

Then

$$v_\varepsilon = \bar{v}_\varepsilon + v_\varepsilon^\perp.$$

It is not difficult to verify that

$$(3.18) \quad \lim J_\varepsilon(\bar{v}_\varepsilon, \varphi) = \|\varphi^\perp\|^2$$

and

$$(3.19) \quad \lim J_\varepsilon(v_\varepsilon, \varphi) = \|\varphi^\perp\|^2.$$

In fact, since $[(\mathcal{S}_T B)\bar{v} - \bar{\varphi}] \in \bar{K}_{(0,T)}$, one has

$$\begin{aligned} \|(\mathcal{S}_T B)\bar{v}_\varepsilon - \bar{\varphi}\|^2 + \|\varphi^\perp\|^2 &= \|(\mathcal{S}_T B)\bar{v}_\varepsilon - \varphi\|^2 \leq J_\varepsilon(\bar{v}_\varepsilon, \varphi) \\ &= \|(\mathcal{S}_T B)\bar{v}_\varepsilon - \varphi\|^2 + \varepsilon \|\bar{v}_\varepsilon(\cdot)\|_{L^2(0,T;U)}^2 \\ (3.20) \quad &= \|(\mathcal{S}_T B)\bar{v}_\varepsilon - \bar{\varphi}\|^2 + \varepsilon \|\bar{v}_\varepsilon(\cdot)\|_{L^2(0,T;U)}^2 + \|\varphi^\perp\|^2 \\ &= J_\varepsilon(\bar{v}_\varepsilon, \bar{\varphi}) + \|\varphi^\perp\|^2. \end{aligned}$$

By Theorem 2.1,

$$(3.21) \quad \lim_{\varepsilon \rightarrow 0} \|(\mathcal{S}_T B)\bar{v}_\varepsilon - \bar{\varphi}\|^2 = \lim_{\varepsilon \rightarrow 0} J_\varepsilon(\bar{v}_\varepsilon; \bar{\varphi}) = 0.$$

Equation (3.18) follows from (3.20) and (3.21). Similarly, since $[(\mathcal{S}_T B)v_\varepsilon - \bar{\varphi}] \in \bar{K}_{(0,T)}$, one has

$$\begin{aligned} \|\varphi^\perp\|^2 &\leq \|(\mathcal{S}_T B)v_\varepsilon - \bar{\varphi}\|^2 + \|\varphi^\perp\|^2 \\ &= \|(\mathcal{S}_T B)v_\varepsilon - \varphi\|^2 \leq J_\varepsilon(v_\varepsilon; \varphi) \leq J_\varepsilon(\bar{v}_\varepsilon; \varphi), \end{aligned}$$

and (3.19) holds by (3.18).

Thus (3.17) follows immediately from (3.19). \square

Remark 3.3. Assume $h \in K_{(0,T)}$. Then the corresponding minimum norm optimal control $u^*(\cdot)$ has the explicit formula (3.3) if and only if $h \in \mathcal{R}(G)$.

Remark 3.4. It is known that $h \in K_{(0,T)}$ if and only if that $\|u_\varepsilon(\cdot)\|$ is uniformly bounded for $0 < \varepsilon < \infty$ or there is strong convergence of some sequence $\{u_{\varepsilon_n}(\cdot); n = 1, 2, \dots\}$ in $L^2(0, T; U)$ under assumption of approximate controllability of the linear system (1.5). It is asked whether the strong convergence of the family $\{u_\varepsilon(\cdot); \varepsilon > 0\}$ with the parameter ε is a necessary and sufficient condition under assumption of approximate controllability of the linear system (1.5). It is a pity that the strong convergence of the family $\{u_\varepsilon(\cdot); \varepsilon > 0\}$ is only necessary and need not be sufficient even under assumption of approximate controllability. (See the example in the Appendix.)

4. Semilinear systems. In this section, we study exact reachability for a class of semilinear control systems described by (1.1) in which there exists some relation between the range of the nonlinear operator $\mathcal{F}_{(t_0, T)}$ (see the definition (4.2)) generated by the nonlinear function F and the range of the linear operator B .

Notation. We use $B_{(t_0, t_1)}$ to denote the linear bounded operator mapping $L^2(t_0, t_1; U)$ into $L^2(t_0, t_1; X)$ generated by B :

$$(4.1) \quad (B_{(t_0, t_1)}u)(t) = Bu(t) \quad \text{for } t_0 \leq t \leq t_1 \leq T.$$

A nonlinear operator $\mathcal{F}_{(t_0, t_1)}$ mapping $L^2(t_0, t_1; U)$ into $L^2(t_0, t_1; X)$ is defined by

$$(4.2) \quad (\mathcal{F}_{(t_0, t_1)}u)(t) = F(y(t; u), u(t)) \quad \text{for } t_0 \leq t \leq t_1 \leq T,$$

where $y(\cdot; u)$ is the solution mapping of (1.1) with $y(t_0; u) = 0$ defined by Lemma 1.1.

The ranges in $L^2(t_0, t_1; X)$ of the linear operator $B_{(t_0, t_1)}$ and the nonlinear operator $\mathcal{F}_{(t_0, t_1)}$ are denoted by $\mathcal{R}[B_{(t_0, t_1)}]$ and $\mathcal{R}[\mathcal{F}_{(t_0, t_1)}]$ respectively.

Define a linear operator $\mathcal{S}_{(t_0, t_1)}B$ and a nonlinear operator $\mathcal{S}_{(t_0, t_1)}\mathcal{F}$, both mapping $L^2(t_0, t_1; U)$ into X as follows:

$$(4.3) \quad \mathcal{S}_{(t_0, t_1)}Bu = \int_{t_0}^{t_1} S(t_1 - s)Bu(s) ds \quad \text{for } u(\cdot) \in L^2(t_0, t_1; U)$$

and

$$(4.4) \quad \mathcal{S}_{(t_0, t_1)}\mathcal{F}u = \int_{t_0}^{t_1} S(t_1 - s)F(y(s; u), u(s)) ds \quad \text{for } u(\cdot) \in L^2(t_0, t_1; U).$$

The ranges in X of the linear operator $\mathcal{S}_{(t_0, t_1)}B$ and the nonlinear operator $\mathcal{S}_{(t_0, t_1)}\mathcal{F}$ are denoted by $\mathcal{R}[\mathcal{S}_{(t_0, t_1)}B]$ and $\mathcal{R}[\mathcal{S}_{(t_0, t_1)}\mathcal{F}]$ respectively.

DEFINITION. The subset $K_{(t_0, t_1)}$ in X defined by

$$(4.5) \quad K_{(t_0, t_1)} = \mathcal{R}[\mathcal{S}_{(t_0, t_1)}B]$$

is called the reachable set at t_1 starting from t_0 of the linear control system (1.5) corresponding to the semilinear control system (1.1), where $0 \leq t_0 \leq t_1 \leq T$.

We need to make some basic assumptions on the reachable set $K_{(t_0, t_1)}$ of the linear control system (1.5) in order to study the exact reachability of the semilinear system (1.1) later.

Assumption (L). The linear control system (1.5) is $[t_0, t_1]$ -null controllable for every pair $0 \leq t_0 < t_1 \leq T$, i.e.

$$(4.6) \quad \mathcal{R}[S(t_1 - t_0)] \subset \mathcal{R}[\mathcal{S}_{(t_0, t_1)}B], \quad \text{for } 0 \leq t_0 < t_1 \leq T.$$

THEOREM 4.1. Assume for some $t_0 \in [0, T)$

$$(4.7) \quad \mathcal{R}[\mathcal{S}_{(t_0, T)}\mathcal{F}] \subset \overline{\mathcal{R}[\mathcal{S}_{(t_0, T)}B]}.$$

Then

$$(4.8) \quad K_{(\tau, T)}(F) \subset \bar{K}_{(t_0, T)}$$

holds for every $\tau \in [0, t_0]$ under Assumption (L). If

$$(4.9) \quad \mathcal{R}[\mathcal{S}_{(t_0, T)}\mathcal{F}] \subset \mathcal{R}[\mathcal{S}_{(t_0, T)}B],$$

then

$$(4.10) \quad K_{(\tau, T)}(F) \subset K_{(t_0, T)}$$

holds for $\tau \in [0, t_0]$ under Assumption (L).

Proof. Let ξ_T be an arbitrary element in $K_{(\tau, T)}(F)$ and $v(\cdot)$ be the corresponding control in $L^2(\tau, T; U)$, such that

$$(4.11) \quad \begin{aligned} \xi_T &= \mathcal{S}_{(\tau, T)}\mathcal{F}v + \mathcal{S}_{(\tau, T)}Bv \\ &= S(T - t_0)[\mathcal{S}_{(\tau, t_0)}\mathcal{F}v + \mathcal{S}_{(\tau, t_0)}Bv] + \mathcal{S}_{(t_0, T)}\mathcal{F}v + \mathcal{S}_{(t_0, T)}Bv. \end{aligned}$$

According to Assumption (L) there exists some $u(\cdot) \in L^2(t_0, T; U)$ such that

$$S(T - t_0)[\mathcal{S}_{(\tau, t_0)}\mathcal{F}v + \mathcal{S}_{(\tau, t_0)}Bv] = \mathcal{S}_{(t_0, T)}Bu.$$

Thus

$$(4.12) \quad \xi_T = \mathcal{S}_{(t_0, T)}B(u + v) + \mathcal{S}_{(t_0, T)}\mathcal{F}v.$$

Therefore (4.8) holds under Assumption (4.7) and (4.10) holds under Assumption (4.9). \square

THEOREM 4.2. *Assume that the constants K_b and K_2 in (1.4) and (1.6) respectively satisfy the following restriction:*

$$(4.13) \quad \sqrt{2}K_bK_2 < 1.$$

Assume that there exists some time t_0 close enough to T satisfying

$$(4.14) \quad T - t_0 < \frac{1 - \sqrt{2}K_bK_2}{\sqrt{2}K_bK_1M_1(\|B\| + K_2)}$$

and

$$(4.15) \quad \mathcal{R}[\mathcal{F}_{(t_0, T)}] \subset \overline{\mathcal{R}[B_{(t_0, T)}]}.$$

Then, for every $\tau \in [0, t_0]$,

$$(4.16) \quad K_{(t_0, T)} \subset K_{(\tau, T)}(F)$$

holds under Assumption (L).

Proof. Let $t_0 \in [0, T]$ satisfy (4.14) and (4.15). Assume that ξ_T is arbitrarily given in $K_{(t_0, T)}$ and let $v(\cdot)$ be an arbitrary control in $L^2(0, t_0; U)$. By Assumption (L) there exists some $\bar{v}(\cdot)$ in $L^2(t_0, T; U)$ such that

$$(4.17) \quad \xi_T - S(T - t_0)[\mathcal{S}_{(\tau, t_0)}\mathcal{F}v + \mathcal{S}_{(\tau, t_0)}Bv] = \mathcal{S}_{(t_0, T)}B\bar{v}.$$

Suppose $u_1(\cdot)$ is any given control in $L^2(t_0, T; U)$. By Assumption (4.15) there exists some $\bar{u}_1(\cdot)$ in $L^2(t_0, T; U)$ such that

$$(4.18) \quad \mathcal{F}_{(t_0, T)}u_1 + B_{(t_0, T)}\bar{u}_1 = e_1,$$

where $e_1(\cdot)$ is in $L^2(t_0, T; X)$ and

$$(4.19) \quad \|e_1(\cdot)\|_{L^2(t_0, T; X)} \leq Q.$$

Here Q is the positive constant defined by

$$(4.20) \quad Q = \sqrt{2}K_bK_2 + \sqrt{2}(T - t_0)K_bK_1M_1(\|B\| + K_2).$$

By inequality (4.14), $Q < 1$. Once $\bar{u}_n(\cdot) \in L^2(t_0, T; U)$, $n = 1, 2, \dots$, has been obtained, define

$$(4.21) \quad u_{n+1}(\cdot) = \bar{u}_n(\cdot) + \bar{v}(\cdot), \quad n = 1, 2, \dots,$$

and determine $\bar{u}_{n+1}(\cdot)$ in $L^2(t_0, T; U)$, $n = 1, 2, \dots$, by Assumption (4.15) as follows.

$$(4.22) \quad \mathcal{F}_{(t_0, T)}u_{n+1} + B_{(t_0, T)}\bar{u}_{n+1} = e_{n+1}, \quad n = 1, 2, \dots,$$

$$(4.23) \quad e_{n+1} \in L^2(t_0, T; X) \quad \text{and} \quad \|e_{n+1}(\cdot)\|_{L^2(t_0, T; X)} < \frac{Q^{n+1}}{(n+1)^2}, \quad n = 1, 2, \dots$$

By Lemma 1.1 and Assumptions (1.4) and (1.6), we have

$$\begin{aligned} \|\bar{u}_{n+1}(\cdot) - \bar{u}_n(\cdot)\|_{L^2(t_0, T; U)} &\leq K_b\|B\bar{u}_{n+1}(\cdot) - B\bar{u}_n(\cdot)\|_{L^2(t_0, T; X)} \\ &< \frac{2K_bQ^n}{n^2} + K_b\|\mathcal{F}_{(t_0, T)}u_{n+1} - \mathcal{F}_{(t_0, T)}u_n\|_{L^2(t_0, T; X)} \\ &\leq \frac{2K_bQ^n}{n^2} + \sqrt{2}K_bK_1\|y(\cdot; u_{n+1}) - y(\cdot; u_n)\|_{L^2(t_0, T; X)} \\ &\quad + \sqrt{2}K_bK_2\|u_{n+1}(\cdot) - u_n(\cdot)\|_{L^2(t_0, T; U)} \end{aligned}$$

$$\begin{aligned}
&\leq \frac{2K_b Q^n}{n^2} + Q \|u_{n+1}(\cdot) - u_n(\cdot)\|_{L^2(t_0, T; U)} \\
&= \frac{2K_b Q^n}{n^2} + Q \|\bar{u}_n(\cdot) - \bar{u}_{n-1}(\cdot)\|_{L^2(t_0, T; U)} \\
&\leq 2K_b Q^n \left[\frac{1}{n^2} + \frac{1}{(n-1)^2} + \cdots + 1 \right] \\
&\quad + Q^n \|\bar{u}_1(\cdot) - \bar{u}_0(\cdot)\|_{L^2(t_0, T; U)} \\
&\leq Q^n \bar{Q}, \quad n = 1, 2, \dots,
\end{aligned}$$

where $\bar{u}_0(\cdot) = u_1(\cdot) - \bar{v}(\cdot)$ and

$$\bar{Q} = 4K_b + \|\bar{u}_1(\cdot) - \bar{u}_0(\cdot)\|_{L^2(t_0, T; U)}.$$

By Assumption (4.14) one has $Q < 1$; therefore the sequence $\{u_n(\cdot); n = 1, 2, \dots\}$ is a Cauchy sequence in $L^2(t_0, T; U)$ and there exists some control $u(\cdot)$ in $L^2(t_0, T; U)$ such that

$$\lim_{n \rightarrow \infty} u_n(\cdot) = u(\cdot) \quad \text{in } L^2(t_0, T; U).$$

Due to the strong continuity of both the linear operator $\mathcal{S}_{(t_0, T)}B$ and the nonlinear operator $\mathcal{S}_{(t_0, T)}\mathcal{F}$ mapping $L^2(t_0, T; U)$ into X one has

$$\begin{aligned}
(4.24) \quad 0 &= \lim_{n \rightarrow \infty} \mathcal{S}_{(t_0, T)}e_n = \lim_{n \rightarrow \infty} [\mathcal{S}_{(t_0, T)}\mathcal{F}u_n + \mathcal{S}_{(t_0, T)}B\bar{u}_n] \\
&= \mathcal{S}_{(t_0, T)}\mathcal{F}u + \mathcal{S}_{(t_0, T)}B\bar{u} = \mathcal{S}_{(t_0, T)}\mathcal{F}u + \mathcal{S}_{(t_0, T)}Bu - \mathcal{S}_{(t_0, T)}B\bar{v} \\
&= S(T - t_0)[\mathcal{S}_{(\tau, t_0)}\mathcal{F}v + \mathcal{S}_{(\tau, t_0)}Bv] + \mathcal{S}_{(t_0, T)}\mathcal{F}u + \mathcal{S}_{(t_0, T)}Bu - \xi_T.
\end{aligned}$$

Define a control $w(\cdot)$ in $L^2(\tau, T; U)$ as follows:

$$(4.25) \quad w(t) = \begin{cases} v(t), & \tau \leq t < t_0, \\ u(t), & t_0 \leq t \leq T. \end{cases}$$

Then $y(T; w) = \xi_T$ follows immediately from (4.24) and (4.25), i.e.

$$\xi_T \in K_{(\tau, T)}(F) \quad \text{for every } \xi_T \in K_{(t_0, T)}.$$

This is (4.16). \square

Remark. Since (4.15) implies (4.7), one has

$$(4.26) \quad K_{(t_0, T)} \subset K_{(\tau, T)}(F) \subset \bar{K}_{(t_0, T)} \quad (\tau \in [0, t_0]),$$

under conditions of Theorem 4.2.

COROLLARY 4.3. (1). Assume that for each $u(\cdot) \in L^2(0, T; U)$

$$(4.27) \quad F(y(t; u), u(t)) \in \overline{\mathcal{R}[B]}, \quad 0 \leq t \leq T.$$

Then there exists some $t_0 \in [0, T)$ such that (4.16) holds under Assumptions (4.13) and (L).

(2). Assume that the nonlinear function F is independent of u i.e. $F(y, u) = F(y)$. Then there exists some $t_0 \in [0, T)$ such that (4.16) holds under Assumptions (L) and (4.27).

Proof. (1). We claim that for any given $t_0 \in [0, T)$ and $u(\cdot) \in C([0, T]; U)$ one has

$$\mathcal{F}_{(t_0, T)}u \in \overline{\mathcal{R}[B_{(t_0, T)}]}$$

under Assumption (4.27). In fact, for any given $t_0 \in [0, T)$ and $u(\cdot) \in C([0, T]; U)$ one has $F(y(\cdot; u), u(\cdot)) \in C([t_0, T]; X)$ by Lemma 1.2. Let $\varepsilon > 0$ be arbitrarily given; then there exists $\delta > 0$ such that

$$\|F(y(t_1; u), u(t_1)) - F(y(t_2; u), u(t_2))\| \leq \frac{\varepsilon}{2\sqrt{T-t_0}}$$

uniformly holds for any pair t_1, t_2 in $[t_0, T]$ with $|t_1 - t_2| < \delta$. Denote

$$\tau_0 = t_0 \quad \tau_{i+1} = \tau_i + \frac{\delta}{2}, \quad i = 1, 2, \dots$$

Then there exists some positive integer N such that $\tau_N \leq T < \tau_{N+1}$. By Assumption (4.27), there exist v_0, v_1, \dots, v_N in U such that

$$\|F(y(\tau_i; u), u(\tau_i)) - Bv_i\| < \frac{\varepsilon}{2\sqrt{T-t_0}}, \quad i = 0, 1, \dots, N.$$

Define

$$v(t) = v_i, \quad \text{for } t \in [\tau_i, \tau_{i+1}), \quad i = 0, 1, \dots, N.$$

Then $v(\cdot) \in L^2(t_0, T; U)$ and

$$\begin{aligned} & \|F(y(t; u), u(t)) - Bv(t)\| \\ & \leq \|F(y(t; u), u(t)) - F(y(\tau_i; u), u(\tau_i))\| + \|F(y(\tau_i; u), u(\tau_i)) - Bv_i\| < \frac{\varepsilon}{\sqrt{T-t_0}} \end{aligned}$$

holds for $t \in [\tau_i, \tau_{i+1}), i = 0, 1, \dots, N$. Thus

$$\|F(y(\cdot; u), u(\cdot)) - Bv(\cdot)\|_{L^2(t_0, T; X)} < \varepsilon,$$

i.e., $F(y(\cdot; u), u(\cdot)) \in \overline{\mathcal{R}[B_{(t_0, T)}]}$ for any $u(\cdot) \in C([t_0, T]; U)$.

Since \mathcal{F} continuously maps $L^2(t_0, T; U)$ into $L^2(t_0, T; X)$ and $C([t_0, T]; U)$ is dense in $L^2(t_0, T; U)$, the inclusion relation (4.15) holds for any $t_0 \in [0, T)$ under Assumption (4.27). Taking t_0 close enough to T such that (4.14) is satisfied we have immediately (4.16) from Theorem 4.2.

(2). Let the nonlinear function F be independent on u ; then $K_2 = 0$. Thus (4.13) is naturally satisfied and (4.16) holds by part (1). \square

COROLLARY 4.4. Assume $F(y, u) = F(y)$ and $\overline{\mathcal{R}(B)} = X$ in (1.1). Then (4.26) holds and $\overline{K_{(\tau, T)}(F)} = X$ ($\tau \in [0, t_0]$), i.e., not only is the system (1.1) approximately controllable but it also has the exact reachability property for $K_{(t_0, T)}$. (There are no restrictions on the range $\mathcal{R}(F)$ or $\mathcal{R}[\mathcal{F}_{(t_0, T)}]$ because of the assumption $\overline{\mathcal{R}(B)} = X$.)

COROLLARY 4.5. Assume that (4.13) holds and

$$(4.28) \quad \mathcal{R}[\mathcal{F}_{(t_0, T)}] \subset \mathcal{R}[B_{(t_0, T)}]$$

holds for some $t_0 \in [0, T)$ satisfying (4.14). Then

$$(4.29) \quad K_{(\tau, T)}(F) = K_{(t_0, T)}$$

holds for every $\tau \in [0, t_0]$ under Assumption (L).

Proof. Since all conditions in Theorem 4.2 are satisfied under Assumptions (4.13), (4.14) and stronger Assumption (4.28), (4.16) holds. Because (4.28) implies (4.9), (4.10) holds. Combining (4.6) and (4.10) gives (4.29). \square

In the rest of this paper, we are going to discuss Assumption (1.4) in § 1 and other assumptions of corollaries in this section with some examples.

Example 1. Consider the finite-dimensional control system (1.1) with $X = R^n$ and $U = R^m$ ($0 < m \leq n < +\infty$). Thus B is an $n \times m$ matrix. If $\text{rank } B = m$, then Assumption (1.4) is satisfied. Without loss of generality assume $|B_m| \neq 0$ where

$$B_m = \begin{pmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mm} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & & \vdots \\ b_{n1} & \cdots & b_{nm} \end{pmatrix}.$$

Since the $m \times m$ matrix B_m is invertible we have

$$\|u\|_{R^m} = \|B_m^{-1} B_m u\|_{R^m} \leq \|B_m^{-1}\| \|B_m u\|_{R^m} \leq \|B_m^{-1}\| \|Bu\|_{R^n}.$$

If $F(y, u)$ is in $\mathcal{R}[B]$ for any $y \in R^n$ and $u \in R^m$, then $K_{(\tau, T)}(F) = K_{(0, T)}$ by Corollary 4.5.

An interesting example is such a control system with m -dimensional nonlinear disturbance of the control:

$$(4.30) \quad \frac{dy}{dt} + Ay = B[\bar{F}(y, u) + u],$$

where $\bar{F}(y, u)$ with m components $F_1(y, u), \dots, F_m(y, u)$ is an arbitrary nonlinear function from $R^n \times R^m$ into R^m satisfying (1.6). Obviously it has the same form (1.1) with $F = B\bar{F}$ and satisfies conditions of Corollary 4.5. Thus $K_{(\tau, T)}(F) = K_{(0, T)}$, $0 \leq \tau < T$, i.e. the reachable set of the linear control system (1.5) is an invariant subset in the state space R^n about any nonlinear disturbance of the control u . \square

For the arbitrary control system (1.1) one of the most typical cases is that there exists a closed subspace X_m in X such that U is isometric to X_m . Denote P_m as the orthogonal projection of X on X_m . Then $P_m B$ is a bounded linear operator mapping U into X_m . Assume that

$$(4.31) \quad \text{the operator } P_m B \text{ has a bounded inverse } (P_m B)^{-1} \text{ mapping } X_m \text{ onto } U.$$

Then Assumption (1.4) holds under Assumption (4.31). In fact,

$$\|u\|_U \leq \|(P_m B)^{-1}(P_m B)u\|_U \leq \|(P_m B)^{-1}\| \|P_m B u\|_{X_m} \leq \|(P_m B)^{-1}\| \|Bu\|_X$$

holds for any given $u \in U$. Obviously Assumption (4.31) is a generalization of the condition $\text{rank } B = m$ in finite dimensional control systems.

Example 2. [9, Example 1]. Let $X = L^2(0, \pi)$, $A = -d^2/dx^2$ with $D(A)$ consisting of all $y \in X$ with $d^2y/dx^2 \in X$ and $y(0) = y(\pi) = 0$. Then $X = \{e_n; n = 1, 2, \dots\}$ where $e_n(x) = \sqrt{2/\pi} \sin nx$, $0 \leq x \leq \pi$, is the eigenfunction corresponding to the eigenvalue $-n^2$ of the operator $-A$, $n = 1, 2, \dots$. Define $U = \{e_2, e_3, \dots\}$. For any $u = \sum_{n=2}^{\infty} u_n e_n$ in U define

$$Bu = u_i e_1 + \sum_{n=2}^{\infty} u_n e_n,$$

where i is some fixed integer larger than 1. In this case, $X_m = U$ and $P_m B$ is the identity operator in U . Hence Assumption (1.4) is satisfied. \square

Example 3. Let $X = \sum_{j=1}^{\infty} \oplus X_j$ where X_j is r_j -dimensional subspace of X ($r_j < \infty$). Denote P_j as the orthogonal projection of X on X_j , $j = 1, 2, \dots$. Thus $B = \sum_{j=1}^{\infty} \oplus B_j$ where $B_j = P_j B$ maps U into X_j , $j = 1, 2, \dots$. Assume that \mathcal{T} is a subset of positive integer set $\{1, 2, \dots\}$ and

$$(4.32) \quad U = \sum_{j' \in \mathcal{T}} \oplus U_{j'} \quad \text{where } U_{j'} \text{ is isometric to } X_{j'}, \quad j' \in \mathcal{T}.$$

Hence the operator B is able to be rewritten as

$$B = \sum_{j' \in \mathcal{T}} \oplus B_{j'} \oplus B_{\text{com}} \quad \text{where } B_{\text{com}} = \sum_{j'' \in \mathcal{T}} \oplus B_{j''}.$$

The restriction of each $B_{j'}$ on $U_{j'}$ defines a finite-dimensional map $\hat{B}_{j'}$ mapping $U_{j'}$ into $X_{j'}$, $j' \in \mathcal{T}$. Assume that each $\hat{B}_{j'}$ is invertible, i.e. $\hat{B}_{j'}$ is equivalent to an invertible $r_{j'}$ by $r_{j'}$ matrix, $j' \in \mathcal{T}$, and

$$\sum_{j' \in \mathcal{T}} \|\hat{B}_{j'}^{-1}\|^2 < +\infty.$$

Then

$$P_m B = \sum_{j' \in \mathcal{T}} \oplus B_{j'}$$

has a bounded inverse $(P_m B)^{-1}$ and

$$\|(P_m B)^{-1}\|^2 = \sum_{j' \in \mathcal{T}} \|\hat{B}_{j'}^{-1}\|^2.$$

Thus (1.4) holds.

Similar to Example 1 the reachable set $K_{(0,T)}$ of the linear control system (1.5) is invariant for any nonlinear disturbance of the control u under suitable assumptions, i.e. $K_{(0,T)} = K_{(\tau,T)}(F)$ where $K_{(\tau,T)}(F)$ is the reachable set of (4.30) in which the nonlinear function \bar{F} maps $X \times U$ into U . \square

Example 4. Assume $U = X$ and $B = I$, the identity operator in X . Then $K_{(\tau,T)}(F) = K_{(0,T)}$ holds for every $\tau \in [0, T)$ under Assumption (4.13). If $F(y, u) = F(y)$ then it holds for every $\tau \in [0, T)$. \square

Appendix. Consider the one-dimensional heat equation with homogeneous boundary condition

$$(A.1) \quad \frac{\partial y(t, x)}{\partial t} = \frac{\partial^2 y(t, x)}{\partial x^2} + u(t, x), \quad 0 < t < T, \quad 0 < x < l.$$

$$(A.2) \quad y(t, 0) = y(t, l) = 0, \quad 0 \leq t \leq T,$$

$$(A.3) \quad y(0, x) = 0, \quad 0 \leq x \leq l.$$

Denote $X = L^2(0, l)$, $U = X$. Suppose $\lambda_n = n^2$, $n = 1, 2, \dots$, are the eigenvalues of the heat equation (A.1) with the boundary condition (A.2) and $e_n(x) = \sqrt{2/l} \sin(n\pi x/l)$, $n = 1, 2, \dots$, are the corresponding eigenfunctions which form a normalized orthogonal base of the space X .

Assume $v(\cdot) \in L^2(0, T; U) = L^2((0, T) \times (0, l))$ has the Fourier expansion

$$v(t) = \sum_{n=1}^{\infty} v_n(t) e_n, \quad \sum_{n=1}^{\infty} \int_0^T v_n^2(t) dt < +\infty,$$

and $\varphi \in X = L^2(0, l)$ has the Fourier expansion

$$\varphi = \sum_{n=1}^{\infty} \varphi_n e_n, \quad \sum_{n=1}^{\infty} \varphi_n^2 < +\infty.$$

Then

$$\begin{aligned}(\mathcal{S}_T B)v &= \sum_{n=1}^{\infty} \int_0^T e^{-\lambda_n(T-t)} v_n(t) dt e_n, \\(\mathcal{S}_T B)^* \varphi &= \sum_{n=1}^{\infty} e^{-\lambda_n(T-t)} \varphi_n e_n, \\\hat{G}v &= \sum_{n=1}^{\infty} e^{-\lambda_n(T-t)} \int_0^T e^{-\lambda_n(T-s)} v_n(s) ds e_n, \\G\varphi &= \sum_{n=1}^{\infty} \frac{1 - e^{-2\lambda_n T}}{2\lambda_n} \varphi_n e_n.\end{aligned}$$

We now consider such an element \bar{h} in X :

$$(A.4) \quad \bar{h} = \sum_{n=1}^{\infty} \bar{h}_n e_n, \quad \bar{h}_n = \frac{1 - e^{-2\lambda_n T}}{2\lambda_n}, \quad n = 1, 2, \dots.$$

It can be verified that

(1) $\bar{h} \in K_{(0,T)}$ and

$$(A.5) \quad \bar{u}(t) = \sum_{n=1}^{\infty} \bar{u}_n(t) e_n, \quad \bar{u}_n(t) = e^{-\lambda_n(T-t)}, \quad n = 1, 2, \dots,$$

is the minimum norm optimal control corresponding to \bar{h} .

(2) $\lim_{\varepsilon \rightarrow 0} u_\varepsilon(\cdot) = \bar{u}(\cdot)$ in $L^2(0, T; U)$.

(3). $\bar{h} \notin \mathcal{R}(G)$ and $\bar{u}(\cdot) \notin \mathcal{U}$.

Proof of (1). It is obvious that $\bar{h} = (\mathcal{S}_T B)\bar{u} \in K_{(0,T)}$. Suppose $u(\cdot) \in L^2(0, T; U)$ is an arbitrary control in $V_{(0,T)}[h]$

$$u(t) = \sum_{n=1}^{\infty} u_n(t) e_n, \quad \sum_{n=1}^{\infty} \int_0^T u_n^2(t) dt < +\infty.$$

Since $\bar{h} = (\mathcal{S}_T B)u$, thus

$$\bar{h}_n = \int_0^T e^{-\lambda_n(T-t)} u_n(t) dt$$

and

$$\int_0^T \bar{u}_n^2(t) dt = \frac{1 - e^{-2\lambda_n T}}{2\lambda_n} = \bar{h}_n \leq \left(\frac{1 - e^{-2\lambda_n T}}{2\lambda_n} \right)^{1/2} \|u_n(\cdot)\|_{L^2(0,T)},$$

i.e., $\|\bar{u}_n(\cdot)\|_{L^2(0,T)} \leq \|u_n(\cdot)\|_{L^2(0,T)}$, $n = 1, 2, \dots$. Hence

$$\|\bar{u}(\cdot)\|_{L^2(0,T;U)} \leq \|u(\cdot)\|_{L^2(0,T;U)}$$

for any $u(\cdot) \in V_{(0,T)}[h]$ and $\bar{u}(\cdot)$ is the unique minimum norm optimal control corresponding to \bar{h} .

Proof of (2). Suppose $\varepsilon > 0$ and $v_\varepsilon = (\varepsilon + \hat{G})^{-1} \bar{u}$ with the Fourier expansion

$$v_\varepsilon(t) = \sum_{n=1}^{\infty} v_{\varepsilon_n}(t) e_n.$$

Then $(\varepsilon + \hat{G})v_\varepsilon = \bar{u}$ is equivalent to that each component $v_{\varepsilon_n}(t)$ ($n = 1, 2, \dots$) of v satisfies

$$(A.6) \quad v_{\varepsilon_n}(t) + \frac{1}{\varepsilon} \int_0^T e^{-\lambda_n(T-t)} e^{-\lambda_n(T-s)} v_{\varepsilon_n}(s) ds = \frac{e^{-\lambda_n(T-t)}}{\varepsilon}, \quad \varepsilon > 0, \quad n = 1, 2, \dots.$$

It is easy to verify that each Fredholm equation (A.6) has a unique solution

$$v_{\varepsilon_n}(t) = \frac{2\lambda_n e^{-\lambda_n(T-t)}}{1 + 2\varepsilon\lambda_n - e^{-2\lambda_n T}}, \quad n = 1, 2, \dots$$

Thus

$$\begin{aligned} \|\varepsilon v_{\varepsilon}(\cdot)\|_{L^2(0,T;U)}^2 &= \varepsilon^2 \sum_{n=1}^{\infty} \int_0^T v_{\varepsilon_n}^2(t) dt \\ &= \varepsilon \sum_{n=1}^{\infty} \frac{2\varepsilon\lambda_n(1 - e^{-2\lambda_n T})}{(1 + 2\varepsilon\lambda_n - e^{-2\lambda_n T})^2} \leq \varepsilon \sum_{n=1}^{\infty} \frac{1 - e^{-2\lambda_n T}}{1 + 2\varepsilon\lambda_n - e^{-2\lambda_n T}}, \quad (\varepsilon > 0). \end{aligned}$$

For each $\varepsilon > 0$ we may uniquely define a positive integer $N(\varepsilon)$ such that

$$\varepsilon N^2(\varepsilon) \leq 1 < \varepsilon [N(\varepsilon) + 1]^2 \quad \text{for } \varepsilon > 0.$$

Thus, by definition of $N(\varepsilon)$, we have

$$\begin{aligned} \lim_{\varepsilon \rightarrow 0} \varepsilon \|v_{\varepsilon}(\cdot)\|_{L^2(0,T;U)}^2 &\leq \lim_{\varepsilon \rightarrow 0} \varepsilon \left(\sum_{n=1}^{N(\varepsilon)} + \sum_{n=N(\varepsilon)+1}^{\infty} \right) \frac{1 - e^{-2\lambda_n T}}{1 + 2\varepsilon\lambda_n - e^{-2\lambda_n T}} \\ &\leq \lim_{\varepsilon \rightarrow 0} \varepsilon N(\varepsilon) + \lim_{\varepsilon \rightarrow 0} \varepsilon \sum_{n=N(\varepsilon)+1}^{\infty} \frac{1 - e^{-2\lambda_n T}}{2\varepsilon\lambda_n} \\ &\leq \lim_{\varepsilon \rightarrow 0} \frac{1}{N(\varepsilon)} + \frac{1}{2} \lim_{\varepsilon \rightarrow 0} \sum_{n=N(\varepsilon)+1}^{\infty} \frac{1}{n^2} = 0. \end{aligned}$$

Thus (2) is proved.

Proof of (3). If $h \in X$ and

$$h = \sum_{n=1}^{\infty} h_n e_n, \quad \sum_{n=1}^{\infty} h_n^2 < +\infty,$$

then $h \in \mathcal{R}(G)$ if and only if that there exists some $\varphi = \sum_{n=1}^{\infty} \varphi_n e_n$ with $\sum_{n=1}^{\infty} \varphi_n^2 < +\infty$ such that

$$h_n = \frac{1 - e^{-2\lambda_n T}}{2\lambda_n} \varphi_n, \quad n = 1, 2, \dots$$

Obviously, the element \bar{h} in $K_{(0,T)}$ defined by (A.4) does not have to be in $\mathcal{R}(G)$. Similarly, $\bar{u}(\cdot) \notin \mathcal{U}$. \square

Acknowledgments. The author would like to thank Professors T. I. Seidman and H. O. Fattorini for their warm invitation to visit UMBC and UCLA successively and to thank Professor H. O. Fattorini for his helpful discussion and his reading of parts of the manuscript.

REFERENCES

- [1] H. O. FATTORINI AND D. L. RUSSELL, *Exact controllability theorems for linear parabolic equations in one space dimension*, Arch. Rational Mech. Anal., 43 (1971), pp. 272–292.
- [2] H. O. FATTORINI, *The time-optimal control problem in Banach spaces*, Appl. Math. Optim., 1 (1974), pp. 163–188.
- [3] T. KOBAYASHI, *Some remarks on controllability for distributed parameter systems*, this Journal, 16 (1978), pp. 733–742.
- [4] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

- [5] D. L. RUSSELL, *Mathematics of Finite-Dimensional Linear Control Systems, Theory and Design*, Marcel Dekker, New York and Basel, 1979.
- [6] T. I. SEIDMAN AND H. X. ZHOU, *Existence and uniqueness of optimal controls for a semilinear parabolic equation*, this Journal, 20 (1982), pp. 747–762.
- [7] K. YOSIDA, *Functional Analysis*, Springer-Verlag, New York, 1980.
- [8] H. X. ZHOU, *A note on approximate controllability for semilinear one-dimensional heat equation*, Appl. Math. Optim., 8 (1982), pp. 275–285.
- [9] ———, *Approximate controllability for a class of semilinear abstract equations*, this Journal, 21 (1983), pp. 551–565.

DUAL APPROXIMATIONS IN OPTIMAL CONTROL*

WILLIAM W. HAGER† AND GEORGE D. IANULESCU‡

Abstract. We analyze a dual approximation for the solution to an optimal control problem. The differential equation is handled with a Lagrange multiplier while other constraints are treated explicitly. An algorithm for solving the dual problem is presented.

Key words. duality, approximation, finite elements, sensitivity, optimal control

TABLE OF CONTENTS

	Page
1. Introduction	423
2. The method	424
3. Bounded variation	426
4. Absolute continuity, I	428
5. Absolute continuity, II	429
6. Interiority	432
7. Pointwise minimization	434
8. Dual formulations	436
9. Fundamental inequalities	442
10. Error estimates	446
11. Algorithms	451
Appendix 1. Existence	458
Appendix 2. Integrand regularity	460
Appendix 3. Exact solutions	462
References	463

1. Introduction. In computing the solution to an optimal control problem, most of the difficulty centers around the differential equation. In this paper, we consider a dual approach where the differential equation is handled with a Lagrange multiplier, while other constraints are treated explicitly. This scheme was first studied by Rockafellar [47], who establishes existence results and optimality conditions for primal and dual solutions. We now analyze the following numerical aspects of the dual procedure:

- (1) Existence of finite dimensional approximations.
- (2) Regularity of dual solutions.
- (3) Relations between dual multipliers and primal solutions.
- (4) Error estimates for piecewise polynomial approximation.
- (5) Techniques for solving the dual problem.

The first error estimate for a dual approximation to a control problem is given by Bosarge and Johnson [5], who study unconstrained problems with quadratic cost and linear system dynamics. For piecewise polynomials of degree k , they show that the \mathcal{L}^2 error in the approximating control and state is order k . In [14] we introduce linear inequality state and control constraints, and analyze a *full dual* scheme where multipliers are attached to each constraint. The dual optimization is related to an

* Received by the editors November 29, 1979, and in revised form February 3, 1983. This research was supported partly by the National Science Foundation under grants MCS 8101892 and MCS 7825526, by the Office of Naval Research under grant N00014-76-C-0369, by the Hertz Foundation, and by the Ford Foundation. The analysis of the dual scheme and the algorithm for solving the dual problem are contributed by the first author. The four numerical examples are contributed by the second author.

† Department of Mathematics, The Pennsylvania State University, University Park, Pennsylvania 16802.

‡ Jet Propulsion Laboratory, Pasadena, California 91109.

energy projection, and the error is at best order 1.5 due to discontinuities in the control's derivative. Later [16] our estimates are extended to general convex problems, and examples are analyzed in [20]. Mathis and Reddien [32] notice that Bosarge and Johnson's estimate for the control error is not optimal, and sharpen this bound using a duality argument [37]. Some dual approximations to systems described by partial differential equations are developed by Mossino [35], [36] and Bosarge, Johnson and Smith [6].

Other techniques for constrained control problems are contained in [54], [25], [3], [10], [24], [56] and [39]. Penalty methods for variational problems were introduced by Courant [9], and applied to control problems by Russell [54], Lasdon, Warren and Rice [25], and others [3], [10]. These methods have wide applicability, although the penalized problem is ill-conditioned as the penalty grows—see Luenberger [29]. Jacobson and Lele [24] note that some state constraints can be removed by Valentine's device [57]. Thompson and Volz [56] show that control problems with linear dynamics, quadratic cost and a single linear inequality state constraint can be solved using a nonsymmetric Riccati equation. Pironneau and Polak [39] present a dual method of centers for problems with inequality endpoint constraints and affine inequality control constraints.

Advantages of the dual scheme are its speed and generality; problems with endpoint, control and state constraints can be handled. Although our convergence theory assumes that the system dynamics is linear, the scheme applies to nonlinear systems. Unfortunately, there are cases [31], [46] where the dual does not solve the primal. A cure for "duality gaps" is the multiplier methods [4], [21], [41], [48], [49] which combine penalty and duality techniques.

2. The method. A control problem is the constrained minimization of a functional $C(x, u)$ over a collection of controls

$$u: \mathcal{T} \rightarrow R^m$$

and a collection of states

$$x: \mathcal{T} \rightarrow R^n$$

where R denotes the real numbers and R^n is the n -fold Cartesian product $R \times R \times \cdots \times R$. For convenience, let us assume that \mathcal{T} is the interval $[0, 1] \subset R$. Throughout this paper, Lebesgue measure is used for \mathcal{T} , and measurable functions are equal if they are equal almost everywhere. Let \mathcal{X} denote the set of pairs (x, u) where x is absolutely continuous and u is summable.

The admissible set for the control problem is described by two types of constraints. First there is the *system dynamics* $M(x, u) = 0$ where $M: \mathcal{X} \rightarrow \mathcal{L}^1$ is a differential operator which we assume is linear:

$$M(x, u)(t) := x'(t) - A(t)x(t) - B(t)u(t);$$

here \mathcal{L}^1 is the space of summable functions $f: \mathcal{T} \rightarrow R^n$, $A(t)$ is an $n \times n$ matrix for each $t \in \mathcal{T}$ whose individual elements are summable, and $B(t)$ is an $n \times m$ matrix for every $t \in \mathcal{T}$ whose elements are essentially bounded. Second, there may be constraints such as

$$\begin{aligned} x(0) &= a && \text{(initial condition),} \\ x(1) &= b && \text{(target),} \\ |u(t)| &\leq 2 && \text{(control constraint), or} \\ x(t) &\geq 0 && \text{(state constraint).} \end{aligned}$$

We assume these conditions are embedded in the cost functional by setting $C(x, u) = \infty$ when the constraint is violated. This convention is discussed in Rockafellar's paper [52]. Hence $C: \mathcal{Z} \rightarrow \bar{R}$ where \bar{R} is the extended reals $R \cup \{\infty\}$, and the control problem takes the form

$$(P) \quad \begin{array}{ll} \text{minimize} & C(z) \\ \text{subject to} & M(z) = 0, \quad z \in \mathcal{Z}. \end{array}$$

Of course, z denotes the pair (x, u) . Since the cost is minimized, we are only concerned with those z for which $C(z)$ is finite. The *effective domain* of C is given by

$$\text{dom } C := \{z \in \mathcal{Z}: C(z) < \infty\}.$$

It is assumed that C *proper* and there exists a *feasible function* for (P); that is, the effective domain of C is nonempty and there exists $z \in \text{dom } C$ such that $M(z) = 0$.

Now we formulate the dual of (P). Letting \mathcal{L}^∞ be the space of essentially bounded functions $f: \mathcal{T} \rightarrow R^n$, the *dual functional* $L: \mathcal{L}^\infty \rightarrow R \cup \{-\infty\}$ is defined by

$$L(p) = \inf \{C(z) + \langle p, M(z) \rangle: z \in \mathcal{Z}\}$$

where $\langle \cdot, \cdot \rangle$ is the usual \mathcal{L}^2 inner product:

$$\langle f, g \rangle := \int_{\mathcal{T}} f(t) \cdot g(t) \, dt$$

for all measurable $f, g: \mathcal{T} \rightarrow R^n$. Here \cdot is the *Euclidean dot product*. The dual problem becomes:

$$(D) \quad \begin{array}{ll} \text{maximize} & L(p) \\ \text{subject to} & p \in \mathcal{L}^\infty. \end{array}$$

Since L is maximized, the effective domain of the dual functional is given by

$$\text{dom } L = \{p \in \mathcal{L}^\infty: L(p) > -\infty\}.$$

Clearly, from the definition of L ,

$$(2.1) \quad \sup \{L(p): p \in \mathcal{L}^\infty\} \leq \inf \{C(z): z \in \mathcal{Z}, M(z) = 0\}.$$

This inequality is sometimes called *weak duality* [31]. The stronger statement, "There exists a solution to (D) and (2.1) is an equality," follows from the

BOUNDEDNESS ASSUMPTION. *There exists $\rho > 0$ such that*

$$\sup_{\substack{w \in \mathcal{L}^1 \\ \|w\|_{\mathcal{L}^1} \leq \rho}} \inf_{\substack{z \in \mathcal{Z} \\ M(z) = w}} C(z) < \infty$$

where

$$\|w\|_{\mathcal{L}^1} = \int_{\mathcal{T}} |w(t)| \, dt$$

and $|\cdot|$ is the *Euclidean norm*.

THEOREM 2.1. *If C is convex and the boundedness hypothesis is satisfied, then there exists a solution p to the dual problem, and*

$$L(p) = \inf \{C(z): z \in \mathcal{Z}, M(z) = 0\}.$$

Notice that the dual problem has a solution even though the primal problem may have no solution. Theorem 2.1 is an elementary application of general duality principles (see Theorem A.3 in Appendix 1).

If p solves the dual problem, any solution to the primal problem minimizes

$$(2.2) \quad C(z) + \langle p, M(z) \rangle$$

over $z \in \mathcal{Z}$. Thus, to solve the primal problem, we can first solve the dual and then find those $z \in \mathcal{Z}$ which attain the minimum in (2.2). This procedure is modified for numerical computations. We replace the dual feasible set by a closed subset \mathcal{S}_h of a finite dimensional space giving us the approximation:

$$(D_h) \quad \begin{array}{ll} \text{maximize} & L(p) \\ \text{subject to} & p \in \mathcal{S}_h. \end{array}$$

An important issue is whether there exists a solution to (D_h) . By Theorem 2.2 below, the boundedness hypothesis assures existence. If p_h solves (D_h) , we take as an approximation to a primal solution any $z_h \in \mathcal{Z}$ for which

$$L(p_h) = C(z_h) + \langle p_h, M(z_h) \rangle.$$

The paper's main focus is the second issue: Is z_h "close" to a primal solution? Under a uniform convexity hypothesis, the answer is yes.

These convergence properties are related to the smoothness of dual solutions. If an optimal p lies just in \mathcal{L}^∞ , the dual problem may be hard since the approximation of essentially bounded functions using standard \mathcal{S}_h is not easy. In the following sections, we observe that p has some smoothness. This section concludes with an existence theorem for (D_h) . Appendix 1 proves a more general result.

THEOREM 2.2. *Suppose that $\mathcal{S}_h \subset \mathcal{L}^\infty$ where \mathcal{S}_h is also a closed subset of a finite dimensional space. If the boundedness hypothesis holds, then (D_h) has a solution.*

3. Bounded variation. Under appropriate assumptions, the classical minimum principle [40], [26] for the control problem

$$\begin{array}{ll} \text{minimize} & \int_{\mathcal{T}} f(x(t), u(t), t) dt \\ \text{subject to} & M(x, u) = 0, \quad (x, u) \in \mathcal{Z}, \quad x(0) = a, \\ & u(t) \in U \subset R^m \quad \text{almost everywhere,} \end{array}$$

states that an optimal solution satisfies

$$h(u(t), t) = \min \{h(v, t): v \in U\} \quad \text{almost everywhere}$$

where

$$h(v, t) = f(x(t), v, t) + q(t)^T B(t)v$$

and

$$\begin{array}{l} q'(t) = -A(t)^T q(t) - \nabla_x f(x(t), u(t), t)^T \quad \text{almost everywhere,} \\ q(1) = 0. \end{array}$$

Above, T denotes *transpose*, ∇_x is the *gradient* with respect to the state argument, and q is called the *costate*.

We expect that dual solutions are related to the costate, but observe that q is differentiable while $\text{dom } L \subset \mathcal{L}^\infty$. However, a fairly weak hypothesis guarantees some smoothness for elements in $\text{dom } L$. Let $\mathcal{C}^\infty \subset \mathcal{L}^\infty$ be the subspace of infinitely differentiable functions. We introduce the sets

$$\mathcal{D}^\gamma = \{y \in \mathcal{C}^\infty: \|y\|_{\mathcal{C}} \leq \gamma\}, \quad \mathcal{E} = \{y \in \mathcal{C}^\infty: y(0) = y(1) = 0\}, \quad \mathcal{E}^\gamma = \mathcal{E} \cap \mathcal{D}^\gamma,$$

where

$$\|y\|_{\mathcal{C}} = \sup \{|y(t)|: t \in \mathcal{T}\},$$

and the

INTERIORITY ASSUMPTION. *There exist $\bar{z} = (\bar{x}, \bar{u}) \in \mathcal{Z}$ and $\gamma > 0$ such that*

$$\sup \{C(\bar{x} + \psi, \bar{u}): \psi \in \mathcal{E}^\gamma\} < \infty.$$

Finally, recall that elements of \mathcal{L}^∞ are equivalence classes of functions equal almost everywhere, and let $\mathcal{B} \subset \mathcal{L}^\infty$ denote the subspace of functions with bounded variation that are right continuous on $(0, 1)$.

THEOREM 3.1. *If $p \in \text{dom } L$ and the interiority assumption holds, then $p \cap \mathcal{B}$ is nonempty.*

Proof. Inserting $z = (\bar{x} - \psi, \bar{u})$ into the relation

$$C(z) + \langle p, M(z) \rangle \geq L(p) \quad \forall z \in \mathcal{Z},$$

and utilizing the interiority hypothesis, we have

$$(3.1) \quad \sup \{\langle p, \psi' \rangle: \psi \in \mathcal{E}^\gamma\} < \infty.$$

Given $\mathcal{S} \subset \mathcal{C}^\infty$ and $f \in \mathcal{L}^1$, let us define

$$f^+(\mathcal{S}) = \sup \{\langle f, \psi' \rangle: \psi \in \mathcal{S}\}.$$

Hence, $p^+(\mathcal{E}^\gamma) < \infty$ by (3.1). Any $\phi \in \mathcal{D}^1$ can be written as

$$\phi = \psi + \delta$$

where δ is the linear function agreeing with $\phi(t)$ at $t = 0$ and 1 , and

$$\psi = (\phi - \delta) \in \mathcal{E}^2.$$

Since $\|\delta'\|_{\mathcal{C}} \leq 2$, it follows that

$$\langle p, \phi' \rangle \leq \langle p, \psi' \rangle + 2\|p\|_{\mathcal{L}^1},$$

and taking the supremum over $\phi \in \mathcal{D}^1$ yields

$$p^+(\mathcal{D}^1) \leq p^+(\mathcal{E}^2) + 2\|p\|_{\mathcal{L}^1} = \frac{2}{\gamma} p^+(\mathcal{E}^\gamma) + 2\|p\|_{\mathcal{L}^1} < \infty.$$

The next lemma completes the proof. \square

LEMMA 3.2. *If $f \in \mathcal{L}^1$ and $f^+(\mathcal{D}^1) < \infty$, then $f \cap \mathcal{B}$ is nonempty.*

Proof. Let \mathcal{C} be the space of continuous functions $y: \mathcal{T} \rightarrow R^n$ and define $\Lambda: \mathcal{C}^\infty \rightarrow R$ by $\Lambda(\phi) = \langle f, \phi' \rangle$. Since

$$\Lambda(\phi) \leq f^+(\mathcal{D}^1) \|\phi\|_{\mathcal{C}},$$

Λ can be extended to a continuous linear functional $\tilde{\Lambda}: \mathcal{C} \rightarrow R$. By the Riesz representation theorem, there exists $g \in \mathcal{B}$ such that $g(1) = 0$ and

$$\tilde{\Lambda}(\phi) = \int_0^1 \phi(t) \cdot dg(t) \quad \forall \phi \in \mathcal{C}.$$

In particular, if $\phi \in \mathcal{C}^\infty$ and $\phi(0) = 0$, integration by parts gives us

$$\tilde{\Lambda}(\phi) = -\langle g, \phi' \rangle = \Lambda(\phi) = \langle f, \phi' \rangle.$$

Therefore, $f(t) = -g(t)$ almost everywhere. \square

4. Absolute continuity, I. Although there exist dual solutions with bounded variation, the following example, called the obstacle problem, shows that a continuous dual solution may not exist:

$$\begin{aligned} &\text{minimize} \quad \int_{\mathcal{T}} u(t)^2 dt \\ &\text{subject to} \\ &\quad x'(t) = u(t) \quad \text{almost everywhere,} \\ &\quad x(t) \geq \alpha(t) \quad \text{for all } t \in \mathcal{T}, \\ &\quad x(0) = x(1) = 0, \quad (x, u) \in \mathcal{X}, \end{aligned}$$

where $\alpha \in \mathcal{A}$ is given data. It turns out that the optimal state is the profile of an elastic string lying in the (t, x) plane with ends fastened at $(0, 0)$ and $(1, 0)$ and stretched over the obstacle $\alpha(t)$; moreover, a solution to the dual problem is the *derivative* of the optimal state. Hence, a dual solution can be discontinuous when the obstacle has discontinuous derivatives.

For problems with “smooth” data, we have already established the existence of a Lipschitz continuous dual solution [15]. On the other hand, the next section refines our earlier work [19] and shows that combinations of multipliers are absolutely continuous even if the data is rough. In this section, the existence of absolutely continuous solutions is established for control constrained problems. We say that the sequence $\{\psi_k\} \subset \mathcal{L}^\infty$ converges pointwise to $\psi \in \mathcal{L}^\infty$ if

$$\lim_k \psi_k(t) = \psi(t) \quad \text{almost everywhere}$$

and the essential supremum of ψ_k over \mathcal{T} is bounded independently of k . In particular, the sequence is called a *0-sequence* if $\psi = 0$. A functional F defined on $\Psi \subset \mathcal{L}^\infty$ is *0-stable* on Ψ if there exists $N < \infty$ such that

$$\overline{\lim}_k F(\psi_k) < N$$

for each 0-sequence $\{\psi_k\} \subset \Psi$. Here $\overline{\lim}_k$ is an abbreviation for $\limsup_{k \rightarrow \infty}$. Finally, let us introduce the

0-STABILITY ASSUMPTION. For some $\bar{z} = (\bar{x}, \bar{u}) \in \mathcal{X}$, $C(\bar{x} + \cdot, \bar{u})$ is 0-stable on \mathcal{E} .

Letting $\mathcal{BE} \subset \mathcal{B}$ be the subspace of functions that are continuous at $t = 0$ and 1, we have:

THEOREM 4.1. If $p \in \mathcal{BE} \cap \text{dom } L$ and the 0-stability hypothesis is satisfied, then p is absolutely continuous.

By Remark 1 in § 8, absolute continuity for a dual solution is also deduced from [47, Thm. 4] in some cases. In his proof of [47, Thm. 4], Rockafellar uses both an “attainability” and an “integrability” assumption. Attainability is related to, but weaker than, our boundedness condition, while integrability implies that the cost functional satisfies a growth condition, a requirement not present in our analysis.

Now let us prove the theorem. Inserting $z = (\bar{x} - \psi_k, \bar{u})$ into the relation

$$C(z) + \langle p, M(z) \rangle \geq L(p) \quad \forall z \in \mathcal{Z},$$

and invoking the 0-stability hypothesis,

$$(4.1) \quad \overline{\lim}_k \langle p, \psi'_k \rangle \leq N - L(p) - \langle p, M(\bar{z}) \rangle$$

for all 0-sequences $\{\psi_k\} \subset \mathcal{E}$. Moreover, if c is a scalar, $\{c\psi_k\}$ is a 0-sequence satisfying (4.1). Since c is arbitrary,

$$(4.2) \quad \lim_k \langle p, \psi'_k \rangle = 0.$$

For convenience, let us assume that $n = 1$ and let λ be the regular Borel measure corresponding to p [53, Thm. 8.14]. We show that λ is absolutely continuous with respect to Lebesgue measure μ ; that is, $\lambda(E) = 0$ for every Lebesgue measurable set E such that $\mu(E) = 0$. Given a closed set $E \subset (0, 1)$, it is well known [22, p. 4] that there exists a sequence $\{\psi_k\} \subset \mathcal{E}$ such that $0 \leq \psi_k \leq 1$ and

$$\lim_k \psi_k(x) = K_E(x)$$

for every $x \in \mathcal{T}$ where K_E is the characteristic function of E . Of course, $\{\psi_k\}$ is a 0-sequence if $\mu(E) = 0$. By the dominated convergence theorem,

$$\lim_k \int_{\mathcal{T}} \psi_k(t) d\lambda(t) = \lambda(E).$$

On the other hand, integrating (4.2) by parts,

$$(4.3) \quad \lim_k \int_{\mathcal{T}} \psi_k(t) d\lambda(t) = 0$$

for all 0-sequences $\{\psi_k\} \subset \mathcal{E}$. Since λ is a regular Borel measure, we conclude that λ is absolutely continuous with respect to μ and hence p is absolutely continuous [53, Thm. 8.16]. \square

Defining the Stieltjes integral

$$\langle p, f \rangle_{\mathcal{E}} = \int_{\mathcal{T}} f(t) \cdot dp(t)$$

for $f \in \mathcal{C}$ and $p \in \mathcal{B}$, observe that (4.3) is equivalent to the statement, “ $\langle p, \cdot \rangle_{\mathcal{E}}$ is 0-stable on \mathcal{E} ,” so we have:

COROLLARY 4.2. *If $p \in \mathcal{B}\mathcal{E}$ and $\langle p, \cdot \rangle_{\mathcal{E}}$ is 0-stable on \mathcal{E} , then p is absolutely continuous.*

5. Absolute continuity, II. Now let us characterize the feasible dual functions for state constrained problems. If $D: \mathcal{X} \rightarrow \bar{R}$ and \mathcal{S} is a set of states, we say that (\mathcal{S}, D) is an *extension* of C if $C = D$ on $\text{dom } C$ and

$$\text{dom } C = \{(x, u) \in \text{dom } D: x \in \mathcal{S}\}.$$

For example, in the obstacle problem,

$$\mathcal{S} = \{x \in \mathcal{A}: x(t) \geq \alpha(t) \forall t \in \mathcal{T}\}$$

where $\mathcal{A} \subset \mathcal{B}$ is the subspace of absolutely continuous functions and

$$D(x, u) = \begin{cases} \langle u, u \rangle & \text{if } x(0) = x(1) = 0, \\ \infty & \text{otherwise} \end{cases}$$

is an extension of C .

The following theorem introduces a multiplier for \mathcal{S} . Defining the norm

$$\|x\|_{\mathcal{A}} = |x(0)| + \int_{\mathcal{T}} |x'(t)| dt \quad \text{for } x \in \mathcal{A},$$

observe that \mathcal{A} and $R^n \times \mathcal{L}^1$ are isomorphic with elements in the respective spaces related by the rule

$$x \rightarrow (x(0), x').$$

Hence, any $f \in \mathcal{A}^*$, the space of bounded linear functionals on \mathcal{A} , can be expressed in the form

$$f(x) = c \cdot x(0) + \langle \omega, x' \rangle := \langle (c, \omega), x \rangle_{\mathcal{A}}$$

where $c \in R^n$ and $\omega \in \mathcal{L}^\infty$.

THEOREM 5.1. *Suppose that $p \in \text{dom } L$, and (\mathcal{S}, D) is an extension of C where \mathcal{S} and D are convex, and $\mathcal{S} \subset \mathcal{A}$ has nonempty interior. Then there exists $\gamma = (c, \omega) \in R^n \times \mathcal{L}^\infty$ such that*

$$(5.1) \quad D(z) + \langle p, M(z) \rangle + \langle \gamma, x - y \rangle_{\mathcal{A}} \geq L(p)$$

for all $z = (x, u) \in \mathcal{Z}$ and $y \in \mathcal{S}$. Furthermore, if D satisfies the 0-stability hypothesis and $(p + \omega) \in \mathcal{BE}$, then $p + \omega$ is absolutely continuous.

Since $D(z) \leq C(z)$ for all $z \in \mathcal{Z}$, (5.1) implies that

$$(5.2) \quad C(z) + \langle p, M(z) \rangle + \langle \gamma, x - y \rangle_{\mathcal{A}} \geq L(p)$$

for all $z = (x, u) \in \mathcal{Z}$ and $y \in \mathcal{S}$. The existence of γ satisfying (5.1) follows directly from Fenchel's duality theorem [46], [28] and the fact that

$$L(p) = \inf \{ D(z) + \langle p, M(z) \rangle : z = (x, u) \in \mathcal{Z}, x \in \mathcal{S} \}.$$

If D satisfies the 0-stability hypothesis and $(p + \omega) \in \mathcal{BE}$, it is easy to deduce from (5.1) that $\langle p + \omega, \cdot \rangle_{\mathcal{C}}$ is 0-stable on \mathcal{E} . By Corollary 4.2, $p + \omega$ is absolutely continuous.

If $x: \mathcal{T} \rightarrow R^n$, we write $x \leq 0$ if $x_i(t) \leq 0$ for every t and i . Similarly, x is nondecreasing if $x(t) - x(s) \leq 0$ for all $t \leq s$. Recall that spaces like \mathcal{C} and \mathcal{A} consist of functions $f: \mathcal{T} \rightarrow R^n$. To denote the corresponding space of functions $f: \mathcal{T} \rightarrow R^s$, we attach the subscript s to the space.

LEMMA 5.2. *Suppose that $K: \mathcal{A} \rightarrow \mathcal{C}_s$ is convex and define the set*

$$\mathcal{S} = \{x \in \mathcal{A} : K(x) \leq 0\}.$$

If there exists $\bar{x} \in \mathcal{A}$ such that $K(\bar{x})_i(t) < 0$ for every t and i , then for each $\gamma \in \mathcal{A}^$, there is a nondecreasing $\nu \in \mathcal{B}_s$ such that*

$$(5.3) \quad \langle \nu, K(x) \rangle_{\mathcal{C}} \geq \inf \{ \langle \gamma, x - y \rangle_{\mathcal{A}} : y \in \mathcal{S} \}$$

for all $x \in \mathcal{A}$.

Combining Theorem 5.1 and Lemma 5.2,

$$(5.4) \quad C(z) + \langle p, M(z) \rangle + \langle \nu, K(x) \rangle_{\mathcal{C}} \geq L(p)$$

for all $z = (x, u) \in \mathcal{Z}$. To prove the lemma, we apply Theorem A.3 from Appendix 1 to the problem

$$\begin{aligned} & \text{maximize} && \langle \gamma, y \rangle_{\mathcal{A}} \\ & \text{subject to} && K(y) \leq 0, \quad y \in \mathcal{A}. \end{aligned}$$

Hence, there exists $\nu \in \mathcal{B}_s$ satisfying (5.3) and

$$(5.5) \quad \langle \nu, f \rangle_{\mathcal{C}} \geq 0$$

for all nonnegative $f \in \mathcal{C}_s$. If λ is the regular Borel measure corresponding to ν , (5.5) implies that λ is positive or ν is nondecreasing. \square

In the last lemma, we generated ν for given γ . Now, let us produce γ for given ν .

PROPOSITION 5.3. *Assume that $K: \mathcal{A} \rightarrow \mathcal{C}_s$ is convex and differentiable, C is convex, $\nu \in \mathcal{B}_s$ is nondecreasing, and (5.4) holds for some $p \in \text{dom } L$. If $\hat{z} = (\hat{x}, \hat{u}) \in \mathcal{Z}$ has the property that $C(\hat{z}) + \langle p, M(\hat{z}) \rangle = L(p)$ and*

$$\hat{x} \in \mathcal{S} := \{x \in \mathcal{A} : K(x) \leq 0\},$$

then the $\gamma \in \mathcal{A}^$ defined by*

$$\langle \gamma, y \rangle_{\mathcal{A}} = \langle \nu, K'[\hat{x}]y \rangle_{\mathcal{C}} \quad \forall y \in \mathcal{A}$$

satisfies (5.2).

Proof. Under our hypotheses, the functional $f(\cdot) := \langle \nu, K(\cdot) \rangle_{\mathcal{C}}$ is differentiable on \mathcal{A} and

$$f'[x](y) = \langle \nu, K'[x]y \rangle_{\mathcal{C}}.$$

Since $\hat{x} \in \mathcal{S}$ and ν is nondecreasing, it also follows that $f(\hat{x}) \leq 0$. The inequality (5.4) and the relations $C(\hat{z}) + \langle p, M(\hat{z}) \rangle = L(p)$ and $f(\hat{x}) \leq 0$ imply that $f(\hat{x}) = 0$ and \hat{z} minimizes $C(z) + \langle p, M(z) \rangle + f(x)$ over $z = (x, u) \in \mathcal{Z}$. Applying Lions' characterization [27, p. 12] for the minimizer of the sum of convex and differentiable functions gives us:

$$(5.6) \quad C(z) + \langle p, M(z) \rangle + f'[x](x - \hat{x}) \geq L(p)$$

for all $z = (x, u) \in \mathcal{Z}$. Since ν is nondecreasing, f is convex and we have the standard inequality [31, p. 84]:

$$(5.7) \quad f(y) \geq f(x) + f'[x](y - x)$$

for all $x, y \in \mathcal{A}$. Inserting $x = \hat{x}$ and recalling that $f(\hat{x}) = 0$, (5.7) yields

$$f'[\hat{x}](y - \hat{x}) \leq 0 \quad \forall y \in \mathcal{S}.$$

Relation (5.6) completes the proof. \square

In some cases, the γ produced by Proposition 5.3 can be described more precisely. Suppose that $G(t)$ is an $s \times n$ matrix for each $t \in \mathcal{T}$, and define $Gx: \mathcal{T} \rightarrow R^s$ by

$$(Gx)(t) = G(t)x(t)$$

where $x: \mathcal{T} \rightarrow R^n$.

LEMMA 5.4. *If the elements of G are absolutely continuous and $\nu \in \mathcal{B}_s$, then*

$$\langle \nu, Gx \rangle_{\mathcal{C}} = \langle G^T \nu, x \rangle_{\mathcal{C}} - \langle G'x, \nu \rangle$$

for every $x \in \mathcal{A}$. Hence, for suitable b and $c \in R^n$, we have

$$\langle \nu, Gx \rangle_{\mathcal{C}} = c \cdot x(0) + \langle \omega, x' \rangle \quad \forall x \in \mathcal{A}$$

where

$$\omega(t) = b - G(t)^T \nu(t) + \int_0^t G'(\sigma)^T \nu(\sigma) d\sigma.$$

These identities are left for exercises. Given more information about the constraints, feasible dual functions can be described more precisely. For example, if the states are unconstrained at $t = 1$, then $p(1) + \omega(1) = 0$ under the hypotheses of Theorem 5.1. Of course, a dual solution may be smoother than a typical feasible element. In an earlier paper [15] we show that when the cost is strictly convex in the control and constraints are smooth enough, there exist optimal Lipschitz continuous functions p , ω , ν , x and u . Moreover, x and $p + \omega$ have Lipschitz continuous derivatives. (To be more precise, ν is only Lipschitz continuous on the open interval $(0, 1)$.)

The analysis of primal and dual solutions is different from the arguments in §§ 3–5. In [15] we start with the control minimum principal and adjoint equation, and use the implicit function theorem to estimate $|\nu(t_1) - \nu(t_2)|$ and $|u(t_1) - u(t_2)|$ in terms of smoother variables, x and $p + \omega$. Greater smoothness for ν and u implies better regularity for x and $p + \omega$. Malanowski [30] extends these results to problems with nonlinear system dynamics. Returning to the question concerning the relation between p and the costate, we show in [19] that $p + \omega$ corresponds to the usual costate.

6. Interiority. Suppose that

$$\mathcal{S} = \{x \in \mathcal{A}: x(t) \in X(t) \forall t \in \mathcal{T}\}$$

where $X(t) \subset \mathbb{R}^n$ for each $t \in \mathcal{T}$. A map such as X from \mathcal{T} to subsets of another space is called a *multifunction* [51]. If $X(t)$ is convex for every $t \in \mathcal{T}$, we say that X is *convex-valued*. In this section, properties of \mathcal{S} are studied under the

POINTWISE INTERIORITY ASSUMPTION. *The interior of $X(t)$ is nonempty for every $t \in \mathcal{T}$ and the set*

$$\dot{X} := \{(t, x): t \in \mathcal{T}, x \in \text{int } X(t)\} \subset \mathcal{T} \times \mathbb{R}^n$$

is open.

Above, “int” denotes interior.

Rockafellar [45, Lemma 2] shows that X is lower semicontinuous when X is convex-valued and pointwise interiority holds, and in proving [45, Thm. 5], it is seen that \mathcal{S} has nonempty interior. Rockafellar’s development utilizes a continuous selection theorem of Michael [34, Thm. 3.2]. The fact that \mathcal{S} has nonempty interior is also deduced from an appropriate partition of unity, as we now demonstrate.

The *support* of a function $f: \mathcal{T} \rightarrow \mathbb{R}^n$ is defined by

$$\text{supp } f = \text{closure } \{t \in \mathcal{T}: f(t) \neq 0\}.$$

Given a collection \mathcal{O} of open sets whose union is \mathcal{T} , there exists [1, p. 51] a finite set $\Psi \subset \mathcal{C}_1^\infty$ of nonnegative functions such that

$$\sum_{\psi \in \Psi} \psi(t) = 1 \quad \forall t \in \mathcal{T},$$

and for every $\psi \in \Psi$,

$$\text{supp } \psi \subset U$$

for some $U \in \mathcal{O}$. The set Ψ is called an *infinitely differentiable partition of unity subordinate to \mathcal{O}* . Defining the set

$$\dot{\mathcal{S}} = \{x \in \mathcal{C}^\infty: (t, x(t)) \in \dot{X} \ \forall t \in \mathcal{T}\},$$

we have:

LEMMA 6.1. *If X is convex-valued and pointwise interiority holds, then $\dot{\mathcal{S}}$ is nonempty.*

Proof. Given $f: \mathcal{T} \rightarrow R^n$, define

$$\mathcal{O}_f = \{t \in \mathcal{T}: f(t) \in \text{int } X(t)\}.$$

By pointwise interiority, \mathcal{O}_f is open when f is continuous. If $\mathcal{F} \subset \mathcal{C}^\infty$ is the collection of constant functions, then

$$\mathcal{T} = \bigcup_{f \in \mathcal{F}} \mathcal{O}_f.$$

Let Ψ be a partition of unity subordinate to $\{\mathcal{O}_f: f \in \mathcal{F}\}$. For each $\psi \in \Psi$, there exists $f(\psi) \in \mathcal{F}$ such that

$$\text{supp } \psi \subset \mathcal{O}_{f(\psi)}.$$

Observe that $x \in \mathcal{C}^\infty$ given by

$$x(t) = \sum_{\psi \in \Psi} f(\psi)\psi(t)$$

is a convex combination of points in the interior of $X(t)$ for every $t \in \mathcal{T}$. \square

LEMMA 6.2. *If pointwise interiority holds, $x \in \mathcal{C}$, and $x(t) \in \text{int } X(t)$ for each $t \in \mathcal{T}$, then there exists $\rho > 0$ such that*

$$\{y \in R^n: |y - x(t)| \leq \rho\} \subset X(t)$$

for every $t \in \mathcal{T}$.

Proof. Since \dot{X}^c , the complement of \dot{X} , and $\{(t, x(t)): t \in \mathcal{T}\}$ are disjoint closed sets, the distance between them is positive. \square

Lemmas 6.1 and 6.2 and the inequality

$$\|x\|_{\mathcal{C}} \leq \|x\|_{\mathcal{A}} \quad \forall x \in \mathcal{A}$$

imply that \mathcal{S} has nonempty interior when X is convex-valued and pointwise interiority holds. Defining the set

$$\mathcal{S}^\infty = \{x \in \mathcal{L}^\infty: x(t) \in X(t) \text{ almost everywhere}\},$$

we have:

THEOREM 6.3. *If pointwise interiority holds and X is convex-valued, then for each $x \in \mathcal{S}^\infty$, there exists a sequence $\{x_k\} \subset \dot{\mathcal{S}}$ converging pointwise to x . Moreover, for any finite set $\{(t_j, a_j)\} \subset \dot{X}$ where the t_j are distinct, it can be arranged so that $x_k(t_j) = a_j$ for every j and k . (Pointwise convergence is defined in § 4.)*

Proof. Given $x \in \mathcal{L}^\infty$ and $\varepsilon > 0$, we exhibit $w \in \dot{\mathcal{S}}$ such that

$$(6.1) \quad \mu\{t \in \mathcal{T}: |w(t) - x(t)| > \varepsilon\} \leq \varepsilon$$

where μ is Lebesgue measure and $\|w\|_{\mathcal{C}}$ is bounded independently of ε . To simplify notation, let x also denote a particular element in its equivalence class for which

$$x(t) \in X(t) \quad \forall t \in \mathcal{T}$$

and $\|x\|_{\mathcal{C}}$ is finite. By Lusin's theorem and regularity properties of Borel measure [53, Thms. 2.23 and 2.17], there exist $y \in \mathcal{C}$ and a closed set $K \subset \mathcal{T}$ such that $\mu(K^c) \leq \varepsilon$, $\|y\|_{\mathcal{C}} \leq \|x\|_{\mathcal{C}}$, and

$$x(t) = y(t) \quad \forall t \in K.$$

Recalling Lemmas 6.1 and 6.2 and the fact that \mathcal{C}^∞ is a dense subset of \mathcal{C} , there is $z \in \mathcal{C}^\infty$ such that

$$\|y - z\|_{\mathcal{C}} \leq \varepsilon,$$

and $z(t) \in \text{int } X(t)$ for every $t \in K$. By pointwise interiority, the set

$$\mathcal{O} = \{t \in \mathcal{T} : z(t) \in \text{int } X(t)\}$$

is open. Let $\{\psi_1, \psi_2\}$ be a partition of unity subordinate to $\{\mathcal{O}, K^c\}$ and define

$$w(t) = \psi_1(t)z(t) + \psi_2(t)\dot{x}(t)$$

where $\dot{x} \in \dot{\mathcal{S}}$. Since $z(t) \in \text{int } X(t)$ on $\text{supp } \psi_1$, and $\psi_1 + \psi_2$ is identically 1, it follows that $w \in \dot{\mathcal{S}}$. Since $\psi_1 = 1$ on K , (6.1) is established.

Next, given $(\sigma, a) \in \dot{X}$, let $\mathcal{O} \subset \mathcal{T}$ be an open interval containing σ such that $\mu(\mathcal{O}) \leq \varepsilon$ and

$$\mathcal{O} \times \{a\} \subset \dot{X}.$$

Letting $\{\phi_1, \phi_2\}$ be a partition of unity subordinate to $\{\mathcal{O}, \{\sigma\}^c\}$, define

$$v(t) = a\phi_1(t) + w(t)\phi_2(t).$$

Observe that $v \in \dot{\mathcal{S}}$, $v(\sigma) = a$, and $v(t) = w(t)$ except on a set of measure $\leq \varepsilon$. The second part of the theorem follows almost immediately. \square

7. Pointwise minimization. The next section provides a convenient representation for the dual functional when the cost and the constraints assume a special form. Here we review some theorems on measurability, drawing on Rockafellar's work [51], and develop preliminary results. Let us consider the following problem:

$$\inf \{I(x) : x \in \mathcal{L}^\infty\}$$

where $I: \mathcal{L}^\infty \rightarrow \bar{\mathcal{R}}$ is defined by

$$I(x) = \int_{\mathcal{T}} f(x(t), t) dt$$

for some $f: R^n \times \mathcal{T} \rightarrow \bar{\mathcal{R}}$. We assume that I is proper, and the integrand is measurable and majorizes a summable function whenever $x \in \mathcal{L}^\infty$.

Classically, $f(x(\cdot), \cdot)$ is measurable when $x(\cdot)$ is measurable if the *Carathéodory conditions* hold; that is, $f(\cdot, t)$ is continuous for each fixed $t \in \mathcal{T}$ and $f(x, \cdot)$ is measurable for each fixed $x \in R^n$. On the other hand, we may wish to embed constraints in the cost functional. For example, the constraint

$$x(t) \in X(t)$$

almost everywhere can be incorporated in the cost through the definition

$$f(x, t) = \infty \quad \text{if } x \notin X(t).$$

The *normal integrand*, introduced by Rockafellar [51], is a natural one-sided extension of the Carathéodory integrand. The integrand $f: R^n \times \mathcal{T} \rightarrow \bar{\mathcal{R}}$ is normal if $f(x, t)$ is

lower semicontinuous in x for each fixed $t \in \mathcal{T}$, and f is measurable on $R^n \times \mathcal{T}$ with respect to the σ -algebra generated by products of Borel sets in R^n and Lebesgue sets in \mathcal{T} . Therefore, it follows that $f(x(\cdot), \cdot)$ is measurable whenever $x(\cdot)$ is measurable. An important property of normal integrands is contained in the following lemma, an immediate consequence of [51, Thm. 2K]:

LEMMA 7.1. *If f is a normal integrand on $R^n \times \mathcal{T}$, then*

$$\inf \{I(x): x \in \mathcal{L}^\infty\} = \int_{\mathcal{T}} \inf \{f(x, t): x \in R^n\} dt$$

and the integrand above is measurable. Furthermore, there exists a measurable function $x: \mathcal{T} \rightarrow R^n$ such that

$$x(t) \in \arg \min \{f(x, t): x \in R^n\}$$

wherever the minimum is attained.

As noted earlier, constraints can be embedded in the cost. Given $X: \mathcal{T} \rightarrow 2^{R^n}$, let us define

$$\bar{f}(x, t) = \begin{cases} f(x, t) & \text{if } x \in X(t), \\ \infty & \text{otherwise.} \end{cases}$$

By [51, Props., 2H and 2L], \bar{f} is normal provided f is normal and X is *closed-valued* and *measurable*; that is, $X(t)$ is closed for each $t \in \mathcal{T}$ and for all closed sets $K \subset R^n$,

$$\{t \in \mathcal{T}: X(t) \cap K \text{ is nonempty}\}$$

is measurable. If X is closed-valued and convex-valued, then X is measurable under the pointwise interiority hypothesis. This follows from Theorem 6.3 and Castaing's characterization of a closed-valued measurable multifunction in terms of the closure of a countable collection of measurable functions [7], [51].

Now consider the problem

$$\hat{C} = \inf \{E(x) + I(x): x \in \mathcal{C}^\infty\}$$

where $E: \mathcal{C}^\infty \rightarrow \bar{R}$ and for some finite set $\Omega \subset \mathcal{T}$,

$$E(x) = E(y)$$

whenever $x, y \in \mathcal{C}^\infty$ and $x(t) = y(t)$ for each $t \in \Omega$. For example, $E(x)$ might be expressed in terms of $x(1)$. Let us define

$$X(t) = \{x \in R^n: f(x, t) < \infty\},$$

and let \mathcal{J} and \mathcal{J}^∞ be the sets defined in § 6.

LEMMA 7.2. *Suppose that X is convex-valued, pointwise interiority holds,*

$$(7.1) \quad \inf \{E(x): x \in \mathcal{J}\} = \inf \{E(x): x \in \mathcal{J}^\infty \cap \mathcal{C}^\infty\},$$

and

$$(7.2) \quad I(x) = \lim_k I(x^k)$$

for each sequence $\{x^k\} \subset \mathcal{J}^\infty$ converging pointwise to some $x \in \text{dom } I$. Then

$$\hat{C} = \inf \{E(x): x \in \mathcal{J}\} + \inf \{I(x): x \in \mathcal{L}^\infty\}.$$

Proof. Since $I(x) = \infty$ if $x \notin \mathcal{J}^\infty$,

$$\hat{C} = \inf \{E(x) + I(x): x \in \mathcal{J}^\infty \cap \mathcal{C}^\infty\}.$$

Given $x \in \mathcal{S}^\infty$ and $y \in \dot{\mathcal{S}}$, Theorem 6.3 provides a sequence $\{x^k\} \subset \dot{\mathcal{S}}$ converging pointwise to x and

$$x^k(t) = y(t) \quad \forall t \in \Omega.$$

If $x \in \text{dom } I$,

$$(7.3) \quad E(y) + I(x) = \lim_k \{E(x^k) + I(x^k)\} \geq \inf \{E(x) + I(x): x \in \dot{\mathcal{S}}\}.$$

Combining (7.1) and (7.3),

$$\begin{aligned} \inf \{E(y): y \in \mathcal{S}^\infty \cap \mathcal{C}^\infty\} + \inf \{I(x): x \in \mathcal{S}^\infty\} \\ &= \inf \{E(y): y \in \dot{\mathcal{S}}\} + \inf \{I(x): x \in \mathcal{S}^\infty\} \\ &\geq \inf \{E(x) + I(x): x \in \dot{\mathcal{S}}\} \\ &\geq \inf \{E(x) + I(x): x \in \mathcal{S}^\infty \cap \mathcal{C}^\infty\} = \hat{C}. \end{aligned}$$

Since the reverse inequalities are trivial, the proof is complete. \square

If $E(y) = e(y(1))$ where $e: R^n \rightarrow \bar{R}$, then (7.1) is satisfied if $\text{dom } e \subset \text{int } X(1)$. Moreover, under these hypotheses,

$$\inf \{E(x): x \in \dot{\mathcal{S}}\} = \inf \{e(a): a \in R^n\}.$$

Relation (7.2) holds if $f(\cdot, t)$ is continuous on $X(t)$ and if for each $\rho > 0$ there is a summable function $g: \mathcal{T} \rightarrow R$ such that

$$g(t) \geq |f(x, t)|$$

whenever $x \in X(t)$ and $|x| \leq \rho$. If f is a normal integrand on $R^n \times \mathcal{T}$, Lemma 7.1 gives us

$$\inf \{I(x): x \in \mathcal{L}^\infty\} = \int_{\mathcal{T}} \inf \{f(x, t): x \in R^n\} dt.$$

8. Dual formulations. Let us evaluate the dual functional when the primal cost has the form

$$C(x, u) = e(x(0), x(1)) + \int_{\mathcal{T}} f(x(t), u(t), t) dt$$

where $e: R^{2n} \rightarrow \bar{R}$ and $f: R^{n+m} \times \mathcal{T} \rightarrow \bar{R}$ is a normal integrand which majorizes a summable function whenever x is essentially bounded and u is summable, and the integral is finite for some $(x, u) \in \mathcal{L}_n^\infty \times \mathcal{L}_m^\infty$. We define

$$H(a, q, z) = e(x(0), x(1)) - a_0 \cdot x(0) - a_1 \cdot x(1) + \int_{\mathcal{T}} [f(z(t), t) - q(t) \cdot z(t)] dt$$

where $z = (x, u) \in \mathcal{L}_n^\infty \times \mathcal{L}_m^1$, $q \in \mathcal{L}_n^1 \times \mathcal{L}_m^\infty$, and $a = (a_0, a_1) \in R^n \times R^n$. Corresponding to e and f , we have the conjugate functions $e^*(a) = \inf \{e(b) - a \cdot b: b \in R^{n+m}\}$ and

$$f^*(q) = \int_{\mathcal{T}} \inf \{f(z, t) - q(t) \cdot z: z \in R^{n+m}\} dt.$$

The integrand of f^* is measurable by Lemma 7.1 and the fact that the sum of normal and Carathéodory integrands is normal [51, Prop. 2M]. By these definitions, the following inequalities are clearly satisfied:

$$e^*(a) + f^*(q) \leq \inf \{H(a, q, z): z \in \mathcal{Z}\} \leq \inf \{H(a, q, x, u): x \in \mathcal{C}^\infty, u \in \mathcal{L}_m^\infty\}.$$

Now set $E(x) = e(x(0), x(1)) - a_0 \cdot x(0) - a_1 \cdot x(1)$ and for fixed $u \in \mathcal{L}_m^\infty$ define

$$I(x) = \int_{\mathcal{T}} f(x(t), u(t), t) dt.$$

We assume that for each fixed $u \in \mathcal{L}_m^\infty$ where the domain of I in \mathcal{L}^∞ is nonempty, there exists an element of u 's equivalence class such that the hypotheses of Lemma 7.2 are satisfied. Referring to the discussion after Lemma 7.2, it is also assumed that for this element of u 's equivalence class, we have the identity $\inf \{E(x): x \in \mathcal{J}\} = e^*(a)$. Then Lemmas 7.1 and 7.2 give us:

$$\begin{aligned} \inf \{H(a, q, x, u): x \in \mathcal{C}^\infty, u \in \mathcal{L}_m^\infty\} &= e^*(a) + \inf \{H(a, q, x, u): x \in \mathcal{L}_n^\infty, u \in \mathcal{L}_m^\infty\} \\ &= e^*(a) + f^*(q). \end{aligned}$$

Combining these relations, it follows that

$$e^*(a) + f^*(q) = \inf \{H(a, q, z): z \in \mathcal{Z}\}.$$

We say that (P) has a *pointwise representation* if this equality holds for every $a \in R^{2n}$ and $q \in \mathcal{L}_n^1 \times \mathcal{L}_m^\infty$.

LEMMA 8.1. *If (P) has a pointwise representation, then for all $p \in \mathcal{A}$,*

$$(8.1) \quad L(p) = e^*(a) + f^*(q)$$

where

$$(8.2) \quad q(t) = \begin{pmatrix} p'(t) + A(t)^T p(t) \\ B(t)^T p(t) \end{pmatrix}$$

and

$$a = \begin{pmatrix} p(0) \\ -p(1) \end{pmatrix}.$$

Proof. Starting with the definition of L and integrating by parts,

$$L(p) = \inf \{H(a, q, z): z \in \mathcal{Z}\}$$

where a and q are given above. The conclusion follows immediately. \square

Remark 1. Rockafellar [47] uses (8.1) to define $L(p)$ when p is absolutely continuous. Since the dual solution may be discontinuous, he shows that the dual function can be extended to the space of functions with bounded variation.

We now examine four problems which will be solved in § 11.

Problem I.

$$\text{minimize } \frac{1}{2} \int_0^1 [x(t)^2 + u(t)^2] dt$$

subject to

$$x'(t) = u(t), \quad u(t) \leq a \quad \text{almost everywhere,}$$

$$x(0) = c, \quad (x, u) \in \mathcal{Z}.$$

Here a and c are given scalars. Defining the functions $f: R^2 \times \mathcal{T} \rightarrow \bar{R}$ and $e: R^2 \rightarrow \bar{R}$ by

$$f(x, u, t) = \begin{cases} \frac{1}{2}(x^2 + u^2) & \text{if } u \leq a, \\ \infty & \text{if } u > a \end{cases}$$

and

$$e(x, y) = \begin{cases} 0 & \text{if } x = c, \\ \infty & \text{if } x \neq c, \end{cases}$$

we can write Problem I as

$$\begin{aligned} &\text{minimize} \quad e(x(0), x(1)) + \int_{\mathcal{T}} f(x(t), u(t), t) dt \\ &\text{subject to} \quad x'(t) = u(t) \quad \text{almost everywhere.} \end{aligned}$$

If a_0 and a_1 are given scalars and u is a fixed element of \mathcal{L}^∞ , let us define

$$E(x) = e(x(0), x(1)) - a_0 \cdot x(0) - a_1 \cdot x(1),$$

and

$$I(x) = \int_{\mathcal{T}} f(x(t), u(t), t) dt.$$

If the domain of I is nonempty, then there exists an element of u 's equivalence class such that $X(t) = R$ for every $t \in \mathcal{T}$ where $X(t)$ is introduced in § 7. Hence (7.1) and the pointwise interiority assumption are satisfied trivially. Likewise, (7.2) holds since $f(\cdot, \cdot, t)$ is continuous on its effective domain. Finally, it is easy to check that $\inf \{E(x) : x \in \mathcal{S}\} = e^*(a)$. Therefore, by the discussion at the start of this section, Problem I has a pointwise representation, and by Lemma 8.1, the dual function is

$$L(p) = e^*(p(0), -p(1)) + f^*(p', p)$$

for every $p \in \mathcal{A}$. The conjugate functions e^* and f^* are easily evaluated:

$$\begin{aligned} e^*(x, y) &= \begin{cases} -cx & \text{if } y = 0, \\ -\infty & \text{if } y \neq 0, \end{cases} \\ f^*(p', p) &= - \int_{\mathcal{T}} l(p'(t), p(t), t) dt, \\ l(x, y, t) &= \begin{cases} \frac{1}{2}[x^2 + y^2] & \text{if } y \leq a, \\ \frac{1}{2}[x^2 + a(2y - a)] & \text{if } y > a. \end{cases} \end{aligned}$$

Although the dual problem is to maximize $L(p)$ over $p \in \mathcal{L}^\infty$, Theorems 3.1 and 4.1 tell us that we only need consider $p \in \mathcal{A}$. Since $e^*(x, y) = -\infty$ when $y \neq 0$, we can also impose the explicit dual constraint $p(1) = 0$. In summary, the dual of Problem I can be written

$$\begin{aligned} &\text{maximize} \quad - \left\{ cp(0) + \int_{\mathcal{T}} l(p'(t), p(t), t) dt \right\} \\ &\text{subject to} \quad p(1) = 0, \quad p \in \mathcal{A}. \end{aligned}$$

Next let us consider

Problem II.

$$\text{minimize} \quad \frac{1}{2} \int_0^1 [x(t)^2 + u(t)^2] dt$$

subject to

$$\begin{aligned} &x'(t) = u(t), \quad u(t) \leq a \quad \text{almost everywhere,} \\ &x(t) \leq b \quad \text{for all } t \in \mathcal{T}, \quad x(0) = c, \quad (x, u) \in \mathcal{X} \end{aligned}$$

where $c < b$. Again, defining the functions $f: R^2 \times \mathcal{T} \rightarrow \bar{R}$ and $e: R^2 \rightarrow \bar{R}$ by

$$f(x, u, t) = \begin{cases} \frac{1}{2}(x^2 + u^2) & \text{if } u \leq a \text{ and } x \leq b, \\ \infty & \text{if } u > a \text{ or } x > b \end{cases}$$

and

$$e(x, y) = \begin{cases} 0 & \text{if } x = c \text{ and } y \leq b, \\ \infty & \text{if } x \neq c \text{ or } y > b, \end{cases}$$

we can cast Problem II in the form

$$\text{minimize } e(x(0), x(1)) + \int_{\mathcal{T}} f(x(t), u(t), t) dt$$

$$\text{subject to } x'(t) = u(t) \text{ almost everywhere.}$$

Let us define E and I as we did for Problem I. If the domain of I is nonempty, then there exists an element of u 's equivalence class such that $X(t) = \{x \in R: x \leq b\}$ for every $t \in \mathcal{T}$. Since $\dot{X} = \{(t, x) \in \mathcal{T} \times R: x < b\}$ is an open subset of $\mathcal{T} \times R$, the pointwise interiority assumption holds. To verify (7.1), suppose that $x \in \mathcal{C}^\infty$, $x(0) = c$, and $x(t) \leq b$ for each $t \in \mathcal{T}$. Then the sequence $\{x_k\}$ defined by

$$x_k(t) = x(t) - \frac{t}{k}$$

lies in \mathcal{S} and $\lim_k E(x_k) = E(x)$. Hence (7.1) holds. Since $f(\cdot, \cdot, t)$ is continuous on its effective domain, (7.2) is satisfied. Again, it is easy to see that $\inf \{E(x): x \in \mathcal{S}\} = e^*(a)$. By the discussion at the start of this section, Problem II has a pointwise representation. Applying Lemma 8.1, the dual function can be expressed

$$L(p) = e^*(p(0), -p(1)) + f^*(p', p)$$

for each $p \in \mathcal{A}$ where

$$e^*(x, y) = \begin{cases} -(cx + by) & \text{if } y \geq 0, \\ -\infty & \text{if } y < 0, \end{cases}$$

$$f^*(p', p) = - \int_{\mathcal{T}} [l_x(p'(t), p(t), t) + l_u(p'(t), p(t), t)] dt,$$

$$l_x(x, y, t) = \begin{cases} \frac{1}{2}x^2 & \text{if } x \leq b, \\ \frac{1}{2}b(2x - b) & \text{if } x > b, \end{cases}$$

$$l_u(x, y, t) = \begin{cases} \frac{1}{2}y^2 & \text{if } y \leq a, \\ \frac{1}{2}a(2y - a) & \text{if } y > a. \end{cases}$$

Although the dual maximization is over $p \in \mathcal{L}^\infty$, it follows from our regularity analysis [15] that there exists a Lipschitz continuous dual solution to Problem II. Consequently, the dual problem reduces to

$$\text{maximize } -\{cp(0) - bp(1) + \int_{\mathcal{T}} [l_x(p'(t), p(t), t) + l_u(p'(t), p(t), t)] dt\}$$

$$\text{subject to } p(1) \leq 0 \quad p \in \mathcal{A}.$$

The derivation of the dual for the final two examples is similar to Problems I and II so we just summarize the conclusions. The primal version of the next problem is found in [24] and [33].

Problem III.

$$\text{minimize } \int_0^1 [x_1(t)^2 + x_2(t)^2 + .005u(t)^2] dt$$

$$\text{subject to } \begin{aligned} x_1'(t) &= x_2(t), & x_2'(t) &= -x_2(t) + u(t) \quad \text{almost everywhere,} \\ x_1(0) &= 0, & x_2(0) &= -1, & (x_1, x_2, u) &\in \mathcal{L}. \end{aligned}$$

In addition, two different state constraints are considered:

$$\text{Case A.} \quad x_2(t) \leq \alpha(t) \quad \text{for all } t \in \mathcal{T},$$

$$\text{Case B.} \quad x_1(t) \leq \alpha(t) \quad \text{for all } t \in \mathcal{T}$$

where $\alpha(t) = 2(1-2t)^2 - \frac{1}{2}$ (see Figs. 1 and 2). In Case A the dual is

$$\text{maximize } \left\{ p_2(0) + \frac{3}{2}p_2(1) - \int_{\mathcal{T}} l(p'(t), p(t), t) dt \right\}$$

$$\text{subject to } p_1(1) = 0, \quad p_2(1) \leq 0, \quad p = (p_1, p_2) \in \mathcal{A}$$

where

$$l(w, x, y, z, t) = \begin{cases} 50z^2 + \frac{1}{4}(w^2 + \beta^2) & \text{if } \beta \leq 2\alpha(t), \\ 50z^2 + \frac{1}{4}w^2 + \alpha(t)(\beta - \alpha(t)) & \text{if } \beta > 2\alpha(t) \end{cases}$$

and $\beta = x + y - z$. In Case B the dual is

$$\text{maximize } \left\{ p_2(0) + \frac{3}{2}p_1(1) - \int_{\mathcal{T}} l(p'(t), p(t), t) dt \right\}$$

$$\text{subject to } p_1(1) \leq 0, \quad p_2(1) = 0, \quad p = (p_1, p_2) \in \mathcal{A}$$

where

$$l(w, x, y, z, t) = \begin{cases} 50z^2 + \frac{1}{4}(w^2 + \beta^2) & \text{if } w \leq 2\alpha(t), \\ 50z^2 + \frac{1}{4}\beta^2 + \alpha(t)(w - \alpha(t)) & \text{if } w > 2\alpha(t) \end{cases}$$

and $\beta = x + y - z$.

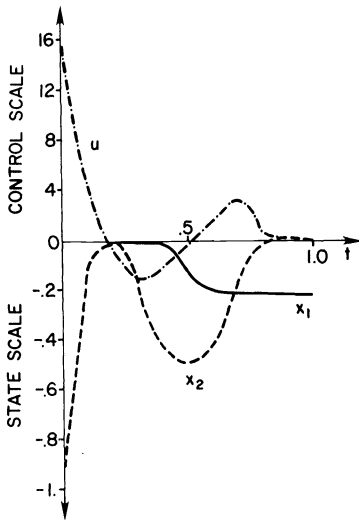


FIG. 1. Solution to Problem IIIA.

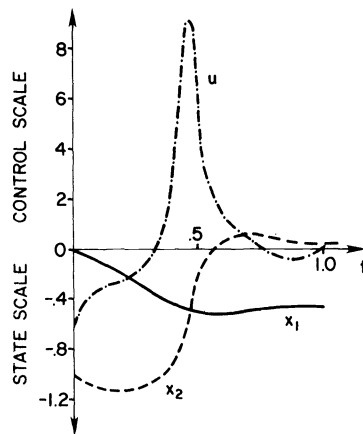


FIG. 2. Solution to Problem IIIB.

The primal version of the following problem is found in [56] (see Fig. 3).

Problem IV.

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} \left\{ x_2(1)^2 + \int_0^1 [x_1(t)^2 + u(t)^2] dt \right\} \\ & \text{subject to} \quad x_1'(t) = x_2(t), \quad x_2'(t) = u(t) \quad \text{almost everywhere,} \\ & \quad x_1(0) = -1, \quad x_2(0) = 0, \\ & \quad x_2(t) \leq \frac{1}{20} \quad \text{for all } t \in \mathcal{T}, \quad (x_1, x_2, u) \in \mathcal{X}. \end{aligned}$$

The dual is

$$\begin{aligned} & \text{maximize} \quad - \left\{ \phi(p_2(1)) - p_1(0) + \int_{\mathcal{T}} l(p'(t), p(t), t) dt \right\} \\ & \text{subject to} \quad p_1(1) = 0, \quad p_1(t) + p_2'(t) \geq 0 \quad \text{almost everywhere,} \\ & \quad p = (p_1, p_2) \in \mathcal{A} \end{aligned}$$

where

$$\begin{aligned} l(w, x, y, z, t) &= \frac{1}{2} [w^2 + z^2 + .05(x + y)], \\ \phi(x) &= \begin{cases} \frac{1}{2}x^2 & \text{if } x \geq -\frac{1}{20}, \\ -\frac{1}{20}(x + \frac{1}{40}) & \text{if } x < -\frac{1}{20}. \end{cases} \end{aligned}$$

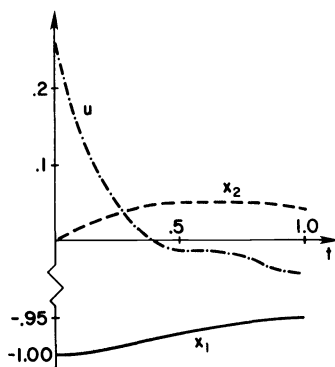


FIG. 3. Solution to Problem IV.

To conclude this section, let us examine the relations between solutions to the primal and the dual problems. If q is related to $p \in \mathcal{A}$ through (8.2), and $z: \mathcal{T} \rightarrow R^{n+m}$ is a measurable function such that

$$f(z(t), t) - q(t) \cdot z(t) = \min \{ f(z, t) - q(t) \cdot z : z \in R^{n+m} \}$$

almost everywhere, we say that (p, z) is a *min-pair*.

THEOREM 8.2. *If $p \in \mathcal{L}^\infty$, $z = (x, u)$ is feasible in (P), and $L(p) = C(z)$, then p is optimal in the dual problem and z is optimal in the primal problem. Moreover, if $p \in \mathcal{A}$ and (P) has a pointwise representation, then (p, z) is a min-pair and*

$$(8.3) \quad e^*(p(0), -p(1)) = e(x(0), x(1)) - \begin{pmatrix} p(0) \\ -p(1) \end{pmatrix} \cdot \begin{pmatrix} x(0) \\ x(1) \end{pmatrix}.$$

Proof. The first part of the theorem, the optimality of p and z , is a standard property of the dual functional. Let us consider the second half. Since $M(z) = 0$ and (P) has a pointwise representation, Lemma 8.1 gives us:

$$e^*(a) + f^*(q) = L(p) = C(z) = C(z) + \langle p, M(z) \rangle = H(a, q, z)$$

where the last equality comes from integrating by parts. Examining the definition of H , (p, z) in a min-pair and (8.3) is satisfied. \square

If z is optimal in the primal problem, p is optimal in the dual problem, and $L(p) = C(z)$, then we say that (p, z) is an *optimal pair*. Hence the preceding theorem states that an optimal pair (p, z) with $p \in \mathcal{A}$ is a min-pair when (P) has a pointwise representation. For p and $\omega \in \mathcal{A}$, let us define

$$q^\omega(t) = q(t) + \begin{pmatrix} \omega'(t) \\ 0 \end{pmatrix}$$

where q is given by (8.2).

THEOREM 8.3. *If (P) has a pointwise representation, p and $\omega \in \mathcal{A}$, and*

$$(8.4) \quad z = \arg \min \{C(\xi) + \langle p, M(\xi) \rangle + \langle (c, \omega), x \rangle_{\mathcal{A}} : \xi = (x, u) \in \mathcal{X}\}$$

for some $c \in R^n$, then

$$(8.5) \quad f(z(t), t) - q^\omega(t) \cdot z(t) = \inf \{f(\xi, t) - q^\omega(t) \cdot \xi : \xi \in R^{n+m}\}$$

almost everywhere.

Proof. Since (P) has a pointwise representation, we integrate by parts to get

$$\min \{C(\xi) + \langle p, M(\xi) \rangle + \langle (c, \omega), x \rangle_{\mathcal{A}} : \xi = (x, u) \in \mathcal{X}\} = e^*(a) + f^*(q^\omega)$$

for some $a \in R^{2n}$. Since z attains the minimum, (8.5) holds. \square

9. Fundamental inequalities. Observe that estimating the error in z_h , the approximation to a primal solution introduced in § 2, is essentially a parametric programming problem in the parameter $p \in \mathcal{L}^\infty$. Defining

$$\Omega(p) = \{z \in \mathcal{X} : L(p) = C(z) + \langle p, M(z) \rangle\},$$

we hope that $\Omega(p)$ approaches a primal solution as p approaches a dual solution. Fiacco and Hutzler [11] and Guddat [13] give good surveys of recent work on parametric programs. Exploiting the structure of the dual functional, we now obtain an estimate for the \mathcal{L}^2 error in z_h when the cost is strictly convex.

First, let us consider parametric programs in finite dimensions. We say that a functional $h: R^n \rightarrow \bar{R}$ is *uniformly convex* if h is convex and there is an $\alpha > 0$ with the following property: For each $x \in \text{dom } h$, there exists $z \in R^n$ such that

$$(9.1) \quad h(w + x) \geq h(x) + z \cdot w + \alpha |w|^2 \quad \forall w \in R^n$$

and

$$(9.2) \quad \lim_{s \downarrow 0} \frac{h(x + sw) - h(x)}{s} = z \cdot w$$

whenever $(x + w) \in \text{dom } h$. The scalar α is called the *modulus of convexity*, and we let $h'(x)$ denote any z satisfying the relations above. Suppose that $\{g_\lambda : \lambda \in \Lambda\}$ is a collection of lower semicontinuous proper functions where $g_\lambda : R^n \rightarrow \bar{R}$ is uniformly convex with

modulus of convexity α independent of $\lambda \in \Lambda$. Under these hypotheses, there is a unique $\xi(\lambda) \in R^n$ for which

$$g_\lambda(\xi(\lambda)) = \inf \{g_\lambda(\xi) : \xi \in R^n\}$$

whenever $\lambda \in \Lambda$.

LEMMA 9.1. *Assume that λ and $\mu \in \Lambda$. If $x := \xi(\lambda)$, then*

$$(9.3) \quad \alpha|x - y|^2 \leq g_\lambda(y) - g_\lambda(x)$$

for every $y \in R^n$. Conversely, if $y := \xi(\mu) \in \text{dom } g_\lambda$, then

$$(9.4) \quad g_\lambda(y) - g_\lambda(x) \leq \frac{1}{4\alpha} |g'_\mu(y) - g'_\lambda(y)|^2$$

for each $x \in \text{dom } g_\mu$.

Proof. Taking $w = y - x$, (9.1) implies that

$$(9.5) \quad g_\lambda(y) - g_\lambda(x) \geq g'_\lambda(x) \cdot w + \alpha|w|^2.$$

Let us assume that $y \in \text{dom } g_\lambda$ since (9.3) is trivial otherwise. Recalling that x minimizes $g_\lambda(\cdot)$, we have the standard inequality [28, p. 178]:

$$g'_\lambda(x) \cdot (y - x) \geq 0.$$

Hence (9.3) follows from (9.5).

Now consider (9.4). Since y minimizes $g_\mu(\cdot)$ and $x \in \text{dom } g_\mu$, we also have the relation

$$g'_\mu(y)(x - y) \geq 0,$$

or equivalently,

$$(9.6) \quad g'_\lambda(y) \cdot w \leq (g'_\lambda(y) - g'_\mu(y)) \cdot w$$

where $w = y - x$. Interchanging y and x in (9.5) and combining with (9.6) gives us

$$g_\lambda(y) - g_\lambda(x) \leq g'_\lambda(y) \cdot w - \alpha|w|^2 \leq (g'_\lambda(y) - g'_\mu(y)) \cdot w - \alpha|w|^2.$$

Finally, utilizing the inequality

$$a \cdot b \leq \frac{1}{4\alpha} |a|^2 + \alpha|b|^2,$$

we get (9.4). \square

Let us return to the cost functional defined at the start of § 8, and impose the following condition on the integrand:

UNIFORM CONVEXITY ASSUMPTION. *For each $t \in \mathcal{T}$, $f(\cdot, t)$ is uniformly convex with modulus of convexity α independent of t .*

We define a function $g : R^{2(n+m)} \times \mathcal{T} \rightarrow \bar{R}$ by the rule

$$(9.7) \quad g(\xi, \lambda, t) = f(\xi, t) - \lambda \cdot \xi.$$

Lemma 7.1 and the uniform convexity hypothesis imply that for each $q \in \mathcal{L}_{n+m}^1$, there is a measurable function $z : \mathcal{T} \rightarrow R^{n+m}$ such that

$$g(z(t), q(t), t) = \min \{g(z, q(t), t) : z \in R^{n+m}\}$$

almost everywhere. If $\|\cdot\|$ denotes the \mathcal{L}^2 norm defined by

$$\|z\| = \langle z, z \rangle^{1/2},$$

we have:

THEOREM 9.2. *Suppose that (P) has a pointwise representation, the uniform convexity hypothesis is satisfied, and (p, z) is an optimal pair where $p \in \mathcal{A}$. Then*

$$\alpha \|z - z_h\|^2 \leq L(p) - L(p_h)$$

for all min-pairs (p_h, z_h) .

Proof. Since (P) has a pointwise representation, Lemma 8.1 yields

$$L(p) = e^*(a) + f^*(q) \quad \text{and} \quad L(p_h) = e^*(a_h) + f^*(q_h).$$

Holding t fixed, we apply Lemma 9.1 to g from (9.7) taking $\lambda = q_h(t)$ and $\mu = q(t)$. Integrating (9.3) over \mathcal{T} and utilizing Theorem 8.2,

$$\begin{aligned} \alpha \|z - z_h\|^2 &\leq \int_{\mathcal{T}} [g(z(t), q_h(t), t) - g(z_h(t), q_h(t), t)] dt \\ &= \int_{\mathcal{T}} [g(z(t), q(t), t) - g(z_h(t), q_h(t), t)] dt + \int_{\mathcal{T}} z(t) \cdot (q(t) - q_h(t)) dt \\ &= f^*(q) - f^*(q_h) + \langle z, q - q_h \rangle. \end{aligned}$$

Integrating the last term by parts,

$$\langle z, q - q_h \rangle = x(0) \cdot (p_h(0) - p(0)) - x(1) \cdot (p_h(1) - p(1))$$

since $M(z) = 0$. By Theorem 8.2,

$$\langle z, q - q_h \rangle \leq e^*(a) - e^*(a_h).$$

Combining these relations, the proof is complete. \square

THEOREM 9.3. *Under the hypotheses of Theorem 9.2, we have:*

$$L(p) - L(p_I) \leq \frac{1}{4\alpha} \|q - q_I\|^2$$

for all $p_I \in \mathcal{A}$ which agree with $p(t)$ at $t=0$ and 1 where q is given by (8.2) and

$$(9.8) \quad q_I(t) = \begin{pmatrix} p'_I(t) + A(t)^T p_I(t) \\ B(t)^T p_I(t) \end{pmatrix}.$$

Proof. As in the last theorem's proof, we hold t fixed and apply Lemma 9.1 to $g(\cdot, \cdot, t)$ taking $\lambda = q_I(t)$ and $\mu = q(t)$. Integrating (9.4) over \mathcal{T} and utilizing Theorem 8.2,

$$\begin{aligned} \frac{1}{4\alpha} \|q - q_I\|^2 &\geq \int_{\mathcal{T}} [g(z(t), q_I(t), t) - g(z(t), q(t), t)] dt \\ &= f^*(q) - f^*(q_I) + \langle z, q - q_I \rangle = f^*(q) - f^*(q_I) = L(p) - L(p_I). \end{aligned}$$

The last step comes from Lemma 8.1 and the fact that $p_I = p$ at the ends of \mathcal{T} . The preceding step utilizes the relation $\langle z, q - q_I \rangle = 0$, which is deduced from the identity $M(z) = 0$. \square

Unfortunately, this upper bound from $L(p) - L(p_I)$ is too coarse for the error estimates in § 10. Since p' appears in the first component of q , and the derivative of

the optimal dual multiplier is often discontinuous for state constrained problems,

$$\|q - q_I\|_{\mathcal{C}} = O(1)$$

when p_I lies in typical piecewise polynomial spaces. Hence the upper bound is expressed in terms of the smoother variable q^ω introduced in § 8. Given $K: R^n \times \mathcal{T} \rightarrow R^s$ and $x: \mathcal{T} \rightarrow R^n$, let $K(x): \mathcal{T} \rightarrow R^s$ be defined by

$$K(x)(t) = K(x(t), t).$$

THEOREM 9.4. *Suppose that (P) has a pointwise representation, the uniform convexity hypothesis holds, K is twice continuously differentiable on $R^n \times \mathcal{T}$, and for each $t \in \mathcal{T}$, $K(\cdot, t)$ is convex and*

$$\text{dom } f(\cdot, t) \subset \{y \in R^n: K(y, t) \leq 0\} \times R^m.$$

If (p, z) is an optimal pair with $p \in \mathcal{A}$ and (8.4) holds for some $\omega \in \mathcal{A}$, then

$$(9.9) \quad L(p) - L(p_I) \leq \frac{1}{4\alpha} \|q^\omega - q_I^\omega\|^2 - \langle \nu_I, K(x) \rangle_{\mathcal{C}}$$

for all $p_I \in \mathcal{A}$ that agree with p at the ends of \mathcal{T} , and for all nondecreasing $\nu_I \in \mathcal{A}_s$ where

$$(9.10) \quad q_I^\omega(t) = q_I(t) - \begin{pmatrix} G(t)^T \nu_I'(t) \\ 0 \end{pmatrix},$$

$G(t) = \nabla_x K(x(t), t)$, and q_I is defined in (9.8).

Proof. By Theorem 8.3,

$$(9.11) \quad f(z(t), t) - q^\omega(t) \cdot z(t) = \inf \{f(\xi, t) - q^\omega(t) \cdot \xi: \xi \in R^{n+m}\}$$

almost everywhere. If (p, z_I) is a min-pair, we have the trivial relation

$$(9.12) \quad f(z_I(t), t) - q_I^\omega(t) \cdot z_I(t) \geq \inf \{f(\xi, t) - q_I^\omega(t) \cdot \xi: \xi \in R^{n+m}\}$$

almost everywhere. Lemma 9.1 with $\lambda = q_I^\omega(t)$ and $\mu = q^\omega(t)$ gives us

$$(9.13) \quad \begin{aligned} & \inf \{f(\xi, t) - q_I^\omega(t) \cdot \xi: \xi \in R^{n+m}\} - \inf \{f(\xi, t) - q^\omega(t) \cdot \xi: \xi \in R^{n+m}\} \\ & \geq (q^\omega(t) - q_I^\omega(t)) \cdot z(t) - \frac{1}{4\alpha} |q^\omega(t) - q_I^\omega(t)|^2 \\ & = (q(t) - q_I(t)) \cdot z(t) - \frac{1}{4\alpha} |q^\omega(t) - q_I^\omega(t)|^2 + (\omega'(t) - \omega_I'(t)) \cdot x(t) \end{aligned}$$

where $\omega_I'(t) := -G(t)^T \nu_I'(t)$. If $\varphi: R^n \rightarrow R$ is convex and differentiable and $\varphi(y) \leq 0$ for some $y \in R^n$, the convexity inequality

$$\varphi(y) \geq \varphi(x) + \varphi'[x](y - x)$$

implies that

$$(9.14) \quad \varphi(x) \leq \varphi'[x](x - y).$$

Subtracting (9.11) from (9.12), utilizing (9.13) and (9.14) and integrating over \mathcal{T} , we get

$$f^*(q_I) - f^*(q) \geq \langle q - q_I, z \rangle + \langle \nu_I, K(x) \rangle_{\mathcal{C}} - \frac{1}{4\alpha} \|q^\omega - q_I^\omega\|^2.$$

Integrating by parts, $\langle q - q_I, z \rangle = 0$ since $M(z) = 0$ and $p_I = p$ at the ends of \mathcal{T} . Finally, by Lemma 8.1,

$$L(p) - L(p_I) = f^*(q) - f^*(q_I).$$

Collecting results, the proof is complete. \square

If $p_h \in \mathcal{S}_h \subset \mathcal{A}$ and

$$L(p_h) = \text{maximum } \{L(p) : p \in \mathcal{S}_h\},$$

we have the trivial relation

$$L(\dot{p}) - L(p_h) \leq L(p) - L(p_I)$$

for all $p_I \in \mathcal{S}_h$ and $p \in \text{dom } L$. Therefore, if (p, z) is an optimal pair, and the hypotheses of Theorems 9.2 and 9.4 hold,

$$(9.15) \quad \alpha \|z - z_h\|^2 \leq L(p) - L(p_h) \leq \frac{1}{4\alpha} \|q^\omega - q_I^\omega\|^2 - \langle \nu_I, K(x) \rangle_{\mathcal{C}}$$

for all $p_I \in \mathcal{S}_h$ which agree with p at the ends of \mathcal{T} , and for all nondecreasing $\nu_I \in \mathcal{A}_s$. Moreover, if

$$q^\omega(t) = q(t) - \begin{pmatrix} G(t)^T \nu'(t) \\ 0 \end{pmatrix},$$

then

$$(9.16) \quad q^\omega(t) - q_I^\omega(t) = \begin{pmatrix} \delta q'(t) + G'(t)^T \delta \nu(t) + A(t)^T \delta p(t) \\ B(t)^T \delta p(t) \end{pmatrix}$$

where

$$\delta p(t) = p(t) - p_I(t), \quad \delta \nu(t) = \nu(t) - \nu_I(t), \quad \delta q(t) = \delta p(t) - G(t)^T \delta \nu(t).$$

10. Error estimates. We now estimate the error in piecewise polynomial approximation. Given an interval $J \subset R$, let $\mathcal{P}^k(J)$ be the space of polynomials defined on J with degree at most k . Associated with a collection of points from \mathcal{T} :

$$0 = t_0 < t_1 < \cdots < t_N = 1,$$

we have the spacing parameter

$$h = \text{maximum } \{t_j - t_{j-1} : j = 1, 2, \dots, N\},$$

and we let \mathcal{P}_h^k denote the n -fold Cartesian product of sets of functions $f : \mathcal{T} \rightarrow R$ whose restriction to each interval $J = (t_{j-1}, t_j)$ lies in $\mathcal{P}^k(J)$. The points $\{t_0, t_1, \dots, t_N\}$ are called the *mesh*.

For any interval $J \subset R$, we let $W^{0,\infty}(J)$ denote the set of essentially bounded functions $f : J \rightarrow R^n$, and for $k \geq 1$, $W^{k,\infty}(J) \subset W^{0,\infty}(J)$ is the subspace of functions with $k-1$ Lipschitz continuous derivatives. The space $W^{s,\infty}(\mathcal{T})$ is abbreviated $W^{s,\infty}$. The main results in this section are stated below:

THEOREM 10.1. *Suppose that (p, z) is an optimal pair, ν satisfies (5.4), and (9.15) holds. If $\mathcal{C} \cap \mathcal{P}_h^1 \subset \mathcal{S}_h$, we have*

$$\alpha \|z - z_h\|^2 \leq L(p) - L(p_h) = O(h^2)$$

provided the following conditions hold:

- (i) x and $(p - G^T \nu) \in W^{2,\infty}$, $G \in W^{2,\infty}$, $A \in \mathcal{L}^2$;
- (ii) $p \in W^{1,\infty}$, $\nu \in W^{1,\infty}$ is nondecreasing;
- (iii) $K(x) \in W^{2,\infty}$ and $K(x) \leq 0$.

Earlier [14] we observe that an optimal pair is often quite smooth except at points where constraints change between binding and nonbinding. Let $\tilde{W}^{k,\infty}$ denote the collection of functions $f \in W^{k-1,\infty}$ for which there is $M > 0$ and scalars $0 = s_0 < s_1 < \dots < s_M = 1$ such that the restriction of f to each interval (s_{j-1}, s_j) has $k-1$ Lipschitz continuous derivatives.

THEOREM 10.2. Suppose that (p, z) is an optimal pair, ν satisfies (5.4) and (9.15) holds. If $\mathcal{C} \cap \mathcal{P}_h^2 \subset \mathcal{S}_h$, we have: .

$$\alpha \|z - z_h\|^2 \leq L(p) - L(p_h) = O(h^3)$$

provided the following conditions hold:

- (i) x and $(p - G^T \nu) \in \tilde{W}^{3,\infty}$, $G \in \tilde{W}^{3,\infty}$, $A \in \mathcal{L}^\infty$;
- (ii) $p \in \tilde{W}^{2,\infty}$, $\nu \in \tilde{W}^{2,\infty}$ is nondecreasing;
- (iii) $K(x) \in W^{2,\infty}$ and $K(x) \leq 0$;
- (iv) the sets

$$T_j = \{t \in \mathcal{T} : K_j(x(t), t) < 0\}, \quad j = 1, 2, \dots, s,$$

are each composed of a finite number of intervals, and there exists $\beta > 0$ such that

$$\nu'_j(t) \geq \beta \quad \forall t \in T_j^c, \quad j = 1, 2, \dots, s.$$

These theorems are based on Lemmas 10.3, 10.4 and 10.5 appearing below. First, let us recall a result concerning polynomial interpolation. For any interval $J \subset \mathbb{R}$ and any $f \in W^{0,\infty}(J)$, let $|f|_J$ denote the essential supremum of $|f(t)|$ over $t \in J$. Then [8] and [55] exhibit various linear maps $I : W_1^{s,\infty}(J) \rightarrow \mathcal{P}^k(J)$ for which

$$(10.1) \quad \left| \frac{d^m}{dt^m}(f - f_I) \right|_J \leq c\mu(J)^{s-m} |f^{(s)}|_J$$

whenever $m \leq s \leq k+1$ and $f \in W_1^{s,\infty}(J)$ where $\mu(J)$ is the measure of J and c is a constant independent of f and J . (Remember that the subscript 1 on the space $W_1^{s,\infty}(J)$ means that the elements of the space map J to \mathbb{R}^1 .) Throughout this section, J is an interval and c denotes a generic constant. The operator I is usually called the *interpolation operator*, and we write f_I rather than $I(f)$; the function f_I is called the *interpolant* of f . For illustration, the following operator satisfies (10.1) when $s \geq 1$: Let f_I be the unique polynomial of degree at most k that agrees with f at $k+1$ evenly spaced points on J .

Suppose that $0 = t_0 < t_1 < \dots < t_N = 1$ is a mesh on \mathcal{T} and $f : \mathcal{T} \rightarrow \mathbb{R}$, and the restriction of f to each interval $J = [t_{j-1}, t_j]$ lies in $W_1^{s,\infty}(J)$. If I is an interpolation operator satisfying (10.1), we let $f_I : \mathcal{T} \rightarrow \mathbb{R}$ be the function composed of interpolants of f over each interval $J = [t_{j-1}, t_j]$. If $f_I \in W_1^{m,\infty}$, (10.1) implies that

$$(10.2) \quad \left| \frac{d^m}{dt^m}(f - f_I) \right|_{\mathcal{T}} \leq ch^{s-m} |f^{(s)}|_{\mathcal{T}}$$

whenever $m \leq s \leq k+1$ and $f \in W_1^{s,\infty}$. Finally, defining

$$\langle f, g \rangle_{\mathcal{C}(J)} = \int_J g(t) \cdot df(t)$$

for $g \in \mathcal{C}$ and $f \in \mathcal{B}$, we have:

LEMMA 10.3. Suppose that $J \subset \mathbb{R}$ is an interval, $f \in W_1^{1,\infty}(J)$, $g \in W_1^{2,\infty}(J)$, $g \leq 0$, and

$$(10.3) \quad f'(t)g(t) = 0 \quad \text{almost everywhere.}$$

If the interpolation operator I satisfies (10.1), then

$$\langle f_I, g \rangle_{\mathcal{C}(J)} \leq c\mu(J)^3 |f^{(1)}|_J |g^{(2)}|_J.$$

Moreover, if $f \in W_1^{2,\infty}(J)$ and $k \geq 1$,

$$\langle f_I, g \rangle_{\mathcal{C}(J)} \leq c\mu(J)^4 |f^{(2)}|_J |g^{(2)}|_J.$$

Proof. If $g(t) > 0$ almost everywhere, (10.3) implies that f is constant. Thus, $f = f_I$ by (10.1) and

$$\langle f_I, g \rangle_{\mathcal{C}(J)} = 0.$$

Now, let us suppose that g vanishes at σ in the interior of J . The relation $g \leq 0$ implies that $g'(\sigma) = 0$. Expanding in a Taylor series about σ yields:

$$|g|_J \leq \frac{1}{2}\mu(J)^2 |g^{(2)}|_J.$$

Utilizing (10.3), we get:

$$\langle f_I, g \rangle_{\mathcal{C}(J)} = \langle f_I - f, g \rangle_{\mathcal{C}(J)} \leq \mu(J) |f'_I - f'|_J |g|_J \leq \frac{1}{2}\mu(J)^3 |g^{(2)}|_J |f'_I - f'|_J.$$

Relation (10.1) completes the proof. \square

LEMMA 10.4. Suppose that $J \subset \mathbb{R}$ is an interval, and the interpolation operator I acts on $f: J \rightarrow \mathbb{R}$ to produce the polynomial of degree at most k that agrees with f at $k+1$ distinct points on J . Then we have:

$$\left| \frac{d^m}{dt^m} [(fg)_I - fg_I] \right|_J \leq c\mu(J)^{k+1-m} \sum_{i=0}^k |f^{(k+1-i)}|_J |g^{(i)}|_J$$

whenever $0 \leq m \leq k+1$ and $f', g \in W_1^{k,\infty}(J)$.

Proof. Since the interpolant is expressed in terms of function values,

$$(fg)_I = (fg_I)_I.$$

Hence (10.1) gives us:

$$\left| \frac{d^m}{dt^m} [(fg_I)_I - fg_I] \right|_J \leq c\mu(J)^{k+1-m} |(fg_I)^{(k+1)}|_J.$$

By Leibniz's formula

$$(fg)^{(m)} = \sum_{i=0}^m \frac{m!}{i!(m-i)!} f^{(i)} g^{(m-i)},$$

we see that

$$|(fg_I)^{(k+1)}|_J \leq c \sum_{i=0}^{k+1} |f^{(k+1-i)}|_J |g_I^{(i)}|_J.$$

Since $g \in W_1^{k,\infty}$, (10.1) implies that

$$|g_I^{(i)}|_J \leq c |g^{(i)}|_J$$

for all $0 \leq i \leq k$. Furthermore, $g_I^{(k+1)}$ is identically zero since g_I is a polynomial of

degree at most k . These relations and the inequality

$$|fg|_J \leq |f|_J |g|_J$$

complete the proof. \square

LEMMA 10.5. *If $J \subset R$ is a closed interval and $f \in W_1^{2,\infty}(J)$, then the quadratic agreeing with f at the two ends and the midpoint of J is nondecreasing if*

$$\mu(J)|f''|_J \leq 2 \text{ minimum } \{f'(t): t \in J\}.$$

Proof. Since a nontrivial interval can be mapped by an affine transformation onto \mathcal{T} , there is no loss of generality in assuming that $J = \mathcal{T}$. Let I be the interpolation operator described by the lemma. Since I is linear and $g_I = g$ if g is constant, we can also assume that $f(0) = 0$. In this case, observe that

$$f_I(t) = 4f(\tfrac{1}{2})t(1-t) + f(1)t(2t-1).$$

The derivative of this quadratic is linear and nonnegative on \mathcal{T} if and only if it is nonnegative at $t=0$ and 1 . Omitting the arithmetic, f_I is nondecreasing if and only if

$$(10.4) \quad \tfrac{3}{4}f(1) \geq f(\tfrac{1}{2}) \geq \tfrac{1}{4}f(1).$$

Since f is continuously differentiable, there exists $\sigma \in \mathcal{T}$ such that

$$f'(\sigma) = f(1).$$

Suppose that $\sigma \leq \frac{1}{2}$ (the case $\sigma > \frac{1}{2}$ is treated in a similar manner). The identity

$$f(\tfrac{1}{2}) = \tfrac{1}{2}f(1) + \int_0^{1/2} \int_\sigma^\tau f''(t) dt d\tau,$$

and the bound

$$\int_0^{1/2} \int_0^\tau |f''(t)| dt d\tau \leq \tfrac{1}{8}|f''|_{\mathcal{T}}$$

imply that

$$(10.5) \quad |f(\tfrac{1}{2}) - \tfrac{1}{2}f(1)| \leq \tfrac{1}{8}|f''|_{\mathcal{T}}.$$

Combining (10.4) and (10.5), f_I is nondecreasing if

$$|f''|_{\mathcal{T}} \leq 2f(1) = 2f'(\sigma),$$

a condition clearly satisfied under the lemma's hypothesis. \square

Now, let us prove Theorem 10.1. Let (p_I, ν_I) be the continuous piecewise linear function which agrees with (p, ν) at each mesh point (except that $\nu_I(0) = \nu(0^+)$ and $\nu_I(1) = \nu(1^-)$ —see the remarks at the end of § 5). Relation (5.4) and the identity

$$L(p) = C(z) + \langle p, M(z) \rangle$$

imply that $\langle \nu, K(x) \rangle_{\mathcal{C}} \geq 0$. Since ν is nondecreasing and $K(x) \leq 0$, we conclude that

$$\nu'(t) \cdot K(x(t), t) = 0 \quad \text{almost everywhere.}$$

Therefore, by Lemma 10.3 and the assumed smoothness properties,

$$\langle \nu_I, K(x) \rangle_{\mathcal{C}} \leq O(h^2).$$

Furthermore, by (10.1),

$$|\delta p|_{\mathcal{T}} = O(h) = |\delta \nu|_{\mathcal{T}}.$$

Finally, observe that δq can be expressed as follows:

$$(10.6) \quad \delta q = (p - G^T \nu) - (p - G^T \nu)_I + G^T \nu_I - (G^T \nu)_I.$$

Since the operator I is linear, Lemma 10.4 and the assumed regularity give us

$$|\delta q'|_{\mathcal{F}} = O(h).$$

Relations (9.15) and (9.16) complete the proof. \square

The proof of Theorem 10.2 is similar. Recall that the sets T_j defined earlier are each composed of a finite number of intervals. Let $\{\sigma_1, \sigma_2, \dots, \sigma_k\}$ denote the union over j of boundary points in T_j , and let $\{\sigma_{k+1}, \dots, \sigma_l\}$ be the points separating intervals where $x^{(3)}, (p - G\nu)^{(3)}, G^{(3)}, p^{(2)}$ and $\nu^{(2)}$ are essentially bounded. We form an interpolant (p_I, ν_I) by pasting together local interpolants of (p, ν) over each grid interval J where the local interpolants are defined as follows:

(1) If $J \cap \{\sigma_j\}$ is nonempty, interpolate linearly between function values at the ends of J .

(2) If $J \cap \{\sigma_j\}$ is empty, use quadratic interpolation based on function values at the ends and middle of J .

By assumption,

$$\nu'_j(t) \geq \beta > 0 \quad \forall t \in T_j^c,$$

$j = 1, 2, \dots, s$. Hence, when h is small enough, Lemma 10.5 asserts that $(\nu_I)_j$ is nondecreasing on all mesh intervals which intersect the complement of T_j . On the other hand, we observed in the proof of Theorem 10.1 that $\langle \nu, K(x) \rangle_{\mathcal{G}} = 0$. Since ν is nondecreasing and $K(x) \leq 0$, it follows that ν_j is constant on intervals contained in T_j . Therefore, $(\nu_I)_j = \nu_j$ on all mesh intervals contained in T_j , and ν_I is nondecreasing if h is small enough.

By (10.1) and the assumed smoothness properties,

$$|\delta p|_J = O(h^2) = |\delta \nu|_J$$

for all mesh intervals J such that $J \cap \{\sigma_j\}$ is empty, and

$$|\delta p|_J = O(h) = |\delta \nu|_J$$

otherwise. Similarly, the identity (10.6) and Lemma 10.4 give us

$$|\delta q'|_J = O(h^2)$$

if $J \cap \{\sigma_j\}$ is empty, and

$$|\delta q'|_J = O(h)$$

otherwise. And by Lemma 10.3,

$$\langle \nu_I, K(x) \rangle_{\mathcal{G}(J)} \leq c\mu(J)^4$$

if $J \cap \{\sigma_j\}$ is empty, and

$$\langle \nu_I, K(x) \rangle_{\mathcal{G}(J)} \leq c\mu(J)^3$$

otherwise. Since the measure of mesh intervals intersecting $\{\sigma_j\}$ is at most lh , relations (9.15) and (9.16) complete the proof. \square

For problems without state constraints, the analysis is much easier. In [18] we give a simple treatment of quadratic cost problems with control constraints. Although smoothness considerations limit the \mathcal{L}^2 convergence rate to 1.5, higher rates are achieved when the grid points are free parameters in the optimization process—see [14].

11. Algorithms. Section 10 establishes the convergence of dual finite element approximations to constrained control problems. We now consider the practical side: How is the dual problem solved? When the dual optimization is unconstrained, steepest descent, conjugate gradient and quasi-Newton methods can be applied, but the cost functional is ill-conditioned, and computing time on an IBM 370 computer can be one hour for simple problems! Our main objective in this section is to present a new algorithm which *quickly* solves the dual problem. We also examine the tightness of the error estimates that were established in § 10.

To illustrate the conditioning problems that can arise when standard optimization techniques are applied to the dual problem, the following experiment is cited: Consider the approximation (D_h) to the dual of Problem II from § 8 where the approximating space \mathcal{S}_h is a space of linear splines on a uniform mesh (see [8], [38], or [55] for a discussion of piecewise polynomial spaces). The time needed to solve this dual problem using: 1000 basis elements (which gives 5-place accuracy), an IBM 370 model 3033 computer, the IMSL conjugate gradient routine, the FORTRAN IV (H) optimizing compiler and the initial guess zero, is 1 hour. We now develop an algorithm which solves this dual problem in 1 second.

For the dual problems in § 8, observe that the dual integrand at time t is chosen from a finite set. For example, the dual integrand in Problem I is $l(x, y, t) = \frac{1}{2}[x^2 + y^2]$ if $y \leq a$ and $l(x, y, t) = \frac{1}{2}[x^2 + a(2y - a)]$ if $y > a$. In general the dual integrand l is expressed in terms of a partition $\{\mathcal{R}_1, \dots, \mathcal{R}_k\}$ of $R^{2n} \times \mathcal{T}$ and integrands l_1, \dots, l_k defined on $R^{2n} \times \mathcal{T}$. And the integrand $l(p'(t), p(t), t)$ of the dual functional satisfies

$$(11.1) \quad l(x, y, t) = l_i(x, y, t)$$

whenever $(x, y, t) \in \mathcal{R}_i$. For Problem I, we have:

$$l_1(x, y, t) = \frac{1}{2}[x^2 + y^2], \quad l_2(x, y, t) = \frac{1}{2}[x^2 + a(2y - a)],$$

$$\mathcal{R}_1 = \{(x, y, t) \in R \times R \times \mathcal{T} : y \leq a\}, \quad \mathcal{R}_2 = \{(x, y, t) \in R \times R \times \mathcal{T} : y > a\}.$$

In formulating our algorithm for the dual problem, we assume that the dual function has the form

$$L(p) = \phi(p) + \int_{\mathcal{T}} l(p'(t), p(t), t) dt$$

where $\phi: \mathcal{A} \rightarrow R \cup \{-\infty\}$ and l satisfies (11.1) for some partition $\{\mathcal{R}_1, \dots, \mathcal{R}_k\}$ of $R^{2n} \times \mathcal{T}$ and integrands l_1, \dots, l_k defined on $R^{2n} \times \mathcal{T}$. Now, given a partition $T = \{T_1, \dots, T_k\}$ of \mathcal{T} into measurable sets, let us define the functional

$$(11.2) \quad M(p, T) = \phi(p) + \sum_{i=1}^k \int_{T_i} l_i(p'(t), p(t), t) dt.$$

Any $p \in \mathcal{A}$ induces a partition $\{T_1, \dots, T_k\}$ of \mathcal{T} where $t \in T_i$ if and only if

$$(p'(t), p(t), t) \in \mathcal{R}_i.$$

Let S be the map that acts on p to produce the associated partition of \mathcal{T} . From these definitions, we see that

$$L(p) = M(p, S(p)).$$

For the examples in § 8, observe that the elements of $S(p)$ are measurable for each $p \in \mathcal{A}$. More generally, it can be shown that the elements of $S(p)$ are measurable if the multifunctions $\mathcal{R}_1, \dots, \mathcal{R}_k$ are measurable—see [51]. Henceforth, we assume that the elements of $S(p)$ are measurable for every $p \in \mathcal{A}$.

Letting $K \subset \mathcal{A}$ denote a convex set of dual feasible functions which contains a solution to the dual problem, our algorithm for solving (D) is the following: Starting from some $p^0 \in K$, we generate a sequence p^1, p^2, \dots (we hope) converging to a dual solution where

$$p^{j+1} = \arg \max \{M(p, T^j) : p \in K\}, \quad T^j := S(p^j).$$

There is an analogous scheme for the dual approximation (D_h) . If $\{\psi_1, \dots, \psi_N\}$ is a basis for the finite element space $\mathcal{S}_h \subset \mathcal{A}$, we define

$$M^h(\alpha, T) = M\left(\sum_{i=1}^N \alpha_i \psi_i, T\right) \quad \text{and} \quad K^h = \left\{\alpha \in R^N : \sum_{i=1}^N \alpha_i \psi_i \in K\right\}.$$

Our scheme for solving (D_h) starts from some $\alpha^0 \in K^h$, and constructs iterations $\alpha^1, \alpha^2, \dots$ by the rule

$$(11.3) \quad \alpha^{j+1} = \arg \max \{M^h(\alpha, T^j) : \alpha \in K^h\}, \quad T^j := S\left(\sum_{i=1}^N \alpha_i^j \psi_i\right).$$

We remark that if $M^h(\cdot, T^j)$ is Gateaux differentiable and α^{j+1} satisfies (11.3), then the following standard inequality holds [27, Thm. I.1.3]:

$$\frac{\partial M^h}{\partial \alpha}[\alpha^{j+1}, T^j](\alpha - \alpha^{j+1}) \leq 0 \quad \forall \alpha \in K^h.$$

This algorithm has been tested on the problems presented in § 8. Experimentally, the convergence is fast; moreover, the iterations seem to converge from any starting point α^0 . For example, starting from the initial guess $\alpha^0 = 0$ in Problem I and using the linear spline basis mentioned earlier, the relative change $|\alpha^{j+1} - \alpha^j|/|\alpha^j|$ is reduced to 10^{-10} after 5 iterations, independent of the number of basis elements; each iteration involves solving a symmetric, tridiagonal system and is easy to implement. The FORTAN code for Problem I has about 60 statements. Later we show under appropriate hypotheses that the scheme (11.3) is quadratically convergent near a solution to the dual problem (D_h) .

First we observe that any fixed point for the iterations (11.3) solves the dual maximization problem (D_h) . This result is based on the rule for differentiating under the integral sign. Below, $W^{1,\infty}$ denotes the space of Lipschitz continuous functions $p: \mathcal{T} \rightarrow R^n$ with the norm

$$\|p\|_{W^{1,\infty}} = \text{essential supremum } \{|p(t)| + |p'(t)| : t \in \mathcal{T}\}.$$

LEMMA 11.1. *Suppose that $T \subset \mathcal{T}$ is measurable and $g: R^{2n} \times \mathcal{T} \rightarrow R$ is continuously differentiable in its first $2n$ arguments on $R^{2n} \times \mathcal{T}$. If $G: W^{1,\infty} \rightarrow R$ is defined by*

$$G(p) = \int_T g(p'(t), p(t), t) dt,$$

then the Fréchet derivative of G is

$$\frac{\partial G}{\partial p}[p](q) = \int_T [\nabla_1 g(p'(t), p(t), t)q'(t) + \nabla_2 g(p'(t), p(t), t)q(t)] dt$$

where $\nabla_1 g$ and $\nabla_2 g$ denote g 's gradient with respect to its first n and second n arguments respectively.

Note that every dual integrand presented in § 8 is continuously differentiable. In Appendix 2 we show that this continuity property holds for a broad class of problems.

To prove that any fixed point of the iterations (11.3) solves the dual maximization problem (D_h) , let us assume that both the integrands l_i and the composite integrand l are continuously differentiable in their first $2n$ arguments on $R^{2n} \times \mathcal{T}$ and the function ϕ in (11.2) is differentiable. Defining the sets

$$\mathcal{R}_i(t) = \{r \in R^{2n} : (r, t) \in \mathcal{R}_i\},$$

we assume moreover that

$$\text{closure } \mathcal{R}_i(t) = \text{closure (interior } \mathcal{R}_i(t))$$

for each $t \in \mathcal{T}$. Hence, if $r \in \mathcal{R}_i(t)$, there exists a sequence $\{r_k\} \subset \text{interior } \mathcal{R}_i(t)$ converging to r , and since $l_i(\cdot, t) = l(\cdot, t)$ near r_k , we have

$$\nabla l_i(r_k, t) = \nabla l(r_k, t).$$

Taking the limit as k goes to infinity, the continuous differentiability assumption implies that

$$(11.4) \quad \nabla l_i(r, t) = \nabla l(r, t)$$

for every $r \in \mathcal{R}_i(t)$. Now, let us define

$$L^h(\alpha) = L\left(\sum_{i=1}^N \alpha_i \psi_i\right).$$

If each ψ_i lies in $W^{1,\infty}$, then Lemma 11.1 and (11.4) yield

$$\frac{\partial L^h}{\partial \alpha}[\beta] = \frac{\partial M^h}{\partial \alpha}[\beta, T] \quad \text{where } T = S\left(\sum_{i=1}^N \beta_i \psi_i\right).$$

This observation, the concavity of the dual function and the following result combine to show that any fixed point of the iterations (11.3) solves (D_h) .

THEOREM 11.2. *Suppose that $G: K \times K \rightarrow R$ where K is a convex subset of a vector space, $y \in K$, and $G(x, x)$ and $G(x, y)$ are Gateaux differentiable functions of x at $x = y$ which satisfy*

$$\left. \frac{\partial G(x, x)}{\partial x} \right|_{x=y} = \left. \frac{\partial G(x, y)}{\partial x} \right|_{x=y}.$$

If $G(x, x)$ is a convex function of $x \in K$ and y minimizes $G(x, y)$ over $x \in K$, then y minimizes $G(x, x)$ over $x \in K$. Conversely, if $G(x, y)$ is a convex function of $x \in K$ and y minimizes $G(x, x)$ over $x \in K$, then y minimizes $G(x, y)$ over $x \in K$.

Proof. First assume that $G(x, x)$ is a convex function of $x \in K$ and y minimizes $G(x, y)$ over $x \in K$. Since $G(\cdot, y)$ is Gateaux differentiable at y and K is convex, we have the standard variational inequality [27, Thm. I.1.3]:

$$(11.5) \quad \left. \frac{\partial G(x, y)}{\partial x} \right|_{x=y} (x - y) \geq 0 \quad \forall x \in K.$$

The hypotheses for the gradient and (11.5) imply that

$$(11.6) \quad \left. \frac{\partial G(x, x)}{\partial x} \right|_{x=y} (x - y) \geq 0 \quad \forall x \in K.$$

Since $G(x, x)$ is convex, it follows from (11.6) and [27, Thm. I.1.3] that y minimizes $G(x, x)$ over $x \in K$. Conversely, let us assume that y minimizes $G(x, x)$ over $x \in K$ and $G(x, y)$ is a convex function of $x \in K$. Again, by [27, Thm. I.1.3], the variational

inequality (11.6) holds, and by the hypotheses for the gradient, we conclude that (11.5) is satisfied. Since $G(\cdot, y)$ is convex on K , (11.5) implies that y minimizes $G(x, y)$ over $x \in K$. \square

Before giving a convergence proof for the iterative scheme (11.3), let us examine Problems I and II to help motivate our theorem's hypotheses. Let \mathcal{S}_h be the space of linear splines defined on a uniform mesh where $h = 1/N$ is the distance between grid points, and let $\{\psi_0, \dots, \psi_N\}$ be the usual basis for \mathcal{S}_h sketched in Fig. 4. Applying Lemma 11.1 to the dual functional for Problem I, we have

$$(11.7) \quad -\frac{\partial L^h(p)}{\partial \alpha_i} = \int_0^1 p'(t) \psi_i'(t) dt + \int_{T_1(\alpha)} p(t) \psi_i(t) dt + a \int_{T_2(\alpha)} \psi_i(t) dt$$

for $i = 1, \dots, N$ where

$$p(t) = \sum_{i=0}^N \alpha_i \psi_i(t), \quad T_1(\alpha) = \{t \in \mathcal{T}: p(t) \leq a\}, \quad T_2(\alpha) = \{t \in \mathcal{T}: p(t) > a\}.$$

The partial derivative of $-L^h$ with respect to α_0 is c plus the terms on the right side of (11.7) where c is the state's initial value in Problem I. If $p(t)$ equals a at just a finite set of $t \in (0, 1)$ and $0 = t_l(\alpha) < t_{l+1}(\alpha) < \dots < t_r(\alpha) = 1$ denote these t where $p(t)$ is a union $\{0, 1\}$, then we can write

$$\int_{T_1(\alpha)} p(t) \psi_i(t) dt + a \int_{T_2(\alpha)} \psi_i(t) dt = \sum_{j \text{ even}} \int_{t_j(\alpha)}^{t_{j+1}(\alpha)} p(t) \psi_i(t) dt + \sum_{j \text{ odd}} a \int_{t_j(\alpha)}^{t_{j+1}(\alpha)} \psi_i(t) dt.$$

Thus for β in a neighborhood of the fixed coefficients $\{\alpha_0, \dots, \alpha_N\}$, the gradient of L^h evaluated at β has the form $g(\beta, T(\beta))$ where $T(\beta)$ is a vector with components $t_l(\beta), \dots, t_r(\beta)$. Our main observation is the following: Since the $\psi_i(t)$ are continuous functions of t and $p(t_j(\alpha)) = a$ for $j = l+1, \dots, r-1$, we have:

$$\left. \frac{\partial g_i(\alpha, T(\beta))}{\partial \beta_k} \right|_{\beta=\alpha} = \sum_{j=l+1}^{r-1} a [\psi_i(t_j) - \psi_i(t_{j+1})] \left. \frac{\partial t_j(\beta)}{\partial \beta_k} \right|_{\beta=\alpha} = 0.$$

More compactly, this result can be stated

$$(11.8) \quad \left. \frac{\partial g(\alpha, T(\beta))}{\partial \beta} \right|_{\beta=\alpha} = 0.$$

This identity also holds for state constrained problems, but the argument is a little different. For Problem II, the terms in the gradient of the dual function corresponding to the state constraint are

$$(11.9) \quad \sum_{j \text{ even}} \int_{t_j(\alpha)}^{t_{j+1}(\alpha)} p'(t) \psi_i'(t) dt + \sum_{j \text{ odd}} b \int_{t_j(\alpha)}^{t_{j+1}(\alpha)} \psi_i'(t) dt$$

where the two sums above correspond to intervals where $p'(t) \leq b$ and $p'(t) > b$ respectively. Since p is a linear spline, p' is piecewise constant. Hence, if $p'(t) \neq b$ for

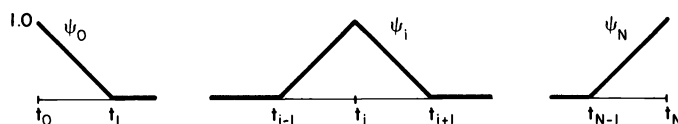


FIG. 4. Linear spline basis.

every $t \in \mathcal{T}$, then $t_j(\beta)$ is independent of β in a neighborhood of α . In summary, even though the integrands in (11.9) are discontinuous functions of t , the identity (11.8) still holds since $t_j(\beta)$ is independent of β in a neighborhood of α . With this motivation, we now present our local quadratic convergence result:

THEOREM 11.3. *Suppose that $g: R^n \times R^n \rightarrow R^n$ and K is a nonempty, closed convex subset of R^n , and consider the problem of finding $\alpha \in K$ such that*

$$(11.10) \quad g(\alpha, \alpha) \cdot (\beta - \alpha) \geq 0$$

for all $\beta \in K$. We assume that there exists a solution α^* to (11.10) and that the following conditions are satisfied:

(1) $g(\alpha, \beta)$ is a continuous function of α and β near α^* , and $\partial g(\alpha, \beta)/\partial \alpha$ exists and is a continuous function of α and β near α^* .

(2) $g(\alpha^*, \cdot)$ is twice continuously differentiable near α^* and the first derivative vanishes at α^* .

(3) Either $K = R^n$ and $\partial g(\alpha, \alpha^*)/\partial \alpha|_{\alpha=\alpha^*}$ is nonsingular, or K is an arbitrary closed convex subset of R^n and $\partial g(\alpha, \alpha^*)/\partial \alpha|_{\alpha=\alpha^*}$ is positive definite.

Then there exists a neighborhood \mathcal{N} of α^* with the following properties: For each $\alpha^0 \in \mathcal{N}$, there is a unique sequence $\{\alpha^1, \alpha^2, \dots\} \subset K \cap \mathcal{N}$ such that

$$(11.11) \quad g(\alpha^{j+1}, \alpha^j) \cdot (\beta - \alpha^{j+1}) \geq 0$$

for all $\beta \in K$ and $j = 0, 1, \dots$, and for some constant c independent of j and $\alpha_0 \in \mathcal{N}$, we have:

$$|\alpha^{j+1} - \alpha^*| \leq c|\alpha^j - \alpha^*|^2.$$

Proof. By Robinson [43, Thms. 2.1 and 3.1], there exist neighborhoods \mathcal{N}_1 and \mathcal{N}_2 of α^* such that the following problem has a unique solution $\alpha \in \mathcal{N}_1$ for each $\gamma \in \mathcal{N}_2$: find $\alpha \in K$ such that

$$(11.12) \quad g(\alpha, \gamma) \cdot (\beta - \alpha) \geq 0$$

for all $\beta \in K$. Shrink \mathcal{N}_2 so it is contained in a bounded region where $g(\alpha^*, \cdot)$ is twice continuously differentiable, and let $\Phi(\gamma) \in \mathcal{N}_1$ denote the solution of (11.12) corresponding to $\gamma \in \mathcal{N}_2$. By [43, Thm. 2.1] there also exists a constant μ such that

$$(11.13) \quad |\Phi(\gamma) - \Phi(\alpha^*)| \leq \mu |g(\alpha^*, \gamma) - g(\alpha^*, \alpha^*)|$$

for all $\gamma \in \mathcal{N}_2$. Expanding $g(\alpha^*, \cdot)$ to first order about α^* and using the integral form for the remainder term, our second hypothesis implies that

$$(11.14) \quad |g(\alpha^*, \gamma) - g(\alpha^*, \alpha^*)| \leq c|\gamma - \alpha^*|^2$$

for some constant c independent of $\gamma \in \mathcal{N}_2$. Combining (11.13) and (11.14), we have for $\alpha^j \in \mathcal{N}_2$:

$$(11.15) \quad |\alpha^{j+1} - \alpha^*| = |\Phi(\alpha^j) - \Phi(\alpha^*)| \leq c|\alpha^j - \alpha^*|^2.$$

Thus if α^0 is sufficiently close to α^* , the entire sequence $\{\alpha^j\}$ given by $\alpha^{j+1} = \Phi(\alpha^j)$ lies in $\mathcal{N}_1 \cap \mathcal{N}_2$ and (11.15) holds. \square

Observe that the inequality (11.10) is essentially the relation (11.6) characterizing the solution to the dual approximation (D_h) while the iterations defined by (11.11) correspond to our scheme (11.3). The inequality (11.13) is a crucial step in our proof of Theorem 11.3. In Robinson's study [43] of the implicit function theorem for inequalities, he establishes this relation in a very general setting whenever the "strong regularity" assumption is satisfied. Moreover, in finite dimensions it follows from his

Theorem 3.1 that the strong regularity assumption holds under hypothesis 3 of our theorem. Hence a more general version of Theorem 11.3 can be established where R^n is replaced by a normed linear space and hypothesis 3 is replaced by the strong regularity assumption.

Now let us study the tightness of the error estimates established in § 10. Since the solutions to Problems I and II from § 8 can be determined analytically (see Appendix 3), the error in finite element approximations can be computed precisely. Taking $a = 1$ and $c = (1 + 3e)/2(1 - e)$ in Problem I, the optimal control is 1 for $t \in [0, \frac{1}{2}]$. Thus the constraint $u(t) \leq a$ is binding for the optimal control when $0 \leq t \leq \frac{1}{2}$. Taking $a = 1$, $b = 2\sqrt{e}/(1 - e)$, and $c = (5e + 3)/4(1 - e)$ in Problem II, the optimal control is a for $t \in [0, \frac{1}{4}]$ and the optimal state is b for $t \in [\frac{3}{4}, 1]$. The solutions for these choices of parameters are shown in Figs. 5 and 6.

In § 10 we give the estimates

$$\|x - x^h\| + \|u - u^h\| = \begin{cases} O(h) & \text{for linear elements,} \\ O(h^{3/2}) & \text{for quadratic elements,} \end{cases}$$

where (x, u) solves the primal problem, (x^h, u^h) is the finite element approximation, and $\|\cdot\|$ is the \mathcal{L}^2 norm. Comparing the exact solution of Problems I and II to the finite element approximations, these estimates are tight. That is, there exists a constant $C > 0$ such that

$$\|x - x^h\| + \|u - u^h\| \cong \begin{cases} Ch & \text{for linear elements,} \\ Ch^{3/2} & \text{for quadratic elements.} \end{cases}$$

Problem IV was solved using linear splines, and we obtained the solution reported in [56]; moreover, the error $\|x - x^h\| + \|u - u^h\|$ was proportional to h . Since Problem IV is not strictly convex, it appears that the convexity assumptions in § 9 can be mildly relaxed.

Although the estimate for $\|x - x^h\| + \|u - u^h\|$ is tight, we also observe in Figs. 7 and 8 that the control converges faster than the state. For linear elements,

$$\|u - u^h\| = O(h^{3/2}),$$

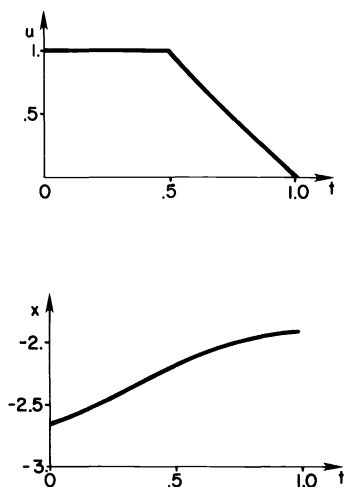


FIG. 5. Solution to Problem I.

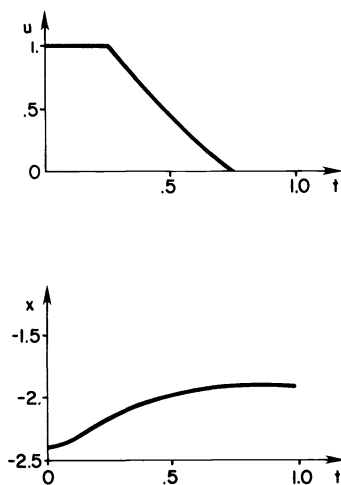
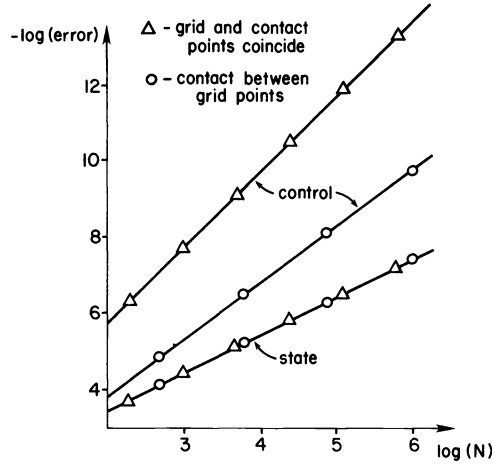
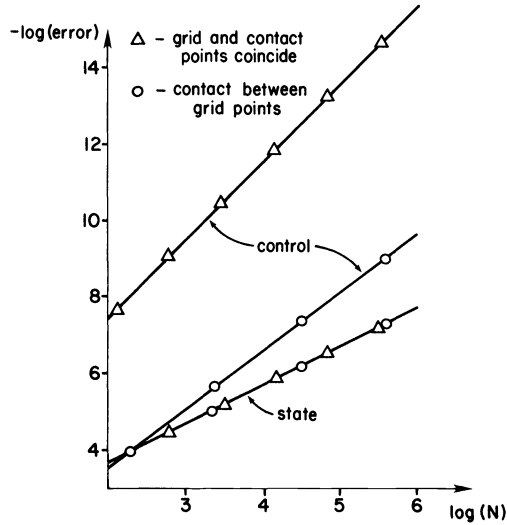


FIG. 6. Solution to Problem II.

FIG. 7. \mathcal{L}^2 error for Problem I and linear elements.FIG. 8. \mathcal{L}^2 error for Problem II and linear elements.

and putting grid points at the contact points (where constraints in the primal problem change between binding and nonbinding) gives us the better result:

$$\|u - u^h\| = O(h^2).$$

For unconstrained problems, Mathis and Reddien [32] use a duality argument to show that the control convergence rate is h times the state rate. The extension of this result to constrained optimization is open. The convergence rate for quadratic elements also improves when the contacts are members of the grid:

$$\|x - x^h\| + \|u - u^h\| = O(h^2).$$

This property of the total error is established in [14] for a full dual scheme.

In summary better approximations to the primal solution are obtained as follows: Solve the dual problem on a fixed mesh and estimate the contacts; then insert grid points at the approximate contacts, and repeat the process. After the contacts converge, generate a better state by integrating forward the system dynamics with the approximate control as input.

Appendix 1. Existence. Suppose that $f: X \rightarrow [-\infty, +\infty)$ and $g: X \rightarrow Y$ where X is a set and Y is a normed vector space that is ordered by a convex cone $N \subset Y$ with vertex at the origin; that is, given a and $b \in Y$, we write $a \geq b$ if $b - a \in N$. If Y^* denotes the space of bounded linear functionals on Y , N induces an ordering on Y^* relative to the convex cone

$$N^* = \{y^* \in Y^*: \langle y^*, y \rangle \geq 0 \text{ for all } y \in N\}.$$

Above $\langle \cdot, \cdot \rangle$ denotes the usual pairing between Y^* and Y . Associated with the primal problem,

$$(P') \quad \begin{array}{ll} \text{maximize} & f(x) \\ \text{subject to} & g(x) \geq 0, \quad x \in X, \end{array}$$

is the dual problem,

$$(D') \quad \begin{array}{ll} \text{minimize} & L(\lambda) \\ \text{subject to} & \lambda \geq 0, \quad \lambda \in Y^*, \end{array}$$

where

$$L(\lambda) = \sup \{f(x) + \langle \lambda, g(x) \rangle : x \in X\}.$$

Although equality constraints are not explicitly stated in the primal problem, the inequality $g(x) \geq 0$ becomes equality when $N = \{0\}$.

Under certain convexity hypotheses and constraint qualifications, a “typical duality theorem” asserts that there exists a solution λ to the dual problem and

$$L(\lambda) = \sup \{f(x) : g(x) \geq 0, x \in X\}.$$

For example, see [28], [50], or Theorem A3 below. First, we observe that dual approximations exist without the convexity hypothesis. If S is a subset of $-N^*$, we consider the following approximation to the dual problem:

$$(D'_S) \quad \text{minimize} \quad \{L(\lambda) : \lambda \in S\}.$$

Let us define the set

$$\Delta(y) = \{x \in X : g(x) \geq y\},$$

and the ball

$$\mathcal{B}^\rho = \{y \in Y : \|y\| \leq \rho\},$$

and let us introduce the following assumption for (P') :

BOUNDEDNESS ASSUMPTION. *There exists $\rho > 0$ such that $\Delta(y)$ is nonempty for all $y \in \mathcal{B}^\rho$ and*

$$M := \inf_{y \in \mathcal{B}^\rho} \sup_{x \in \Delta(y)} f(x) > -\infty.$$

THEOREM A.1. *If S is a closed subset of a finite dimensional space and the boundedness hypothesis is satisfied, there exists a solution to (D'_S) .*

Proof. If $L(\lambda)$ is ∞ for all $\lambda \in S$, the theorem is trivial, so let us assume that $S \cap \text{dom } L$ is nonempty. Beginning with the definition of the dual functional and utilizing the boundedness assumption,

$$L(\lambda) = \sup \{f(x) + \langle \lambda, g(x) \rangle : x \in X\} \geq \sup \{f(x) + \langle \lambda, y \rangle : x \in \Delta(y)\} \geq M + \langle \lambda, y \rangle$$

for each $y \in \mathcal{B}^p$. Maximizing over $y \in \mathcal{B}^p$, it follows that

$$(A.1) \quad \|\lambda\|_{Y^*} := \sup \{\langle \lambda, y \rangle : y \in \mathcal{B}^1\} \leq (L(\lambda) - M)/\rho.$$

By the next lemma, L is lower semicontinuous. Since S is a closed subset of a finite dimensional space, (A.1) implies that the level sets

$$\{\lambda \in S : L(\lambda) \leq \alpha\}$$

are compact. Hence, there exists a solution to (D'_S) . \square

LEMMA A.2. L is lower semicontinuous with respect to both the norm topology of Y^* and the weak topology induced on Y^* by Y .

Proof. This result is essentially contained in Rockafellar's work [50, Thm. 5] or [46, p. 104]. Consider the epigraph set

$$\text{epi } L = \{(\alpha, \lambda) \in R \times Y^* : \alpha \geq L(\lambda)\}.$$

Alternatively, we can view this set as the intersection of half spaces which are closed in the weak topology induced on Y^* by Y ; in particular,

$$\text{epi } L = \bigcap_{x \in X} \{(\alpha, \lambda) \in R \times Y^* : \alpha \geq f(x) + \langle \lambda, g(x) \rangle\}.$$

Therefore, $\text{epi } L$ is closed in both the norm and the weak topologies. Since lower semicontinuity of L is equivalent to the epigraph being closed, the proof is complete. \square

Suppose that X is a convex subset of a vector space. We say that g is concave if

$$g(\alpha x_1 + (1 - \alpha)x_2) \geq \alpha g(x_1) + (1 - \alpha)g(x_2)$$

for all $x_1, x_2 \in X$ and $0 \leq \alpha \leq 1$.

THEOREM A.3. Suppose that X is a convex subset of a vector space and both f and g are concave. Under the boundedness hypothesis, there exists a solution λ to (D') and

$$L(\lambda) = v := \supremum \{f(x) : g(x) \geq 0, x \in X\}.$$

Moreover, for any solution x of (P') , we have $\langle \lambda, g(x) \rangle = 0$.

(This last result is called the complementary slackness condition.)

Proof. If $v = \infty$, the result is trivial. Since $v \geq M > -\infty$ by the boundedness hypothesis, assume that v is finite. Let us define the convex sets:

$$E = \{(\alpha, y) \in R \times Y : \exists x \in \Delta(y) \text{ with } \alpha \leq f(x)\}$$

and

$$D = \{(\alpha, y) \in R \times Y : \alpha = v, y \geq 0\}.$$

By the boundedness assumption, $(M - 1, 0) \in R \times Y$ is an interior point of E . Separating D from E with a hyperplane gives us $r \in R$ and $\mu \in Y^*$ such that

$$(A.2) \quad rv + \langle \mu, z \rangle \geq r\alpha + \langle \mu, y \rangle$$

for all $(v, z) \in D$ and $(\alpha, y) \in E$. Clearly, $\mu \geq 0$. Since $(M - 1, 0)$ lies in the interior of E and $(\alpha, 0) \in E$ for all $\alpha < M$, we see that $r > 0$. Dividing by r , setting $\lambda = \mu/r$, and

inserting $(\alpha, y) = (f(x), g(x))$ and $z = 0$, (A.2) gives us

$$v \geq f(x) + \langle \lambda, g(x) \rangle$$

for all $x \in X$, or equivalently, $L(\lambda) \leq v$. Since $L(\lambda) \geq v$ by weak duality, it follows that $L(\lambda) = v$. Finally, if (P) has a solution $x \in X$, then the inequality $v \geq f(x) + \langle \lambda, g(x) \rangle = v + \langle \lambda, g(x) \rangle$ implies that $\langle \lambda, g(x) \rangle \leq 0$. Since $\lambda \geq 0$ and $g(x) \geq 0$, we conclude that $\langle \lambda, g(x) \rangle = 0$. \square

Appendix 2. Integrand regularity. In § 11, we note that the dual integrands $l(x, y, t)$ for Problems I–IV are continuously differentiable in x and y . In these examples, this result amounts to showing that the optimal cost of the quadratic program

$$\begin{aligned} &\text{minimize} && x^T Q x + q^T x \\ &\text{subject to} && A x \leq a, \quad x \in R^n \end{aligned}$$

depends smoothly on q . Here Q and A are matrices and q and a are vectors of the appropriate dimensions. Let us study the more general class of problems

$$(A.3) \quad \text{minimize} \quad \{f(x, \xi): g(x, \xi) \leq 0, h(x, \xi) = 0, x \in R^n\}$$

where $f: R^n \times R^p \rightarrow R$, $g: R^n \times R^p \rightarrow R^m$, and $h: R^n \times R^p \rightarrow R^l$. Above, $\xi \in R^p$ is a fixed parameter, and the minimization is over $x \in R^n$. Our development is based on Lipschitz properties established earlier for the solution and the multiplier of (A.3). In [15] these properties are verified for quadratic programs, and in [15, Appendix], we indicate that these results extend to more general programs. This extension is now presented; as a corollary, we show that the optimal cost of (A.3) depends smoothly on ξ .

Suppose that (A.3) has a unique solution $x(\xi)$ for ξ near 0. Under fairly weak assumptions, it has been shown [42] that the feasible set

$$\{x \in R^n: g(x, \xi) \leq 0, h(x, \xi) = 0\}$$

is stable with respect to perturbations in ξ and hence by [44] $x(\xi)$ is a continuous function of ξ . Defining $z = x(0)$, we assume that $f(x, \xi)$, $g(x, \xi)$, and $h(x, \xi)$ are continuously Fréchet differentiable in x near $(x, \xi) = (z, 0)$. If $g_B(x, \xi)$ is the vector composed of g 's components satisfying $g_i(x, \xi) = 0$, we also assume that the rows of

$$\begin{bmatrix} \nabla_1 g_B(x(\xi), \xi) \\ \nabla_1 h(x(\xi), \xi) \end{bmatrix}$$

are linearly independent for ξ near 0. Under these hypotheses, there exist unique multipliers $\lambda(\xi) \in R^m$ and $\mu(\xi) \in R^l$ satisfying the Kuhn–Tucker conditions [29, p. 233]:

$$(A.4) \quad \nabla_1 \mathcal{L}(x(\xi), \xi) = 0, \quad \lambda(\xi) \geq 0, \quad \lambda(\xi)^T g(x(\xi), \xi) = 0$$

where

$$\mathcal{L}(x, \xi) = f(x, \xi) + \lambda(\xi)^T g(x, \xi) + \mu(\xi)^T h(x, \xi).$$

LEMMA A.4. *If $x(\xi)$ is a continuous function of ξ near zero and the differentiability and independence assumptions stated above hold, then $\lambda(\xi)$ and $\mu(\xi)$ are continuous functions of ξ near zero.*

Proof. Let ξ be a fixed parameter near zero, let $I \subset \{1, \dots, m\}$ be the set of indices i for which $g_i(x(\xi), \xi) = 0$, and let g_I be the vector with components g_i , $i \in I$. Consider the system

$$(A.5) \quad \nabla_1 f(x, \eta) + \nabla_1 g_I(x, \eta)^T \lambda_I + \nabla_1 h(x, \eta)^T \mu = 0$$

in the unknowns x , λ_I , and μ . Since $x(\eta)$ depends continuously on η , $g_i(x(\eta), \eta) < 0$ if $i \notin I$ and η is near ξ . Assume that $|\eta - \xi|$ is so small that $g_i(x(\eta), \eta) < 0$ if $i \notin I$. By the Kuhn–Tucker conditions (A.4), $\lambda_i(\eta) = 0$ if $i \notin I$ and $(x(\eta), \lambda_1(\eta), \mu(\eta))$ satisfies (A.5). Since the rows of $\nabla_1 g_B(x(\xi), \xi)$ and $\nabla_1 h(x(\xi), \xi)$ are linearly independent and $x(\eta)$ is a continuous function of η near zero, it follows that the rows of $\nabla_1 g_I(x(\eta), \eta)$ and $\nabla_1 h(x(\eta), \eta)$ are uniformly independent of η near ξ . Hence (A.5) implies that $\lambda_I(\eta)$ and $\mu(\eta)$ are continuous functions of η near ξ . Since $\lambda_i(\eta) = 0$ if $i \notin I$, we conclude that

$$\lim_{\eta \rightarrow \xi} (x(\eta), \lambda(\eta), \mu(\eta)) = (x(\xi), \lambda(\xi), \mu(\xi)). \quad \square$$

THEOREM A.5. *In addition to the hypotheses of Lemma A.4, we assume:*

- (i) *f, g and h have partial derivatives $\partial^2/\partial x^2, \partial^2/\partial x \partial \xi$ and $\partial/\partial \xi$ which are continuous near $(z, 0)$, and*
- (ii) *for each ξ near zero, we have*

$$(A.6) \quad y^T \nabla_{xx} \mathcal{L}(x(\xi), \xi) y > 0$$

for every nonzero vector y such that

$$\nabla_1 g_B(x(\xi), \xi) y = 0 = \nabla_1 h(x(\xi), \xi) y.$$

Then $(x(\xi), \lambda(\xi), \mu(\xi))$ is a Lipschitz continuous function of ξ near zero.

Proof. Again, let ξ be a fixed parameter near zero, let $I \subset \{1, \dots, m\}$ be the set of indices i for which $g_i(x(\xi), \xi) = 0$, and let g_I be the vector with components g_i , $i \in I$. Consider the system

$$(A.7) \quad \begin{aligned} \nabla_1 f(x, \eta) + \nabla_1 g_I(x, \eta)^T \lambda_I + \nabla_1 h(x, \eta)^T \mu &= 0, \\ g_I(x, \eta) &= 0, \\ h(x, \eta) &= 0, \end{aligned}$$

in the unknowns x , λ_I , and μ . Since (A.6) holds and the rows of $\nabla_1 g_B(x(\xi), \xi)$ and $\nabla_1 h(x(\xi), \xi)$ are linearly independent, it follows from [15, Lemma 3.2] that the Jacobian of the system (A.7) with respect to (x, λ_I, μ) is nonsingular at $\eta = \xi$ and $(x, \lambda_I, \mu) = (x(\xi), \lambda_I(\xi), \mu(\xi))$. By the implicit function theorem, (A.7) has a unique solution $(x, \lambda_I, \mu)(\eta)$ for η in a neighborhood of ξ which is a continuously differentiable function of η . Now, as ξ ranges over a neighborhood of zero, the solution $x(\xi)$ and the multipliers $\lambda(\xi)$ and $\mu(\xi)$ for the program (A.3) satisfy (A.7) for different choices of I . As in [15, § 3], it follows from [15, Thm. 2.3] that $(x(\xi), \lambda(\xi), \mu(\xi))$ is a Lipschitz continuous function of ξ near zero. \square

In a related paper [43], Robinson shows that the Kuhn–Tucker conditions (A.4) have a solution depending Lipschitz continuously on a parameter. His assumptions are similar to ours except that (A.6) is strengthened slightly while the assumption that (A.3) has a unique solution is dropped. Now consider the optimal cost $f(x(\xi), \xi)$. Since $x(\xi)$ depends Lipschitz continuously on ξ , we might expect that $f(x(\xi), \xi)$ depends just Lipschitz continuously on ξ . But the cost is smoother than expected:

COROLLARY A.6. *Under the hypotheses of Theorem A.5, $f(x(\xi), \xi)$ is a continuously differentiable function of ξ near 0. Moreover, if f, g and h have continuous second partial derivatives near $(z, 0)$, then the derivative of $f(x(\xi), \xi)$ is Lipschitz continuous.*

Proof. Since $x(\cdot)$ is differentiable almost everywhere, the chain rule gives us

$$\begin{aligned} \frac{\partial}{\partial \xi} f(x(\xi), \xi) &= \frac{\partial}{\partial \xi} \mathcal{L}(x(\xi), \xi) \\ &= \nabla_1 \mathcal{L}(x(\xi), \xi) \frac{\partial x(\xi)}{\partial \xi} + g(x(\xi), \xi)^T \frac{\partial \lambda(\xi)}{\partial \xi} + h(x(\xi), \xi)^T \frac{\partial \mu(\xi)}{\partial \xi} \\ &\quad + \frac{\partial}{\partial \xi} \{f(x, \xi) + \lambda^T g(x, \xi) + \mu^T h(x, \xi)\} \Bigg|_{\substack{x=x(\xi) \\ \lambda=\lambda(\xi) \\ \mu=\mu(\xi)}}. \end{aligned}$$

By the Kuhn–Tucker conditions, $\nabla_1 \mathcal{L}(x(\xi), \xi) = 0$, and since $x(\xi)$ is feasible in (A.3), $h(x(\xi), \xi) = 0$. Suppose that $g_i(x(\xi), \xi) < 0$. Since $x(\eta)$ depends continuously on η , $g_i(x(\eta), \eta) < 0$ for η near ξ . Hence the Kuhn–Tucker conditions also tell us that $\lambda_i(\eta) = 0$ for η near ξ , and

$$\frac{\partial \lambda_i(\xi)}{\partial \xi} = 0.$$

Combining these observations,

$$(A.8) \quad \frac{\partial f}{\partial \xi}(x(\xi), \xi) = \frac{\partial}{\partial \xi} \{f(x, \xi) + \lambda^T g(x, \xi) + \mu^T h(x, \xi)\} \Bigg|_{\substack{x=x(\xi) \\ \lambda=\lambda(\xi) \\ \mu=\mu(\xi)}}.$$

Theorem A.5 completes the proof. \square

Gauvin and Tolle [12] obtain (A.8) under weaker assumptions, although the feasible set is required to satisfy a uniform compactness condition. Also Armacost and Fiacco [2] give (A.8), but require the so-called strict complementary slackness condition which is not satisfied in our applications to the dual integrand.

Appendix 3. Exact solutions. Consider the problem

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} \int_0^1 [x(t)^2 + u(t)^2] dt \\ &\text{subject to} \quad x'(t) = u(t), \quad u(t) \leq 1 \quad \text{almost everywhere,} \\ &\quad \quad \quad x(0) = \frac{1+3e}{2(1-e)}, \quad (x, u) \in \mathcal{X}. \end{aligned}$$

This problem's solution, computed in [23], is given below.

Region 1. $0 \leq t \leq \frac{1}{2}$.

$$x(t) = t + \frac{1+3e}{2(1-e)}, \quad u(t) = 1, \quad p(t) = \frac{t^2}{2} + \frac{(1+3e)}{2(1-e)}t + \frac{13e-5}{8(e-1)}.$$

Region 2. $\frac{1}{2} \leq t \leq 1$.

$$x(t) = \frac{e^t + e^{2-t}}{\sqrt{e(1-e)}}, \quad u(t) = p(t) = \frac{e^t - e^{2-t}}{\sqrt{e(1-e)}}.$$

The optimal cost is

$$\frac{55e^2 - 2e - 5}{48(e-1)^2}.$$

Next, consider the problem

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} \int_0^1 [x(t)^2 + u(t)^2] dt \\ & \text{subject to} \quad x'(t) = u(t), \quad u(t) \leq 1 \quad \text{almost everywhere,} \\ & \quad \quad \quad x(t) \leq \frac{2\sqrt{e}}{1-e} \quad \text{for all } t \in [0, 1], \\ & \quad \quad \quad x(0) = \frac{5e+3}{4(1-e)}, \quad (x, u) \in \mathcal{X}. \end{aligned}$$

This problem's solution, computed in [23], is given below.

Region 1. $0 \leq t \leq \frac{1}{4}$.

$$x(t) = t - \frac{1}{4} + \frac{1+e}{1-e}, \quad u(t) = 1, \quad p(t) = \frac{33}{32} + \frac{1+e}{1-e} \left(t - \frac{1}{4} \right) + \frac{1}{4} t(2t-1).$$

Region 2. $\frac{1}{4} \leq t \leq \frac{3}{4}$.

$$x(t) = \frac{e^{t-1/4}}{1-e} (1 + e^{3/2-2t}), \quad u(t) = p(t) = \frac{e^{t-1/4}}{1-e} (1 - e^{3/2-2t}).$$

Region 3. $\frac{3}{4} \leq t \leq 1$.

$$x(t) = \frac{2\sqrt{e}}{1-e}, \quad u(t) = p(t) = 0.$$

The optimal cost is

$$\frac{49}{384} + \frac{e+1}{2(e-1)} + \frac{x(0)}{32} + \frac{x(0)^2 + b^2}{8}$$

where $b = 2\sqrt{e}/(1-e)$.

Acknowledgments. The authors are grateful for many helpful comments from the referee. In particular, a streamlined proof of Lemma 3.2 and Theorem 4.1 is mainly due to the referee. Experiments with the IMSL conjugate gradient routine reported at the start of § 11 were conducted by Y. Ting and E. Yeh, graduate students at the Pennsylvania State University.

REFERENCES

- [1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] R. L. ARMACOST AND A. V. FIACCO, *Sensitivity analysis for parametric nonlinear programming using penalty methods*, in *Computers and Mathematical Programming*, Special Publication 502, National Bureau of Standards, Washington, DC, pp. 261-269.
- [3] A. V. BALAKRISHNAN, *On a new computing technique in optimal control*, this Journal, 6 (1968), pp. 149-173.
- [4] D. P. BERTSEKAS, *Multiplier methods: A survey*, in *Proc. IFAC 6th Triennial World Congress*, Boston, 1975.
- [5] W. E. BOSARGE, JR. AND O. G. JOHNSON, *Error bounds of high order accuracy for the state regulator problem via piecewise polynomial approximations*, this Journal, 9 (1971), pp. 15-28.
- [6] W. E. BOSARGE, JR., O. G. JOHNSON AND C. L. SMITH, *A direct method approximation to the linear parabolic regulator problem over multivariate spline bases*, SIAM J. Numer. Anal., 10 (1973), pp. 35-49.

- [7] C. CASTAING, *Sur les multi-applications mesurables*, Thèse, Univ. de Caën, France, 1967.
- [8] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1980.
- [9] R. COURANT, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc., 49 (1943), pp. 1–23.
- [10] J. CULLUM, *Penalty functions and nonconvex continuous optimal control problems*, in *Computing Methods in Optimization Problems*, 2, L. A. Zadeh, L. W. Neustadt and A. V. Balakrishnan, eds., Academic Press, New York, 1969, pp. 55–67.
- [11] A. V. FIACCO and W. P. HUTZLER, *Basic results in the development of sensitivity and stability analysis in nonlinear programming*, Computers and Operations Research, to appear.
- [12] J. GAUVIN and J. W. TOLLE, *Differential stability in nonlinear programming*, this Journal, 15 (1977), pp. 294–311.
- [13] J. GUDDAT, *On some actual questions in parametric optimization*, in *Mathematical Methods in Operations Research*, Bulgarian Academy of Sciences, Sofia, 1981, pp. 39–54.
- [14] W. W. HAGER, *The Ritz–Trefftz method for state and control constrained optimal control problems*, SIAM J. Numer. Anal., 12 (1975), pp. 854–867.
- [15] ———, *Lipschitz continuity for constrained processes*, this Journal, 17 (1979), pp. 321–338.
- [16] ———, *Convex control and dual approximations*, Control Cybernet., 8 (1979), Part I: pp. 5–22, Part II: pp. 73–86.
- [17] ———, *Inequalities and approximation*, in *Constructive Approaches to Mathematical Models*, C. V. Coffman and G. J. Fix, eds., Academic Press, New York, 1979, pp. 189–202.
- [18] W. W. HAGER AND G. D. IANCULESCU, *Semi-dual approximations in optimal control: Quadratic cost*, in *Free Boundary Problems*, Vol. II (Pavia, 1979), Ist. Naz. Alta Mat. Francesco Severi, Rome, 1980, pp. 321–332.
- [19] W. W. HAGER AND S. K. MITTER, *Lagrange duality theory for convex control problems*, this Journal, 14 (1976), pp. 843–856.
- [20] W. W. HAGER AND G. STRANG, *Free boundaries and finite elements in one dimension*, Math. Comp., 29 (1975), pp. 1020–1031.
- [21] M. R. HESTENES, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.
- [22] L. HÖRMANDER, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1969.
- [23] G. D. IANCULESCU, *Semi-dual approximations for convex optimal control problems*, Ph.D. dissertation, Carnegie–Mellon Univ., Pittsburgh, PA, 1979.
- [24] D. H. JACOBSON AND M. M. LELE, *A transformation technique for optimal control problems with a state variable inequality constraint*, IEEE Trans. Automat. Control, 14 (1969), pp. 457–464.
- [25] L. S. LASDON, A. D. WARREN AND R. K. RICE, *An interior penalty method for inequality constrained optimal control problems*, IEEE Trans. Automat. Control, 12 (1967), pp. 388–395.
- [26] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control*, John Wiley, New York, 1967.
- [27] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, S. K. Mitter, transl., Springer-Verlag, New York, 1971.
- [28] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
- [29] ———, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.
- [30] K. MALANOWSKI, *On the regularity of solutions to optimal control problems for systems with control appearing linearly*, Arch. Automat. Telemekh., 23 (1978), pp. 227–242.
- [31] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.
- [32] F. H. MATHIS AND G. W. REDDIEN, *Ritz–Trefftz approximations in optimal control*, this Journal, 17 (1979), pp. 307–310.
- [33] R. K. MEHRA AND R. E. DAVIS, *A generalized gradient method for optimal control problems with inequality constraints and singular arcs*, IEEE Trans. Automat. Control, 17 (1972), pp. 69–79.
- [34] E. MICHAEL, *Continuous selections*, I, Ann. of Math., 63 (1956), pp. 361–382.
- [35] J. MOSSINO, *An application of duality to distributed optimal control problems with constraints on the state and the control*, J. Math. Anal. Appl., 50 (1975), pp. 223–242.
- [36] ———, *Approximation numérique de problèmes de contrôle optimal avec contrainte sur le contrôle et sur l'état*, Calcolo, 13 (1976), pp. 21–62.
- [37] J. A. NITSCHKE, *Ein Kriterium für die Quasi-optimalität des Ritzschen Verfahrens*, Numer. Math., 11 (1968), pp. 346–348.
- [38] J. T. ODEN AND J. N. REDDY, *An Introduction to the Mathematical Theory of Finite Elements*, Interscience, New York, 1976.
- [39] O. PIRONNEAU AND E. POLAK, *A dual method for optimal control problems with initial and final boundary constraints*, this Journal, 11 (1973), pp. 534–549.
- [40] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1965.

- [41] M. J. D. POWELL, *A method of nonlinear constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, New York, 1972.
- [42] S. M. ROBINSON, *Stability theory for systems of inequalities, Part II: Differentiable nonlinear systems*, SIAM J. Numer. Anal., 12 (1976), pp. 497–513.
- [43] ———, *Strongly regular generalized equations*, Mathematics Research Center Report 1877, Univ. of Wisconsin, Madison, WI, 1978.
- [44] S. M. ROBINSON AND R. H. DAY, *A sufficient condition for continuity of optimal sets in mathematical programming*, J. Math. Anal. Appl., 45 (1974), pp. 506–511.
- [45] R. T. ROCKAFELLAR, *Integrals which are convex functionals*, II, Pacific J. Math., 39 (1971), pp. 439–469.
- [46] ———, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1972.
- [47] ———, *State constraints in convex control problems of Bolza*, this Journal, 10 (1972), pp. 691–715.
- [48] ———, *A dual approach to solving nonlinear programming problems by unconstrained optimization*, Math. Programming, 5 (1973), pp. 354–373.
- [49] ———, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, this Journal, 12 (1974), pp. 268–285.
- [50] ———, *Conjugate Duality and Optimization*, CBMS Regional Conference Series in Applied Mathematics 16, Society for Industrial and Applied Mathematics, Philadelphia, 1974.
- [51] ———, *Integral functionals, normal integrands and measurable selections*, in Nonlinear Operators and the Calculus of Variations, Lucien Waelbroeck, ed., Springer-Verlag, New York, 1976, pp. 157–207.
- [52] ———, *Duality in optimal control*, in Mathematical Control Theory, W. A. Coppel, ed., Springer-Verlag, New York, 1978, pp. 219–257.
- [53] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
- [54] D. L. RUSSELL, *Penalty functions and bounded phase coordinate control*, this Journal, 2 (1965), pp. 409–422.
- [55] G. STRANG AND G. J. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [56] D. D. THOMPSON AND R. A. VOLZ, *The linear quadratic cost problem with linear state constraints and the nonsymmetric Riccati equation*, this Journal, 13 (1975), pp. 110–145.
- [57] F. A. VALENTINE, *The problem of Lagrange with differential inequalities as added side conditions*, in Contributions to the Calculus of Variations, Univ. of Chicago Press, Chicago, 1937, pp. 407–448.

OPTIMAL CONTROL IN SOME VARIATIONAL INEQUALITIES*

F. MIGNOT† AND J. P. PUEL†

Abstract. We study a nonconvex, nondifferentiable problem of optimal control where the state of the system is defined by an elliptic variational inequality with obstacle, and where the cost function is quadratic with respect to the state and the control. We show the existence of an optimal control u with which is associated a state $y(u)$. When there is no constraint on the control variable, we give necessary conditions of first order satisfied by the triple $(u, y(u), p)$ where p is an adjoint state associated with the problem.

Key words. optimal control, nonlinear optimality system, conical derivative, variational inequalities

Introduction. We are going to consider an optimal control problem in which the state of the system is defined as the (unique) solution of a stationary variational inequality.

The main difficulty comes from the fact that the mapping between the control and the state is not differentiable but only Lipschitz-continuous and so it is not easy to get optimality conditions of first order which make sense and which correctly describe the situation.

This problem has been already considered from both the theoretical and numerical points of view by many people, for example Yvon [9], Mignot [6], Barbu [1], [2], Saguez [7] and Zrikem [10]. They have used either an approximation of the variational inequality by penalization, or the differentiability almost everywhere for Lipschitz continuous mappings or the generalized gradient. Here, using the conical derivative (cf. Mignot [6]), in the case where there is no constraint on the control, we shall obtain necessary conditions of first order including strictly the ones obtained by Barbu [2] for example.

In § 1 we describe the problem; the main results are given in § 2; in § 3 we give auxiliary results and prove the main theorems; in § 4 we make some complementary remarks and state some open problems.

1. Statement of the problem. In the interests of clarity, we shall not consider the most general abstract situation, and we leave to the reader the possibility of adapting the proofs to some connected problems.

Let Ω be a bounded domain of R^n and let Γ be its boundary.

We consider a Hilbert space V such that

$$H_0^1(\Omega) \hookrightarrow V \hookrightarrow H^1(\Omega)$$

and such that if $u \in V$, $u^+ \in V$.

We denote by $((\cdot, \cdot))$ and $\|\cdot\|$ the scalar product and the associated norm in V .

Let us consider the bilinear form $a(\cdot, \cdot)$ defined on $V \times V$ by

$$(1.1) \quad a(\phi, \psi) = \sum_{i,j=1}^n \int_{\Omega} a_{ij} \frac{\partial \phi}{\partial x_j} \frac{\partial \psi}{\partial x_i} dx + \sum_{i=1}^n \int_{\Omega} b_i \frac{\partial \phi}{\partial x_i} \psi dx + \int_{\Omega} c \phi \psi dx,$$

where a_{ij} , b_i , c belong to $L^\infty(\Omega)$. The bilinear form $a(\cdot, \cdot)$ is continuous on $V \times V$ and we shall assume it is coercive, i.e.,

$$(1.2) \quad \exists \alpha > 0, \forall \phi \in V, \quad a(\phi, \phi) \geq \alpha \|\phi\|^2.$$

* Received by the editors January 25, 1983, and in revised form March 10, 1983.

† Laboratoire D'Analyse Numérique (LA 189), Université P. et M. Curie, Tour 55-65—5ème étage, 4 place Jussieu, 75230 Paris Cedex 05, France.

If $\langle \cdot, \cdot \rangle$ is the duality between V' and V , we have

$$(1.3) \quad \forall \phi, \psi \in V, \quad a(\phi, \psi) = \langle A\phi, \psi \rangle, \quad \text{where } A \in \mathcal{L}(V, V').$$

Now define

$$(1.4) \quad K = \{\phi \mid \phi \in V, \phi \geq 0 \text{ a.e. in } \Omega\}.$$

The set K is closed, convex and nonempty in V .

We are now able to define correctly the control problem. Let f be given in V' , and let U_{ad} be a closed convex subset of $L^2(\Omega)$. For each $v \in U_{\text{ad}}$ we define $y = y(v)$ (the state of the system) as the solution of the variational inequality

$$(1.5) \quad a(y, \phi - y) \geq \langle f + v, \phi - y \rangle \quad \forall \phi \in K, \quad y \in K.$$

We can interpret (1.5) as follows:

$$(1.6) \quad Ay = f + v + \xi, \quad y \geq 0, \quad \xi \geq 0, \quad \langle \xi, y \rangle = 0.$$

We know by classical arguments [3], [5] that (1.5) has a unique solution.

Now, for $z_d \in L^2(\Omega)$ and $N > 0$, we define the cost function J by

$$(1.7) \quad J(v) = \frac{1}{2} \int_{\Omega} (y(v) - z_d)^2 dx + \frac{N}{2} \int_{\Omega} (v)^2 dx,$$

and we look for v_0 (optimal control) such that

$$(1.8) \quad v_0 \in U_{\text{ad}}, \quad J(v_0) = \min_{v \in U_{\text{ad}}} J(v).$$

Remark 1.1. We could have considered various examples of convex sets, of control and of cost functions in which we can obtain analogous results without additional difficulty in the proofs.

In particular we can consider the following examples.

Example 1.1.

$$V = H^1(\Omega), \quad K = \{\phi \mid \phi \in V, \phi \geq 0 \text{ a.e. on } \Gamma\}.$$

Then for the same type of bilinear form and the same control, we get the Signorini problem, and we may consider the following cost function (with $z_d \in L^2(\Gamma)$):

$$(1.9) \quad J(v) = \frac{1}{2} \int_{\Gamma} (y(v) - z_d)^2 d\Gamma + \frac{N}{2} \int_{\Omega} (v)^2 dx.$$

Example 1.2. If $y(v)$ is defined as in Example 1.1 we can consider another cost function (with $z_d \in H^{-1/2}(\Gamma)$)

$$(1.10) \quad J(v) = \frac{1}{2} \left| \frac{\partial y}{\partial \nu_A}(v) - z_d \right|_{H^{-1/2}(\Gamma)}^2 + \frac{N}{2} \int_{\Omega} (v)^2 dx,$$

where $\partial/\partial \nu_A$ denotes the conormal derivative associated with A .

Example 1.3. If Ω is a bounded regular open set of R^n such that its boundary Γ is the union of two connected components Γ_0 and Γ_1 , for $v \in L^2(\Gamma_1)$ we consider $y(v)$ the solution of

$$(1.11) \quad \begin{aligned} Ay(v) &= f \quad \text{in } \Omega, \\ y(v) &\geq 0, \quad \frac{\partial y(v)}{\partial \nu_A} \geq 0, \quad \frac{\partial y(v)}{\partial \nu_A} \cdot y(v) = 0 \quad \text{on } \Gamma_0, \\ y(v) &= v \quad \text{on } \Gamma_1, \end{aligned}$$

and the cost function ($z_d \in H^{-1/2}(\Gamma_0)$)

$$(1.12) \quad J(v) = \frac{1}{2} \left| \frac{\partial y(v)}{\partial \nu_A} - z_d \right|_{H^{-1/2}(\Gamma_0)}^2 + \frac{N}{2} \int_{\Gamma_1} (v)^2 d\Gamma.$$

In this case, we have to define carefully what we mean by a solution of (1.11) with $v \in L^2(\Gamma_1)$.

Example 1.4. For $f \in R$, $v \in R$, $z_d \in R$, we consider

$$(1.13) \quad y(v) = (f + v)^+,$$

$$(1.14) \quad J(v) = \frac{1}{2} (y(v) - z_d)^2 + \frac{N}{2} (v)^2.$$

All that follows can be adapted to this very simple interesting situation, which contains the main difficulties and which will give rise to some counterexamples.

2. Main results. First we get a simple existence result for an optimal control.

THEOREM 2.1. *There exists an optimal control $v_0 \in U_{ad}$ (and, in general, there is no uniqueness).*

In order to get optimality conditions of first order, we shall assume that $U_{ad} = L^2(\Omega)$.

If y is solution of (1.5) we can define:

$$(2.1) \quad Zy = \{x | x \in \Omega, y(x) = 0\} \quad (\text{defined up to a set of zero capacity}).$$

$$(2.2) \quad Sy = \{\phi | \phi \in V, \phi \geq 0 \text{ on } Zy, \langle \xi, \phi \rangle = 0\},$$

where $\xi = Ay - f - v$ is given by (1.6).

THEOREM 2.2. *An optimal control v_0 satisfies the following:*

- (i) $v_0 \in V$.
- (ii) *If $y_0 = y(v_0)$, there exists p_0 such that:*

$$(2.3) \quad \begin{aligned} p_0 &\in Sy_0, \\ \forall \psi \in Sy_0, \quad a(\psi, p_0) &\leq \int_{\Omega} (y_0 - z_d) \psi dx, \\ p_0 + Nv_0 &= 0. \end{aligned}$$

Remark 2.1. If we define $(Sy_0^{a*})^0$ (polar cone of Sy_0 with respect to the adjoint form $a^*(\cdot, \cdot)$) by

$$(2.4) \quad (Sy_0^{a*})^0 = \{\phi | \phi \in V, \forall \psi \in Sy_0, a(\psi, \phi) \leq 0\},$$

we can write (2.3) as follows:

$$(2.5) \quad p_0 \in Sy_0, \quad p_0 - A^{*-1}(y_0 - z_d) \in (Sy_0^{a*})^0, \quad p_0 + Nv_0 = 0.$$

Now, eliminating the adjoint state p_0 we obtain:

COROLLARY 2.3. *There exists at least one solution (y, v) of the following system:*

$$(2.6) \quad \begin{aligned} a(y, \phi - y) &\geq \int_{\Omega} (f + v)(\phi - y) dx \quad \forall \phi \in K, \quad y \in K, \\ a(\psi, v) &\geq -\frac{1}{N} \int_{\Omega} (y - z_d) \psi dx \quad \forall \psi \in s_y, \quad -v \in Sy, \end{aligned}$$

and (y_0, v_0) is one such solution.

3. Proofs of the results.

3.1. Proof of Theorem 2.1. We know that $J(v) \geq 0 \forall v \in U_{\text{ad}}$.

Let j be the infimum value of $J(v)$ for $v \in U_{\text{ad}}$ and let $(v_n)_{n \in N}$ be a minimizing sequence. We then have

$$\lim_{n \rightarrow \infty} J(v_n) = j = \inf_{v \in U_{\text{ad}}} J(v).$$

As N is strictly positive, $(v_n)_{n \in N}$ is a bounded sequence in $U_{\text{ad}} \subset L^2(\Omega)$ and we can extract a weakly convergent subsequence $(v_{n_k})_{k \in N}$ such that

$$v_{n_k} \rightarrow v_0 \quad \text{in } L^2(\Omega) \text{ weakly, as } k \rightarrow +\infty.$$

Then $v_0 \in U_{\text{ad}}$ because U_{ad} is closed and convex.

As Ω is bounded, the injection from $L^2(\Omega)$ into V' is compact and so

$$v_{n_k} \rightarrow v_0 \quad \text{in } V' \text{ strongly, as } k \rightarrow +\infty.$$

Then we have

$$y(v_{n_k}) \rightarrow y(v_0) = y_0 \quad \text{in } V, \quad \text{as } k \rightarrow \infty.$$

Using the lower semicontinuity for the weak topology of $L^2(\Omega)$ of $v \rightarrow \int_{\Omega} (v)^2 dx$, we get

$$j = \liminf_{k \rightarrow \infty} J(v_{n_k}) \geq J(v_0) \quad \text{and} \quad J(v_0) = \min_{v \in U_{\text{ad}}} J(v).$$

3.2. Proof of Theorem 2.2. We first give the results obtained by approximating the variational inequality by a penalized equation. This method has been used by Barbu [1], [2] and Mignot and Tartar [8], but as we shall see, it does not give the result of Theorem 2.2. Nevertheless it shows the important fact that the optimal control v_0 belongs to V .

For $\delta > 0$, let us consider

$$\beta^{\delta}(r) = \begin{cases} r + \frac{\delta}{2} & \text{if } r \leq -\delta, \\ -\frac{1}{2\delta} r^2 & \text{if } -\delta \leq r \leq 0, \\ 0 & \text{if } r \geq 0. \end{cases}$$

For $\varepsilon > 0$, we denote by $y_{\varepsilon}(v)$ the unique solution (which does exist) of the penalized equation

$$(3.1) \quad Ay_{\varepsilon}(v) + \frac{1}{\varepsilon} \beta^{\delta}(y_{\varepsilon}(v)) = f + v, \quad y_{\varepsilon}(v) \in V.$$

Using a trick of Barbu [2], we define an adapted cost function

$$(3.2) \quad J_{\varepsilon}(v) = \frac{1}{2} \int_{\Omega} (y_{\varepsilon}(v) - z_d)^2 dx + \frac{N}{2} \int_{\Omega} (v)^2 dx + \frac{1}{2} \int_{\Omega} (v - v_0)^2 dx,$$

where v_0 is a solution of (1.8), given by Theorem 2.1.

We can now obtain easily the following result (the proof is classical):

THEOREM 3.1. *For each $\varepsilon > 0$, there exists $v_{\varepsilon} \in L^2(\Omega)$ such that*

$$(3.3) \quad J_{\varepsilon}(v_{\varepsilon}) = \min_{v \in L^2(\Omega)} J_{\varepsilon}(v).$$

Moreover we have

$$\begin{aligned}
 (3.4) \quad & Ay_\varepsilon + \frac{1}{\varepsilon} \beta^\delta(y_\varepsilon) = f + v_\varepsilon, \quad y_\varepsilon \in V, \\
 & A^* p_\varepsilon + \frac{1}{\varepsilon} \beta^{\delta'}(y_\varepsilon) \cdot p_\varepsilon = (y_\varepsilon - z_d), \quad p_\varepsilon \in V, \\
 & p_\varepsilon + Nv_\varepsilon + (v_\varepsilon - v_0) = 0.
 \end{aligned}$$

Using Theorem 3.1 we can derive some estimates and convergence results when $\varepsilon \rightarrow 0$.

THEOREM 3.2. When $\varepsilon \rightarrow 0$, we have

$$(3.5) \quad v_\varepsilon \rightarrow v_0 \text{ in } L^2(\Omega) \text{ strongly,} \quad y_\varepsilon \rightarrow y_0 \text{ in } V \text{ strongly,} \quad p_\varepsilon \rightarrow p_0 \text{ in } V \text{ weakly,}$$

with

$$(3.6) \quad p_0 + Nv_0 = 0$$

and

$$\begin{aligned}
 (3.7) \quad & Ay_0 = f + v_0 + \xi_0, \\
 & y_0 \geq 0, \quad \xi_0 \geq 0, \quad \langle \xi_0, y_0 \rangle = 0, \\
 & A^* p_0 = (y_0 - z_d) + \eta_0, \\
 & \langle \eta_0, y_0 \rangle = \langle \xi_0, p_0 \rangle = 0, \\
 & \langle \eta_0, p_0 \rangle \leq 0.
 \end{aligned}$$

Remarks 3.1. 1) In the following we shall not use (3.7) directly, but we shall use (3.6) which shows that $v_0 \in V$.

2) In fact we shall obtain directly strictly more than (3.7) as will be shown in § 4 via a counterexample.

Proof. We know that for v fixed in $L^2(\Omega)$, when $\varepsilon \rightarrow 0$, $y_\varepsilon(v) \rightarrow y(v)$ in V strongly (because $(1/\varepsilon)\beta^\delta(\cdot)$ is a penalization adapted to the convex set K).

From (3.3), we have

$$J_\varepsilon(v_\varepsilon) \leq J_\varepsilon(v_0) = \frac{1}{2} \int_\Omega (y_\varepsilon(v_0) - z_d)^2 dx + \frac{N}{2} \int_\Omega (v_0)^2 dx.$$

Then, $J_\varepsilon(v_0) \rightarrow J(v_0)$ when $\varepsilon \rightarrow 0$, and

$$(3.8) \quad \limsup_{\varepsilon \rightarrow 0} J_\varepsilon(v_\varepsilon) \leq J(v_0).$$

Moreover $(v_\varepsilon)_{\varepsilon > 0}$ is bounded in $L^2(\Omega)$ (independently of ε) and we can extract a subsequence (still denoted by v_ε) such that

$$v_\varepsilon \rightarrow \overline{v_0} \text{ in } L^2(\Omega) \text{ weakly if } \varepsilon \rightarrow 0.$$

Then $v_\varepsilon \rightarrow \overline{v_0}$ in V' strongly, if $\varepsilon \rightarrow 0$, and we could easily show that

$$y_\varepsilon(v_\varepsilon) \rightarrow y(\overline{v_0}) \text{ in } V \text{ if } \varepsilon \rightarrow 0.$$

Therefore,

$$\begin{aligned}\liminf_{\varepsilon \rightarrow 0} J_\varepsilon(v_\varepsilon) &\geq \frac{1}{2} \int_{\Omega} (y(\bar{v}_0) - z_d)^2 dx + \frac{N}{2} \int_{\Omega} (\bar{v}_0)^2 dx + \frac{1}{2} \int_{\Omega} (\bar{v}_0 - v_0)^2 dx \\ &= J(\bar{v}_0) + \frac{1}{2} \int_{\Omega} (\bar{v}_0 - v_0)^2 dx \\ &\geq J(v_0) + \frac{1}{2} \int_{\Omega} (\bar{v}_0 - v_0)^2 dx \quad \text{from (1.8)}.\end{aligned}$$

From (3.8) and (3.2) we obtain

$$\begin{aligned}J_\varepsilon(v_\varepsilon) &\rightarrow J(v_0) \quad \text{if } \varepsilon \rightarrow 0, \quad \bar{v}_0 = v_0, \\ v_\varepsilon &\rightarrow v_0 \quad \text{in } L^2(\Omega) \text{ strongly} \quad \text{if } \varepsilon \rightarrow 0.\end{aligned}$$

Then $y_\varepsilon = y_\varepsilon(v_\varepsilon) \rightarrow y_0 = y(v_0)$ in V strongly, if $\varepsilon \rightarrow 0$. Multiplying the second equation of (3.4) by p_ε , and using the fact that $\beta^{\delta'}(y_\varepsilon) \geq 0$, we obtain that p_ε is bounded in V , independently of ε . After extraction of a subsequence, we have

$$p_\varepsilon \rightarrow p_0 \quad \text{in } V \text{ weakly} \quad \text{if } \varepsilon \rightarrow 0.$$

From the last equation of (3.4) we get

$$p_0 + Nv_0 = 0,$$

and then the whole sequence p_ε converges to $p_0 = -Nv_0$.

This gives the first part of Theorem 3.2. As already mentioned, the second part of Theorem 3.2 will be a consequence of our general result, but we shall prove it directly here, assuming that $f \in L^2(\Omega)$.

Let us call

$$\xi_\varepsilon = -\frac{1}{\varepsilon} \beta^\delta(y_\varepsilon) = Ay_\varepsilon - (f + v_\varepsilon), \quad \eta_\varepsilon = -\frac{1}{\varepsilon} \beta^{\delta'}(y_\varepsilon) \cdot p_\varepsilon = A^*p_\varepsilon - (y_\varepsilon - z_d).$$

From (3.5), we know that if $\varepsilon \rightarrow 0$,

$$\begin{aligned}\xi_\varepsilon &\rightarrow \xi_0 \quad \text{in } V', \quad \text{where } \xi_0 = Ay_0 - (f + v_0), \\ \eta_\varepsilon &\rightarrow \eta_0 \quad \text{in } V' \text{ weakly, where } \eta_0 = A^*p_0 - (y_0 - z_d).\end{aligned}$$

We have $\langle \eta_\varepsilon, y_\varepsilon^+ \rangle = 0$, because of the definition of β^δ .

When $\varepsilon \rightarrow 0$,

$$\eta_\varepsilon \rightarrow \eta_0 \quad \text{in } V' \text{ weakly,} \quad y_\varepsilon^+ \rightarrow y_0^+ = y_0 \quad \text{in } V \text{ strongly.}$$

Then $\langle \eta_\varepsilon, y_\varepsilon^+ \rangle \rightarrow \langle \eta_0, y_0 \rangle$ in R , and

$$\langle \eta_0, y_0 \rangle = 0.$$

Now we have

$$\begin{aligned}\langle \xi_\varepsilon, p_\varepsilon \rangle &= \frac{1}{\varepsilon} \int_{\Omega} \beta^\delta(y_\varepsilon) p_\varepsilon dx \\ &= \frac{1}{\varepsilon} \left[\int_{\{y_\varepsilon \leq -\delta\}} \left(y_\varepsilon + \frac{\delta}{2} \right) p_\varepsilon dx - \frac{1}{2\delta} \int_{\{-\delta \leq y_\varepsilon \leq 0\}} p_\varepsilon y_\varepsilon^2 dx \right],\end{aligned}$$

and

$$\langle \eta_\varepsilon, y_\varepsilon \rangle = \frac{1}{\varepsilon} \int_{\Omega} \beta^{\delta'}(y_\varepsilon) \cdot p_\varepsilon \cdot y_\varepsilon dx = \frac{1}{\varepsilon} \left[\int_{\{y_\varepsilon \leq -\delta\}} p_\varepsilon y_\varepsilon dx - \frac{1}{\delta} \int_{\{-\delta \leq y_\varepsilon \leq 0\}} y_\varepsilon^2 p_\varepsilon dx \right].$$

Then

$$\langle \xi_\varepsilon, p_\varepsilon \rangle - \frac{1}{2} \langle \eta_\varepsilon, y_\varepsilon \rangle = \frac{1}{2\varepsilon} \int_{\{y_\varepsilon \leq -\delta\}} (y_\varepsilon + \delta) p_\varepsilon dx,$$

and

$$\left| \langle \xi_\varepsilon, p_\varepsilon \rangle - \frac{1}{2} \langle \eta_\varepsilon, y_\varepsilon \rangle \right| \leq \frac{1}{\varepsilon} \left(\int_{\{y_\varepsilon \leq -\delta\}} ((y_\varepsilon)^2 + \delta^2) dx \right)^{1/2} \cdot \left(\int_{\{y_\varepsilon \leq -\delta\}} (p_\varepsilon)^2 dx \right)^{1/2}.$$

Multiplying the first equation of (3.4) by $(1/\varepsilon)\beta^\delta(y_\varepsilon)$, we see that $(1/\varepsilon)|\beta^\delta(y_\varepsilon)|_{L^2(\Omega)}$ is bounded, and so is $(1/\varepsilon)(\int_{\{y_\varepsilon \leq -\delta\}} ((y_\varepsilon)^2 + \delta^2) dx)^{1/2}$.

As $V \subset L^q(\Omega)$ with $q > 2$, we have

$$\begin{aligned} \left(\int_{\{y_\varepsilon \leq -\delta\}} (p_\varepsilon)^2 dx \right)^{1/2} &\leq \left(\int_{\{y_\varepsilon \leq -\delta\}} (p_\varepsilon)^q dx \right)^{1/q} \cdot [\text{meas } \{y_\varepsilon \leq -\delta\}]^{q/2(q-2)} \\ &\leq C \cdot \|p_\varepsilon\| [\text{meas } \{y_\varepsilon \leq -\delta\}]^{q/2(q-2)}. \end{aligned}$$

As $\|p_\varepsilon\|$ is bounded, if we show that $[\text{meas } \{y_\varepsilon \leq -\delta\}] \rightarrow 0$ when $\varepsilon \rightarrow 0$, we have

$$\langle \xi_\varepsilon, p_\varepsilon \rangle - \frac{1}{2} \langle \eta_\varepsilon, y_\varepsilon \rangle \rightarrow 0 \quad \text{and} \quad \langle \xi_\varepsilon, p_\varepsilon \rangle \rightarrow 0 \quad \text{if } \varepsilon \rightarrow 0,$$

because $\langle \eta_\varepsilon, y_\varepsilon \rangle \rightarrow 0$ if $\varepsilon \rightarrow 0$.

We know that

$$\frac{1}{\varepsilon^2} \int_{\{y_\varepsilon \leq -\delta\}} y_\varepsilon^2 dx \leq M.$$

So

$$\frac{\delta^2}{\varepsilon^2} \int_{\{y_\varepsilon \leq -\delta\}} dx \leq M \quad \text{and} \quad [\text{meas } \{y_\varepsilon \leq -\delta\}] \leq \frac{M}{\delta^2} \varepsilon^2,$$

so we have

$$\langle \xi_\varepsilon, p_\varepsilon \rangle \rightarrow 0 \quad \text{if } \varepsilon \rightarrow 0 \quad \text{and therefore } \langle \xi_0, p_0 \rangle = 0.$$

Multiplying the second equation in (3.4) by p_ε , we get

$$a(p_\varepsilon, p_\varepsilon) - \int_{\Omega} (y_\varepsilon - z_d) p_\varepsilon dx = -\frac{1}{\varepsilon} \int_{\Omega} \beta^{\delta'}(y_\varepsilon) \cdot p_\varepsilon^2 dx \leq 0.$$

When $\varepsilon \rightarrow 0$, $p_\varepsilon \rightarrow p_0$ in V weakly, and $y_\varepsilon \rightarrow y_0$ in V strongly. Then

$$a(p_0, p_0) - \int_{\Omega} (y_0 - z_d) p_0 dx \leq \liminf_{\varepsilon \rightarrow 0} \left[a(p_\varepsilon, p_\varepsilon) - \int_{\Omega} (y_\varepsilon - z_d) p_\varepsilon dx \right] \leq 0$$

and

$$\langle \eta_0, p_0 \rangle \leq 0.$$

This finishes the proof of Theorem 3.2.

Now, using the information $v_0 \in V$ which is a regularity result on the optimal control, we are going to give a direct proof of Theorem 2.2.

We know (cf. Mignot [6]) that the mapping $v \rightarrow y(v)$ possesses at each point v a conical derivative $w \rightarrow Dy_v(w)$ such that, for all $w \in V'$, we have

$$(3.9) \quad \begin{aligned} Dy_v(w) &\in S_{y(v)}, \\ \forall \phi \in S_{y(v)}, \quad a(Dy_v(w), \phi - Dy_v(w)) &\geq \langle w, \phi - Dy_v(w) \rangle, \end{aligned}$$

where $S_{y(v)}$ is defined by (2.2).

Therefore, the mapping $v \rightarrow J(v)$ possesses at each point v a conical derivative $w \rightarrow DJ_v(w)$ defined by

$$(3.10) \quad DJ_v(w) = \int_{\Omega} (y(v) - z_d) Dy_v(w) dx + N \int_{\Omega} v \cdot w dx.$$

LEMMA 3.1. *If v_0 is an optimal control, we have*

$$(3.11) \quad \forall w \in V', \quad DJ_{v_0}(w) \geq 0.$$

Proof. It is evident that

$$\forall w \in L^2(\Omega), \quad DJ_{v_0}(w) \geq 0.$$

Then it is easy to prove (3.11), because $L^2(\Omega)$ is dense in V' and because $w \rightarrow DJ_{v_0}(w)$ is continuous from V' into R .

Remark 3.2. The condition $DJ_{v_0}(w) \geq 0$ means that at the point v_0 in each half direction w , the functional $J(\cdot)$ does not decrease strictly, up to the first order. So it seems to be a "good" optimality condition.

THEOREM 3.3. *If $v_0 \in V$, the optimality condition (3.11) holds at the point v_0 if and only if there exists p_0 such that*

$$(3.12) \quad p_0 \in S_{y(v_0)}, \quad p_0 - A^{*-1}(y(v_0) - z_d) \in (S_{y(v_0)}^{a*})^0, \quad p_0 + Nv_0 = 0.$$

Remarks 3.3. 1) Theorem 2.2 follows immediately from Theorem 3.3 and Lemma 3.1.

2) Equation (3.12) and the definition of $y(v_0)$ include (3.7).

Proof of Theorem 3.3. For $\xi \in V$ and $y \in K$, let us call $P_y(\xi)$ the solution of

$$(3.13) \quad a(P_y(\xi), \phi - P_y(\xi)) \geq a(\xi, \phi - P_y(\xi)) \quad \forall \phi \in S_y, \quad P_y(\xi) \in S_y,$$

and $P_y^*(\xi)$ the solution of

$$(3.14) \quad a(\phi - P_y^*(\xi), P_y^*(\xi)) \geq a(\phi - P_y^*(\xi), \xi) \quad \forall \phi \in S_y, \quad P_y^*(\xi) \in S_y.$$

Then we have, for all $\xi \in V$,

$$(3.15) \quad \xi = P_y(\xi) + Q_y(\xi),$$

$$(3.16) \quad \xi = P_y^*(\xi) + Q_y^*(\xi),$$

where

$$Q_y(\xi) \in (S_y^a)^0 \quad (\text{polar cone of } S_y \text{ with respect to } a),$$

$$Q_y^*(\xi) \in (S_y^{a*})^0 \quad (\text{polar cone of } S_y \text{ with respect to } a^*),$$

with

$$(3.17) \quad a(Q_y(\xi), P_y(\xi)) = 0, \quad a(P_y^*(\xi), Q_y^*(\xi)) = 0.$$

Notice that we have

$$(3.18) \quad \forall \phi \in S_y \quad \forall \psi \in (S_y^a)^0, \quad a(\psi, \phi) \leq 0,$$

$$(3.19) \quad \forall \phi \in S_y \quad \forall \psi^* \in (S_y^{a*})^0, \quad a(\phi, \psi^*) \leq 0.$$

Now from (3.9) we have, if $y = y(v)$

$$D_{y_v}(w) = P_y(A^{-1}w),$$

and, if $Nv_0 \in V$, we can write

$$\begin{aligned} DJ_{v_0}(w) &= \int_{\Omega} (y_0 - z_d) D_{y_{v_0}}(w) dx + N \int_{\Omega} v_0 w dx \\ &= a(P_{y_0}(A^{-1}w), A^{*-1}(y_0 - z_d)) + a(A^{-1}w, Nv_0) \\ &= a(P_{y_0}(A^{-1}w), A^{*-1}(y_0 - z_d) + Nv_0) + a(Q_{y_0}(A^{-1}w), Nv_0). \end{aligned}$$

Set

$$\xi_0 = -A^{*-1}(y_0 - z_d) - Nv_0, \quad \xi_1 = -Nv_0.$$

We have

$$DJ_{v_0}(w) = -a(P_{y_0}(A^{-1}w), \xi_0) - a(Q_{y_0}(A^{-1}w), \xi_1),$$

and therefore

$$\begin{aligned} (3.20) \quad DJ_{v_0}(w) &= -a(P_{y_0}(A^{-1}w), P_{y_0}^*(\xi_0)) - a(P_{y_0}(A^{-1}w), Q_{y_0}^*(\xi_0)) \\ &\quad - a(Q_{y_0}(A^{-1}w), Q_{y_0}(\xi_1)) - a(Q_{y_0}(A^{-1}w), P_{y_0}(\xi_1)). \end{aligned}$$

Suppose that (3.11) holds at the point $v_0 \in V$, so that

$$DJ_{v_0}(w) \geq 0 \quad \forall w \in V'.$$

Take $w_0 = AP_{y_0}^*(\xi_0)$, so that $A^{-1}w_0 = P_{y_0}^*(\xi_0) \in S_{y_0}$, and

$$P_{y_0}(A^{-1}w_0) = P_{y_0}^*(\xi_0), \quad Q_{y_0}(A^{-1}w_0) = 0.$$

Then

$$\begin{aligned} DJ_{v_0}(w_0) &= -a(P_{y_0}^*(\xi_0), P_{y_0}^*(\xi_0)) - a(P_{y_0}^*(\xi_0), Q_{y_0}^*(\xi_0)) \\ &= -a(P_{y_0}^*(\xi_0), P_{y_0}^*(\xi_0)) \geq 0, \end{aligned}$$

and we must have

$$(3.21) \quad P_{y_0}^*(\xi_0) = 0.$$

Now take $w_1 = AQ_{y_0}(\xi_1)$, so that $A^{-1}w_1 = Q_{y_0}(\xi_1)$, and

$$P_{y_0}(A^{-1}w_1) = 0, \quad Q_{y_0}(A^{-1}w_1) = Q_{y_0}(\xi_1).$$

Then

$$\begin{aligned} DJ_{v_0}(w_1) &= -a(Q_{y_0}(\xi_1), Q_{y_0}(\xi_1)) - a(Q_{y_0}(\xi_1), P_{y_0}(\xi_1)) \\ &= -a(Q_{y_0}(\xi_1), Q_{y_0}(\xi_1)) \geq 0, \end{aligned}$$

and we must have

$$(3.22) \quad Q_{y_0}(\xi_1) = 0.$$

We have shown that (3.11) implies $\xi_0 \in (S_{y_0}^{a*})^0$, $\xi_1 \in S_{y_0}$, which is equivalent to (3.12).

Suppose now that we have (3.12) and so that $P_{y_0}^*(\xi_0) = 0$, $Q_{y_0}(\xi_1) = 0$. Then, from (3.20), (3.18) and (3.19), we have, for all $w \in V'$,

$$DJ_{v_0}(w) = -a(P_{y_0}(A^{-1}w), Q_{y_0}^*(\xi_0)) - a(Q_{y_0}(A^{-1}w), P_{y_0}(\xi_1)) \geq 0$$

and this finishes the proofs of Theorems 2.2 and 3.3.

4. Some comments and open problems. Let us show briefly that once we know that the optimal control v_0 belongs to V , Theorem 2.2 implies Theorem 3.2, and, in fact, that (2.3) together with the definition of $y(v_0)$ implies (3.7).

If we set

$$\eta_0 = A^* p_0 - (y_0 - z_d),$$

we get from (2.3)

$$p_0 \in S_{y_0}, \quad \langle \eta_0, \psi \rangle \leq 0 \quad \forall \psi \in S_{y_0}.$$

So, $\langle \eta_0, p_0 \rangle \leq 0$ and $\langle \xi_0, p_0 \rangle = 0$. Now $y_0 \in S_{y_0}$ and $-y_0 \in S_{y_0}$, so $\langle \eta_0, y_0 \rangle = 0$, and this proves (3.7).

We are now going to show by means of a simple counterexample that Theorem 2.2 is strictly stronger than Theorem 3.2.

Take $V = R$; and for $v \in R$,

$$(4.1) \quad y(v) = (-1 + v)^+$$

($y(v)$ is solution of a variational inequality in R), with the cost function

$$(4.2) \quad J(v) = (y(v) - 1)^2 + v^2.$$

Then

$$J(v) = \begin{cases} 2v^2 - 4v + 4 & \text{if } v \geq 1, \\ v^2 + 1 & \text{if } v \leq 1. \end{cases}$$

The optimal control v_0 is here unique, and we have $v_0 = 0$. But it is easy to show that the point $v_1 = 1$, to which $y(v_1) = 0$ and $p(v_1) = -1$ correspond, satisfies (3.7), but does not satisfy (2.3). Here the only solution of (2.3) is $v_0 = 0$, with $y(v_0) = 0$ and $p(v_0) = 0$.

So we see that (2.3) is strictly stronger than (3.7). Notice also that, in our case, 0 belongs to the generalized gradient of J at the point v_1 .

We therefore see that in our type of problem, the optimality condition

$$DJ_{v_0}(w) \geq 0 \quad \forall w \in V'$$

appears as a “good” optimality condition.

Unfortunately we have not been able, till now, to say exactly what this condition means when the set U_{ad} is not the whole space $L^2(\Omega)$ and this is an open problem.

Let us mention three other important open problems:

- How can we solve directly the optimality system (2.6)? This would be important for numerical applications.
- What can we say when we replace the convex set K by more general convex sets such as for example.

$$K' = \{v \mid v \in H_0^1(\Omega), |\text{Grad } v(x)| \leq 1 \text{ a.e. in } \Omega\}.$$

In this situation we do not know that the mapping $v \rightarrow y(v)$ admits at each point v a conical derivative.

- What can we say for the evolution case, even with the convex set K ? Here again we do not know whether $v \rightarrow y(v)$ has a conical derivative at each point v .

REFERENCES

- [1] V. BARBU, *Necessary conditions for distributed control problems governed by parabolic variational inequalities*, this Journal, 19 (1981), pp. 64–86.
- [2] ———, *Necessary conditions for nonconvex distributed control problems governed by elliptic variational inequalities*, J. Math. Anal. Appl., 80 (1981), pp. 566–598.
- [3] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Gauthier-Villars, Paris, 1969.
- [4] ———, *Sur le contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Gauthier-Villars, Paris, 1968.
- [5] J. L. LIONS AND G. STAMPACCHIA, *Variational inequalities*, Comm. Pure Applied Math., 20 (1967), pp. 493–519.
- [6] F. MIGNOT, *Contrôle dans les inéquations variationnelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130–185.
- [7] CH. SAGUEZ, *Contrôle de problèmes à frontière libre*, Thèse d'Etat, Université de Compiègne, 1980.
- [8] F. MIGNOT AND L. TARTAR, personal communication, 1980.
- [9] J. P. YVON, *Contrôle optimal de systèmes gouvernés par des inéquations variationnelles*, Rapport Laboria, IRIA, Rocquencourt, France, February 1974.
- [10] A. ZRIKEM, *Un problème de contrôle frontière dans une inéquation variationnelle*, Thèse de 3ème cycle, Université Paris, 6, 1979.

RITZ-GALERKIN APPROXIMATION OF THE TIME OPTIMAL BOUNDARY CONTROL PROBLEM FOR PARABOLIC SYSTEMS WITH DIRICHLET BOUNDARY CONDITIONS*

IRENA LASIECKA†

Abstract. This paper deals with Ritz-Galerkin approximations of the following two problems:

- (i) boundary-value problems with L_2 -boundary data given in the form of Dirichlet boundary conditions;
- (ii) time optimal control problems for parabolic systems with control acting on the boundary.

In both cases, optimal rates of convergence of approximation are obtained. Also, the specialization is considered to approximating with nonconformal elements which are not required to satisfy boundary conditions. Corresponding rates of convergence are also proved.

Key words. Ritz-Galerkin approximations, finite element approximations, bang-bang principle, analytic semigroups, Dirichlet map

1. Introduction.

1.1. Formulation of the problem. Let $A(\xi, \partial)$ be a second order uniformly elliptic operator

$$A(\xi, \partial)u = - \sum_{j,k=1}^n \frac{\partial}{\partial \xi_j} \left(a_{jk}(\xi) \frac{\partial u}{\partial \xi_k} \right) + \sum_{j=1}^n b_j(\xi) \frac{\partial u}{\partial \xi_j} + c(\xi)u$$

with real, smooth coefficients a_{jk} , b_j , c defined on $\xi \in \Omega$, Ω a bounded open domain in R^n with a sufficiently smooth boundary Γ . Consider the following parabolic equation:

$$\begin{aligned} \frac{\partial y(t, \xi)}{\partial t} &= -A(\xi, \partial)y(t, \xi) \quad \text{on } (0, T] \times \Omega, \\ (1.1.1) \quad y(0, \xi) &= y_0(\xi), \\ y|_{\Gamma} &= u(t, \sigma) \quad \text{on } (0, T] \times \Gamma \end{aligned}$$

with control function (boundary input) $u(t, \sigma)$ acting within the Dirichlet boundary conditions (B.C.).

Define the following sets:

$$\begin{aligned} U_A &= \{u \in L_\infty[0T; L_2(\Gamma)]; \|u(t)\|_{L_2(\Gamma)} \leq 1 \text{ a.e.}\}, \\ U_B &= \{u \in L_\infty[0T; L_2(\Gamma)]; |u(x, t)| \leq 1 \text{ a.e. in } x, t\}. \end{aligned}$$

We shall indicate by $y(u, t)$ the solution $y(t, \cdot)$ in $L_2(\Omega)$ of problem (1.1.1) at time t . Suppose that the point $y_1 \in L_2(\Omega)$ is approximately controllable in the following sense: Given $\sigma > 0$, there exist $T > 0$ and $u = u_A$ (resp. $u = u_B$) in U_A (resp. U_B) such that the solution $y(u, T)$ of (1.1.1) at time T , due to the control function u , satisfies

$$(1.1.2) \quad \|y(u, T) - y_1\|_{L_2(\Omega)} < \sigma.$$

We can now formulate the two time optimal control problems of interest in the present paper.

Problem P_A . Find $T_A^0 = \min \{T > 0; \exists u \in U_A \text{ with } \|y(u, T) - y_1\| \leq \sigma\}$.

Problem P_B has the same formulation except that U_A is now replaced by U_B .

* Received by the editors June 29, 1981, and in revised form March 22, 1983. This research was supported in part by the National Science Foundation under grant MCS81-02-837.

† Mathematics Department, University of Florida, Gainesville, Florida 32611.

As is well known [F1], [S1], the above control problems admit unique solutions u_A^0, T_A^0 (resp. u_B^0, T_B^0) with the following bang-bang property:

$$(1.1.3) \quad \|u_A^0(t)\|_{L_2(\Gamma)} \equiv 1 \quad \text{a.e. in } t \in [0, T_A^0],$$

$$(1.1.4) \quad |u_B^0(x, t)| \equiv 1 \quad \text{a.e. } x \in \Gamma, \quad t \in [0, T_B^0],$$

$$(1.1.5) \quad \|y(u_\alpha^0, T_\alpha^0) - y_1\|_{L_2(\Gamma)} = \sigma, \quad \alpha = A \text{ or } B.$$

The major goal of the present paper is to introduce a Galerkin approximation (in particular, a finite element approximation) of the above time optimal control problems, and to prove convergence of the resulting schemes. This will require the following two preliminary steps: (i) an approximation of the parabolic equation with nonhomogeneous (Dirichlet type) boundary conditions; and (ii) an estimate of the rate of convergence of the approximating solutions in the $L_\infty(0, T; L_2(\Omega))$ -norm, for boundary inputs u in $L_\infty(0, T; L_2(\Gamma))$. As to (i), a standard method of approximating nonhomogeneous Dirichlet boundary problems is to subtract the boundary data and to consider the resulting homogeneous equation [D1]. This technique, however, requires a regularity of the boundary data higher than the one available here (i.e., it requires $u(t) \in H^{1/2}(\Gamma)$ instead of $u(t) \in L_2(\Gamma)$). Therefore it fails in our present situation.

In order to overcome this difficulty, we propose a radically different approach. This is based on replacing the parabolic problem (1.1.1) with a functional analytic model (see § 1.2), which gives an effective input-solution formula: $u \rightarrow y$. This formula, recently introduced in the study of nonhomogeneous parabolic problems, is expressed in terms of three quantities: (i) the generator $-A$ of the differential operator $-A(\xi, \partial)$ with zero Dirichlet B.C.; (ii) its corresponding C_0 -analytic semigroup $S(t)$; and (iii) the Dirichlet map D of the corresponding elliptic problem, responsible for the action of u into the interior Ω . Our approach, therefore, will consist in approximating this functional analytic model of (1.1.1), i.e., in approximating A , $S(t)$ and D . We have already applied with success this approach to a quadratic cost optimal control problem in [L1], where we have obtained error estimates in the $L_2((0, T) \otimes \Omega)$ -norm. In the present case, however, estimates in the $L_\infty(0, T; L_2(\Gamma))$ -norm are needed. Thus the arguments of [L1] cannot be applied directly, as different technical difficulties now arise. In short, the present problem requires a new, ad hoc treatment. As a matter of fact, a major portion of the paper (§ 2) is devoted to the approximation, in the way described, of the nonhomogeneous (Dirichlet type) parabolic problem per se, with no reference at all to the control problems. This will provide the desired rates of convergence in terms of the $L_\infty(0, T; L_2(\Gamma))$ -norm for u , which are not available in the literature and which are needed in completing the study of the control problems (in § 3 for problem P_A and in § 4 for problem P_B). In fact, the way of choosing an approximation to the Dirichlet map D will depend on the type of control problem under consideration. More precisely, in the case of the control problem P_A , in order to obtain an optimal rate of convergence, it will suffice to restrict our attention to approximation schemes based on *piecewise linear splines*. This technique fails, however, in the case of control problem P_B : Here, by contrast, different techniques based on the Rayleigh-Ritz method [B2] will be used to approximate the Dirichlet map. This way we shall still obtain an optimal rate of convergence for problem P_B ; however, to this end, *cubic splines*, rather than linear splines, will be needed.

In the general study of (Dirichlet) nonhomogeneous parabolic problems of § 2, two approximating strategies will be pursued. The first (§ 2.1) will be based on finite dimensional subspaces of $H_0^r(\Omega)$, with $r > 1$, whereby the approximating elements have low differentiability requirements, but must also satisfy zero boundary conditions.

This last condition is, however, stringent and not easy to accomplish, particularly for nonpolygonal domains (it is implemented in practice by the use of elements [Z1], [Z2]). Thus, there is the need for different discretization schemes possessing the desirable feature of being based on nonconforming elements which are *not required to vanish on the boundary*. A study of such a scheme is then presented in § 2.2, and the price we pay is that the approximating elements possess higher differentiability properties (i.e. are in $H^r(\Omega)$, $r \geq 2$). We will extend to the parabolic case some results from elliptic theory given in [N1] (see also [B3]). Sections 2.1 and 2.2 are each organized as follows: A general statement of the main result is given first, in terms of some "abstract conditions" (part A); this is followed by technical proofs (in part B), and by examples of approximations (for the operators A and D) which do satisfy the abstract conditions (part C). It should be also pointed out that, in contrast with much of the literature [B4], [D1], [Z1], etc., our approach on the input-solution formula described above does *not* require the operator A ($A(\xi, \partial)$ plus zero Dirichlet B.C.) to be selfadjoint.

Notation. In the sequel we shall use the following notation:

$$\begin{aligned} \|\cdot\|, |\cdot| & \quad \text{norms in } L_2(\Omega), L_2(\Gamma), \text{ respectively,} \\ \|\cdot\|_s, |\cdot|_s & \quad \text{norms in } H^s(\Omega), H^s(\Gamma), \text{ respectively,} \end{aligned}$$

where H^s is the usual Sobolev space of order s ;

$$\begin{aligned} (\cdot; \cdot) & \quad \text{scalar product in } L_2(\Omega), \\ \langle \cdot, \cdot \rangle & \quad \text{scalar product in } L_2(\Gamma), \\ \mathcal{L}(X \rightarrow Y) & \quad \text{space of linear bounded transformation from } X \text{ into } Y, \\ H_0^s(\Omega) & \quad \text{Sobolev space of order } s \text{ with zero boundary conditions.} \end{aligned}$$

1.2. Semigroup model for the parabolic problem (1.1.1). Let $A: L_2(\Omega) \supset \mathcal{D}(A) \rightarrow L_2(\Omega)$ be the operator $Af = A(\xi, \partial)f$ for $f \in \mathcal{D}(A)$ where $\mathcal{D}(A) = \{y \in L_2(\Omega); Ay \in L_2(\Omega); y|_\Gamma = 0\}$.

It is well known that $-A$ generates a strongly continuous, analytic semigroup $S(t)$ on $L_2(\Omega)$, such that $\|S(t)\| \leq C e^{-wt}$ for some w , $t \geq 0$, where C will stand for the generic constant.

Let D be the Dirichlet map defined by

$$(1.2.1) \quad v = Du \quad \text{iff} \quad -A(\xi, \partial)v = 0 \quad \text{in } \Omega, \quad v|_\Gamma = u \quad \text{on } \Gamma.$$

It is well known [L3] that

$$(1.2.2) \quad D \in \mathcal{L}(H^s(\Gamma) \rightarrow H^{s+1/2}(\Omega)), \quad s \text{ real},$$

and we shall be concerned mostly with the case $s = 0$. Without loss of generality, we can assume that the spectrum of A is on the right of the complex plane, so that the fractional powers of A are well defined [P1]. By analyticity of the semigroup, we have

$$(1.2.3) \quad \|A^\alpha S(t)x\| \leq \frac{C\|x\|}{t^{1-\alpha}}, \quad 0 < t, \quad x \in L_2(\Omega), \quad 0 \leq \alpha \leq 1.$$

The following identification (set theoretical and topological) will be freely used in the sequel:

$$(1.2.4) \quad D(A^\alpha) = H_0^{2\alpha}(\Omega), \quad \alpha < \frac{3}{4}, \quad \alpha \neq \frac{1}{4},$$

(where of course $H_0^s(\Omega) = H^s(\Omega)$ $s < \frac{1}{2}$) with topology given by:

$$(1.2.5a) \quad |x|_{D(A^\alpha)} = \|A^\alpha x\|, \quad \alpha < \frac{1}{4}$$

(see [F2], [L2]). Moreover, in general

$$(1.2.5b) \quad \|x\|_{2\beta} \leq C \|A^\beta x\|, \quad x \in D(A^\beta), \quad \beta > 0.$$

By (1.2.2) with $s = 0$ and by (1.2.4), we then have

$$(1.2.6) \quad \text{range of } D = DL_2(\Gamma) \subset D(A^{1/4-\rho}), \quad \rho > 0.$$

With this preliminary background, we can now give the abstract semigroup model [L1] for the parabolic problem (1.1.1). It is given by

$$(1.2.7) \quad \begin{aligned} y(u, t) &= y(t) = S(t)y_0 + (Lu)(t), \\ (Lu)(t) &= A \int_0^t S(t-z) Du(z) dz = \int_0^t A^{3/4+\rho} S(t-z) A^{1/4-\rho} Du(z) dz. \end{aligned}$$

It follows directly (or by the convolution theorem) that the operator

$$(1.2.8) \quad L: L_\infty(0T; L_2(\Gamma)) \rightarrow C([0T]; H^{1/2-\varepsilon}(\Omega)), \quad \varepsilon > 0,$$

is linear and bounded. In fact, by (1.2.5) and (1.2.3), we get from (1.2.7)

$$\begin{aligned} \|(Lu)(t)\|_{1/2-\varepsilon} &= \left\| A^{1/4-\varepsilon/2} A \int_0^t S(t-z) Du(z) dz \right\| \\ &= \left\| \int_0^t A^{1-\varepsilon/4} S(t-z) A^{1/4-\varepsilon/4} Du(z) dz \right\| \\ &\leq C \int_0^t \frac{dz}{(t-z)^{1-\varepsilon/4}} |u|_{L_\infty(0T; L_2(\Gamma))} \\ &\leq C_T |u|_{L_\infty(0T; L_2(\Gamma))}. \end{aligned}$$

Thus, $y_0 \in H^{1/2-\varepsilon}(\Omega)$ implies, via (1.2.6), $y \in C([0T]; H^{1/2-\varepsilon}(\Omega))$. Moreover, the operator

$$L_T u = A \int_0^T S(T-z) Du(z) dz$$

is linear and bounded: $L_\infty[0T; L_2(\Gamma)] \rightarrow H^{1/2-\varepsilon}(\Omega)$.

2. Approximation of solutions to parabolic equations with nonhomogeneous boundary data of Dirichlet type. In this section, we present a general approach for approximating solutions $y(u, \cdot)$ to parabolic equations like (1.1.1) with Dirichlet boundary input u . It is based on the (input-solution) functional analytic formula (1.2.7) and consists, therefore, in providing appropriate approximations for the generator $-A$, its semigroup $S(t)$ and the Dirichlet map D . We shall establish in this way *optimal rates* of convergence of solutions.

This section is divided into two subsections. Subsection 2.1 deals with approximating spaces which are subspaces of $H_0^r(\Omega)$, $r > 1$. Thus, the approximating elements have the advantage of low differentiability requirements combined with the disadvantage of their vanishing on the boundary. This situation is corrected and reversed in § 2.2, where the approximating elements will not be required to satisfy zero boundary conditions. However, they will have to possess higher differentiability properties, as members, in fact, of $H^r(\Omega)$, $r \geq 2$. Each subsection contains the general statement of

results (in part A) in terms of some “abstract assumptions”, followed by their proofs (in part B), while part C provides specific approximating schemes for the operators A and D , which fulfill those abstract assumptions.

2.1. Approximation by subspaces with zero boundary conditions.

2.1.A. Statement of results. Henceforth, h will denote the parameter of discretization, with $h \downarrow 0$.

Assumptions.

Approximation of A . Let V_h^0 be an $N(h)$ -dimensional subspace of $H_0^r(\Omega)$, $r > 1$ such that:

- (i) V_h^0 is an $S_{h^1}^{2,1}(\Omega)$ -system (see [B1]),
 (2.1.1) (ii) V_h^0 satisfies the so-called “inverse assumption”

$$\|v_h\|_1 \leq Ch^{-1}\|v_h\|, \quad v_h \in V_h^0.$$

As a consequence of (i) and (ii), one can prove that

$$(2.1.2) \quad \|R_h^0 y - y\|_s \leq Ch^{\alpha-s}\|y\|_\alpha, \quad 0 \leq s \leq 1, \quad \alpha \leq 2, \quad \alpha - s \geq 0, \quad y \in H^\alpha(\Omega)$$

where R_h^0 is the orthogonal projection of $L_2(\Omega)$ onto V_h^0 , orthogonality being in the $L_2(\Omega)$ -inner product¹. Let $A_h: V_h^0 \rightarrow V_h^0$ be a finite dimensional positive² operator (for each $h > 0$) which approximates the positive operator A in the following sense:

$$(H1) \quad \|A_h^\beta (R_h^0 A^{-1} - A_h^{-1} R_h^0) A x\| \leq Ch^{\alpha-2\beta} \|A^{\alpha/2} x\|, \quad x \in \mathcal{D}(A^{\alpha/2}),$$

for all $0 \leq \beta \leq 1, 1 \leq \alpha \leq 2$.

Assume, furthermore³, that

- (i) either A_h^2 is also positive,
 (H2) (ii) or else

$$((A_h - A_h^*)v_h, y_h) \leq C(A_h v_h, v_h)^{1/2} \|y_h\|, \quad v_h \in V_h^0, \quad y_h \in V_h^0.$$

Approximation of $S(t)$. Naturally enough, we take

$$(2.1.3) \quad S_h(t) = e^{-A_h t}, \quad t \in \mathbb{R}, \quad \text{on } V_h^0$$

as the corresponding approximation of $S(t)$ on V_h^0 .

Approximation of D . Let $U_h^1 \subset L_2(\Gamma)$ be an approximation subspace of $L_2(\Gamma)$ with the property that

$$(2.1.4) \quad |P_h u - u|_{-1/2+\varepsilon} \leq Ch^{1/2-\varepsilon} |u|, \quad u \in L_2(\Gamma), \quad \varepsilon > 0,$$

where P_h is the orthogonal projection of $L_2(\Gamma)$ onto U_h^1 . Let $D_h: U_h^1 \rightarrow V_h^0 \subset H_0^1(\Omega)$ be an approximation of the Dirichlet map D (see (1.2.1)), which satisfies the following properties (approximation and stability) for all $u_h \in U_h^1$, $0 \leq \varepsilon \leq \frac{1}{2}$,

$$(2.1.5) \quad \|(D - D_h)u_h\|_\varepsilon \leq Ch^{1/2-\varepsilon} |u_h|,$$

$$(2.1.6)^4 \quad \|D_h u_h\|_{1/2-\varepsilon} \leq C |u_h|.$$

¹ There are many examples of spaces satisfying this condition (see [S3], [Z1], [Z2]), e.g., the space of piecewise linear functions for polygonal domains, and that of curvilinear splines for arbitrary domains.

² A is positive iff $(Ax, x) > 0$ for $x \in D(A)$.

³ This assumption, which automatically holds in the selfadjoint case, is needed only in the proof of Lemma 2.1.B.1. Assumption (H2)(ii) is of the same nature as in [H1].

⁴ (2.1.6) follows from (2.1.5) and (1.2.2).

Having introduced approximation for A , $S(t)$ and D , we can now define a natural approximation of the solution to the parabolic problem (1.1.1) given by (1.2.7).

Approximation of parabolic solutions. We take

$$(2.1.7) \quad \begin{aligned} y_h(u_h, t) &= y_h(t) = S_h(t)R_h^0 y + (L_h u_h)(t), \quad y \in L_2(\Omega), \\ (L_h u_h)(t) &= A_h \int_0^t S_h(t-z) D_h u_h(z) dz. \end{aligned}$$

Notice that (2.1.7) corresponds to an $N(h)$ -dimensional system of linear ordinary differential equations.

Main results.

THEOREM 2.1.1. *Let A_h satisfy hypotheses (H1)–(H2), and let D_h satisfy (2.1.5)–(2.1.6). Then for $y_0 \in H^{1/2-\varepsilon}(\Omega)$ and every $u_h \in L_\infty(0T; U_h^1)$ we have*

$$\|y(u, t) - y_h(u_h, t)\| \leq Ch^{1/2-\varepsilon} \{\|y_0\|_{1/2-\varepsilon} + \|u_h\|_{L_\infty[0T; L_2(\Gamma)]}\} + C \|u - u_h\|_{L_\infty(0T; H^{-1/2+\varepsilon}(\Gamma))}$$

for all $t \in [0, T]$ where $\varepsilon > 0$ is arbitrarily small and where $y(u, t)$ and $y_h(u_h, t)$ are given by (1.2.7) and (2.1.7) respectively. Here C is a constant which does not depend on h .

From this, if $u_h(t)$ are approximations in U_h^1 of $u(t)$, we shall derive

COROLLARY 2.1.2. *If $u_h(t) \in U_h^1$ approximates $u(t) \in L_2(\Gamma)$ in the sense described by (2.1.4) with $u_h(t) = P_h u(t)$, then*

$$\|y(u, t) - y_h(u_h, t)\| \leq Ch^{1/2-\varepsilon} [\|y_0\|_{1/2-\varepsilon} + \|u\|_{L_\infty[0T; L_2(\Gamma)]}].$$

2.1.B. Proof of Theorem 2.1.1 and Corollary 2.1.2. This is achieved through a series of lemmas.

LEMMA 2.1.B.1. *Let $x_h \in V_h^0$. Then*

$$\|A_h^\beta S_h(t) x_h\| \leq \frac{C_T}{t^\beta} \|x_h\|, \quad 0 \leq \beta \leq 1, \quad 0 < t \leq T.$$

Proof. By interpolation, it is enough to prove the two cases $\beta = 0$ and $\beta = 1$. If we take the inner product with $S_h(t)x_h$ on the identity (from (2.1.3)),

$$(2.1.8) \quad \frac{dS_h(t)}{dt} x_h + A_h S_h(t) x_h \equiv 0,$$

and integrate, we arrive at

$$(2.1.9) \quad \|S_h(t) x_h\|^2 + 2 \int_0^t (A_h S_h(z) x_h, S_h(z) x_h) dz = 2 \|x_h\|^2.$$

This proves the case $\beta = 0$, since A_h is positive. Next, if we differentiate in t identity (2.1.8) and take the inner product with $(d/dt)S_h(t)x_h$, we arrive at

$$\frac{1}{2} \frac{d}{dt} \|A_h S_h(t) x_h\|^2 + \left(A_h \frac{d}{dt} S_h(t) x_h, \frac{d}{dt} S_h(t) x_h \right) \equiv 0.$$

Multiplying this identity by $2t^2$ and integrating by parts the result from 0 to t_1 gives (we replace t_1 with t)

$$(2.1.10) \quad \begin{aligned} t^2 \|A_h S_h(t) x_h\|^2 + 2 \int_0^t z^2 (A_h^2 S_h(z) x_h, A_h S_h(z) x_h) dz \\ = 2 \int_0^t z \|A_h S_h(z) x_h\|^2 dz. \end{aligned}$$

Also, if we apply to (2.1.8) the inner product with $t \cdot (d/dt)S_h(t)x_h$, we obtain

$$t\|A_h S_h(t)x_h\|^2 + t\left(A_h S_h(t)x_h, \frac{d}{dt}S_h(t)x_h\right) = 0,$$

which can be rewritten equivalently as

$$(2.1.11) \quad t\|A_h S_h(t)x_h\|^2 + \frac{d}{dt}[t(A_h S_h(t)x_h, S_h(t)x_h)] \\ + t[(A_h^2 S_h(t)x_h, S_h(t)x_h)] = (A_h S_h(t)x_h, S_h(t)x_h).$$

Now, if A_h^2 is positive (assumption (H2)(i)), then we drop the third term and get

$$(2.1.12)(i) \quad \int_0^t z\|A_h S_h(z)x_h\|^2 dz \leq \int_0^t (A_h S_h(z)x_h, S_h(z)x_h) dz.$$

On the other hand, if A_h satisfies assumption (H2)(ii), we add and subtract the same quantity to (2.1.11) and write

$$\int_0^t z\|A_h S_h(z)x_h\|^2 dz + t(A_h S_h(t)x_h, S_h(t)x_h) + \int_0^t z(A_h S_h(z)x_h, A_h S_h(z)x_h) dz \\ = \int_0^t z((A_h - A_h^*)S_h(z)x_h, A_h S_h(z)x_h) dz + \int_0^t (A_h S_h(z)x_h, S_h(z)x_h) dz.$$

On the left-hand side of the above equality, we drop the second and third terms, which are positive, while on the first term on the right-hand side we use assumption (H2)(ii). We obtain

$$\int_0^t z\|A_h S_h(z)x_h\|^2 dz \\ \leq C \int_0^t z^{1/2}(A_h S_h(z)x_h, S_h(z)x_h)^{1/2} z^{1/2}\|A_h S_h(z)x_h\| dz \\ + \int_0^t (A_h S_h(z)x_h, S_h(z)x_h) dz \\ \leq \frac{C}{2} \left[\gamma \int_0^t z(A_h S_h(z)x_h, S_h(z)x_h) dz + \frac{1}{\gamma} \int_0^t z\|A_h S_h(z)x_h\|^2 dz \right] \\ + \int_0^t (A_h S_h(z)x_h, S_h(z)x_h) dz.$$

Thus, choosing γ suitably large yields the inequality

$$(2.1.12)(ii) \quad \int_0^t z\|A_h S_h(z)x_h\|^2 dz \leq C_T \int_0^t (A_h S_h(z)x_h, S_h(z)x_h) dz, \quad 0 \leq t \leq T,$$

which is of the same form as (2.1.12)(i). The statement in the lemma for $\beta = 1$ now follows from (2.1.10), (2.1.12)(i)–(ii), and (2.1.9) after dropping the (positive) second term in (2.1.10) (since A_h is positive) and the first term in (2.1.9). The proof of Lemma 2.1.B.1 is now complete. \square

LEMMA 2.1.B.2. *For any $x \in L_2(\Omega)$ we have*

$$S_h(t)R_h^0 x - R_h^0 S(t)x = - \int_0^t A_h S_h(t-z)[R_h^0 A^{-1} - A_h^{-1} R_h^0] A S(z)x dz.$$

Proof. For $t > 0$ and $\phi^0 \in V_h^0$, we subtract the identity

$$\left(\frac{d}{dt} R_h^0 S(t)x, \phi_h^0 \right) + (A_h R_h^0 S(t)x, \phi_h^0) = (A_h R_h^0 S(t)x - AS(t)x, \phi_h^0)$$

from the identity

$$\left(\frac{d}{dt} S_h(t) R_h^0 x, \phi_h^0 \right) + (A_h S_h(t) R_h^0 x, \phi_h^0) \equiv 0,$$

and we integrate the result, using $R_h^0 A_h = A_h$. We thus obtain

$$S_h(t) R_h^0 x - R_h^0 S(t)x = - \int_0^t S_h(t-z) R_h^0 [A_h R_h^0 S(z) - AS(z)] x \, dz,$$

which is equivalent to the desired assertion in the lemma. \square

The next lemma deals with some specific approximation properties of $S_h(t)$.

LEMMA 2.1.B.3. For $y \in \mathcal{D}(A^{s/2})$ we have:

- (i) $\|S_h(t) R_h^0 y - R_h^0 S(t)y\| \leq Ch^s \|A^{s/2} y\|, \quad 0 \leq s < 2,$
- (ii) $\|A_h S_h(t) R_h^0 y - R_h^0 AS(t)y\| \leq \frac{Ch^{\alpha-2\varepsilon}}{t^{1-\varepsilon}} \|A^{\alpha/2} y\| \quad \text{for } 1 < \alpha \leq 2, \quad t > 0.$

Here C is a constant which does not depend on h .

Proof. (i) By Lemma 2.1.B.2 and Lemma 2.1.B.1 applied with $\beta = 1 - \varepsilon$, we have

$$\begin{aligned} \|S_h(t) R_h^0 y - R_h^0 S(t)y\| &\leq \int_0^t \|A_h^{1-\varepsilon} S_h(t-z) A_h^\varepsilon [R_h^0 A^{-1} - A_h^{-1} R_h^0] AS(z)y\| \, dz \\ &\leq \int_0^t \frac{C}{(t-z)^{1-\varepsilon}} \|A_h^\varepsilon [R_h^0 A^{-1} - A_h^{-1} R_h^0] AS(z)y\| \, dz \\ &\quad \text{(by (H1) with } \beta = \varepsilon \text{ and } 1 \leq \alpha \leq 2) \\ &\leq Ch^{\alpha-2\varepsilon} \int_0^t \frac{1}{(t-z)^{1-\varepsilon}} \|A^{\alpha/2} S(z)y\| \, dz \\ &\leq Ch^{\alpha-2\varepsilon} \int_0^t \frac{1}{(t-z)^{1-\varepsilon}} \|A^\varepsilon S(z) A^{\alpha/2-\varepsilon} y\| \, dz \\ &\leq Ch^{\alpha-2\varepsilon} \int_0^t \frac{1}{(t-z)^{1-\varepsilon}} \frac{1}{z^\varepsilon} \, dz \|A^{\alpha/2-\varepsilon} y\|. \end{aligned}$$

Let $\beta(p, q) = \int_0^1 z^{p-1} (1-z)^{q-1} \, dz$. Then

$$(2.1.13) \quad \|S_h(t) R_h^0 y - R_h^0 S(t)y\| \leq Ch^{\alpha-2\varepsilon} \|A^{\alpha/2-\varepsilon} y\| \beta(1-\varepsilon, \varepsilon) \quad \text{for all } 1 \leq \alpha \leq 2.$$

Since we also have (by Lemma 2.1.B.1)

$$(2.1.14) \quad \|S_h(t) R_h^0 y - R_h^0 S(t)y\| \leq C \|y\|,$$

then by interpolating between (2.1.13) and (2.1.14) we arrive at the desired conclusion for part (i).

(ii) To prove (ii), we differentiate in t the identity of Lemma 2.1.B.2 to get for $y \in \mathcal{D}(A)$

$$\begin{aligned} (2.1.15) \quad A_h S_h(t) R_h^0 y - R_h^0 AS(t)y &= A_h S_h(t) [R_h^0 A^{-1} - A_h^{-1} R_h^0] A y \\ &\quad + \int_0^t A_h S_h(t-z) [R_h^0 A^{-1} - A_h^{-1} R_h^0] A^2 S(z)y \, dz. \end{aligned}$$

We now estimate the two terms (let us call them ① and ②) on the right of (2.1.15). As to the first, we have, from Lemma 2.1.B.1 with $\beta = 1 - \varepsilon$, and assumption (H1) for $1 \leq \alpha \leq 2$ and $\beta = \varepsilon$,

$$(2.1.16) \quad \begin{aligned} \textcircled{1} &= \|A_h^{1-\varepsilon} S_h(t) A_h^\varepsilon [R_h^0 A^{-1} - A_h^{-1} R_h^0] A y\| \\ &\leq \frac{C_T}{t^{1-\varepsilon}} \|A_h^\varepsilon [R_h^0 A^{-1} - A_h^{-1} R_h^0] A y\| \leq \frac{C_T h^{\alpha-2\varepsilon}}{t^{1-\varepsilon}} \|A^{\alpha/2} y\|. \end{aligned}$$

As to the second term, we have similarly, with (H1) applied with $\beta = \varepsilon/2$ and α replaced by $\alpha - \varepsilon$,

$$\begin{aligned} \textcircled{2} &\leq \int_0^t \frac{C}{(t-z)^{1-\varepsilon/2}} \|A_h^{\varepsilon/2} [R_h^0 A^{-1} - A_h^{-1} R_h^0] A^2 S(z) y\| dz \\ &\leq Ch^{\alpha-2\varepsilon} \int_0^t \frac{1}{(t-z)^{1-\varepsilon/2}} \|A^{1+\alpha/2-\varepsilon/2} S(z) y\| dz \\ &\leq Ch^{\alpha-2\varepsilon} \int_0^t \frac{1}{(t-z)^{1-\varepsilon/2}} \|A^{1-\varepsilon/2} S(z)\| \|A^{\alpha/2} y\| dz \\ &\leq Ch^{\alpha-2\varepsilon} \int_0^t \frac{1}{(t-z)^{1-\varepsilon/2}} \frac{1}{z^{1-\varepsilon/2}} dz \|A^{\alpha/2} y\| \\ &= Ch^{\alpha-2\varepsilon} \frac{1}{t^{1-\varepsilon}} \|A^{\alpha/2} y\| \beta(\varepsilon/2, \varepsilon/2). \end{aligned}$$

This inequality, with (2.1.16), once applied to (2.1.15), yields the desired part (ii). Lemma 2.1.B.3 is thus proved. \square

We are now in a position to proceed with the proof of Theorem 2.1.1.

By (1.2.7) and (2.1.7), the difference between y and y_h can be written as

$$(2.1.17) \quad \begin{aligned} y(u, t) - y_h(u_h, t) &= S(t) y_0 - S_h(t) R_h^0 y_0 \quad \textcircled{1} \\ &\quad + \int_0^t [AS(t-z) - A_h S_h(t-z)] D_h u_h(z) dz \quad \textcircled{2} \\ &\quad - \int_0^t AS(t-z) [D_h u_h(z) - Du(z)] dz \quad \textcircled{3}. \end{aligned}$$

①, ② and ③ denote the three different terms in (2.1.17). We now estimate the first difference ① by means of (2.1.2) with $\alpha = \frac{1}{2}$ and $s = 0$, and Lemma 2.1.B.3(i) with $s = \frac{1}{2} - \varepsilon$:

$$(2.1.18) \quad \begin{aligned} \textcircled{1} &\leq \|S(t) y_0 - R_h^0 S(t) y_0\| + \|R_h^0 S(t) y_0 - S_h(t) R_h^0 y_0\| \\ &\leq Ch^{1/2-\varepsilon} e^{wt} \|y_0\|_{1/2-\varepsilon}. \end{aligned}$$

The second difference ② can be estimated as follows:

$$\begin{aligned} \textcircled{2} &\leq \int_0^t \|[R_h^0 AS(t-z) - AS(t-z)] D_h u_h(z)\| dz \\ &\quad + \int_0^t \|[R_h^0 AS(t-z) - A_h S_h(t-z)] D_h u_h(z)\| dz \end{aligned}$$

(by (2.1.2) with $\alpha = \frac{1}{2} - 4\varepsilon$, $s = 0$, and Lemma 2.1.B.3(ii) with $\alpha = 1 + \varepsilon$)

$$(2.1.19a) \quad \begin{aligned} &\leq Ch^{1/2-4\varepsilon} \int_0^t \|A^{1-\varepsilon} S(t-z) A^{1/4-\varepsilon} D_h u_h(z)\| dz \\ &\quad + Ch^{1-\varepsilon} \int_0^t \frac{\|A^{1/2+\varepsilon/2} D_h u_h\|}{(t-z)^{1-\varepsilon}}. \end{aligned}$$

We now use the fact that $D_h u_h \in V_h^0 \subset H_0^{1+\varepsilon}(\Omega) = \mathcal{D}(A^{1/2+\varepsilon/2})$ (see (1.2.4)) so that $\|A^{1/2+\varepsilon/2} D_h u_h\| \leq C \|D_h u_h\|_{1+\varepsilon}$. Then by the inverse property enjoyed by the space V_h^0 (assumption (2.1.1)) we finally have

$$\|A^{1/2+\varepsilon/2} D_h u_h\| \leq Ch^{-1/2-\varepsilon} \|D_h u_h\|_{1/2} \leq Ch^{-1/2-\varepsilon} |u_h|$$

where in the last step we have used (2.1.6). From here and (2.1.19a) it is then easy to conclude via (1.2.3) that

$$(2.1.19b) \quad \textcircled{2} \leq Ch^{1/2-4\varepsilon} t^\varepsilon |u_h|_{L_\infty(0,T;L_2(\Gamma))}.$$

As to the third difference $\textcircled{3}$ in (2.1.17), we have by (1.2.3) and (1.2.5a)

$$\begin{aligned} \textcircled{3} &= \left\| \int_0^t A^{1-\varepsilon} S(t-z) A^\varepsilon [D_h u_h(z) - Du(z)] dz \right\| \\ &\leq \int_0^t \frac{C}{(t-z)^{1-\varepsilon}} \|D_h u_h(z) - Du(z)\|_{2\varepsilon} dz \\ &\leq \int_0^t \frac{C}{(t-z)^{1-\varepsilon}} \{ \|D(u_h - u)(z)\|_{2\varepsilon} + \|(D - D_h)u_h(z)\|_{2\varepsilon} \} dz. \end{aligned}$$

We now use property (1.2.2) with $s = -\frac{1}{2} + 2\varepsilon$, and inequality (2.1.5) to get finally

$$(2.1.20) \quad \textcircled{3} \leq C \{ |u_h - u|_{L_\infty(0,T;H^{-1/2+2\varepsilon}(\Gamma))} + h^{1/2-2\varepsilon} |u_h|_{L_\infty(0,T;L_2(\Gamma))} \}.$$

Using now inequalities (2.1.18)–(2.1.20) in (2.1.17) yields the desired conclusion and proves Theorem 2.1.1. \square

Corollary 2.1.2 follows from Theorem 2.1.1 via (2.1.4).

2.1.C. Examples. Approximations A_h and D_h which satisfy assumptions (H1)–(H2), and (2.1.5)–(2.1.6), respectively.

We now provide examples of specific approximation schemes for A and D , which fulfill the requirement postulated in § 2.1.A.

Approximation A_h which satisfies assumptions (H1)–(H2). Define the (positive) operator A_h approximating the (positive) operator A on V_h^0 by

$$(2.1.21a) \quad (A_h y_h, v_h) = a(y_h, v_h) \quad \forall y_h, v_h \in V_h^0$$

where $a(y, v) = (Ay, v)$ is the bilinear form associated with A , continuous on $H_0^1(\Omega) \otimes H_0^1(\Omega)$. Now let R_h^1 be the orthogonal projection of $H_0^1(\Omega)$ onto V_h^0 (orthogonality in the $H_0^1(\Omega)$ -inner product), defined by

$$(2.1.21b) \quad a(R_h^1 y, v_h) = a(y, v_h) \quad \forall y \in H_0^1(\Omega), \quad v_h \in V_h^0.$$

It is well known [B1] that R_h^1 satisfies the inequality

$$(2.1.22) \quad \|R_h^1 y - y\|_s \leq Ch^{\alpha-s} \|y\|_\alpha, \quad 1 \leq \alpha \leq 2, \quad 0 \leq s \leq 1, \quad \alpha - s \geq 0, \quad y \in H^\alpha(\Omega),$$

which is analogous to (2.1.2) (but is more restricted in the α -range). By virtue of its definition, $(R_h^0 y, v_h) = (y, v_h)$, $y \in L_2(\Omega)$, $v_h \in V_h^0 \subset H_0^1(\Omega)$, the projection R_h^0 can be

extended to act on $H^{-1}(\Omega) = [\mathcal{D}(A^{1/2})]'$ (by (1.2.4)) where X' denotes the dual of X , and thus $R_h^0 A$ acts on $H_0^1(\Omega) = \mathcal{D}(A^{1/2})$.

CLAIM. *For the approximation A_h defined in (2.1.21a) we have*

$$A_h^{-1} R_h^0 A = R_h^1 \quad \text{on } (A^{1/2})' = H_0^1(\Omega).$$

In fact, if $y \in \mathcal{D}(A^{1/2})$ and $v_h \in V_h^0$, we have from (2.1.21a)

$$a(A_h^{-1} R_h^0 A y, v_h) = (A_h A_h^{-1} R_h^0 A y, v_h) = (R_h^0 A y, v_h) = (A y, v_h) \equiv a(y, v_h),$$

and the claim follows from the definition of R_h^1 in (2.1.21b). Thus, in view of the claim, assumption (H1) in § 2.1.A can be equivalently rewritten in our present example as

$$(H1') \quad \|A_h^\beta (R_h^0 - R_h^1) x\| \leq C h^{\alpha-2\beta} \|A^{\alpha/2} x\|, \quad x \in \mathcal{D}(A^{\alpha/2}),$$

for all $0 \leq \beta \leq 1, 1 \leq \alpha \leq 2$.

We now want to show that A_h defined in (2.1.21a) satisfies (H1'). To this end, we need the following

LEMMA 2.1.C.1. *For $y_h \in V_h^0$ and A_h as in (2.1.21a) we have*

$$\|A_h^\beta y_h\| \leq C h^{-\beta} \|A^{\beta/2} y_h\|, \quad 0 \leq \beta \leq 1.$$

Proof. We compute directly

$$\|A_h y_h\|^2 = (A_h y_h, A_h y_h) = a(y_h, A_h y_h) \leq C \|A^{1/2} y_h\| \|A^{1/2} A_h y_h\| \leq C h^{-1} \|A^{1/2} y_h\| \|A_h y_h\|$$

where in the last step we have used the “inverse assumption” on V_h^0 . Hence the case $\beta = 1$ follows and by interpolation the lemma is proved.

Now, for $x \in \mathcal{D}(A^{\alpha/2})$ for $1 \leq \alpha \leq 2$ (so that R_h^1 is well defined on x) and $0 \leq \beta < \frac{1}{2}$, the above lemma gives via identification (1.2.5a):

$$\|A_h^\beta (R_h^0 - R_h^1) x\| \leq C h^{-\beta} \|A^{\beta/2} (R_h^0 - R_h^1) x\| = C h^{-\beta} \|R_h^0 x - x + x - R_h^1 x\|_\beta$$

(by the approximating properties (2.1.2) for R_h^0 and (2.1.22) for R_h^1)

$$\leq C h^{\alpha-2\beta} \|x\|_\alpha \leq C h^{\alpha-2\beta} \|A^{\alpha/2} x\|,$$

where in the last step we have used (1.2.5b): thus assumption (H1') (equivalently (H1)) is proved in the present example.

As to assumption (H2), this can be satisfied by simply requiring $a_{ij} = a_{ji}$ for the coefficients of the principal part of the differential operator $A(\xi, \partial)$. (See [H1].)⁵

Approximation D_h which satisfies assumptions (2.1.5)–(2.1.6). We shall first provide an approximating subspace $U_h^1 \subset L_2(\Gamma)$ which satisfies the required property (2.1.4). D_h will then be defined as $U_h^1 \rightarrow V_h^0 \subset H_0^1(\Omega)$.

Let $V_h \supset V_h^0$ be the $\tilde{N}(h)$ -dimensional subspace of $H^1(\Omega)$ with the usual approximation properties that guarantee that V_h is an $S_h^{2,1}(\Omega)$ system (see [B1]) (example: subspace of piecewise linear functions). Construct the $M(h)$ -dimensional subspace $U_h^1 \subset L_2(\Gamma)$ as follows ($M(h) < \tilde{N}(h)$):

$$(2.1.23) \quad U_h^1 = \text{span} \{ \phi_h|_\Gamma; \phi_h \in V_h \}$$

where ϕ_h is a local basis in V_h . By [B1, Thm. 4.2.1, p. 101] we know that such U_h^1

⁵ Assumption (H2) was used only to prove Lemma 2.1.B.1. However, the result of Lemma 2.1.B.1 remains true provided that: $(A_n u_n, v_n) \leq C \|u_n\| \|v_n\|$ and $(A_n u_n, u_n) \geq \rho \|u_n\| \|u_n\|$, which conditions do not require that $a_{ij} = a_{ji}$. (For details see I. Lasiecka, “Convergence estimates for nonselfadjoint parabolic equations”, SIAM J. Numer. Anal., 21 (1984), to appear.)

can be viewed as an $S_h^{3/2,1/2}(\Gamma)$ -system. Hence by virtue of [B1, Thm. 4.1.1, p. 84] we have in particular that

$$|P_h u - u|_{-1/2+\varepsilon} \leq Ch^{1/2-\varepsilon} |u| \quad \text{for } \varepsilon \in [0, \tfrac{1}{2}],$$

P_h being the orthogonal projection of $L_2(\Gamma)$ onto U_h^1 , as postulated in (2.1.4). In the sequel, we shall also assume that U_h^1 satisfies the inverse property in the following form:

$$(2.1.24) \quad |u_h|_{1/2} \leq Ch^{-1/2} |u_h|, \quad u_h \in U_h^1.$$

Now let $U_h^1 \rightarrow u_h = b \cdot \phi_h|_\Gamma$, for some vector $b \in R^{M(h)}$, where \cdot is the dot product, and define an extension operator $\eta: U_h^1 \rightarrow V_h$ by $\eta u_h = b \cdot \phi_h$. Next, define the operator $D_h^1: U_h^1 \rightarrow V_h^0 + V_h$ by $D_h^1 u_h = C_h + \eta u_h$, where $C_h = A_h^{-1}(A(x, \partial) \eta u_h)$ is determined via the bilinear form: $a(C_h, \phi_h^0) = -a(\eta u_h, \phi_h^0)$, $\phi_h^0 \in V_h^0$. We can see that D_h^1 can be identified with the finite dimensional operator (matrix) from $R^{M(h)}$ into $R^{N(h)} \oplus R^{\tilde{N}(h)}$ given by

$$D_h^1 u_h = (l(u_h))^T M_h^{-1} \phi_h^0 + b \cdot \phi_h$$

where $M_h = [a(\phi_{h,j}^0, \phi_{h,i}^0)]$, $i, j = 1, \dots, \tilde{N}(h)$, $-l(u_h) = [a(\eta u_h, \phi_{h,j}^0)]$, $j = 1, \dots, \tilde{N}(h)$. We observe also that since $\eta u_h|_\Gamma = u_h$ we have

$$(2.1.25) \quad a(D_h^1 u_h, \phi_h^0) = 0, \quad \phi_h^0 \in V_h^0,$$

$$(2.1.26) \quad D_h^1 u_h|_\Gamma = u_h.$$

(2.1.25) and (2.1.26) indicate in which sense the operator D_h^1 is an approximation of the Dirichlet map D . In fact, recall that D satisfies

$$(2.1.27) \quad a(Du, \phi^0) = 0, \quad \phi^0 \in H_0^1(\Omega),$$

$$(2.1.28) \quad Du|_\Gamma = u.$$

Therefore (2.1.25) and (2.1.26) can be viewed as a discrete version of (2.1.27), (2.1.28).

We prove now that the following estimates hold:

$$(2.1.29) \quad \|(D - D_h^1)u_h\|_\varepsilon \leq Ch^{1/2-\varepsilon} |u_h|, \quad 0 \leq \varepsilon \leq \tfrac{1}{2},$$

$$(2.1.30) \quad \|D_h^1 u_h\|_{1/2-\varepsilon} \leq C |u_h|.$$

Proof of (2.1.29) and (2.1.30). It can be readily seen that

$$(D - D_h^1)u_h = A^{-1}(A(x, \partial) \eta u_h) - A_h^{-1}(A(x, \partial) \eta u_h).$$

Now by [B1, Thm. 6.3.3, p. 195] with $\alpha = \varepsilon$, $l = 1$, $\mu = 1 - \varepsilon$, by (2.1.22) and by [L3, Props. 12.1], we obtain

$$(2.1.31) \quad \begin{aligned} \|(D - D_h^1)u_h\|_\varepsilon &\leq Ch^{1-\varepsilon} \|A(x, \partial) \eta u_h\|_{-1} \\ &\leq Ch^{1-\varepsilon} \|\eta u_h\|_1 \leq Ch^{1-\varepsilon} |u_h|_{1/2} \leq Ch^{1/2-\varepsilon} |u_h| \end{aligned}$$

(since η can be viewed as a right inverse of the trace operator). As for (2.1.30), by (1.2.2) and [B1, Thm. 6.3.3, p. 195] with $\alpha = \frac{1}{2} - \varepsilon$, $l = 1$ we have:

$$\begin{aligned} \|D_h^1 u_h\|_{1/2-\varepsilon} &\leq \|(D - D_h^1)u_h\|_{1/2-\varepsilon} + \|Du_h\|_{1/2-\varepsilon} \\ &\leq \|A^{-1}(A(x, \partial) \eta u_h) - A_h^{-1}(A(x, \partial) \eta u_h)\|_{1/2-\varepsilon} + C |u_h| \\ &\leq Ch^{1/2+\varepsilon} \|A(x, \partial) \eta u_h\|_{-1} + C |u_h| \\ &\leq Ch^{1/2+\varepsilon} |u_h|_{1/2} + C |u_h| \leq C |u_h|, \end{aligned}$$

which completes the proof of (2.1.30).

Finally, as an approximation D_h of the Dirichlet map we take

$$(2.1.32) \quad D_h u_h = R_h^0 D_h^1 u_h.$$

It is readily seen that in view of (2.1.29) and (2.1.30)—or D_h^1 and the approximation properties of R_h^0 in (2.1.2)—map $D_h: V_h^1 \rightarrow V_h^0$ complies with all the requirements postulated by (2.1.5) and (2.1.6).

2.2. Approximation by nonconforming elements, which are not required to satisfy homogeneous Dirichlet B.C. In the previous subsection, an approximating scheme for the generator $-A$ and its semigroup $S(t)$ was given by means of the subspaces $V_h^0 \subset H_0^r(\Omega)$, $r > 1$. In the analysis, use of these was crucial on the level of estimates (2.1.19a) and (2.1.19b). Thus the approximating elements were required to satisfy *homogeneous Dirichlet* B.C. Generally, this requirement is not easy to meet, and its implementation in practice is achieved by use of curvilinear elements (see [Z1], [Z2]). It is precisely this difficulty that has stimulated research in the area of elliptic equations with the aim of approximating elliptic (Dirichlet) problems by means of spaces which do not satisfy zero B.C. (see [B1], [N1], [B2]). Our goal in this subsection is two-fold:

- (i) we wish to extend the aforementioned elliptic results to parabolic equations with nonhomogeneous (Dirichlet) boundary conditions; and
- (ii) we wish to find an estimate in the $C([0, T]; L_2(\Omega))$ -norm on the rate of convergence of the solutions subject to boundary inputs $u \in L_\infty(0, T; L_2(\Gamma))$.

Remark. Note that the question of using spaces with (or without) boundary conditions becomes irrelevant in the Neumann case. In fact, in order to approximate the generator corresponding to zero Neumann boundary condition, we use subspaces of $H^1(\Omega)$ (no boundary conditions).

2.2.A. Statement of results.

Assumptions. Let V_h be a finite dimensional approximating subspace of $H^k(\Omega)$ for some $k \geq 1$ such that

- (a) V_h is an $S_h^{r,k}(\Omega)$ system;
- (2.2.1) (b) V_h satisfies the inverse property

$$\|V_h\|_1 \leq Ch^{-1} \|V_h\|, \quad V_h \in V_h.$$

As a consequence of (2.2.1) we have:

$$(2.2.2) \quad \|R_h y - y\|_s \leq Ch^{\alpha-s} \|y\|_\alpha, \quad \alpha - s \geq 0 \quad \text{for all } 0 \leq s \leq k, \quad \alpha \leq r,$$

where R_h is the orthogonal projection of $L_2(\Omega)$ on V_h . We explicitly note that elements of V_h are not required to satisfy zero Dirichlet B.C.

Let $A_h: V_h \rightarrow V_h$ be a family of approximations of A having the following properties:

$$(F1) \quad \|A_h^\beta x_h\| \leq Ch^{-\beta} \|x_h\|_\beta, \quad 0 \leq \beta \leq 1;$$

$$(F2) \quad \|(R_h A^{-1} - A_h^{-1} R_h) x\| \leq Ch^{2+\alpha} \|A^{\alpha/2} x\|, \quad x \in D(A^{\alpha/2});$$

and for $0 \leq \alpha + 2 \leq r$, $0 \leq \alpha \leq \frac{1}{2}$

- (F3) (i) either A_h^2 is positive
- (ii) or else $((A_h - A_h^*) V_h, y_h) \leq C(A_h V_h, y_h)^{1/2} \|y_h\|$.

As an *approximation* of a *parabolic equation* we take as before

$$(2.2.3) \quad y_h(t) = S_h(t)R_h y_0 - A_h \int_0^t S_h(t-z)D_h u_h(z) dz$$

where

$$S_h(t) = e^{-A_h t}, \quad t \in \mathbb{R}, \quad \text{on } V_h$$

and $D_h: U_h^1 \rightarrow V_h$ is an approximation of the Dirichlet map while possessing properties (2.1.5) and (2.1.6).

We shall prove the following result:

THEOREM 2.2.1. *Let A_h satisfy hypothesis (F1)–(F3), and let D_h satisfy (2.1.5)–(2.1.6). Then for $y_0 \in H^{1/2-\varepsilon}(\Omega)$ and for any $u_h \in U_h^1$ we have:*

$$\|y(u, t) - y_h(u_h, t)\| \leq Ch^{1/2-\varepsilon} \{\|y_0\|_{1/2-\varepsilon} + \|u_h\|_{L_\infty[0T; L_2(\Gamma)]} + C\|u - u_h\|_{L_\infty[0T; H^{-1/2+\varepsilon}(\Gamma)]}\}.$$

As in § 2.1 for $u_h(t) = P_h u(t)$, we have similarly Corollary 2.1.2.

2.2.B. Proof of Theorem 2.2.1. Similarly as in the case of Theorem 2.2.1, we formulate a series of lemmas which are conceptual counterparts of Lemmas 2.1.B.1, 2.1.B.2 and 2.1.B.3.

LEMMA 2.2.B.1. *Let $x_h \in V_h$. Then*

$$\|A_h^\beta S_h(t)x_h\| \leq \frac{C_T}{t^\beta} \|x_h\|, \quad 0 \leq \beta \leq 1.$$

LEMMA 2.2.B.2. *For any $x \in L_2(\Omega)$ we have:*

$$S_h(t)R_h x - R_h S(t)x = - \int_0^t A_h S_h(t-z)[R_h A^{-1} - A_h^{-1} R_h] A S(z)x dz.$$

The proofs of Lemmas 2.2.B.1 and 2.2.B.2 go along the same lines as the proofs of Lemmas 2.1.B.1 and 2.1.B.2.

The next lemma deals with the approximation properties of $S(t)$. Its proof is a somewhat technical variation of the proof of Lemma 2.1.B.3.

LEMMA 2.2.B.3. *For $x \in D(A^{\alpha/2})$, we have*

$$(i) \quad \|S_h(t)R_h x - R_h S(t)x\| \leq Ch^s \|A^{s/2}x\|, \quad 0 \leq s < \alpha + 2,$$

$$(ii) \quad \|A_h S_h(t)R_h x - R_h A S(t)x\| \leq \frac{C}{t^{1-\varepsilon}} h^{\alpha-2\varepsilon} \|A^{\alpha/2}x\|$$

for $0 < \alpha \leq \frac{1}{2}$ and ε arbitrarily small.

Proof. First let us notice that due to (F1) and the inverse property (F2) implies that:

$$(F2)' \quad \|A_h^\beta [R_h A^{-1} - A_h^{-1} R_h]x\| \leq Ch^{2+\alpha-2\beta} \|A^{\alpha/2}x\| \quad \text{for } 0 \leq \beta \leq 1, \quad 0 \leq \alpha \leq \frac{1}{2}.$$

In fact by (F1) and inverse approximation property

$$\|A_h^\beta [R_h A^{-1} - A_h^{-1} R_h]x\| \leq Ch^{-2\beta} \|(R_h A^{-1} - A_h^{-1} R_h)x\|$$

(by F2)

$$\leq Ch^{2+\alpha-2\beta} \|A^{\alpha/2}x\|.$$

□

Now we proceed with the proof of Lemma 2.2.B.3.

Part (i). By Lemma 2.2.B.2,

$$\begin{aligned}
 \|S_h(t)R_h x - R_h S(t)x\| &\leq \int_0^t \|A_h S_h(t-z)[R_h A^{-1} - A_h^{-1} R_h] \cdot AS(z)x\| dz \\
 &= \int_0^t \|A_h^{1-\varepsilon} S_h(t-z) A_h^\varepsilon [R_h A^{-1} - A_h^{-1} R_h] AS(z)x\| dz \\
 &\quad \text{(by Lemma 2.2.B.1 and (F2)' with } \varepsilon \in (0, 1)) \\
 &\leq Ch^{2+\alpha-2\varepsilon} \int_0^t \frac{1}{(t-z)^{1-\varepsilon}} \|A^{1+\alpha/2} S(z)x\| dz \\
 &\quad \text{(by analytic estimates of } S(t)) \\
 &\leq Ch^{2+\alpha-2\varepsilon} \int_0^t \frac{1}{(t-z)^{1-\varepsilon}} \frac{dz}{z^\varepsilon} \|A^{1+\alpha/2-\varepsilon} x\| \\
 &= Ch^{2+\alpha-2\varepsilon} \|A^{1+\alpha/2-\varepsilon} x\| \beta(1-\varepsilon, \varepsilon).
 \end{aligned}$$

Letting ε go from 0 to 1 we obtain part (i).

Proof of part (ii). To prove (ii) we differentiate in t the identity of Lemma 2.2.B.2 to get for $x \in D(A)$

$$\begin{aligned}
 A_h S_h(t) R_h x - R_h AS(t)x &= A_h [R_h A^{-1} - A_h^{-1} R_h] AS(t)x \\
 (2.2.4) \quad &+ \int_0^t A_h S_h(t-z) A_h [R_h A^{-1} - A_h^{-1} R_h] AS(z)x dz.
 \end{aligned}$$

We now estimate the two terms in (2.2.4). As for the first term we apply (F2') with $\beta = 1$ and $\alpha = \alpha - 2\varepsilon$ to get

$$\|A_h [R_h A^{-1} - A_h^{-1} R_h] AS(t)x\| \leq Ch^{\alpha-2\varepsilon} \|A^{\alpha/2-\varepsilon} AS(t)x\| \leq \frac{Ch^{\alpha-2\varepsilon}}{t^{1-\varepsilon}} \|A^{\alpha/2} x\|.$$

As for the second term, we have similarly

$$\begin{aligned}
 &\int_0^t \|A_h S_h(t-z) A_h [R_h A^{-1} - A_h^{-1} R_h] AS(z)x\| dz \\
 &\leq \int_0^t \|A_h^{1-\varepsilon} S_h(t-z) A_h^{1+\varepsilon} [R_h A^{-1} - A_h^{-1} R_h] AS(z)x\| dz \\
 &\quad \text{(by Lemma 2.2.B.1 and (F1))} \\
 &\leq \int_0^t \frac{1}{(t-z)^{1-\varepsilon}} h^{-2\varepsilon} \|A_h [R_h A^{-1} - A_h^{-1} R_h] AS(z)x\| dz \\
 &\quad \text{(by (F2') applied with } \beta = 1 \text{ and } \alpha = \alpha - 2\varepsilon) \\
 &\leq Ch^{\alpha-4\varepsilon} \int_0^t \frac{1}{(t-z)^{1-\varepsilon}} \|A^{\alpha/2-\varepsilon} AS(z)x\| dz \\
 &\leq Ch^{\alpha-4\varepsilon} \int_0^t \frac{1}{(t-z)^{1-\varepsilon}} \frac{1}{z^{1-\varepsilon}} dz \|A^{\alpha/2} x\| \\
 &\leq \frac{Ch^{\alpha-4\varepsilon}}{t^{1-2\varepsilon}} \|A^{\alpha/2} x\| \beta(1-\varepsilon, 1-\varepsilon),
 \end{aligned}$$

which completes the proof of part (ii).

We are now in a position to proceed with the proof of Theorem 2.2.1.

By (1.2.7) and (2.2.3) the difference between y and y_h can be written as

$$(2.2.5) \quad \begin{aligned} y(u, t) - y_h(u_h, t) &= S(t)y_0 - S_h(t)R_h y_0 - \int_0^t [AS(t-z) - A_h S_h(t-z)] D_h u_h(z) \, dz \\ &\quad + \int_0^t AS(t-z) [D_h u_h(z) - Du(z)] \, dz. \end{aligned}$$

The estimates of the first and the third terms on the right-hand side of (2.2.5) follow along the same lines as for the proof of Theorem 2.1, with the use of Lemma 2.2.B.3 part (i) and the properties of D_h postulated by (2.1.5) and (2.1.6).

As for the second term in (2.2.5) we have:

$$(2.2.6) \quad \begin{aligned} &\left\| \int_0^t (AS(t-z) - A_h S_h(t-z)) D_h u_h(z) \, dz \right\| \\ &\leq \int_0^t \|(R_h - I)AS(t-z) D_h u_h(z)\| \, dz \\ &\quad + \int_0^t \|(R_h AS(t-z) - A_h S_h(t-z) R_h) D_h u_h(z)\| \, dz \\ &\hspace{15em} (\text{by (2.2.2) and Lemma 2.2.B.3 (ii)}) \\ &\leq Ch^{1/2-4\varepsilon} \int_0^t \|AS(t-z) D_h u_h(z)\|_{1/2-4\varepsilon} \, dz \\ &\quad + Ch^{1/2-4\varepsilon} \int_0^t \frac{1}{(t-z)^{1-\varepsilon}} \|A^{1/4-\varepsilon} D_h u_h(z)\| \, dz \\ &\leq Ch^{1/2-4\varepsilon} \int_0^t \frac{1}{(t-z)^{1-\varepsilon}} \|D_h u_h(z)\|_{1/2-2\varepsilon} \, dz \\ &\hspace{15em} (\text{by 2.1.6}) \\ &\leq Ch^{1/2-4\varepsilon} \|u_h\|_{L_\infty[0,T; L_2(\Gamma)]}, \end{aligned}$$

which completes the proof of the theorem. \square

Remark 2.2. Notice that assumption (F2) requires that $0 \leq \alpha \leq \frac{1}{2}$; hence r must be greater than $2\frac{1}{2}$. Thus one has to use a higher order method than in § 2.1. As a consequence of this we obtain Lemma 2.2.B.3 part (ii), valid for all $0 \leq \alpha < \frac{1}{2}$ (compare with Lemma 2.1.B.3 where $\alpha \geq 1$), and in the proof of the major estimate, we are in a position to apply the term $D_h u_h$ with $\alpha = \frac{1}{2} - 2\varepsilon$. This fact constitutes the main difference between the approaches in §§ 2.1 and 2.2. Using lower (say second) order methods, we had an estimate in Lemma 2.1.B.3(ii) valid for $\alpha > 1$ and hence we needed to estimate the term $\|A^{1/2+\varepsilon/2} D_h u_h\|$, which makes sense only for conforming elements.

To conclude, at the expenses of using a higher order method, we can use elements which do not satisfy zero boundary conditions.

2.2.C. Examples. Approximations A_h and D_h which satisfy assumptions (F1)–(F3) and (2.1.5), (2.1.6) respectively. We will provide examples of approximations A_h and D_h which comply with the requirements (F1)–(F3) and (2.1.5) and (2.1.6).

Approximations of A_h . An obvious example of an approximation to A_h which satisfies assumptions (F1)–(F3) is an already considered standard Galerkin approxima-

tion $A_h: V_h^0 \rightarrow V_h^0$:

$$(2.2.7) \quad (A_h y_h, v_h) = a(y_h, v_h) \quad \forall y_h, v_h \in V_h^0$$

with V_h^0 , this time, being the system $S^{r,k}$ for $r \geq 2\frac{1}{2}$ (in order to satisfy condition (F2)). This requirement ($r \geq 2\frac{1}{2}$) precludes the use of piecewise linear functions. However, the advantage of the approach in § 2.2 is that these techniques *do not require subspaces* V_h to satisfy boundary conditions. As an example of such discretization we quote approximation A_h provided by Nitsche [N1].

Let $V_h \subset H^1(\Omega)$ be a family of finite dimensional subspaces with the following properties:

$$(2.2.8) \quad \begin{aligned} & (a) \quad V_h|_{\Gamma} \subset H^1(\Gamma), \\ & (b) \quad \inf_{v_h \in V_h} \{ \|v - v_h\| + h\|v - v_h\|_1 + h^{1/2}|v - v_h| + h^{3/2}|v - v_h|_1 \} \leq Ch^s \|v\|_s, \\ & \hspace{25em} 2 \leq s \leq r, \\ & (c) \quad \left| \frac{\partial v_h}{\partial \eta} \right| \leq Ch^{-1/2} \|v_h\|_1, \\ & (d) \quad |v_h| \leq Ch^{1/2} \|v_h\|_1, \quad v_h \in V_h, \\ & (e) \quad \|v_h\|_1 \leq Ch^{-1} \|v_h\|. \end{aligned}$$

Let $A_h: V_h \rightarrow V_h$ be defined as follows:

$$(2.2.9) \quad (A_h u_h, v_h) = a(u_h, v_h) - \left\langle u_h, \frac{\partial}{\partial \eta} v_h \right\rangle - \left\langle \frac{\partial u_h}{\partial \eta}, v_h \right\rangle + \beta h^{-1} \langle u_h, v_h \rangle, \quad u_h, v_h \in V_h.$$

It is readily seen that for a suitable choice of $\beta > 0$ the approximation A_h is positive, and moreover,

$$\begin{aligned} \|A_h u_h\|^2 &= a(u_h, A_h u_h) - \left\langle u_h, \frac{\partial}{\partial \eta} A_h u_h \right\rangle - \left\langle A_h u_h, \frac{\partial}{\partial \eta} u_h \right\rangle + \beta h^{-1} \langle u_h, A_h u_h \rangle \\ &\hspace{25em} \text{(by (2.2.8c, d))} \\ &\leq C \|u_h\|_1 \|A_h u_h\|_1 + Ch^{1/2} \|u_h\|_1 h^{-1/2} \|A_h u_h\|_1 \\ &\quad + C \beta h^{-1} h^{1/2} \|u_h\|_1 h^{1/2} \|A_h u_h\|_1 \\ &\leq C \|u_h\|_1 \|A_h u_h\|_1 \\ &\hspace{25em} \text{(by (2.2.8e))} \\ &\leq Ch^{-1} \|u_h\|_1 \|A_h u_h\|. \end{aligned}$$

Hence $\|A_h u_h\| \leq Ch^{-1} \|u_h\|_1$ which by interpolation yields (F1).

As for condition (F2), this is a standard error estimate claimed to hold in [N1]. The validity of hypothesis (F3)(ii) can be asserted as in [H1].

Therefore all the requirements (F1)–(F3) are satisfied and Theorem 2.2 applies.

Other examples of an approximation A_h which can be applied to nonconforming elements and can satisfy our requirements are *Babuska's method* formulated in [B2], [H1] and the *method of interpolated boundary conditions* (see [S5]). Both methods are formally defined through relation (2.2.6); therefore, justification of hypothesis (F1) goes as for the standard Galerkin method; condition (F2) is a result of error estimates given in [B2] and [S1] respectively. As for verification of condition (F3)(ii) we refer to [H1].

Now, let us turn to approximations of D_h .

Approximations of D_h . As an example of approximation $D_h: U_h^1 \rightarrow V_h$ one can obviously take $D_h = R_h D_h^1$ with D_h^1 defined as in § 2.1.C. This construction requires however that $V_h^1 = \text{span} \{\text{traces of } V_h\}$. In order to be able to define D_h on a larger subspace of $L_2(\Gamma)$ (say, piecewise constant functions defined on Γ , used later in the approximation of our control problem), we resort to another technique introduced in [B3] and referred to as the Rayleigh–Ritz–Galerkin method:

Let $U_h^1 = \text{space of piecewise constant functions on } \Gamma$.

Let V_h be an $S^{4,2}(\Omega)$ system (for example cubic splines). Define $D_h: U_h^1 \rightarrow V_h$ by

$$(2.2.10) \quad (-A(x, \partial) D_h u_h, A(x, \partial) v_h) + h^{-3} \langle u_h - D_h u_h, v_h \rangle_\Gamma = 0, \quad v_h \in V_h.$$

It was shown in [B3, Thm. 4.1 with $r=4$ and $\gamma=\frac{3}{2}$] that

$$\|D_h u_h - Du_h\|_l \leq Ch^{1/2-l} |u_h|, \quad 0 \leq l \leq \frac{1}{2}.$$

Hence in particular we obtain

$$\|D_h u_h - Du_h\|_\varepsilon \leq Ch^{1/2-\varepsilon} |u_h| \quad \text{and} \quad \|D_h u_h\|_{1/2-\varepsilon} \leq C |u_h| \quad \text{for } \varepsilon \in [0, \frac{1}{2}].$$

Remark. Note that the space V_h is not required to satisfy zero boundary conditions, and the algorithm (2.9) can be applied to *any function* $u_h \in L_2(\Gamma)$. On the negative side, in order to obtain the optimal rate of convergence, we have to use higher order splines (say cubic). Observe also that D_h no longer has the property of being a harmonic function on a finite dimensional subspace (i.e. $(A_h D_h' u_h, \phi_h^0) = 0$) and $D_h' u_h|_\Gamma = u_h$, as opposed to the definition of D_h^1 given in § 2.1.C.

3. Approximation of control problem P_A . In the present section we formulate a discrete version of our control problem, and we prove the corresponding optimal rate of convergence. To accomplish this, we shall apply the theoretical results of the approximation of nonhomogeneous (boundary) parabolic equations obtained in § 2.

3.1. Statement of results. In order to formulate a discrete version of control problem P_A , let us define:

$$(3.1.1) \quad U_h^A = U_A \cap U_h$$

where $U_h = L_\infty[0T; U_h^1]$ with U_h^1 as in § 2.1.C. An approximation of the control problem P_A can now be stated as follows:

$$P_h^A. \text{ Find } \min \{T_h; \exists u_h \in U_h^A; \|y_h(u_h, T_h) - R_h^0 y_1\| \leq \delta\}$$

with y_h given by (2.1.7), with A_h defined by (2.1.21) and D_h by (2.1.32). Since the assumptions (H1), (H2), (2.1.5) and (2.1.6) are satisfied by the aforementioned approximations A_h and D_h , Theorem 2.1 and its corollary apply. As a consequence of assumption (1.1.2), we have the following:

CLAIM 3.1. *There exist $T > 0$ and $u_h \in U_h^A$ such that*

$$\|y_h(u_h, T) - R_h^0 y_1\| \leq \delta.$$

Proof. Let T and \bar{u}_A be as in (1.1.2). Let $u_h(t) = P_h \bar{u}_A(t) \in U_h^A$ where P_h is defined as in § 2.1.C. By the triangle inequality

$$(3.1.2) \quad \|y_h(u_h, T) - R_h^0 y_1\| \leq \|y(\bar{u}_A, T) - y_1\| + \|y_h(u_h, T) - y(\bar{u}_A, T)\| + \|y_1 - R_h^0 y_1\|.$$

Since the last two terms on the right-hand side of (3.1.2) go to zero via (2.1.4), Theorem 2.1 and (2.1.2) combined with (1.1.2), Claim 3.1 follows. Since P_h^A is a standard time optimal control problem for o.d.e.'s, it possesses, in view of Claim 3.1, the unique solution $u_{h,A}^0, T_{h,A}^0$ with a bang-bang property, i.e.,

$$(3.1.3) \quad |u_{h,A}^0(t)| = 1, \quad t \in [0, T_{h,A}^0],$$

$$(3.1.4) \quad \|y_h^0(u_{h,A}^0, T_h) - R_h^0 y_1\| = \delta.$$

The major goal of this section is to establish convergence properties of the control problem P_h^A . Toward this end, we have

THEOREM 3.1. *Let u_A^0, T_A^0 (resp. $u_{h,A}^0, T_{h,A}^0$) be an optimal solution of the control problem P_A (resp. $P_{h,A}$). Then for $y_1, y_0 \in H^{1/2-\varepsilon}(\Omega)$ we have*

- (i) $\|y(u_{h,A}^0, T_{h,A}^0) - y_1\| \leq \delta + O(h^{1/2-\varepsilon})$,
- (ii) $T_{h,A}^0 \rightarrow T_A^0$,
- (iii) $u_{h,A}^0 \rightarrow u_A^0$ in $L_2[0T \times \Gamma]$ and $y_h(u_{h,A}^0, T_{h,A}^0) \rightarrow y(u^0, T^0)$ in $L_2[\Omega]$.

Remark 1. Observe that the rate of convergence in (i) is optimal, in view of the regularity of $y(u^0)$ (see (1.2.7)).

Remark 2. Notice, that in order to approximate the control problem P_A , we use linear (curvilinear) splines (spaces V_h^0 and V_h introduced in § 2).

3.2. Proof of Theorem 3.1. Observe first that by optimality of $T_{h,A}^0$, and by Claim 3.1 we have:

$$(3.2.1) \quad T_{h,A}^0 \leq T \quad \text{for all } h > 0.$$

We proceed now with a proof of (i), which in fact is almost an immediate consequence of Theorem 2.1. For, by the triangle inequality, we estimate

$$(3.2.2) \quad \|y(u_{h,A}^0, T_{h,A}^0) - y_1\| \leq \|y(u_{h,A}^0, T_{h,A}^0) - y_h(u_{h,A}^0, T_{h,A}^0)\| \\ + \|y_1 - R_h^0 y_1\| + \|y_h(u_{h,A}^0, T_{h,A}^0) - R_h^0 y_1\|.$$

By Theorem 2.1, (3.2.1), and (2.1.2), the first two terms on the right-hand side of (3.2.2) can be estimated by

$$Ch^{1/2-\varepsilon}[\|y_0\|_{1/2-\varepsilon} + \|y_1\|_{1/2-\varepsilon} + |u_{h,A}^0|_{L_\infty[0T, L_2(\Gamma)]}] = O(h^{1/2-\varepsilon}),$$

which in view of (3.1.4) completes the proof of (i). In order to prove (ii), we shall first show that:

$$(3.2.3) \quad \forall \varepsilon > 0 \quad \exists H, \quad \forall h > H \quad T_{h,A}^0 - T_A^0 < \varepsilon.$$

For, since $\|y(u_A^0, T_A^0) - y_1\| = \delta$, due to the approximate controllability of the system, we can extend u_A^0 for $t \in [T_A^0, T_A^0 + \varepsilon]$ in such a way that $u_{\text{ext}}^0 \in U_A$ and that

$$(3.2.4) \quad \|y(u_{\text{ext}}^0, T_A^0 + \varepsilon) - y_1\| < \delta.$$

Now we shall show that $P_h u_{\text{ext}}^0 \in U_h^A$ steers $R_h^0 y_0$ to the ball $\|y_h - R_h^0 y_1\| \leq \delta$ in time $T_A^0 + \varepsilon$. In fact

$$(3.2.5) \quad \|y_h(P_h u_{\text{ext}}^0, T_A^0 + \varepsilon) - R_h^0 y_1\| \\ \leq \|y_h(P_h u_{\text{ext}}^0, T_A^0 + \varepsilon) - y(u_{\text{ext}}^0, T_A^0 + \varepsilon)\| + \|y_1 - R_h^0 y_1\| \\ + \|y(u_{\text{ext}}^0, T_A^0 + \varepsilon) - y_1\|.$$

By Theorem 2.1, (2.1.4) and (2.1.2), the first two terms on the right-hand side of (3.2.5) tend to zero. Therefore, by virtue of (3.2.4), we have

$$(3.2.6) \quad \|y_h(P_h u_{\text{ext}}^0, T_A^0 + \varepsilon) - R_h^0 y_1\| \leq \delta,$$

which together with the optimality of T_h^0 proves (3.2.3).

Let us now show that for all $\varepsilon > 0$ there is an H such that for all $h > H$

$$(3.2.7) \quad T_{hA}^0 - T_A^0 > -\varepsilon.$$

As before, using (3.1.4), we extend u_h^0 for $t \in [T_h^0, T_h^0 + \varepsilon] \subset [T_h^0, T + \varepsilon]$ (by 3.2.1), in such a way that $u_{h \text{ ext}}^0 \in U_A$ and

$$(3.2.8) \quad \|y_h(u_{h \text{ ext}}^0, T_{hA}^0 + \varepsilon) - R_h^0 y_1\| < \delta.$$

Since $u_{h \text{ ext}}^0$ is uniformly bounded in $L_\infty[0, T + \varepsilon, L_2(\Gamma)]$, we can subtract a weakly star convergent subsequence, say $u_{h \text{ ext}}^0 \rightarrow u^*$ weak star, where u^* by the Alcoulu theorem [D2, p. 424] belongs to U_A . We prove that u^* steers y_0 to the ball $\|y - y_1\| \leq \delta$ in time $T_{hA}^0 + \varepsilon$ for h small enough. In fact, let us write

$$(3.2.9) \quad \begin{aligned} \|y(u^*, T_{hA}^0 + \varepsilon) - y_1\| &\leq \|y(u^*, T_{hA}^0 + \varepsilon) - y(u_{h \text{ ext}}^0, T_{hA}^0 + \varepsilon)\| \\ &+ \|y_h(u_{h \text{ ext}}^0, T_{hA}^0 + \varepsilon) - y(u_{h \text{ ext}}^0, T_{hA}^0 + \varepsilon)\| \\ &+ \|R_h^0 y_1 - y_1\| + \|y_h(u_{h \text{ ext}}^0, T_{hA}^0 + \varepsilon) - R_h^0 y_1\|. \end{aligned}$$

By Theorem 2.1 and by (3.2.1) and (2.1.2), the second and the third terms on the right-hand side of (3.2.9) go to zero. The first term goes to zero by virtue of $u_{h \text{ ext}}^0 \rightarrow u^*$ weak star and by compactness of the map L_T (see the Appendix). Hence, due to (3.2.8), we have $\|y(u^*, T_{hA}^0 + \varepsilon) - y_1\| \leq \delta$ for h small, which in view of the optimality of T^0 , implies (3.2.7). Thus, (3.2.3) together with (3.2.7) implies (ii).

The proof of Theorem 3.1 will be completed as soon as we prove (iii).

As before, by the uniform boundedness of u_{hA}^0 in $L_\infty[0T; L_2(\Gamma)]$ we have $u_{hA}^0 \rightarrow u^*$ weak star. By the Alcoulu theorem, $u^* \in U_A$. We shall show that

$$(3.2.10) \quad y(u^*, T^0) - y_1 = \lim_{h \rightarrow 0} (y_h(u_{hA}^0, T_{hA}^0) - R_h^0 y_1).$$

Let us write

$$\begin{aligned} y_h(u_{hA}^0, T_{hA}^0) - R_h^0 y_1 - y(u^*, T^0) + y_1 \\ = [y_h(u_{hA}^0, T_{hA}^0) - y(u_{hA}, T_{hA}^0)] + [y_1 - R_h^0 y_1] + [y(u_{hA}^0, T^0) - y(u^*, T^0)] \\ + [y(u_{hA}^0, T_{hA}^0) - y(u_{hA}^0, T^0)]. \end{aligned}$$

By virtue of Theorem 2.1, (3.2.1), (2.1.2) and the Appendix, the expressions in the first three brackets go to zero in the $L_2(\Omega)$ norm. As for the fourth term, we have

$$\begin{aligned} \|y(u_{hA}^0, T_{hA}^0) - y(u_{hA}^0, T^0)\| \\ \leq \int_{T_{hA}^0}^{T_A^0} \|AS(T_{hA}^0 - z)Du_{hA}^0(z)\| + \int_{T_{hA}^0 - T_A^0}^{T_A^0} \|A(S(z) - S(T^0 - T_{hA}^0 + z))Du_{hA}^0(z)\| dz. \end{aligned}$$

Both integrals go to zero, since $T_{hA}^0 \rightarrow T_A^0$ and the integrands are in $L^1[0T]$ ([D2, p. 632]). Thus (3.2.10) is established.

Since $\|y_h(u_{hA}^0, T_{hA}^0) - R_h^0 y_1\| \leq \delta$, by closedness of the ball we get

$$(3.2.11) \quad \|y(u^*, T_A^0) - y_1\| \leq \delta.$$

Therefore, by uniqueness of the optimal control u^0 and by (3.2.11), $u^* = u_A^0$. Since $|u^0(t)| = 1$; $t \in [0, T_A^0]$ and $|u_h^0(t)| = 1$, $t \in [0, T_{hA}^0]$, then

$$(3.2.12) \quad |u^0|_{L_2[0T_A^0, L_2(\Gamma)]}^2 = T_A^0,$$

$$(3.2.13) \quad |u_h^0|_{L_2[0T_{hA}^0, L_2(\Gamma)]}^2 = T_{hA}^0.$$

Consequently, weak convergence of u_{hA}^0 to u_A^0 implies strong convergence in $L_2[0T; L_2(\Gamma)]$. In fact, after extending u_A^0 to zero for $t > T_A^0$ and u_{hA}^0 for $t > T_{hA}^0$, we obtain:

$$\begin{aligned} |u_A^0 - u_{hA}^0|_{L_2[0T; L_2(\Gamma)]} &= (u_A^0 - u_{hA}^0, u_A^0 - u_{hA}^0)_{L_2[0T; L_2(\Gamma)]} \\ &= |u_A^0|_{L_2[0T; L_2(\Gamma)]}^2 + |u_{hA}^0|_{L_2[0T; L_2(\Gamma)]}^2 - 2(u_{hA}^0, u_A^0). \end{aligned}$$

By weak convergence of u_{hA}^0 to u_A^0 and by (ii), $|u_A^0 - u_{hA}^0|_{L_2[0T; L_2(\Gamma)]}^2$ converges to

$$T_A^0 + \lim_{h \rightarrow 0} T_{hA}^0 - 2T_A^0 = \lim_{h \rightarrow 0} T_{hA}^0 - T_A^0 = 0,$$

which proves that $u_{hA}^0 \rightarrow u_A^0$.

Thus by virtue of Theorem 2.1.1 we also have

$$(3.2.14) \quad y(u_A^0, T^0) - y_h(u_{hA}^0, T_{hA}^0) \rightarrow 0, \quad h \rightarrow 0 \quad \text{in } L_2(\Omega).$$

which completes the proof of (iii).

4. Approximation of the control problem P_B . Recall that the optimal control corresponding to problem P_B is bang-bang in space and time; i.e.,

$$|u_B^0(x, t)| = 1 \quad \text{a.e. } x, t \in \Gamma \times [0, T_B^0].$$

In order to make the discrete problem well posed, the space of approximating controls U_h must contain bang-bang functions (in space and time). This is certainly not the case with $U'_h = \text{traces } V_h$, as defined in the previous section. Therefore, for the control problem P_B , we are forced to redefine the space of discrete controls in such a way that it still contains the functions $u(x, t) = \pm 1$. The most natural way of proceeding is to set:

$$U'_h = \text{piecewise constant functions on } \Gamma \in L_2(\Gamma),$$

$$(4.1) \quad U_h = L_\infty[0T; U'_h].$$

However, in this case, we can no longer apply the definition of D_h given by (2.2.5) (since its construction relies on selecting $U'_h = \text{space of traces}$). In order to cope with the above difficulty we apply the Rayleigh-Ritz method to approximating the Dirichlet map. Since this method requires the use of higher order subspaces (say $S_h^{4,2}$), it is advantageous to also approximate A_h with $V_h \subset S_h^{4,2}$ *without* requiring zero boundary conditions.

Therefore, let $V_h \subset S_h^{4,2}$ be such that conditions (2.2.1) and (2.2.8) are satisfied with $r=4$ and $k=2$. Let $A_h: V_h \rightarrow V_h$ be defined as in (2.2.9) and let $D_h: U_h^1 \rightarrow V_h$ be defined by (2.2.10). As an approximation of the trajectory, we take as usual

$$(4.2) \quad y_h(t) = S_h(t)R_h^0 y_0 + A_h \int_0^t S_h(t-z)D_h u_h(z) dz.$$

In this case, we have shown that all the assumptions (F1)–(F3) and (2.1.5) and (2.1.6) are satisfied. Thus Theorem 2.2 applies and gives, for $y_0 \in H^{1/2-\epsilon}(\Omega)$, $t \in [0, T]$,

$$(4.3) \quad \|y(u, t) - y_h(u_h, t)\| \leq Ch^{1/2-\epsilon} [\|y_0\|_{1/2-\epsilon} + |u_h|_{L_\infty[0T; L_2(\Gamma)]}] + C|u - u_h|_{L_\infty[0T; H^{-1/2+\epsilon}(\Gamma)]}$$

where $y(u, t)$ is a solution of (1.1.1), y_h is given by (4.2) and constant C does not depend on h .

Now we are in a position to formulate an approximation of the control problem P_B , say P_h^B

$$P_h^B. \text{ Find } \min \{T_h; u_h \in U_h \cap U_B = U_h^B; \|y_h(u_h, T_h) - R_h y_1\| \leq \delta\}$$

with y_h defined by (4.2). Let P_h be an orthogonal projection on U_h^1 . In particular we have

$$|P_h u - y|_{-1/2-\varepsilon} \leq Ch^{1/2-\varepsilon}|u|.$$

Note also that $P_h(U_B) \subset U_B$, hence $P_h(U_B) \subset U_h^B$. As in the previous case, we have

CLAIM 4.1. *There exist $T > 0$ and $u_h \in U_h^B$ such that*

$$\|y_h(u_h, T) - R_h^0 y_1\| \leq \delta.$$

The proof goes the same way as that of Claim 3.1, after setting $u_h = P_h u$ with u given in (1.1.2) and using (4.3) instead of Theorem 2.1.

Therefore, by applying standard separation arguments, we can establish the existence of an optimal solution to problem P_h^B , say u_{hB}^0, T_{hB}^0 , such that

$$(4.4) \quad |u_{hB}^0(x, t)| = 1, \quad x, t \in \Gamma \times [0, T_h^0],$$

$$(4.5) \quad \|y_h^0(u_{hB}^0, T_h^0) - R_h^0 y_1\| = \delta.$$

Finally, we have the following convergence result:

THEOREM 4.1. *Let $y_0, y_1 \in H^{1/2-\varepsilon}(\Omega)$, u_B^0 , (resp. u_{hB}^0) be the optimal solution of the problem P_B (resp. P_{hB}). Then*

- (i) $\|y(u_{hB}^0, T_{hB}^0) - R_h^0 y_1\| \leq \delta + O(h^{1/2-\varepsilon})$,
- (ii) $T_{hB}^0 \rightarrow T_B^0$,
- (iii) $u_{hB}^0 \rightarrow u_B^0$ in $L_2[0T; L_2(\Gamma)]$ and $y_{hB}^0(u_{hB}^0, T_{hB}^0) \rightarrow y^0(u_B^0, T^0)$ in $L_2(\Omega)$.

A proof of Theorem 4.1 parallels the treatment given in § 3.2, but instead of Theorem 2.1 we now use Theorem 2.2.

Remarks. 1. Notice that in the case of control problem P_B we still obtain the optimal rate of convergence, at the expense, however, of using higher order (cubic) splines.

2. Observe that the scheme (4.2) can also be applied to approximating problem P_B .

Conclusions. 1. The approach of approximating the semigroup model allows us to treat separately the approximation of the semigroup and that of the Dirichlet map (responsible for the elliptic part of the problem). This enables us to combine independently the desirable features of both approximations in order to best fit our needs. In fact,

(i) An approximation of the control problem P_A can be accomplished using only first order splines (curvilinear to approximate both the Dirichlet map and the semigroup), with an optimal rate of convergence $O(h^{1/2-\varepsilon})$.

(ii) Approximation of the problem P_B with the same (optimal) rate of convergence can be obtained using *cubic polynomials* for the Dirichlet map and *linear* (curvilinear) functions for the semigroup.

(iii) For nonpolygonal shaped domains, the best strategy is to use subspaces which are not required to satisfy zero boundary conditions. To obtain in these cases the optimal rate of convergence, we approximate the Dirichlet map with *cubic splines*.

2. Another advantage of using semigroup methods is that they do not require the operator $A(x, \partial)$ to be selfadjoint.

3. It is natural to consider the above control problems with $\delta = 0$ (i.e., $y(u, T) = y_1$). Under some technical conditions on y_1 it was proved in [S4] that the optimal

controls for such problems are unique and possess the bang-bang property. There is, however, one major difference between those two cases: namely, the optimal solution of a control problem with $\delta = 0$ does *not depend continuously on the final state* y . Therefore, any attempt to discretize does not have to produce a convergent algorithm. The way to overcome this difficulty is through a regularization approach (i.e., we first consider the problems P_A (resp. P_B) with $\delta > 0$ and then let $\delta \rightarrow 0$). It was proved in [S2] that the optimal solution (u^0, T^0) for $\delta = 0$ is a limit of the regularized solutions. Therefore, by combining these two procedures (regularization and approximation of a control problem with $\delta = 0$), we also produce a convergence algorithm for a control problem with $\delta = 0$.

Appendix.

CLAIM A.1. *If $u_h \rightarrow u$ weak star then $L_T u_h \rightarrow L_T u$ in $L_2(\Omega)$ strongly.*

Proof. By the definition of weak star convergence, we have

$$(A.1) \quad \int_0^T (f(z), u_h(z) - u(z)) \, dz \rightarrow 0, \quad f \in L_1[0, T, L_2(\Gamma)].$$

Therefore

$$(A.2) \quad (A^\varepsilon(L_T u_h - L_T u), w) \rightarrow 0, \quad w \in L_2(\Omega).$$

In fact,

$$\begin{aligned} (A^\varepsilon(L_T u_h - L_T u), w) &= \left(\int_0^T A^{3/4+2\varepsilon} S(T-z) A^{1/4-\varepsilon} D(u_h - u)(z) \, dz, w \right) \\ &= \int_0^T (A^{3/4+2\varepsilon} S(T-z) A^{1/4-\varepsilon} D(u_h - u)(z), w) \, dz. \end{aligned}$$

Since

$$\|A^{3/4+2\varepsilon} S(T-z) A^{1/4-\varepsilon} D u(z)\| \leq \frac{C}{(T-z)^{1/4-2\varepsilon}} |u(z)| \quad (\text{by (1.2.3), (1.2.4)}),$$

then by (A.1) (with $f(z) = (A^{1+\varepsilon} S(T-z) D)^* w$) we obtain (A.2). Since $A^{-\varepsilon}$ is a compact operator, (A.2) implies that $L_T u_h \rightarrow L_T u$ strongly in $L_2(\Omega)$. \square

Note added in proof.

After this paper was submitted for publication, an article entitled, *Finite element approximation of parabolic time optimal control problems* by Greg Knowles appeared in SIAM J. Control and Optimization, 20 (1982), pp. 414–427. This article treats the companion problem of approximation of the time optimal control in the case where the control enters Neumann boundary conditions and A is selfadjoint. It should be emphasized that there is a substantial difference between Neumann boundary conditions and Dirichlet boundary conditions (considered in the present paper) particularly on the level of approximation of boundary value problems with $L_2(\Gamma)$ boundary terms. While the *standard* Galerkin method is perfectly suitable for Neumann problems and provides the optimal notes of convergence with $H'(\Omega)$ elements, this is not the case for Dirichlet problems as explained in the introduction of this paper. For this reason, the major part of the present paper is devoted to the proof of the optimal rate of convergence for several schemes approximating Dirichlet L_2 -nonhomogeneous boundary conditions (*modified* Galerkin methods) including those which allow the use of nonconformal elements.

REFERENCES

- [B1] I. BABUSKA AND A. AZIZ, *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press, New York, 1972.
- [B2] I. BABUSKA, *The finite element methods with Lagrangian multipliers*, Numer. Math., 20 (1973), pp. 179–192.
- [B3] J. BRAMBLE AND A. SCHATZ, *Rayleigh–Ritz–Galerkin methods for Dirichlet’s problem using subspaces without boundary conditions*, Comm. Pure Appl. Math., 23 (1970), pp. 653–675.
- [B4] J. BRAMBLE, A. SCHATZ, V. THOMÉE AND L. WAHLBIN, *Some convergence estimates for semidiscrete Galerkin type approximations for parabolic equations*, SIAM J. Numer. Anal., 14 (1977), pp. 218–241.
- [D1] J. DOUGLAS AND T. DUPONT, *Galerkin methods for parabolic equations*, SIAM J. Numer. Anal., 7 (1970), pp. 575–626.
- [D2] N. DUNFORD AND J. SCHWARTZ, *Linear Operators, Part 1*, Interscience–John Wiley, New York, 1958.
- [F1] M. FATTORINI, *The time optimal problem for boundary control of the heat equation*, in Calculus of Variations and Control Theory, D. L. Russel, ed., Academic Press, New York, 1976, pp. 305–320.
- [F2] D. FUJIVARA, *Concrete characterizations of the domains of fractional powers of some elliptic differential operators of the second order*, Proc. Japan Acad., 43 (1967), pp. 82–86.
- [H1] M. HUANG AND V. THOMÉE, *Some convergence estimates for semidiscrete type schemes for time-dependent nonselfadjoint parabolic equations*, Math. Comp., 37 (1981), pp. 327–345.
- [K1] S. G. KREIN, *Linear Differential Equations in Banach Space*, American Mathematical Society Press, Providence, RI, 1971.
- [L1] I. LASIECKA, *Boundary control of parabolic systems: Finite-element approximation*, Appl. Math. Optim., 6 (1980), pp. 31–62.
- [L2] ———, *Unified theory for abstract parabolic boundary problems—A semigroup approach*, Appl. Math. Optim., 6 (1980), pp. 287–333.
- [L3] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. I, II, Springer-Verlag, New York, 1972.
- [N1] J. NITSCHKE, *Über ein variationsprinzip der Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die Keinen Randbedingungen unterworfen sind*, Abh. Math. Sem. Univ. Hamburg, 36 (1971), pp. 9–15.
- [P1] A. PAZY, *Semi-groups of linear operators and applications to partial differential equations*, Lecture Notes, 10, Dept. of Mathematics, Univ. of Maryland, College Park, MD, 1974.
- [S1] G. SCHMIDT AND N. WECK, *On the boundary behavior of solutions to elliptic and parabolic equations with applications to boundary control for parabolic equations*, this Journal, 16 (1978), pp. 593–597.
- [S2] T. SEIDMAN, *Approximation methods for distributed systems*, Research Report 79-18, Math. Programs at UMBC, Univ. of Maryland, College Park, MD, 1979.
- [S3] G. STRANG AND G. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [S4] G. SCHMIDT, *The bang-bang principle for the time-optimal problem in boundary control of the heat equation*, this Journal, 18 (1980), pp. 101–107.
- [S5] R. SCOTT, *Interpolated boundary conditions in the finite element method*, SIAM J. Numer. Anal., 12 (1975), pp. 404–427.
- [Z1] M. ZLAMAL, *Curved elements in the finite element method I*, SIAM J. Numer. Anal., 10 (1973), pp. 229–240.
- [Z2] ———, *The finite element method in domains with curved boundary*, Numer. Meth. Engrg., 5 (1973), pp. 367–373.

CONTROLLABILITY PROPERTIES OF AFFINE SYSTEMS*

VELIMIR JURDJEVIC† AND GAUTHIER SALLET‡

Abstract. This paper deals with transitivity (controllability) of affine families of vector fields on a finite dimensional vector space V . In particular we focus on affine families whose corresponding families of linear fields are transitive on $V - \{0\}$, and which in addition have no fixed points in V . We show that such families are necessarily transitive on V , and we also show that they remain transitive under small perturbations. In general, however, affine families need not remain transitive under small perturbations—for example, small affine perturbations of transitive linear systems are not necessarily transitive. Since any affine system \mathcal{F} naturally defines a system \mathcal{F}_r of right-invariant vector fields on the semi-direct product of V with $GL(V)$ we also investigate transitivity properties of \mathcal{F}_r . Our result is that if \mathcal{F} is an affine family satisfying the preceding conditions then \mathcal{F}_r generates the full Lie algebra on the semi-direct product.

Key words. controllability, transitivity, affine fields, Lie algebras, semi-direct products, Lie groups, convexity

Introduction. In essence this paper gives transitivity (controllability) conditions for control systems of the form:

$$(1) \quad \frac{dx}{dt} = (A_0x + a_0) + \sum_{i=1}^m u_i(t)(A_i x + a_i),$$

where A_0, \dots, A_m are $n \times n$ matrices with real entries and where a_0, \dots, a_m are vectors in \mathbb{R}^n . The controls u_1, \dots, u_m are real valued functions of time t defined on the interval $[0, \infty)$, while the state vector x belongs to \mathbb{R}^n .

A vector field $X(x) = Ax + a$, where A is an $n \times n$ matrix and where a is a vector on \mathbb{R}^n , is an affine vector field. Rather than working directly with the differential system (1), we consider arbitrary families \mathcal{F} of affine vector fields in $V = \mathbb{R}^n$, as is commonly done in the literature, and we call such families \mathcal{F} affine.

If $X(x) = Ax + a$ is an affine field, then \tilde{X} denotes the corresponding linear field $x \rightarrow Ax$ for all $x \in V$. Any affine family \mathcal{F} defines the corresponding linear family $\tilde{\mathcal{F}} = \{\tilde{X} : X \in \mathcal{F}\}$. At the risk of adding yet another name to the already existing melange of terminology in the literature we will refer to systems of the form

$$(2) \quad \frac{dx}{dt} = Ax + \sum_{i=1}^m u_i(t)a_i,$$

where A is an $n \times n$ matrix and where a_1, \dots, a_m are vectors in \mathbb{R}^n , as *Kalman systems*.

Our main results then are the following:

(A) If \mathcal{F} is an affine family on a vector space V such that the corresponding linear family $\tilde{\mathcal{F}}$ is transitive on $V - \{0\}$, and if \mathcal{F} has no fixed points in V , then \mathcal{F} is transitive on V . \mathcal{F} is said to have no fixed points in V if for each $x \in V$ there exists $X \in \mathcal{F}$ such that $X(x) \neq 0$ (Theorems 1 and 2). This result generalizes a result of B. Bonnard contained in [B].

(B) Affine systems considered in (A) constitute an open subset of transitive affine systems: i.e., they remain transitive under small perturbations (Theorem 4).

(C) Transitive affine systems in general need not remain transitive under small perturbations (Example 2). In fact, we show that in every neighborhood of a transitive Kalman system there exists an affine system which is not transitive.

* Received by the editors November 15, 1982, and in revised form February 21, 1983.

† Department of Mathematics, University of Toronto, Toronto, Ontario, Canada M5S 1A1.

‡ Department of Mathematics, Université de Metz, Metz, France.

(D) Finally, we make a connection between these results and those obtained in our previous paper [BJKS]. Any affine system \mathcal{F} on a vector space V naturally defines a system \mathcal{F}_r of right-invariant vector fields on a semi-direct product of V and a Lie sub-group G of $GL(V)$. G is defined as the group generated by $\bigcup_{X \in \mathcal{F}} \{\exp t\tilde{X} : t \in \mathbb{R}\}$, and an element $X(x) = Ax + a$ in \mathcal{F} defines a right-invariant vector field \tilde{X} , whose value at the identity is equal to (a, A) . If we denote by $V * G$ the semi-direct product of V and G then we have the following (Theorem 3):

Suppose that \mathcal{F} is an affine system which has no fixed points in V and such that the corresponding linear system $\tilde{\mathcal{F}}$ is transitive on $V - \{0\}$. Then, the right-invariant system \mathcal{F}_r is such that the Lie algebra which it generates is equal to the Lie algebra of all right-invariant vector fields on $V * G$. Equivalently, if Γ is the subset of the Lie algebra of $V * G$ defined by the values of the elements of \mathcal{F}_r at the identity, then the Lie algebra generated by Γ is equal to the Lie algebra of $V * G$.

This result is a generalization of [B, Prop. 4.2] and [BJKS, Thm. 2].

Finally we give an example showing an affine system \mathcal{F} , which has no fixed points in \mathbb{R}^n and whose linear part $\tilde{\mathcal{F}}$ is transitive on $\mathbb{R}^n - \{0\}$ but for which the set of accessibility of the corresponding right-invariant system is a proper semi-group of $\mathbb{R}^n * SL_n(\mathbb{R})$ (Example 1).

1. Definitions and basic concepts. Throughout this paper V will denote an n -dimensional real vector space, $\text{End}(V)$ and $GL(V)$ will respectively denote the set of all linear endomorphisms on V and the set of all linear automorphisms of V . A vector field X on V is an *affine vector field* if $X(x) = Ax + a$ for all $x \in V$, where $A \in \text{End}(V)$ and where $a \in V$. In general, if X is any (complete) vector field on V we will write $\exp tX$ for the one parameter group of diffeomorphisms generated by X . If X is affine, then, as is well known,

$$(\exp tX)x = \exp tA \left(x + \int_0^t \exp(-\tau A)a \, d\tau \right) \quad \text{for all } x \in V.$$

We will denote by $\text{Aff}(V)$ the set of all affine vector fields on V . Since the Lie bracket of affine vector fields is affine, it follows that $\text{Aff}(V)$ is a finite dimensional Lie sub-algebra of the set of analytic vector fields on V . Alternatively, $\text{Aff}(V)$ can be viewed as the algebra of vector fields generated by the action of the semi-direct product of $GL(V)$ and V on V . If G is any Lie group which acts on V , then we write $V * G$ for the semi-direct product of V and G . As described in [BJKS], $V * G$ is a Lie group. If $L(G)$ is the Lie algebra of G , then we denote by $V * L(G)$ the Lie algebra of $V * G$. The connection between $\text{Aff}(V)$ and $V * GL(V)$ is as follows: $V * GL(V)$ acts on V by $(v, g)x = gx + v$ for all $x \in V$ and all $(v, g) \in V * GL(V)$. If $L = (a, A)$ and $L \in V * \text{End}(V)$ then the action of $\exp tL$ on V gives a one-parameter group of diffeomorphisms on V . Its infinitesimal generator is the affine field X defined by $X(x) = Ax + a$ for all $x \in V$. If $X(x) = Ax + a$ for all $x \in V$, then we will denote by \tilde{X} , the right-invariant vector field on $V * GL(V)$ whose value at the identity is $L = (a, A)$. The orbit of \tilde{X} , through the identity is $\{\exp tL : t \in \mathbb{R}\}$. We will also denote by \tilde{X} the linear field corresponding to X , i.e., $\tilde{X}(x) = Ax$ for all $x \in V$.

If \mathcal{F} is any family of (complete) vector fields on a manifold M then as usual $S(\mathcal{F})$ will stand for the semi-group generated by $\bigcup_{X \in \mathcal{F}} \{\exp tX : t \geq 0\}$. For each $x \in M$, $S(\mathcal{F})(x)$ will denote the orbit of $S(\mathcal{F})$ through x . \mathcal{F} is said to be transitive if $S(\mathcal{F})(x) = M$ for all $x \in M$.

In particular if $\mathcal{F} \subset \text{Aff}(V)$, then \mathcal{F}_r is the family of right-invariant fields on $V * GL(V)$, where \mathcal{F}_r is defined by $\mathcal{F}_r = \{\tilde{X}_r : X \in \mathcal{F}\}$. Similarly, $\tilde{\mathcal{F}} = \{\tilde{X} : X \in \mathcal{F}\}$. The connection between the various semi-groups associated with \mathcal{F} is as follows: the

projection of $S(\mathcal{F}_r)$ on $GL(V)$ is equal to $S(\tilde{\mathcal{F}})$, while the action of $S(\mathcal{F}_r)$ through x is equal to the orbit $S(\mathcal{F})(x)$. Naturally, \mathcal{F} is transitive on V if \mathcal{F}_r is transitive on $V * GL(V)$. As we mentioned in the introduction, in [BJKS] we studied transitivity properties of \mathcal{F}_r on a semi-direct product of V with a Lie group G and, as a by-product, we obtained transitivity results on V . To study the converse it is necessary to consider the restriction of \mathcal{F}_r to a sub-group of $V * GL(V)$. The most natural way is as follows: Let G be the group generated by $S(\tilde{\mathcal{F}})$. We call G the integral group of $\tilde{\mathcal{F}}$. G is a connected Lie sub-group of $GL(V)$. The orbit of \mathcal{F}_r through the identity is contained in $V * G$, and in the sequel we will examine the connection between transitivity properties of \mathcal{F} on V and \mathcal{F}_r on $V * G$.

We conclude this section with a few elementary facts from affine geometry which we use in the next section. If $Q \subset V$, then $A(Q)$ will stand for the affine hull of Q . That is, $v \in A(Q)$ if and only if $v = \sum_{i=1}^p \lambda_i q_i$ for some points q_1, \dots, q_p in Q and some scalars $\lambda_1, \dots, \lambda_p$ such that $\sum_{i=1}^p \lambda_i = 1$. We shall denote by $\overline{A(Q)}$ the tangent space of $A(Q)$. $\overline{A(Q)}$ is the vector space spanned by all the differences of elements in Q . Alternatively, $v \in \overline{A(Q)}$ if and only if $v = \sum_{i=1}^p \lambda_i q_i$ for some points q_1, \dots, q_p in Q and some scalars $\lambda_1, \dots, \lambda_p$ with $\sum_{i=1}^p \lambda_i = 0$.

2. The main results. In this section we assume that V is an inner product space. $\langle \cdot, \cdot \rangle$ denotes this inner product. Then, S^n stands for the unit sphere in V . If $A \in \text{End}(V)$, $\text{sp}(A)$ is the spectrum of A , and A^* is the adjoint of A relative to $\langle \cdot, \cdot \rangle$.

Our first result concerns the transitivity properties of linear vector fields and is as follows:

THEOREM 1. *Let $\tilde{\mathcal{F}}$ be a family of linear vector fields on V which is transitive on $V - \{0\}$. Then there exists a finite sub-family $\tilde{\mathcal{F}}_0$ of $\tilde{\mathcal{F}}$ which is also transitive on $V - \{0\}$.*

Before proving this theorem we give several lemmas which are of relevance.

LEMMA 1. *If $\tilde{\mathcal{F}}$ is a family of linear vector fields on V , then $\tilde{\mathcal{F}}$ is transitive on $V - \{0\}$ if and only if the following conditions hold:*

- a) $S^n \subset S(\tilde{\mathcal{F}})(x)$ for each $x \in S^n$, and
- b) there exist X_1 and X_2 in $\tilde{\mathcal{F}}$ such that

$$\text{Min sp}(\tfrac{1}{2}(X_1 + X_1^*)) < 0 \quad \text{and} \quad \text{Max sp}(\tfrac{1}{2}(X_2 + X_2^*)) > 0.$$

Proof. If $\tilde{\mathcal{F}}$ is transitive on $V - \{0\}$ we need to show only that condition b) holds for condition a) is evidently satisfied. There exists $X_1 \in \tilde{\mathcal{F}}$ and $x_1 \in S^n$ such that $\langle X_1 x_1, x_1 \rangle < 0$. For if not, then the exterior of S^n would be invariant under $\tilde{\mathcal{F}}$ which is contrary to the transitivity assumption. Then, the minimum value of the map $F: x \rightarrow \langle X_1 x, x \rangle$ from S^n into \mathbb{R} is less than zero. It is well known that this minimum value is equal to $\text{Min sp}(\tfrac{1}{2}(X_1 + X_1^*))$. The argument concerning the existence of X_2 such that $\text{Max sp}(\tfrac{1}{2}(X_2 + X_2^*)) > 0$ is similar (where the exterior of S^n is replaced by the interior of S^n), and will therefore be omitted.

Conversely, assume now that conditions a) and b) are satisfied. It follows from the preceding paragraph that there exist x_1 and x_2 on S^n such that $\langle X_1 x_1, x_1 \rangle < 0$ and $\langle X_2 x_2, x_2 \rangle > 0$. Hence, the same holds for all λx_1 and λx_2 with λ positive. If S_r^n denotes the sphere in V of radius r , then the above implies that, for each $r > 0$ there exists $\varepsilon > 0$ such that $\{\exp tX_1((r+\varepsilon)x_1): t \leq 0\}$ and $\{\exp tX_2((r-\varepsilon)x_2): t \geq 0\}$ cut each sphere S_λ^n for all λ with $r-\varepsilon \leq \lambda \leq r+\varepsilon$. Since for each $x \in V$ and each $\lambda > 0$, $S(\tilde{\mathcal{F}})(\lambda x) = \lambda S(\tilde{\mathcal{F}})x$, it follows from condition a) that $S_\lambda^n \subset S(\tilde{\mathcal{F}})(x)$ for each $x \in S_\lambda^n$. Thus, for each $r > 0$ there exists $\varepsilon > 0$ such that the annulus $A_{r,\varepsilon} = \{x: r-\varepsilon \leq \|x\| \leq r+\varepsilon\}$ is contained in $S(\tilde{\mathcal{F}})(x)$ for each $x \in A_{r,\varepsilon}$. This shows that $\tilde{\mathcal{F}}$ is transitive on $V - \{0\}$, and our proof is now complete.

Before stating the next lemma, it will be convenient to recall the notion of normal accessibility of H. J. Sussmann [SU]. If \mathcal{F} is any family of smooth vector fields on a manifold M then y is said to be *normally accessible* from x if there exists a finite set X_1, \dots, X_k of vector fields in \mathcal{F} such that:

A). The mapping F from $\mathbb{R}_+^k = \{(t_1, \dots, t_k): t_i > 0, i = 1, 2, \dots, k\}$ into M defined by $(t_1, \dots, t_k) \rightarrow \exp t_k X_k \circ \dots \circ \exp t_1 X_1(x)$ contains y in its image.

B). If $F(\hat{t}_1, \dots, \hat{t}_k) = y$, then the rank of F at $(\hat{t}_1, \dots, \hat{t}_k)$ is equal to $\dim(M)$.

It will be convenient for the next lemma to denote the above finite sub-family $\{X_1, \dots, X_k\}$ by $\mathcal{F}_{x,y}$. It is clear that y belongs to the interior of $S(\mathcal{F}_{x,y})(x)$.

LEMMA 2. *Let $\tilde{\mathcal{F}}$ be a family of linear vector fields which is transitive on $V - \{0\}$. If K is any compact subset of $V - \{0\}$, then there exists a finite sub-family $\tilde{\mathcal{F}}_0$ of $\tilde{\mathcal{F}}$ such that $K \subset S(\tilde{\mathcal{F}}_0)(x)$ for any $x \in K$.*

Proof. Since $\tilde{\mathcal{F}}$ is transitive on $V - \{0\}$, it follows from [SU] that for any $x \in K$ every point y of K is normally accessible from x by the elements of $\tilde{\mathcal{F}}$. We let \bar{x} be a fixed point in K . Since K is compact there exist points y_1, \dots, y_m in K such that $K \subset S(\bigcup_{i=1}^m \tilde{\mathcal{F}}_{\bar{x}, y_i})(\bar{x})$. We let $\tilde{\mathcal{F}}_1 = \bigcup_{i=1}^m \tilde{\mathcal{F}}_{\bar{x}, y_i}$. It is clear that $-\tilde{\mathcal{F}}$ is also transitive on $V - \{0\}$. By an argument identical to the preceding one, but applied to $-\tilde{\mathcal{F}}$ instead of to $\tilde{\mathcal{F}}$, we conclude that there exists a finite sub-family $\tilde{\mathcal{F}}_2$ of $\tilde{\mathcal{F}}$ such that $K \subset S(-\tilde{\mathcal{F}}_2)(\bar{x})$. If $\tilde{\mathcal{F}}_0 = \tilde{\mathcal{F}}_1 \cup \tilde{\mathcal{F}}_2$, then $K \subset S(\tilde{\mathcal{F}}_0)(x)$ for each $x \in K$. The proof of the lemma is therefore finished.

Remark. The statement of Lemma 1 remains true if $\tilde{\mathcal{F}}$ is replaced by an arbitrary family of smooth vector fields and $V - \{0\}$ by a smooth manifold M in which \mathcal{F} is transitive.

We now turn to the proof of Theorem 1.

Proof. Let $\tilde{\mathcal{F}}$ be transitive on $V - \{0\}$. By Lemma 1 this is equivalent to conditions A) and B). Let $\tilde{\mathcal{F}}_1$ be a finite sub-family of $\tilde{\mathcal{F}}$ such that $S^n \subset S(\tilde{\mathcal{F}}_1)(x)$ for each $x \in S^n$. The existence of such a set follows from Lemma 2. Let X_1 and X_2 be the elements of $\tilde{\mathcal{F}}$ which satisfy condition b) of Lemma 1. Then, $\tilde{\mathcal{F}}_0 = \tilde{\mathcal{F}}_1 \cup \{X_1, X_2\}$ satisfies conditions a) and b) of Lemma 1 and hence is transitive on $\mathbb{R}^n - \{0\}$. Our proof is therefore complete.

COROLLARY 1. *Let \mathcal{F} be a family of affine vector fields on V such that a) $\tilde{\mathcal{F}}$ is transitive on $V - \{0\}$ and b) there exists no $x \in V$ such that $X(x) = 0$ for all $X \in \mathcal{F}$. Then, there exists a finite sub-family \mathcal{F}_0 of \mathcal{F} which satisfies properties a) and b).*

Proof. The set of all affine vector fields $\text{Aff}(V)$ is a finite dimensional vector space. Hence, if \mathcal{F}_1 is a basis for the vector space spanned by \mathcal{F} , then \mathcal{F}_1 is a finite set. Moreover, for each $x \in V$ there exists an element X in \mathcal{F}_1 such that $X(x) \neq 0$. Hence, \mathcal{F}_1 has no fixed point in \mathbb{R}^n .

Let $\mathcal{F}_2 \subset \mathcal{F}$ be such that $\tilde{\mathcal{F}}_2$ is a finite sub-family of $\tilde{\mathcal{F}}$ which is transitive on $V - \{0\}$. Then $\mathcal{F}_0 = \mathcal{F}_1 \cup \mathcal{F}_2$ is a finite sub-family which satisfies conditions a) and b). The proof is now complete.

If \mathcal{F} is any family of affine fields on V which satisfies condition b) of Corollary 1 then we say that \mathcal{F} has no fixed points in V . Our next result is an improvement on a result in [B].

THEOREM 2. *Let \mathcal{F} be an affine family of vector fields on V such that*

- a) $\tilde{\mathcal{F}}$ is transitive on $V - \{0\}$; and,
- b) \mathcal{F} has no fixed points in V .

Then, \mathcal{F} is transitive on V .

Before we give a proof of this theorem it will be convenient to recall certain basic facts from affine geometry which we assemble in several lemmas.

LEMMA 3. Let \mathcal{T} be a family of affine transformations which leaves a subset $Q \subset V$ invariant. Then,

- a) $A(Q)$ is invariant under \mathcal{T} ;
- b) $\overline{A(Q)}$ is invariant under $\tilde{\mathcal{T}}$.

The proof is elementary and we omit it.

LEMMA 4. Let \mathcal{T} be a family of affine transformations, and let K be a compact subset of V with a nonempty interior. If $\mathcal{T}(K) \subset K$, then there exists a number B (depending on K) such that $\|\tilde{T}\| \leq B(K)$ for all $\tilde{T} \in \tilde{\mathcal{T}}$.

Proof. Let w be a point in the interior of K . Then, $K - w$ contains the origin in its interior. Let B_ε be the ball of radius ε centered at the origin which is contained in $K - w$. If $T \in \mathcal{T}$ then for any $x \in K$.

$$T(x - w) = T(x) - T(w) \in K - T(w).$$

In particular, $\tilde{T}(B_\varepsilon) \subset K - T(w)$. If $A = \sup \{\|x - y\| : x \in K, y \in K\}$ then let $B(K) = A/\varepsilon$. The statement of Lemma 4 now follows.

The next lemma concerns orbits of affine vector fields, and is interesting in its own right.

LEMMA 5. Let \mathcal{F} be a family of affine vector fields such that

- a) $\tilde{\mathcal{F}}$ is transitive on $V - \{0\}$; and,
- b) \mathcal{F} has no fixed points in V .

Then, for each $x \in V$, $S(\mathcal{F})(x)$ is unbounded.

Proof. Let $x \in V$ be fixed, and assume that $S(\mathcal{F})(x)$ is bounded. Let K be the closure of $S(\mathcal{F})(x)$. Then K is compact and invariant under the semi-group $S(\mathcal{F})$. By Lemma 3, $\overline{A(K)}$ is invariant under $\tilde{S}(\mathcal{F})$. Since $\tilde{S}(\mathcal{F}) = S(\tilde{\mathcal{F}})$, and since $A(K)$ is a vector space, it follows from our transitivity assumption on $\tilde{\mathcal{F}}$ that either $\overline{A(K)} = V$ or that $\overline{A(K)} = \{0\}$. Since $\overline{A(K)} = A(K) - x$, the latter assumption implies that $S(\mathcal{F})(x) = \{x\}$ which is contrary to assumption b) of Lemma 5. Therefore, $\overline{A(K)}$ is V and hence $\text{CO}(K)$, the convex hull of K , has a nonempty interior in V . But then by Lemma 5 the elements of $S(\mathcal{F})$ are uniformly bounded in the norm which would exclude the transitivity of $\tilde{\mathcal{F}}$ on $V - \{0\}$. Thus, in either case it is impossible for $S(\mathcal{F})(x)$ to be bounded, and our proof is finished.

Remark 2. It is evident that under the assumptions of Lemma 5 the orbits of $S(-\mathcal{F})$ are also unbounded.

We are now ready to give a proof of Theorem 2.

Proof. We assume that \mathcal{F} satisfies conditions a) and b) of Theorem 2 and we will next prove that these assumptions imply that all orbits of $S(\mathcal{F})$ are open in V . Let $w \in V$ be a fixed point. For each $\lambda > 0$ we define the mapping $h_{\lambda, w}$ by $h_{\lambda, w}(x) = w + \lambda(x - w)$ for all $x \in V$. $h_{\lambda, w}(w) = w$ and $h_{\lambda, w}^{-1}(x) = w + (1/\lambda)(x - w)$ for all $x \in V$.

If $X \in \mathcal{F}$ then $h_{\lambda, w} \circ \exp tX \circ h_{\lambda, w}^{-1}$ is a one-parameter group of diffeomorphisms. Its infinitesimal generator is $dh_{\lambda, w} \circ X \circ h_{\lambda, w}^{-1}(x)$ for each $x \in V$. If $X(x) = Ax + a$ for all $x \in V$, then $dh_{\lambda, w} \circ X \circ h_{\lambda, w}^{-1}(x) = \lambda(A(w + 1/\lambda(x - w)) + a) = A(x - w) + \lambda(A(w) + a) = \tilde{X}(x - w) + \lambda X(w)$.

Let \mathcal{F}^k be a k -finite sub-family of \mathcal{F} such that $\tilde{\mathcal{F}}^k$ is transitive on $V - \{0\}$, and such that \mathcal{F}^k has no fixed points in V . (That this is possible follows from Corollary 1.) If we denote by $\mathcal{F}_{\lambda, w}^k$ the family $dh_{\lambda, w} \circ \mathcal{F}^k \circ h_{\lambda, w}^{-1}$, then, $\lim_{\lambda \rightarrow 0} \tilde{\mathcal{F}}_{\lambda, w}^k(x) = \tilde{\mathcal{F}}_w^k$, where $\tilde{\mathcal{F}}_w^k$ is defined by $\tilde{\mathcal{F}}_w^k(x) = \tilde{\mathcal{F}}^k(x - w)$. $\tilde{\mathcal{F}}_w^k$ is transitive on $V - \{w\}$. Hence, if S_w^n is the sphere of radius one centered at w , then $S_w^n \subset S(\tilde{\mathcal{F}}_{0, w}^k)(x)$ for each $x \in S_w^n$.

It now follows from [SU] that for λ sufficiently small $S_w^n \subset S(\mathcal{F}_{\lambda, w}^k)(x)$ for $x \in S_w^n$. We let $\lambda > 0$ be such a number which we subsequently regard as fixed.

If $X \in \mathcal{F}$, then $dh_{\lambda,w} \circ X \circ h_{\lambda,w}^{-1}(x) = 0$ implies that $X(h_{\lambda,w}^{-1}(x)) = 0$. Hence $\mathcal{F}_{\lambda,w}$ has no fixed points in V whenever \mathcal{F} has no fixed points. Moreover, $\overrightarrow{\mathcal{F}_{\lambda,w}}$ is equal to $\overrightarrow{\mathcal{F}}$ centered at w . Therefore, Lemma 5 is applicable, and we conclude that the orbits of $S(\mathcal{F}_{\lambda,w})$ and $S(-\mathcal{F}_{\lambda,w})$ cannot be bounded.

It now easily follows that for each x and y in $B_w^n = \{x: \|x - w\| \leq 1\}$, $y \in S(\mathcal{F}_{\lambda,w})(x)$. For, then $S(\mathcal{F}_{\lambda,w})(x)$, being unbounded, intersects S_w^n . By the same argument $S(-\mathcal{F}_{\lambda,w})(y)$ also intersects S_w^n . Since $S_w^n \subset S(\mathcal{F}_{\lambda,w})(x)$ for each $x \in S_w^n$, we get that $y \in S(\mathcal{F}_{\lambda,w})(x)$. Equivalently, $B_w^n \subset S(\mathcal{F}_{\lambda,w})(w) \cap S(-\mathcal{F}_{\lambda,w})(w)$.

The final step in our proof is to reinterpret this last fact in terms of $S(\mathcal{F})$.

$$S(\mathcal{F}_{\lambda,w}) = h_{\lambda,w} \circ S(\mathcal{F}) \circ h_{\lambda,w}^{-1},$$

hence

$$h_{\lambda,w}^{-1} B_w^n \circ h_{\lambda,w}(w) \subset S(\mathcal{F})(w) \cap S(-\mathcal{F})(w).$$

But $h_{\lambda,w}^{-1} \circ B_w^n \circ h_{\lambda,w}(w) = h_{\lambda,w}^{-1} \circ B_w^n = (1/\lambda) B_w^n$. Thus, $(1/\lambda) B_w^n \subset S(\mathcal{F})(w) \cap S(-\mathcal{F})(w)$. Thus, the orbits of $S(\mathcal{F})$ and $S(-\mathcal{F})$ are all open and this implies transitivity. Our proof is now finished.

3. Related results and applications.

3.1. Connection with the semi-direct product. In our previous paper [BJKS] we gave transitivity conditions for a family \mathcal{F}_r of right-invariant vector fields to be transitive on a semi-direct product of a Lie group G and the vector space V on which this group acts. As we mentioned in § 1, any such family \mathcal{F}_r defines an affine family \mathcal{F} on V . If \mathcal{F}_r is transitive on $V * G$ then necessarily \mathcal{F} is transitive on V . Our first example shows that the converse need not be true, and hence it shows that the results in § 2 are different from those in [BJKS].

Example 1. Let $V = \mathbb{R}^2$ and let $G = GL_2^+(V)$ be the group of all nonsingular transformations of V with a positive determinant. We let

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -2 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

If

$$a = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = b,$$

then define $\mathcal{F} = \{(a, A), (b, B)\}$.

It is evident that \mathcal{F} has no fixed points in V , and moreover, it is equally evident that $\overrightarrow{\mathcal{F}}$ is transitive on $V - \{0\}$. Thus, by Theorem 2, \mathcal{F} is transitive on V . However, \mathcal{F}_r is not transitive on $V * G$ because the semi-group generated by $\{\exp tA: t \geq 0\}$ and $\{\exp tB: t \geq 0\}$ is contained in the set of elements in G with determinant less than or equal to one.

The connection between the affine systems considered in this paper and the associated right-invariant systems on the semi-direct product is explained by the following theorem.

THEOREM 3. *Let \mathcal{F} be an affine family of vector fields on a real vector space V . If \mathcal{F} is such that $\overrightarrow{\mathcal{F}}$ is transitive on $V - \{0\}$ and has no fixed points in V , then the corresponding right-invariant system \mathcal{F}_r generates the Lie algebra of $V * G$, where G is the integral group of $\overrightarrow{\mathcal{F}}$.*

Proof. Let G be the integral group of $\overrightarrow{\mathcal{F}}$, i.e., G is the group generated by $\bigcup_{A \in \overrightarrow{\mathcal{F}}} \{\exp tA: t \in \mathbb{R}\}$. G acts transitively on $V - \{0\}$, hence its Lie algebra $L(G)$ is

irreducible. Thus, $L(G) = L' \oplus \mathcal{C}$, where L' is the derived algebra of $L(G)$, and where \mathcal{C} is the center of $L(G)$ [J, p. 46].

Let $\Gamma = \{(a, A): x \rightarrow Ax + a \text{ belongs to } \mathcal{F}\}$. We want to show that $\text{Lie}(\Gamma) = V * L(G)$. Let $\pi: \text{Lie}(\Gamma) \rightarrow L(G)$ be the projection $(a, A) \rightarrow A$. Since G is the integral group of \mathcal{F} the projection of $\text{Lie}(\Gamma)$ is onto $L(G)$. If we identify $\ker \pi$ with a subspace of V then such a subspace is invariant under $L(G)$. Since $L(V)$ acts irreducibly on V it follows that $\ker \pi$ is either V or $\{0\}$.

In the first case it immediately follows that $\text{Lie}(\Gamma) = V * L(G)$, which yields the conclusion of our theorem, and so we assume the second. In such a case π is an isomorphism. If $\mathcal{L} = \text{Lie}(\Gamma)$ then $\mathcal{L} = \mathcal{L}' \oplus \mathcal{C}$, where $\mathcal{L}' = [\mathcal{L}, \mathcal{L}]$ and, where \mathcal{C} is the center of \mathcal{L} .

We first show that $\mathcal{L}' = \exp v_0 L' \exp(-v_0)$ for some $v_0 \in V$. This fact follows from the Levi theorem [J, p. 91] and the conjugacy of Levi factors [J, p. 93]. We first note that V is a commutative ideal of $V * L'$. Thus, $V \subset \text{Rad}(V * L')$ where the latter stands for the radical of $V * L'$. On the other hand, $\pi(\text{Rad}(V * L')) \subset \text{Rad } L' = \{0\}$. Thus, V is the radical of $V * L'$, and hence L' is a Levi factor of $V * L'$. Since \mathcal{L}' is isomorphic to L' it follows that \mathcal{L}' is also a Levi factor of $V * L'$. It now follows that $\mathcal{L}' = \exp v_0 L' \exp(-v_0)$. Let $E = \{v: \mathcal{L}'(v) = 0\}$. $\mathcal{L}'(v_0) = (\exp v_0 L' \exp(-v_0))(v_0) = 0$. Thus $v_0 \in E$. If $v_1 \in E$ and $v_1 \neq v_0$ then $L'(v_1 - v_0) = 0$. Since V is an irreducible L' module it follows that this is impossible. Thus $E = \{v_0\}$.

Now let $z \in \mathcal{C}$. Then $z = (w, \lambda I)$ for some $\lambda \in \mathbb{R}$ and some $w \in V$. For any $(a, A) \in \mathcal{L}'$, $[(a, A), (w, \lambda I)] = 0$. Thus, $Aw = \lambda a$. But $(a, A)v_0 = 0$ and so $Av_0 = -a$. Hence, $A(w + \lambda v_0) = 0$, or $L'(w + \lambda v_0) = 0$. This implies that $w = -\lambda v_0$. We have thus shown that $\mathcal{C} = \{(-\lambda v_0, \lambda I): \lambda \in \mathbb{R}\}$.

This shows that $\ker \pi = \{0\}$ is not possible; for then $\text{Lie } \Gamma(v_0) = 0$ which would be contrary to our hypothesis concerning the fixed points of Γ . This concludes the proof of the theorem.

3.2. Stability of transitivity under small perturbations. We first show by an example that transitivity of affine systems is not stable under small perturbations. As before we regard V as a metric space with norm $\|\cdot\|$, and hence $\text{Aff}(V)$ has a natural metric induced by $V \times \text{End}(V)$. For simplicity of exposition we restrict our considerations to finite systems. If $\mathcal{F}_1 \subset \text{Aff}(V)$ then we say that $\mathcal{F}_2 \subset \text{Aff}(V)$ is in an ε -neighbourhood of \mathcal{F}_1 if $\text{card } \mathcal{F}_1 = \text{card } \mathcal{F}_2$ and if when $\mathcal{F}_1 = \{X_1, \dots, X_m\}$ and $\mathcal{F}_2 = \{Y_1, \dots, Y_m\}$ then $\|X_i - Y_i\| < \varepsilon$ for each $i = 1, 2, \dots, m$.

Example 2. Let $V = \mathbb{R}^2$, and let $\mathcal{F} = \{X, Y, -Y\} \subset \text{Aff}(V)$, where $X(x) = Ax$ and where $Y(x) = b$ for all $x \in V$. Such a family is induced by a linear, scalar control system of the form $dx/dt = Ax + u(t)b$, where the control u is not constrained in magnitude.

As is well known, \mathcal{F} is transitive on V if and only if b and Ab are linearly independent. We assume that A is a diagonal matrix with negative eigenvalues α_1 and α_2 with $\alpha_1 \neq \alpha_2$, and we assume that the vector b has unequal coordinates, neither of which is zero. Then \mathcal{F} is certainly transitive on V .

However, we next show that in every ε -neighbourhood of \mathcal{F} there exists $\mathcal{F}_\varepsilon \subset \text{Aff}(V)$ which is not transitive on V . Let $\varepsilon > 0$ be fixed. Let B be a diagonal matrix with distinct eigenvalues. Let $Y_\varepsilon(x) = \varepsilon Bx + b$ for all $x \in V$. The trajectories of Y_ε are the same as those of $x \rightarrow Bx$, except that they are centered at the new origin $c = -1/\varepsilon B^{-1}b$. If $c = (c_1, c_2)$, then the line $l = \{(r, c_2): r \in \mathbb{R}\}$ is invariant under Y_ε . Moreover, the strip $\{(x_1, x_2): 0 \leq x_2 \leq |c_2|\}$ is positively invariant under X . Hence, such a strip is invariant under $\mathcal{F}_\varepsilon = \{X, Y_\varepsilon, -Y_\varepsilon\}$, and hence \mathcal{F}_ε is not transitive (see Fig. 1).

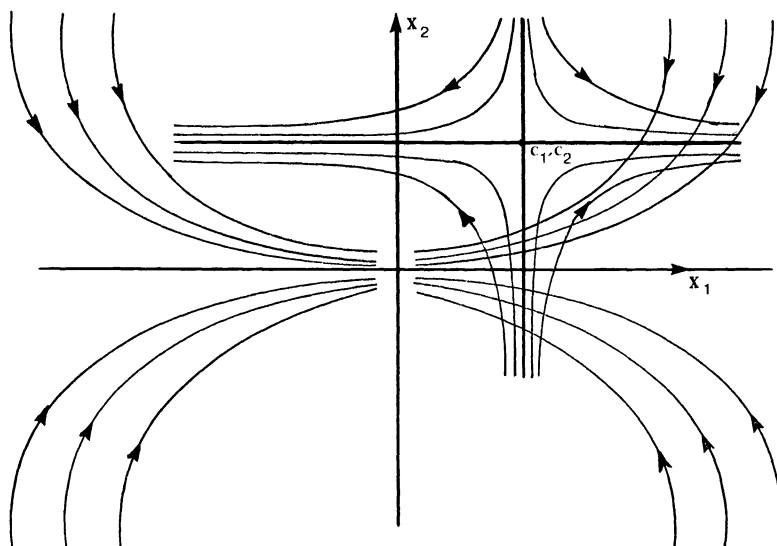


FIG. 1

On the other hand, the affine systems considered in this paper are stable under small perturbations. We express this more precisely in the following:

THEOREM 4. *Let \mathcal{F} be a family of affine vector fields on V such that $\tilde{\mathcal{F}}$ is transitive on $V - \{0\}$ and such that \mathcal{F} has no fixed points in V . Then, for sufficiently small perturbations of the elements of \mathcal{F} the perturbed system remains transitive.*

Proof. Let \mathcal{F}_0 be the finite sub-family of \mathcal{F} such that $\tilde{\mathcal{F}}_0$ is transitive on $V - \{0\}$ and such that \mathcal{F}_0 has no fixed points in V (Corollary 1). It then follows by Lemma 1 and by standard normal accessibility arguments that any small perturbation $\mathcal{F}_{0\varepsilon}$ of \mathcal{F}_0 will be such that $\tilde{\mathcal{F}}_{0\varepsilon}$ remains transitive on $V - \{0\}$. The claim of the theorem is now evident, since for sufficiently small ε , \mathcal{F}_ε does not have a fixed point in V .

Acknowledgments. The proof in the text of Theorem 3 is essentially due to Professor I. Kupka. We are grateful that we have been able to use it, rather than to use a somewhat lengthy argument involving the Boothby list of groups which act transitively on $V - \{0\}$ ([BT] and [BW]).

REFERENCES

- [B] B. BONNARD, *Controlabilité des systèmes bilinéaires*, Math. Systems Theory, 15 (1981), pp. 79–92.
- [BJKS] B. BONNARD, V. JURDJEVIC, I. KUPKA AND G. SALLET, *Transitivity of families of invariant vector fields on the semi-direct product of Lie groups*, Trans. Amer. Math. Soc., 271 (1982), pp. 525–535.
- [BT] W. BOOTHBY, *A transitivity problem from control theory*, J. Differential Equations, 17 (1975), pp. 296–307.
- [BW] W. BOOTHBY AND E. N. WILSON, *Determination of the transitivity of bilinear systems*, this Journal, 17 (1979), pp. 212–221.
- [D] J. DIEUDONNÉ, *Elements d'analyse*, Tôme 4, Hermann, Paris.
- [J] N. JACOBSON, *Lie Algebras*, John Wiley, New York, London, 1962.
- [JK] V. JURDJEVIC AND I. KUPKA, *Control systems on semi-simple Lie groups and their homogeneous spaces*, Ann. Inst. Fourier, 31 (1981), pp. 151–179.
- [SA] G. SALLET, *Thèse de 3^{ème} cycle*, Université, Metz.
- [SU] H. SUSSMANN, *Some properties of vector fields not altered by small perturbations*, J. Differential Equations, 20 (1976), pp. 292–315.

FEEDBACK STABILIZATION OF PARABOLIC DISTRIBUTED PARAMETER SYSTEMS BY DISCRETE-TIME INPUT-OUTPUT DATA*

TOSHIHIRO KOBAYASHI†

Abstract. In this paper we investigate feedback stabilizability of continuous-time parabolic distributed parameter systems by discrete-time input-output data. We construct a stabilizer using the concepts of state feedback and output feedback through an observer. The design procedure is basically a modal approach which is realized in finite-dimensional theories and algorithms. It is shown that any initial state is reduced with an arbitrary decay rate. Explicit sufficient conditions are given for the convergence of the design scheme.

Key words. feedback stabilization, parabolic distributed parameter system, discrete-time data, finite-dimensional algorithm, finite-dimensional observer

1. Introduction. Many control components deliver their outputs in discrete, or sampled-data, form. Whenever a digital computer or a micro-computer constitutes a part of a control system, continuous signals must be discretized in order to be digestible by the computer. Discrete-time control theory is of great interest because of its application in computer control.

There have been several articles on stabilization of distributed parameter systems by continuous-time input-output data [2], [3], [4], [5], [8], [9], [10], [11]. Stabilizability only by discrete-time input-output data for continuous-time distributed parameter systems is an interesting problem.

In this paper we investigate stabilizability of parabolic distributed parameter systems by discrete-time input-output data. We construct a stabilizer using the concepts of a state feedback and a state observer. The key to stabilizability for parabolic systems is a decomposition of the state space based on the modes of the system. We present a design procedure which can be realized in finite-dimensional theories and techniques from a practical point of view. Explicit sufficient conditions are given for state feedback stabilizability and output feedback stabilizability through a finite-dimensional discrete-time observer. We show that any initial state is reduced with an arbitrary decay rate.

2. System description. Boundary or pointwise controls must first be formulated as a state equation. In this chapter the system description is following Lions [7].

Let V and H be Hilbert spaces with V , H and V' (the dual space of V) satisfying the inclusion relation

$$(2.1) \quad V \subset H \subset V'$$

with each space dense in the following with continuous injection.

We consider the system described by

$$(2.2) \quad \frac{du(t)}{dt} = Au(t) + Bf(t), \quad 0 < t < t_1 = jT, \quad u(0) = u_0 \in H,$$

where B is a bounded linear operator from a p -dimensional Euclidean space E^p to V' (we denote $B \in L(E^p, V')$), T is a sampling period and j is any positive integer. The control function $f(t) \in E^p$ is assumed to be constant between each sampling period

* Received by the editors December 22, 1982, and in revised form May 30, 1983.

† Department of Control Engineering, Kyushu Institute of Technology, Tobata, Kitakyushu 804, Japan.

$[kT, (k+1)T)$, that is,

$$(2.3) \quad f(t) = f_k, \quad kT \leq t < (k+1)T, \quad k = 0, 1, \dots, j-1$$

such that $\sum_{k=0}^{j-1} \|f_k\|_{E^p}^2 < \infty$. Then $f \in L^2(0, t_1; E^p)$ for $t_1 = jT$. The operator A is a bounded linear operator from V into V' and $-A$ is coercive; that is,

$$(2.4) \quad \begin{aligned} &\text{there exist } \beta \text{ and } \alpha > 0 \text{ such that} \\ &(-Au, u) + \beta \|u\|_H^2 \geq \alpha \|u\|_V^2, \quad u \in V. \end{aligned}$$

In this case the operator A is the generator of a strongly continuous semigroup $U(t)$ on H .

There exists the unique solution $u(t)$ of the system (2.2) such that

$$(2.5) \quad u(t) = U(t)u_0 + \int_0^t U(t-s)Bf(s) ds$$

and $u \in L^2(0, t_1; V) \cap C(0, t_1; H)$. Moreover the solution u depends continuously on u_0 and f [7, p. 102].

In physical situations, the output space is finite-dimensional. Suppose that outputs of the system are given at discrete times in the form

$$(2.6) \quad y_k = Cu(kT), \quad k = 0, 1, \dots,$$

where $C \in L(H, E')$.

Since $U(t)$ is a strongly continuous semigroup on H , we can obtain the system state at each instant as follows;

$$(2.7) \quad \begin{aligned} u(\overline{k+1}T) &= U(T)u(kT) + \int_{kT}^{(k+1)T} U(\overline{k+1}T-s)Bf_k ds \\ &= U(T)u(kT) + \int_0^T U(T-s)Bf_k ds. \end{aligned}$$

Defining an operator $D \in L(E^p, H)$ by

$$(2.8) \quad Df_k = \int_0^T U(T-s)Bf_k ds \quad \text{for } f_k \in E^p,$$

we have the discrete-time system

$$(2.9) \quad u(\overline{k+1}T) = U(T)u(kT) + Df_k, \quad k = 0, 1, \dots, \quad u(0) = u_0$$

$$(2.10) \quad y_k = Cu(kT), \quad k = 0, 1, \dots.$$

Moreover we can obtain the system state between each sampling period as follows. For $0 \leq m \leq 1$

$$(2.11) \quad \begin{aligned} u(kT+mT) &= U(mT)u(kT) + \int_{kT}^{kT+mT} U(kT+mT-s)Bf_k ds \\ &= U(mT)u(kT) + \int_0^{mT} U(mT-s)Bf_k ds \\ &= U(mT)u(kT) + D_m f_k, \quad k = 0, 1, \dots, \end{aligned}$$

where the operator $D_m \in L(E^p, H)$ is defined by

$$(2.12) \quad D_m f_k = \int_0^{mT} U(mT-s) B f_k ds \quad \text{for } f_k \in E^p.$$

In the next chapter we shall investigate stabilization of the system (2.2) by discrete-time state feedback.

3. Stabilization by discrete-time state feedback. In this chapter we investigate stabilization of the continuous-time system (2.2) and of the discrete-time system (2.9) using discrete-time state feedback.

We shall be interested in bounded feedback controls

$$(3.1) \quad f_k = -Fu(kT), \quad k = 0, 1, \dots,$$

that is,

$$(3.2) \quad f(t) = -Fu(kT), \quad kT \leq t < (k+1)T, \quad k = 1, 2, \dots,$$

where $F \in L(H, E^p)$. This results in the closed-loop systems

$$(3.3) \quad u(\overline{k+1}T) = (U(T) - DF)u(kT), \quad k = 0, 1, \dots$$

and

$$(3.4) \quad u(kT + mT) = (U(mT) - D_m F)u(kT), \quad k = 0, 1, \dots$$

for the system (2.9) and the system (2.11), respectively. Then we have

$$(3.5) \quad u(kT) = (U(T) - DF)^k u_0, \quad k = 1, 2, \dots$$

and

$$(3.6) \quad u(kT + mT) = (U(mT) - D_m F)(U(T) - DF)^k u_0, \quad 0 \leq m \leq 1, \quad k = 1, 2, \dots$$

Now we define stabilizability of the system (2.2) and of the system (2.9).

DEFINITION 1. The system (2.2) is stabilizable if there exists a feedback control (3.2) such that $\lim_{t \rightarrow \infty} \|u(t)\|_H = 0$.

DEFINITION 2. The system (2.9) is stabilizable if there exists a feedback control (3.1) such that $\lim_{k \rightarrow \infty} \|u(kT)\|_H = 0$.

It follows from (3.5) and (3.6) that $u(kT)$ and $u(kT + mT)$ have the same decay rate. Thus if the discrete-time system (2.9) is stabilizable by a feedback control (3.1), then the continuous-time system (2.2) is also stabilizable by a feedback control (3.2). We obtain the following theorem.

THEOREM 1. *The continuous-time system (2.2) is stabilized by a feedback control (3.2) if and only if the discrete-time system (2.9) is stabilized by a feedback control (3.1).*

The key to stabilization of the discrete-time system (2.9) is a decomposition of the state space H . We assume that the operator A satisfies the spectrum decomposition assumption [2, p. 75]; then there exists the orthogonal projection P such that

$$(3.7) \quad H = PH + QH, \quad Q = I - P$$

and PH , QH form A invariant subspaces of H . From the viewpoints of system analysis and synthesis, it is practical and interesting to take PH as a finite-dimensional space. We shall assume henceforth that PH is an N -dimensional subspace.

Let A_P and A_Q be the restrictions of A on PH and QH , respectively. We denote by $U_P(t)$ and $U_Q(t)$ the strongly continuous semigroups on PH and QH generated by A_P and A_Q , respectively. Actually A_P is bounded on PH and $U_P(t)$ is a uniformly continuous analytic semigroup.

In this case we obtain from (2.9) and (2.10)

$$(3.8) \quad Pu(\overline{k+1}T) = U_P(T)Pu(kT) + D_P f_k, \quad Pu(0) = Pu_0,$$

$$(3.9) \quad Qu(\overline{k+1}T) = U_Q(T)Qu(kT) + D_Q f_k, \quad Qu(0) = Qu_0,$$

$$(3.10) \quad y_k = C_P Pu(kT) + C_Q Qu(kT), \quad k = 0, 1, \dots$$

$$(3.11) \quad u(kT) = Pu(kT) + Qu(kT),$$

where D_P, D_Q are the restrictions of D on PH and QH , respectively. C_P, C_Q are the restrictions of C on PH and QH .

We also assume that the semigroup $U_O(t)$ satisfies the condition

$$(3.12) \quad \|U_O(kT)\| \leq Lq^k, \quad k = 1, 2, \dots$$

for constants $L \geq 1$ and $0 < q < 1$. Then the system (3.9) is stable without feedback controls.

We refer to the state Pu governed by the N -dimensional system (3.8) as the stabilized modes of the system (2.9) and the state Qu governed by the infinite-dimensional system (3.9) as the residual modes of the system (2.9).

Remark 1. If A is a symmetric operator with compact resolvent and lower semibounded spectrum, then there exists a sequence $\{\lambda_n, \phi_n; n = 1, 2, \dots\}$ of eigenvalues and corresponding orthonormal eigenfunctions such that for a constant c

$$c > \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq \dots, \quad \lim_{n \rightarrow \infty} \lambda_n = -\infty$$

and

$$A\phi_n = \lambda_n \phi_n, \quad n = 1, 2, \dots$$

Every vector u in H has a unique representation

$$u = \sum_{n=1}^{\infty} u_n \phi_n, \quad u_n = (u, \phi_n)_H.$$

The semigroup $U(t)$ is given by

$$U(t)u = \sum_{n=1}^{\infty} u_n \exp(\lambda_n t) \phi_n \quad \text{if } u \in H.$$

If we define the orthogonal projections P and Q by

$$Pu = \sum_{n=1}^N u_n \phi_n, \quad Qu = \sum_{n=N+1}^{\infty} u_n \phi_n \quad \text{if } u \in H,$$

then P and Q decompose the space H as $H = PH + QH$. Here N is chosen such that $\lambda_{N+1} < 0$ and $\lambda_N \neq \lambda_{N+1}$. The subspace PH is N -dimensional. The semigroup $U_O(t)$ is given by

$$U_O(t)u = \sum_{n=N+1}^{\infty} u_n \exp(\lambda_n t) \phi_n \quad \text{if } u \in H.$$

In this case the eigenvalues of $U_O(T)$ are $\exp(\lambda_{N+1}T), \exp(\lambda_{N+2}T), \dots$, and $U_O(T)$ satisfies $\|U_O(T)\| \leq \exp(\lambda_{N+1}T) < 1$, since $0 > \lambda_{N+1} \geq \lambda_{N+2} \geq \dots$. We get

$$(3.13) \quad \|U_O(kT)\| \leq (\exp(\lambda_{N+1}T))^k, \quad k = 1, 2, \dots$$

Thus we can take $L = 1$ and $q = \exp(\lambda_{N+1}T)$ in (3.12).

Now we shall construct a stabilizing feedback operator F for the decomposed subsystem (3.8). Let us consider a feedback operator $F = F_0 P$, $F_0 \in L(PH, E^P)$; that is, we use the feedback

$$(3.14) \quad f_k = -F_0 P u(kT), \quad k = 0, 1, \dots$$

Then from (3.8) and (3.9) we have the closed-loop system

$$(3.15) \quad P u(\overline{k+1}T) = (U_P(T) - D_P F_0) P u(kT),$$

$$(3.16) \quad Q u(\overline{k+1}T) = U_Q(T) Q u(kT) - D_Q F_0 P u(kT).$$

The eigenvalues of $U_{PF} = U_P(T) - D_P F_0$ can be arbitrarily located in the complex plane (within the restriction that complex eigenvalues occur in complex conjugate pairs) by choosing F_0 suitably if and only if the N -dimensional system (3.8) is controllable [6, p. 488]. Thus if the system (3.8) is controllable, the system is stabilizable by F_0 ; that is, all the eigenvalues of U_{PF} have moduli strictly less than 1.

A case of special interest occurs when all closed-loop eigenvalues are assigned to the origin. According to the Cayley-Hamilton theorem we have $U_{PF}^N = 0$. The state $P u(kT)$ can be expressed as

$$(3.17) \quad P u(kT) = U_{PF}^k P u_0.$$

If $U_{PF}^N = 0$, any initial state $P u_0$ is reduced to the zero state at or before the instant NT , that is, in N step or less. Thus we have at least

$$(3.18) \quad P u(kT) = 0, \quad k = N, N+1, \dots$$

Next let us estimate $Q u(kT)$. From (3.16)

$$(3.19) \quad Q u(kT) = U_Q^k(T) Q u_0 - \sum_{i=0}^{k-1} (U_Q(T))^{k-1-i} D_Q F_0 P u(iT).$$

Moreover from (3.18) we obtain for $k = N+1, N+2, \dots$

$$(3.20) \quad Q u(kT) = U_Q^k(T) Q u_0 - \sum_{i=0}^{N-1} (U_Q(T))^{k-1-i} D_Q F_0 P u(iT).$$

Using (3.12) we can estimate $Q u(kT)$ for $k = N+1, N+2, \dots$

$$(3.21) \quad \begin{aligned} \|Q u(kT)\| &\leq \|U_Q^k(T) Q u_0\| + \sum_{i=0}^{N-1} \|(U_Q(T))^{k-1-i} D_Q F_0 P u(iT)\| \\ &\leq L q^k \|Q u_0\| + L \|D_Q\| \|F_0\| \sum_{i=0}^{N-1} q^{k-1-i} \|P u(iT)\| \end{aligned}$$

(since there exists a constant L_N such that $\|P u(iT)\| \leq L_N$, $i = 0, \dots, N-1$)

$$\leq L \|Q u_0\| q^k + L L_N \|D_Q\| \|F_0\| \frac{1-q^N}{1-q} q^{k-N}.$$

From (3.11), (3.18) and (3.21) we have for $k = N+1, N+2, \dots$

$$(3.22) \quad \|u(kT)\| \leq \|P u(kT)\| + \|Q u(kT)\| = \|Q u(kT)\| \leq \bar{L}_N q^k,$$

where

$$\bar{L}_N = L \|Q u_0\| + L L_N \|D_Q\| \|F_0\| \frac{q^{-N}-1}{1-q}.$$

However, the control law that assigns all the eigenvalues of U_{PF} to the origin may lead to excessively large input amplitudes or to an undesirable transient behavior [6, p. 489].

If a feedback operator F_0 is chosen such that

$$(3.23) \quad \|U_{PF}^k\| \leq K_1 \gamma^k, \quad k = 1, 2, \dots$$

for constants $K_1 \geq 1$ and $0 < \gamma < 1$; then we can estimate $Pu(kT)$ and $Qu(kT)$ as follows.

$$(3.24) \quad \|Pu(kT)\| \leq K_1 \|Pu_0\| \gamma^k, \quad k = 1, 2, \dots,$$

$$(3.25) \quad \begin{aligned} \|Qu(kT)\| &\leq Lq^k \|Qu_0\| + L\|D_O\| \|F_0\| \sum_{i=0}^{k-1} q^{k-1-i} \|Pu(iT)\| \\ &\leq L\|Qu_0\| q^k + LK_1 \|D_O\| \|F_0\| \|Pu_0\| \sum_{i=0}^{k-1} q^{k-1-i} \gamma^i \\ &= L\|Qu_0\| q^k + LK_1 \|D_O\| \|F_0\| \|Pu_0\| \frac{q^k - \gamma^k}{q - \gamma}, \quad k = 1, 2, \dots, \end{aligned}$$

since we can choose γ such that $\gamma \neq q$. Estimates (3.24) and (3.25) give

$$(3.26) \quad \begin{aligned} \|u(kT)\| &\leq \|Pu(kT)\| + \|Qu(kT)\| \\ &\leq K_1 \|Pu_0\| \gamma^k + L\|Qu_0\| q^k + LK_1 \|D_O\| \|F_0\| \|Pu_0\| \frac{q^k - \gamma^k}{q - \gamma}, \\ &\leq \left[K_1 + L + LK_1 \|D_O\| \|F_0\| \frac{1}{|q - \gamma|} \right] \|u_0\| \delta^k, \\ &= K_2 \delta^k \|u_0\|, \quad \delta = \max(q, \gamma), \quad k = 1, 2, \dots \end{aligned}$$

from which we have

$$(3.27) \quad \|(U(T) - DF)^k\| \leq K_2 \delta^k, \quad k = 1, 2, \dots$$

The estimate (3.22) and (3.27) show that any initial state will be reduced with an arbitrary decay rate.

Consequently we have obtained

THEOREM 2. *Suppose that A satisfies the spectrum decomposition assumption and (3.12) holds. Then the discrete-time system (2.9) is stabilizable by a feedback control $F = F_0 P$, $F_0 \in L(PH, E^p)$, if the N -dimensional system (3.8) is controllable.*

Remark 2. The feedback operator F_0 can be determined by pole allocation or by optimal regulator design using known finite-dimensional algorithms [6, p. 490].

Theorem 1 and Theorem 2 say that the continuous-time system (2.2) is stabilizable by a feedback control (3.2), if the N -dimensional system (3.8) is controllable.

4. Stabilization by discrete-time output feedback. Theorem 2 gives the basic solution for stabilizability of the system (2.9). However we assumed the knowledge of $Pu(kT)$ in the feedback control law (3.14). In this chapter we shall show that the system (2.2) can be stabilized by discrete-time output feedback, if we construct a finite-dimensional discrete-time observer.

First we construct a discrete-time identity observer

$$(4.1) \quad v(k+1) = U(T)v(k) + Df_k + G(y_k - Cv(k)), \quad v(0) = 0$$

where $G \in L(E', H)$. From (2.9), (2.10) and (4.1) we have the error system

$$(4.2) \quad e(k+1) = (U(T) - GC)e(k), \quad e(0) = -u_0$$

where $e(k) = v(k) - u(kT)$.

We can prove the following theorem.

THEOREM 3. *Suppose that A satisfies the spectrum decomposition assumption and (3.12). Then the error system (4.2) is stabilized by a feedback $G \in L(E', H)$ if the N -dimensional system $(U_P(T), C_P)$ is observable.*

Proof. Let us choose G such that $QG = 0$, that is,

$$(4.3) \quad Gy = \begin{cases} G_0 y & \text{on } PH, \\ 0 & \text{on } QH, \end{cases} \quad \text{for } y \in E',$$

where $G_0 \in L(E', PH)$. Decomposing e by Pe and Qe , we have

$$(4.4) \quad Pe(k+1) = (U_P(T) - G_0 C_P)Pe(k) - G_0 C_Q Qe(k), \quad Pe(0) = -Pu_0,$$

$$(4.5) \quad Qe(k+1) = U_Q(T)Qe(k), \quad Qe(0) = -Qu_0.$$

From the finite-dimensional theory, the eigenvalues of $W_P = U_P(T) - G_0 C_P$ can be arbitrarily located in the complex plane by choosing G_0 suitably, if the N -dimensional system $(U_P(T), C_P)$ is observable. Thus there are constants $K_3 \geq 1$ and $0 < \sigma < 1$ such that

$$(4.6) \quad \|W_P^k\| \leq K_3 \sigma^k \quad k = 1, 2, \dots$$

In this case from (4.5) and (3.12)

$$(4.7) \quad \|Qe(k)\| = \|U_Q^k(T)Qe_0\| \leq L \|Qe_0\| q^k.$$

From (4.4)

$$(4.8) \quad Pe(k) = W_P^k Pe_0 - \sum_{i=0}^{k-1} W_P^{k-1-i} G_0 C_Q Qe(i).$$

Equations (4.6) and (4.7) imply

$$(4.9) \quad \begin{aligned} \|Pe(k)\| &\leq K_3 \|Pe_0\| \sigma^k + L K_3 \|G_0\| \|C_Q\| \|Qe_0\| \sum_{i=0}^{k-1} \sigma^{k-1-i} q^i \\ &= K_3 \|Pe_0\| \sigma^k + L K_3 \|G_0\| \|C_Q\| \|Qe_0\| \frac{q^k - \sigma^k}{q - \sigma}, \end{aligned}$$

since we can choose σ such that $\sigma \neq q$. It follows from (4.7) and (4.9) that

$$(4.10) \quad \begin{aligned} \|e(k)\| &\leq \|Pe(k)\| + \|Qe(k)\| \\ &\leq \left[K_3 + L K_3 \|G_0\| \|C_Q\| \frac{1}{|q - \sigma|} + L \right] \|e_0\| \mu^k \\ &= K_4 \|e(0)\| \mu^k, \quad k = 1, 2, \dots, \end{aligned}$$

where $\mu = \max(q, \sigma) < 1$. From (4.10) the following estimate holds.

$$(4.11) \quad \|(U(T) - GC)^k\| \leq K_4 \mu^k, \quad k = 1, 2, \dots$$

Thus we have proved the theorem.

Remark 3. If we specially choose G_0 such that all the eigenvalues of W_P are assigned to the origin, we have $W_P^N = 0$. In this case (4.8) becomes for $k = N, N+1, \dots$

$$(4.12) \quad Pe(k) = - \sum_{i=k-N}^{k-1} W_P^{k-1-i} G_0 C_0 Qe(i).$$

Estimating $\|Pe(k)\|$, we have $\|Pe(k)\| \leq \text{Const. } q^k \|e_0\|$. Thus

$$(4.13) \quad \|e(k)\| \leq \text{Const. } q^k \|e(0)\|$$

is obtained from (4.7).

Remark 4. Although the feedback operator G in (4.1) can be determined using finite-dimensional algorithms, the infinite-dimensional observer (4.1) is not so easy to realize. If D has its range in PH , we can construct an N -dimensional observer

$$(4.14) \quad Pv(k+1) = U_P(T)Pv(k) + D_P f_k + G_0 y_k - G_0 C_P Pv(k),$$

since $C_Q Qv(k) = 0$, $k = 1, 2, \dots$ are implied from $Qv(0) = 0$.

Now we investigate stabilization of the system by output feedback through an observer (4.1). In place of a feedback control law (3.14), let us use the feedback

$$(4.15) \quad f_k = -F_0 Pv(k), \quad k = 0, 1, \dots$$

Then we get the following closed-loop system

$$(4.16) \quad \begin{aligned} u(\overline{k+1}T) &= U(T)u(kT) - DF_0 Pv(k), \\ v(k+1) &= U(T)v(k) - DF_0 Pv(k) + G(y_k - Cv(k)), \end{aligned}$$

that is,

$$(4.17) \quad \begin{aligned} u(\overline{k+1}T) &= (U(T) - DF)u(kT) - DFe(k), \\ e(k+1) &= (U(T) - GC)e(k). \end{aligned}$$

If the N -dimensional system (3.8) is controllable and the N -dimensional system $(U_P(T), C_P)$ is observable, there exist operators F and G such that the estimates (3.27) and (4.11) hold. Then we have

$$(4.18) \quad \begin{aligned} \|u(kT)\| &= \left\| (U(T) - DF)^k u_0 + \sum_{i=0}^{k-1} (U(T) - DF)^{k-1-i} DFe(i) \right\| \\ &\leq K_2 \|u_0\| \delta^k + K_4 \|D\| \|F\| \|e(0)\| \sum_{i=0}^{k-1} \delta^{k-1-i} \mu^i \end{aligned}$$

(since we can choose γ and σ such that $\delta \neq \mu$)

$$\begin{aligned} &\leq K_2 \|u_0\| \delta^k + K_4 \|D\| \|F_0\| \|e_0\| \frac{\delta^k - \mu^k}{|\delta - \mu|} \\ &\leq K_5 \|u_0\| \eta^k, \quad k = 1, 2, \dots, \end{aligned}$$

where $K_5 = K_2 + K_4 \|D\| \|F_0\| / |\delta - \mu|$ and $\eta = \max(\delta, \mu) = \max(q, \gamma, \sigma) < 1$. Furthermore

$$\begin{aligned} \|v(k)\| &= \|e(k) + u(kT)\| \\ &\leq \|e(k)\| + \|u(kT)\| \leq K_4 \mu^k \|e_0\| + K_5 \eta^k \|u_0\| \\ &\leq (K_4 + K_5) \eta^k \|u_0\|. \end{aligned}$$

From this and (4.18) we get

$$(4.19) \quad \left\| \begin{pmatrix} u(kT) \\ v(k) \end{pmatrix} \right\| = (\|u(kT)\|^2 + \|v(k)\|^2)^{1/2} \\ \leq \sqrt{K_5^2 + (K_4 + K_5)^2} \eta^k \|u_0\| = K_6 \eta^k \|u_0\|, \quad k = 1, 2, \dots$$

The estimate (4.19) shows that we can obtain an arbitrary decay rate by suitable choice of q , γ and σ . Therefore we have

THEOREM 4. *Suppose that A satisfies the spectrum decomposition assumption and (3.12) holds. Then the discrete-time system (2.9) is stabilizable by a feedback law (4.15), if the system (3.8) is controllable and the system $(U_P(T), C_P)$ is observable.*

From (2.11), (4.15) and (4.19) Theorem 4 says that the continuous-time system (2.2) is stabilizable by discrete-time output feedback through an identity observer (4.1), if the system (3.8) is controllable and the system $(U_P(T), C_P)$ is observable.

However, the infinite-dimensional observer (4.1) is not so easy to realize. Unfortunately, the assumption that D has its range in PH excludes boundary or pointwise controls. This assumption is also unrealistic and one could not expect to be so lucky in practice.

In the case where A satisfies the conditions of Remark 1, we can show that the system (2.2) is stabilized by discrete-time output feedback through a finite-dimensional observer.

Define the other orthogonal projections P_M and Q_M by

$$P_M u = \sum_{n=1}^M u_n \phi_n, \quad Q_M u = \sum_{n=M+1}^{\infty} u_n \phi_n \quad \text{if } u \in H,$$

where $M \geq N$. Let us consider the following system:

$$(4.20) \quad \begin{aligned} u(\overline{k+1}T) &= U(T)u(kT) - DF_0 P v(k), \\ v(k+1) &= (U(T) - GC)v(k) - P_M DF_0 P v(k) + G y_k. \end{aligned}$$

This system adds a bounded perturbation

$$(4.21) \quad \bar{Q} = \begin{pmatrix} 0 & 0 \\ 0 & Q_M DF_0 P \end{pmatrix}$$

to the system (4.16).

For a bounded perturbation the following lemma holds.

LEMMA 1. *If A is a linear bounded operator on a Banach space X such that for $0 < \eta < 1$, $C \geq 1$ $\|A^k\| \leq C\eta^k$, $k = 1, 2, \dots$ and B is a bounded linear operator on X , then $A + B$ is bounded with*

$$(4.22) \quad \|(A + B)^k\| \leq \frac{C(\eta + \|B\|)}{\eta + C\|B\|} (\eta + C\|B\|)^k, \quad k = 1, 2, \dots$$

To show Lemma 1, the next lemma is necessary.

LEMMA 2. *If the sequence $\{u_k\}_{k=0}^{\infty}$ of positive reals satisfies the recurrent inequality*

$$u_{k+1} \leq b + a \sum_{i=0}^k u_i, \quad k = 0, 1, \dots$$

where $a, b > 0$ and $u_0 = 1$, then

$$u_k \leq \frac{b+a}{1+a} (1+a)^k, \quad k = 1, 2, \dots$$

Proof of Lemma 2. If we define $H_k = a \sum_{i=0}^k u_i$, then

$$H_{k+1} - H_k = au_{k+1} \leq ab + aH_k.$$

Hence

$$H_{k+1} \leq ab + (1+a)H_k.$$

Let $\{y_k\}_{k=0}^\infty$ be the solution of the difference equation

$$y_{k+1} = ab + (1+a)y_k \quad \text{with } y_0 = H_0.$$

By induction we get that $H_k \leq y_k$ for all k . The solution y_k is easily obtained

$$y_k = (1+a)^k y_0 + \{(1+a)^k - 1\}b.$$

Thus we get

$$u_{k+1} \leq b + H_k \leq b + (1+a)^k a + \{(1+a)^k - 1\}b = (1+a)^k (a+b) = \frac{b+a}{1+a} (1+a)^{k+1}.$$

Proof of Lemma 1. Consider the difference equation

$$u_{k+1} = Au_k + Bu_k, \quad k = 0, 1, \dots$$

Then $u_k = (A+B)^k u_0$, $k = 1, 2, \dots$. An alternative representation for the solution is

$$u_k = A^k u_0 + \sum_{i=0}^{k-1} A^{k-1-i} B u_i, \quad k = 1, 2, \dots$$

Thus $(A+B)^k$ must satisfy

$$(4.23) \quad (A+B)^k = A^k + \sum_{i=0}^{k-1} A^{k-1-i} B (A+B)^i, \quad k = 1, 2, \dots,$$

from which we get

$$\begin{aligned} \|(A+B)^k\| &\leq \|A^k\| + \sum_{i=0}^{k-1} \|A^{k-1-i}\| \|B\| \|(A+B)^i\| \\ &\leq C\eta^k + \sum_{i=0}^{k-1} C\eta^{k-1-i} \|B\| \|(A+B)^i\|. \end{aligned}$$

Therefore

$$\left\| \frac{1}{\eta^k} (A+B)^k \right\| \leq C + \frac{C}{\eta} \sum_{i=0}^{k-1} \|B\| \left\| \frac{1}{\eta^i} (A+B)^i \right\|.$$

From Lemma 2 we have

$$\left\| \frac{1}{\eta^k} (A+B)^k \right\| \leq \frac{C + (C/\eta) \|B\|}{1 + (C/\eta) \|B\|} \left(1 + \frac{C}{\eta} \|B\| \right)^k = \frac{C(\eta + \|B\|)}{\eta + C\|B\|} \left(1 + \frac{C}{\eta} \|B\| \right)^k,$$

so that

$$\|(A+B)^k\| \leq \frac{C(\eta + \|B\|)}{\eta + C\|B\|} (\eta + C\|B\|)^k.$$

Now for any small $\varepsilon > 0$ there exists some large M such that $\|\bar{Q}\| = \|Q_M D F_0 P\| < \varepsilon/K_6$. Lemma 1 says that for any small $\varepsilon > 0$ there exists sufficiently large M such that the perturbed system (4.20) has a decay rate $(\eta + \varepsilon)^k$.

Moreover restricting the system (4.20) to $H \times P_M H$, we have

$$(4.24) \quad u(\overline{k+1}T) = U(T)u(kT) - DF_0 P v(k),$$

$$(4.25) \quad P_M v(k+1) = (U(T) - GC)_M P_M v(k) - P_M DF_0 P v(k) + P_M G y_k,$$

where $(U(T) - GC)_M$ is the restriction of $U(T) - GC$ to $P_M H$. The system (4.25) is M -dimensional observer. For the system (4.24) and (4.25) the estimate:

$$(4.26) \quad \left\| \begin{pmatrix} u(kT) \\ P_M v(k) \end{pmatrix} \right\| \leq \text{Const. } (\eta + K_6 \|Q_M DF_0 P\|)^k \|u_0\|, \quad k = 1, 2, \dots$$

holds corresponding to the estimate (4.19).

The estimate (4.26) says that if we choose an appropriately large M , we can stabilize the system (2.9) (moreover the system (2.2)) by output feedback through an M -dimensional observer (4.25). Thus for a finite-dimensional observer (4.25) it has been shown that Theorem 4 still holds.

5. Example. In this chapter we will give a simple example to illustrate the presented theory.

Let us consider the system

$$(5.1) \quad \begin{aligned} \frac{\partial u(t, x)}{\partial t} &= \frac{\partial^2 u(t, x)}{\partial x^2} + 5\pi^2 u(t, x) + \delta(x - 0.3)f(t) \quad \text{on } 0 \leq x \leq 1, \\ u(t, 0) &= u(t, 1) = 0, \quad u(0, x) = u_0(x). \end{aligned}$$

For this system we take $H = L^2(0, 1)$ and $Au = \Delta u + 5\pi^2 u$. Here Δ is the Laplacian with Dirichlet conditions at $x = 0$ and $x = 1$. For a pointwise control f the operator B becomes $B = \delta(x - 0.3)$. Thus we can take $V = H_0^1(0, 1) = \{v \in L^2(0, 1) \text{ such that } dv/dx \in L^2(0, 1) \text{ and } v(0) = v(1) = 0\}$. Then $V' = H^{-1}(0, 1)$ [7] and $B \in L(E, H^{-1}(0, 1))$.

In this case since

$$\begin{aligned} (-Au, u) &= -\int_0^1 \frac{d^2 u(x)}{dx^2} u(x) dx - 5\pi^2 \int_0^1 u^2(x) dx \\ &= \int_0^1 \left(\frac{du(x)}{dx} \right)^2 dx - 5\pi^2 \int_0^1 u^2(x) dx, \end{aligned}$$

we have $A \in L(H_0^1(0, 1), H^{-1}(0, 1))$ and we can take $\alpha = 1$, $\beta = 1 + 5\pi^2$ in (2.4).

The discrete-time output y_k is given

$$(5.2) \quad y_k = Cu(kT) = \int_0^1 c(x)u(kT, x) dx, \quad k = 0, 1, \dots$$

where $c \in L^2(0, 1)$. Then the operator $C \in L(L^2(0, 1), E)$.

The eigenvalues of A are

$$\lambda_n = (5 - n^2)\pi^2, \quad n = 1, 2, \dots$$

with the eigenfunctions

$$\phi_n(x) = \sqrt{2} \sin(n\pi x), \quad n = 1, 2, \dots$$

which constitute an orthonormal basis for $L^2(0, 1)$. Thus the operator A satisfies the spectrum decomposition assumption. The semigroup $U(t)$ is given by

$$(U(t)u_0)(x) = \sum_{n=1}^{\infty} e^{\lambda_n t} \phi_n(x) \int_0^1 u_0(y) \phi_n(y) dy.$$

Choose

$$PH = \text{span} \{ \phi_n(x); n = 1, 2 \}, \quad QH = \text{span} \{ \phi_n(x); n = 3, 4, \dots \},$$

then $N = 2$ and

$$(U_P(t)u_0)(x) = \sum_{n=1}^2 e^{\lambda_n t} \phi_n(x) \int_0^1 u_0(y) \phi_n(y) dy,$$

$$(U_Q(t)u_0)(x) = \sum_{n=3}^{\infty} e^{\lambda_n t} \phi_n(x) \int_0^1 u_0(y) \phi_n(y) dy.$$

In this case

$$(5.3) \quad \|U_Q(T)\| \leq e^{\lambda_3 T} = e^{-4\pi^2 T}$$

which implies that $L = 1$ and $q = e^{-4\pi^2 T}$ in (3.12).

Relative to the basis ϕ_1, ϕ_2 for PH , we have

$$U_P(T) = \begin{pmatrix} e^{\lambda_1 T} & 0 \\ 0 & e^{\lambda_2 T} \end{pmatrix}, \quad C_P = [c_1 \quad c_2], \quad c_n = \int_0^1 c(x) \phi_n(x) dx, \quad n = 1, 2.$$

Moreover since

$$Df = \int_0^T U(T-s) \delta(x-0.3) f ds = \sum_{n=1}^{\infty} \frac{e^{\lambda_n T} - 1}{\lambda_n} \phi_n(0.3) \phi_n(x) f,$$

we have

$$D_P = \begin{pmatrix} \frac{e^{\lambda_1 T} - 1}{\lambda_1} \phi_1(0.3) \\ \frac{e^{\lambda_2 T} - 1}{\lambda_2} \phi_2(0.3) \end{pmatrix},$$

relative to the basis ϕ_1, ϕ_2 for PH .

It is easily checked that the system $(U_P(T), D_P)$ is controllable. The system $(U_P(T), C_P)$ is observable, if $c_n \neq 0$, $n = 1, 2$.

So an output feedback through an identity observer for our system is given by

$$(5.4) \quad f_k = -F_0 P v(k), \quad k = 0, 1, \dots,$$

$$(5.5) \quad v(k+1) = U(T) v(k) + D f_k + G y_k - G C v(k),$$

where

$$F_0 P v(k) = [F_{01} \quad F_{02}] \begin{pmatrix} v_1(k) \\ v_2(k) \end{pmatrix}, \quad v_n(k) = \int_0^1 v(k, x) \phi_n(x) dx,$$

$$G C v(k) = \sum_{n=1}^2 g_n \left(\int_0^1 c(y) v(k, y) dy \right) \phi_n(x),$$

$$G y_k = \sum_{n=1}^2 g_n y_k \phi_n(x), \quad G_0 = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}.$$

If we construct the matrices F_0 and G_0 such that the eigenvalues of $U_P(T) - D_P F_0$ and $U_P(T) - G_0 C_P$ are distinct and contained in $|\lambda| \leq e^{-2\pi^2 T}$, then from (4.18) $\eta = e^{-2\pi^2 T}$ and the reduced rate of $\|u(kT)\|$ is of the order of $e^{-2\pi^2 T k}$.

Moreover restricting the system (5.5) to $P_M H = \text{span} \{\phi_n(x); n = 1, \dots, M\}$ and constructing an M -dimensional observer (4.25), we get

$$(5.6) \quad P_M v(k+1) = (U(T))_M P_M v(k) + P_M D F_0 P v(k) + P_M G y_k - (G C)_M P_M v(k)$$

where

$$\begin{aligned} (U(T))_M P_M v(k) &= \sum_{n=1}^M e^{\lambda_n T} \phi_n(x) \int_0^1 \left[\sum_{i=1}^M v_i(k) \phi_i(y) \right] \phi_n(y) dy, \\ P_M D F_0 P v(k) &= \sum_{n=1}^M \frac{e^{\lambda_n T} - 1}{\lambda_n} \left[\phi_n(0.3) \sum_{i=1}^2 F_{0i} v_i(k) \right] \phi_n(x), \\ P_M G y_k &= \sum_{n=1}^2 g_n y_k \phi_n(x), \\ (G C)_M P_M v(k) &= \sum_{n=1}^2 g_n \left(\int_0^1 c(y) \sum_{i=1}^M v_i(k) \phi_i(y) dy \right) \phi_n(x). \end{aligned}$$

Then, relative to the basis $\phi_1, \phi_2, \dots, \phi_M$ for $P_M H$, we have for (5.6)

$$\begin{aligned} (5.7) \quad \begin{pmatrix} v_1(k+1) \\ v_2(k+1) \\ \vdots \\ v_M(k+1) \end{pmatrix} &= \begin{pmatrix} e^{\lambda_1 T} & & 0 \\ & e^{\lambda_2 T} & \\ 0 & & \ddots \\ & & & e^{\lambda_M T} \end{pmatrix} \begin{pmatrix} v_1(k) \\ v_2(k) \\ \vdots \\ v_M(k) \end{pmatrix} \\ &+ \begin{pmatrix} (e^{\lambda_1 T} - 1) \phi_1(0.3) / \lambda_1 \\ (e^{\lambda_2 T} - 1) \phi_2(0.3) / \lambda_2 \\ \vdots \\ (e^{\lambda_M T} - 1) \phi_M(0.3) / \lambda_M \end{pmatrix} \begin{bmatrix} F_{01} & F_{02} & 0 & \dots & 0 \end{bmatrix} \begin{pmatrix} v_1(k) \\ v_2(k) \\ \vdots \\ v_M(k) \end{pmatrix} \\ &+ \begin{pmatrix} g_1 \\ g_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} y_k - \begin{pmatrix} g_1 \\ g_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{bmatrix} c_1 & c_2 & \dots & c_M \end{bmatrix} \begin{pmatrix} v_1(k) \\ v_2(k) \\ \vdots \\ v_M(k) \end{pmatrix}. \end{aligned}$$

Next estimate the constants K_1, K_2, \dots, K_6 . Let the eigenvalues of $U_P(T) - D_P F_0$ and $U_P(T) - G_0 C_P$ be ξ_1, ξ_2, ξ_3 and ξ_4 such that $1 > \xi_1 > \xi_2 > \xi_3 > \xi_4 > q = e^{-4\pi^2 T}$. Then we take $K_1 = K_3 = 1$ in (3.24) and (4.6). Thus the constant K_2 in (3.26) is given

$$K_2 = 2 + \frac{\|D_Q\| \|F_0\|}{|q - \gamma|} = 2 + \frac{\|D_Q\| \|F_0\|}{\xi_1 - q},$$

since $\gamma = \xi_1$. The constant K_4 in (4.10) is given

$$K_4 = 2 + \frac{\|G_0\| \|C_Q\|}{|q - \sigma|} = 2 + \frac{\|G_0\| \|C_Q\|}{\xi_3 - q},$$

since $\sigma = \xi_3$. Moreover we have

$$K_5 = K_2 + \frac{K_4 \|D\| \|F_0\|}{\xi_1 - \xi_3},$$

since $\delta = \max(q, \gamma) = \xi_1$ and $\mu = \max(q, \sigma) = \xi_3$.

On the other hand the feedback matrices F_0 and G_0 are obtained as follows:

$$\begin{aligned} F_{01} &= \frac{(\beta_1 - \xi_1)(\beta_1 - \xi_2)}{d_1(\beta_1 - \beta_2)}, & F_{02} &= \frac{(\beta_2 - \xi_1)(\beta_2 - \xi_2)}{d_2(\beta_2 - \beta_1)}, \\ g_1 &= \frac{(\beta_1 - \xi_3)(\beta_1 - \xi_4)}{c_1(\beta_1 - \beta_2)}, & g_2 &= \frac{(\beta_2 - \xi_3)(\beta_2 - \xi_4)}{c_2(\beta_2 - \beta_1)}, \end{aligned}$$

where $\beta_i = e^{\lambda_i T}$, $d_i = (D_P)_i$ for $i = 1, 2$.

Let us consider the case where $T = 0.02$, $\xi_1 = e^{-0.04\pi^2}$, $\xi_2 = e^{-0.05\pi^2}$, $\xi_3 = e^{-0.06\pi^2}$, $\xi_4 = e^{-0.07\pi^2}$ and

$$c(x) = \begin{cases} 0, & 0 \leq x \leq 0.6, \quad 0.8 \leq x \leq 1 \\ 1, & 0.6 \leq x \leq 0.8. \end{cases}$$

Then we have the following numerical results.

$$\begin{aligned} F_0 &= [70.9507 \quad -11.3034], & G &= [12.6667 \quad 1.92562]^T, \\ \|D\| &= 2.25021 \times 10^{-3}, & \|C\| &= 2, \\ \|D_Q\| &= 1.51439 \times 10^{-4}, & \|C_Q\| &= 8.58104 \times 10^{-2}, \\ K_2 &= 6.02272, & K_4 &= 39.8793, & K_5 &= 1132.02, & K_6 &= 1629.36, \\ \eta &= \xi_1 = 0.673824. \end{aligned}$$

In this case, if M is larger than 667, the sufficient condition for stability $\eta + K_6 \|Q_M D\| \|F_0\| < 1$ holds. For example when $M = 800$, that is, we construct 800-dimensional observer, we have $\eta + K_6 \|Q_M D\| \|F_0\| = 0.885238$. In the numerical simulation we approximated the space $L^2(0, 1)$ by $\text{span}\{\phi_n(x); n = 1, 2, \dots, 1000\}$.

6. Conclusions. In this paper we have investigated feedback stabilizability of continuous-time distributed parameter systems by discrete-time input-output data. First we have shown that the continuous-time system is stabilizable by discrete-time input-output data if the corresponding discrete-time system is stabilizable. The key to stabilizability for parabolic systems is a decomposition of the state space based on the modes of the system. We have assumed that the spectrum of A can be decomposed into two parts, a finite, possibly unstable part, and an infinite stable part. In this case we have constructed a stabilizer and a state observer in discrete-time form using finite-dimensional theories and algorithms. We have given explicit sufficient conditions for state feedback stabilizability and output feedback stabilizability through a finite-dimensional observer. We have also shown that any initial state is reduced with an arbitrary decay rate.

REFERENCES

- [1] R. F. CURTAIN AND A. J. PRITCHARD, *Functional Analysis in Modern Applied Mathematics*, Academic Press, London, 1977.
- [2] ———, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, Berlin, 1978.

- [3] R. F. CURTAIN, *Finite-dimensional compensator design for parabolic distributed systems with point sensors and boundary input*, IEEE Trans. Automat. Control, AC-23 (1982), pp. 98–104.
- [4] N. FUJII, *Feedback stabilization of distributed parameter systems by a functional observer*, this Journal, 18 (1980), pp. 108–120.
- [5] T. KOBAYASHI, *Controllability and stabilizability of sensitivity combined systems for distributed parameter systems*, Int. J. Control, 35 (1982), pp. 309–321.
- [6] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.
- [7] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
- [8] T. NAMBU, *Feedback stabilization of diffusion equations by a functional observer*, J. Differential Equations, 43 (1982), pp. 257–280.
- [9] A. J. PRITCHARD AND J. ZABCZYK, *Stability and stabilizability of infinite-dimensional systems*, SIAM Rev., 23 (1981), pp. 25–52.
- [10] Y. SAKAWA AND T. MATSUSHITA, *Feedback stabilization of a class of distributed systems and construction of a state estimator*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 748–753.
- [11] R. TRIGGIANI, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383–403.
- [12] T. NAMBU, *Feedback stabilization of diffusion equations*, 1982 Joint Automatic Control Conference in Japan. (In Japanese.)

SECOND ORDER NECESSARY CONDITIONS IN OPTIMIZATION*

J. WARGA†

Abstract. It is known that if a restricted minimization problem satisfies first order necessary conditions for minimum at some point with multiple choices of Lagrange multiplier vectors (or linear functionals) then, in general, second order conditions for different critical variations may require different Lagrange multipliers. We present here a relatively simple derivation of new second order necessary conditions in which different critical variations share a common Lagrange multiplier if they are “pairwise critical”. The problems that we consider contain restrictions in the form of finitely many equalities and of (possibly infinite-dimensional) inclusions involving arbitrary convex bodies. These new conditions generalize in some respects previous results of Dennis S. Bernstein (*A systematic approach to higher-order necessary conditions in optimization theory*, SIAM J. Control Optim., 22 (1984), pp. 211–238).

Key words. equality restrictions, infinite-dimensional inclusion restrictions, nonunique Lagrange multipliers, nonlinear programming, optimal control

The purpose of the present note is to provide a rather simple and self-contained derivation of new second order conditions for optimization problems, more general than those of [4, Thm. 2.3, p. 294]. These new conditions also generalize certain recent results of Bernstein [2] (specifically [2, Thm. 5.2] which is an improvement of [4, Thm. 2.3, p. 294]) in two respects. The set C defining the infinite-dimensional restrictions may be an arbitrary convex body and not only a cone; and the Lagrange “multiplier” l may be common to all elements of a set Y of “critical variations” satisfying certain conditions [(a) and (b) of Theorem A below] and need not be chosen separately for each critical variational direction. For a listing of other related work we refer the reader to the extensive bibliography in Bernstein’s paper [2].

We might add that our proof (closely patterned after a derivation of first-order conditions [3, Lemma V.2.2, pp. 301–303]) can be simply adapted to derive third—and higher-order conditions (such as [2, Thm. 6.1]) for problems with an arbitrary convex body C but with the Lagrange multiplier l possibly dependent on individual critical variations.

Let Q be a convex subset of a real vector space, \mathcal{X} a normed vector space with its topological dual denoted by \mathcal{X}^* , C a convex body in \mathcal{X} (i.e. a closed convex set with a nonempty interior), and $\phi \triangleq (\phi_0, \phi_1, \phi_2): Q \rightarrow \mathbb{R} \times \mathbb{R}^m \times \mathcal{X}$. Let

$$\mathcal{T}_k \triangleq \{(\theta_1, \dots, \theta_k) \in \mathbb{R}^k \mid \theta_j \geq 0, \sum \theta_j \leq 1\},$$

and assume that \bar{q} yields the minimum of $\phi_0(q)$ on the set $\{q \in Q \mid \phi_1(q) = 0, \phi_2(q) \in C\}$ and that, for every choice of a positive integer k and of $q_1, \dots, q_k \in Q$, the function

$$\theta \rightarrow \Phi(\theta) \triangleq \phi\left(\bar{q} + \sum_{j=1}^k \theta_j(q_j - \bar{q})\right): \mathcal{T}_k \rightarrow \mathbb{R} \times \mathbb{R}^m \times \mathcal{X}$$

admits first derivatives relative to \mathcal{T}_k in some neighborhood of 0 in \mathcal{T}_k and admits a second derivative relative to \mathcal{T}_k at 0. (Here, and in the sequel, we define the derivative $\psi'(\bar{\theta})$ relative to \mathcal{T}_k of a function ψ on \mathcal{T}_k by the relation

$$\lim |h|^{-1}[\psi(\bar{\theta} + h) - \psi(\bar{\theta}) - \psi'(\bar{\theta})h] = 0 \quad \text{as } h \rightarrow 0 \text{ while } h \in \mathcal{T}_k - \bar{\theta}.$$

* Received by the editors March 28, 1983, and in revised form June 13, 1983. This research was supported in part by the National Science Foundation under grant MCS 8102079.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.

The second derivative is similarly defined.) We observe that $\Phi'(0)(\theta_1, \dots, \theta_k)$ and $\Phi''(0)(\theta_1, \dots, \theta_k)(\omega_1, \dots, \omega_k)$ depend only on the choice of q_j corresponding to $\theta_j \neq 0$ or $\omega_j \neq 0$. We can write, therefore,

$$\begin{aligned} & \phi'(\bar{q})(q_1 - \bar{q}) \quad \text{for } \Phi'(0)(1, 0, \dots, 0); \\ & \phi''(\bar{q})(q_1 - \bar{q})(q_1 - \bar{q}) \quad \text{or} \quad \phi''(\bar{q})(q_1 - \bar{q})^2 \quad \text{for } \Phi''(0)(1, 0, \dots, 0)(1, 0, \dots, 0); \\ & \phi''(\bar{q})(q_1 - \bar{q})(q_2 - \bar{q}) \quad \text{for } \Phi''(0)(1, 0, \dots, 0)(0, 1, \dots, 0), \end{aligned}$$

where the “Gâteaux derivatives” $\phi'(\bar{q})$ and $\phi''(\bar{q})$ are linear respectively bilinear operators restricted to $Q - \bar{q}$.

THEOREM A. *Let $Y \subset Q - \bar{q}$ be such that*

(a) *$y \in Y$ implies*

$$\phi'_0(\bar{q})y \leq 0, \quad \phi'_1(\bar{q})y = 0, \quad \phi'_2(\bar{q})y \in C - \phi_2(\bar{q})$$

and

(b) *$y_1, y_2 \in Y, y_1 \neq y_2$ implies*

$$\phi''_0(\bar{q})y_1y_2 \leq 0, \quad \phi''_1(\bar{q})y_1y_2 = 0, \quad \phi''_2(\bar{q})y_1y_2 \in C - \phi_2(\bar{q}).$$

Then there exists $l = (l_0, l_1, l_2) \in [0, \infty) \times \mathbb{R}^m \times \mathcal{X}^*$ such that

$$(c) \quad l \neq 0, \quad l\phi'(\bar{q})h \geq 0 \quad \forall h \in Q - \bar{q}, \quad l_2\phi_2(\bar{q}) = \max_{c \in C} l_2c$$

and

$$(d) \quad l\phi''(\bar{q})y^2 \geq 0 \quad \forall y \in Y.$$

Example. There are examples (see e.g. Ben-Tal's Example 2.1 [1, p. 150]) showing that Theorem A is not valid without assumption (b). In such examples there exists no Lagrange multiplier l satisfying conditions (c) and (d) of Theorem A with Y replaced by the set of all y satisfying assumption (a) alone. However, the following simple example shows that Theorem A, involving a set Y with possibly several elements, may yield useful information in some cases when other second order conditions fail.

We first observe that if the optimization problem involves no restriction of the form $\phi_2(q) \in C$ then Theorem A is applicable with $l = (l_0, l_1)$. This can be verified by adding the “restriction” $\phi_2(q) \triangleq 0 \in C \triangleq \mathbb{R}$ which, together with the last relation in (c), yields $l_2 = 0$.

Let

$$Q \triangleq \mathbb{R}^2, \quad q \triangleq (x_1, x_2), \quad \phi_0(q) = x_2, \quad \phi_1(q) = x_1x_2, \quad \bar{q} = (0, 0).$$

Then $l = (l_0, l_1)$ is a “first order” Lagrange multiplier for the problem if it satisfies condition (c), i.e. if

$$l \neq 0, \quad l_0 \geq 0, \quad l\phi'(\bar{q}) = (0, l_0) = (0, 0);$$

hence $l_0 = 0, l_1 \neq 0$. Now let $y = (a, b)$ be a “critical variation” i.e. let y satisfy condition (a). Then we have

$$l\phi''(\bar{q})y^2 = l_1\phi''_1(\bar{q})y^2 = 2l_1ab.$$

Thus, by choosing $l_1 = ab$ if $ab \neq 0$ and otherwise $l_1 = 1$, we have $l\phi''(\bar{q})y^2 \geq 0$. This shows that the “old” second order conditions (with Y a singlet) can be satisfied, leaving open the possibility that $(0, 0)$ yields a minimum.

Now we apply Theorem A with

$$Y = \{y_1, y_2\}, \quad y_1 = (-1, -1), \quad y_2 = (1, -1).$$

We verify that Y satisfies assumptions (a) and (b). If $(0, 0)$ yields a restricted minimum then, by Theorem A, there exists $l = (l_0, l_1) \neq 0$ satisfying (c) and (d). As we have seen above, (c) implies $l_0 = 0$, $l_1 \neq 0$ and therefore (d) implies

$$l\phi''(\bar{q})y_1^2 = 2l_1 \geq 0 \quad \text{and} \quad l\phi''(\bar{q})y_2^2 = -2l_1 \leq 0.$$

This contradicts $l_1 \neq 0$, thus showing that $(0, 0)$ does not yield a restricted minimum.

In proving Theorem A we shall use the following lemma which is a slight extension of the classical theorem on the separation of convex sets.

LEMMA [3, Lemma V.2.1, p. 299]. *Let W be a convex subset of $\mathbb{R} \times \mathbb{R}^m \times \mathcal{X}$, C' an open convex subset of \mathcal{X} , $0 \in W$, and $0 \in C'$. Then either there exists $l \triangleq (l_0, l_1, l_2) \in [0, \infty) \times \mathbb{R}^m \times \mathcal{X}^*$ such that $l \neq 0$,*

$$lw = l_0 w_0 + l_1 w_1 + l_2 w_2 \geq 0 \quad \forall w \triangleq (w_0, w_1, w_2) \in W$$

and

$$l_2 c \leq 0 \quad \forall c \in \overline{C'}$$

or there exist points $\xi^i \triangleq (\xi_0^i, \xi_1^i, \xi_2^i) \in W$ and numbers $\beta_i > 0$ ($i = 0, \dots, m$) such that $\sum_{i=0}^m \beta_i = 1$, $\xi_2^i \in C'$, the set $\{(\xi_0^0, \xi_1^0), \dots, (\xi_0^m, \xi_1^m)\}$ is linearly independent, $\xi_0^i < 0$ and $\sum_{i=0}^m \beta_i \xi_1^i = 0$.

Proof of Theorem A. Let C° denote the interior of C ,

$$C' \triangleq C^\circ - \phi_2(\bar{q})$$

and

$$W \triangleq \{\phi''(\bar{q}) \sum' \tau_y^2 y^2 + 2\phi'(\bar{q})h | y \in Y, \tau_y \in [0, 1], h \in Q - \bar{q}\},$$

where \sum' denotes finite sums in which different terms contain distinct elements y . Clearly, W is convex and $0 \in W$ and $0 \in \overline{C'}$. Thus, by the lemma above, either there exists $l \triangleq (l_0, l_1, l_2) \in [0, \infty) \times \mathbb{R}^m \times \mathcal{X}^*$ such that

$$l \neq 0, \quad lw \geq 0 \quad \forall w \in W, \quad l_2(c) \leq 0 \quad \forall c \in \overline{C'}$$

or there exist $\xi^i \triangleq (\xi_0^i, \xi_1^i, \xi_2^i) \in W$ and β_i ($i = 0, \dots, m$) such that the set $\{(\xi_0^i, \xi_1^i) | i = 0, \dots, m\}$ is linearly independent, $\xi_0^i < 0$, $\xi_2^i \in C'$, $\beta_i > 0$, $\sum_{i=0}^m \beta_i = 1$ and $\sum_{i=0}^m \beta_i \xi_1^i = 0$. If the first alternative holds then

$$l_2 \phi_2(\bar{q}) = \max_{c \in \overline{C'}} l_2 c$$

and

$$lw = l(\sum' \tau_y^2 \phi''(\bar{q})y^2 + 2\phi'(\bar{q})h) \geq 0 \quad \forall w \in W \quad \text{and} \quad h \in Q - \bar{q}.$$

Then relation (c) is obtained by setting $w = 2\phi'(\bar{q})h$ and relation (d) by setting $w = \phi''(\bar{q})y^2$.

Now assume, by way of contradiction, that the second alternative holds, and let

$$\xi^i = \phi''(\bar{q}) \sum_{j=1}^{k_i} \tau_{i,j}^2 y_{i,j}^2 + 2\phi'(\bar{q})h_i \quad (i = 0, \dots, m).$$

If we represent all the distinct $y_{i,j}$ as y_1, \dots, y_k and redefine $\tau_{i,j}$ appropriately, we can write that

$$\xi^i = \phi''(\bar{q}) \sum_{j=1}^k \tau_{i,j}^2 y_j^2 + 2\phi'(\bar{q})h_i.$$

We observe that for $|\theta|_1 \triangleq \sum_{j=0}^m |\theta_j|$ and for

$$\begin{aligned}\phi_{0,1} &\triangleq (\phi_0, \phi_1), \quad \alpha \in [0, 1], \quad \alpha \sum_{j=1}^k \left(\sum_{i=0}^m \tau_{i,j}^2 \right)^{1/2} + \alpha^2 \leq 1, \\ \theta &\triangleq (\theta_0, \dots, \theta_m), \quad \theta_i \geq 0, \quad |\theta|_1 = 1,\end{aligned}$$

we have

$$\tilde{q} \triangleq \bar{q} + \alpha \sum_{j=1}^k \left(\sum_{i=0}^m \theta_i \tau_{i,j}^2 \right)^{1/2} y_j + \alpha^2 \sum_{i=0}^m \theta_i h_i \in Q$$

and

$$\begin{aligned}\phi_{0,1}(\tilde{q}) &= \phi_{0,1}(\bar{q}) + \alpha \phi'_{0,1}(\bar{q}) \sum_{j=1}^k \left(\sum_{i=0}^m \theta_i \tau_{i,j}^2 \right)^{1/2} y_j \\ &\quad + \frac{1}{2} \alpha^2 \left[\phi''_{0,1}(\bar{q}) \left(\sum_{j=1}^k \left(\sum_{i=0}^m \theta_i \tau_{i,j}^2 \right)^{1/2} y_j \right)^2 + 2 \phi'_{0,1}(\bar{q}) \sum_{i=0}^m \theta_i h_i \right] + \psi_{0,1}(\alpha, \theta),\end{aligned}$$

where $\psi_{0,1}(\alpha, \theta) = o(\alpha^2)$ uniformly for all θ as $\alpha \rightarrow 0+$. (This last assertion can be verified by expanding $\phi_{0,1}(\tilde{q})$ as a function of (α, u, v) about $(0, 0, 0)$, where $u = (u_1, \dots, u_k)$, $v = (v_1, \dots, v_m)$, u_j is the coefficient of y_j in \tilde{q} and $v_i \triangleq \alpha \theta_i$). Since

$$\phi'_0(\bar{q}) y_j \leq 0, \quad \phi'_1(\bar{q}) y_j = 0, \quad \phi''_0(\bar{q}) y_i y_j \leq 0, \quad \phi''_1(\bar{q}) y_i y_j = 0 \quad \text{if } i \neq j,$$

we have

$$\begin{aligned}\tilde{a}_0 &\triangleq \alpha \phi'_0(\bar{q}) \sum_{j=1}^k \left(\sum_{i=0}^m \theta_i \tau_{i,j}^2 \right)^{1/2} y_j \\ &\quad + \frac{1}{2} \alpha^2 \phi''_0(\bar{q}) \sum_{p,r=1, p \neq r}^k \left(\sum_{i=0}^m \theta_i \tau_{i,p}^2 \right)^{1/2} \left(\sum_{i=0}^m \theta_i \tau_{i,r}^2 \right)^{1/2} y_p y_r \leq 0\end{aligned}$$

and

$$\begin{aligned}\phi_{0,1}(\tilde{q}) &= \phi_{0,1}(\bar{q}) + \frac{1}{2} \alpha^2 \left[\phi''_{0,1}(\bar{q}) \sum_{j=1}^k \left(\sum_{i=0}^m \theta_i \tau_{i,j}^2 \right) y_j^2 + 2 \phi'_{0,1}(\bar{q}) \sum_{i=0}^m \theta_i h_i \right] \\ &\quad + \psi_{0,1}(\alpha, \theta) + (\tilde{a}_0, 0) \\ (1) \quad &= \phi_{0,1}(\bar{q}) + \frac{1}{2} \alpha^2 \sum_{i=0}^m \theta_i \left[\phi''_{0,1}(\bar{q}) \sum_{j=1}^k \tau_{i,j}^2 y_j^2 + 2 \phi'_{0,1}(\bar{q}) h_i \right] \\ &\quad + \psi_{0,1}(\alpha, \theta) + (\tilde{a}_0, 0) \\ &= \phi_{0,1}(\bar{q}) + \frac{1}{2} \alpha^2 H \theta + \psi_{0,1}(\alpha, \theta) + (\tilde{a}_0, 0),\end{aligned}$$

where H is the (nonsingular) $(m+1) \times (m+1)$ -matrix with columns (ξ_0^i, ξ_1^i) ($i = 0, \dots, m$).

Let $p_{\max}(p_{\min})$ denote the maximum (minimum) of $\{p_0, \dots, p_m\}$. We observe that there exists $\bar{\alpha} > 0$ such that

$$(2) \quad |2H^{-1}\psi_{0,1}(\alpha, \theta)| \leq \frac{\alpha^2}{3} \beta_{\min} \quad \text{if } |\theta|_1 = 1 \text{ and } 0 \leq \alpha \leq \bar{\alpha}.$$

We shall choose $\bar{\alpha}$ sufficiently small so that

$$\bar{\alpha} \sum_{j=1}^k \left(\sum_{i=0}^m \tau_{i,j}^2 \right)^{1/2} + \bar{\alpha}^2 \leq 1.$$

Let

$$\gamma \triangleq \frac{2}{3}\bar{\alpha}, \quad X \triangleq \{x \triangleq (x_0, \dots, x_m) \in \mathbb{R}^{m+1} \mid |x_i - \gamma\beta_i| \leq \frac{1}{2}\gamma\beta_{\min} (i=0, \dots, m)\}.$$

The set X is compact and convex and

$$(3) \quad x \in X \text{ implies } 0 < \frac{1}{2}\gamma\beta_{\min} \leq x_i, \quad |x|_1 \leq \bar{\alpha}.$$

Furthermore, the function

$$x \rightarrow \gamma\beta - 2H^{-1}|x|_1^{-1}\psi_{0,1}(|x|_1, |x|_1^{-1}x) : X \rightarrow \mathbb{R}^{m+1}$$

is continuous and, by (2) and (3), this function maps X into itself. Therefore this function has a fixed point $\hat{x} = \hat{\alpha}\hat{\theta} \in X$, where $\hat{\alpha} = |\hat{x}|_1$.

We have

$$\frac{1}{2}\hat{\alpha}^2 H\hat{\theta} + \psi_{0,1}(\hat{\alpha}, \hat{\theta}) = \frac{1}{2}\hat{\alpha}\gamma H\beta$$

and therefore, if we define \hat{q} , $\hat{\alpha}_0$ the same way as \tilde{q} , $\tilde{\alpha}_0$ but with $\hat{\alpha}$, $\hat{\theta}$ replacing α , θ , we have, by (1),

$$\begin{aligned} \phi_{0,1}(\hat{q}) &= \phi_{0,1}(\tilde{q}) + \frac{1}{2}\hat{\alpha}^2 H\hat{\theta} + \psi_{0,1}(\hat{\alpha}, \hat{\theta}) + (\hat{\alpha}_0, 0) \\ &= \phi_{0,1}(\tilde{q}) + \frac{1}{2}\hat{\alpha}\gamma H\beta + (\hat{\alpha}_0, 0). \end{aligned}$$

Thus

$$\phi_0(\hat{q}) = \phi_0(\tilde{q}) + \frac{1}{2}\hat{\alpha}\gamma \sum_{i=0}^m \beta_i \xi_0^i + \hat{\alpha}_0 < \phi_0(\tilde{q}),$$

$$\phi_1(\hat{q}) = \phi_1(\tilde{q}) + \frac{1}{2}\hat{\alpha}\gamma \sum_{i=0}^m \beta_i \xi_1^i = 0.$$

Furthermore, for all α and θ with $\alpha \leq \bar{\alpha}$, an expansion similar to the one in (1) yields

$$\phi_2(\tilde{q}) = \phi_2(\tilde{q}) + \frac{1}{2}\alpha^2 \sum_{i=0}^m \theta_i \xi_2^i + \tilde{\alpha}_2 + \psi_2(\alpha, \theta),$$

where $\tilde{\alpha}_2$ is defined like $\tilde{\alpha}_0$ but with ϕ_2 replacing ϕ_0 and where $\psi_2(\alpha, \theta) = o(\alpha^2)$ uniformly for all θ as $\alpha \rightarrow 0+$. We may choose $\bar{\alpha}$ small enough so that $\bar{\alpha} \leq \frac{1}{2}$ and

$$\frac{1}{2}\alpha \xi_2^i + \alpha^{-1}\psi_2(\alpha, \theta) \in C' \quad \forall i=0, \dots, m, 0 < \alpha < \bar{\alpha}.$$

Since, by assumption, $\phi_2'(\tilde{q})y_i$ and $\phi_2''(\tilde{q})y_p y_r$, belong to $C - \phi_2(\tilde{q})$ if $p \neq r$, it follows that $\tilde{\alpha}_2 \in \frac{1}{2}[C - \phi_2(\tilde{q})]$ if $\bar{\alpha}$ is chosen sufficiently small and then

$$\phi_2(\hat{q}) \in C.$$

Thus

$$\phi_0(\hat{q}) < \phi_0(\tilde{q}), \quad \phi_1(\hat{q}) = 0, \quad \phi_2(\hat{q}) \in C,$$

contrary to the assumption that \tilde{q} minimizes $\phi_0(q)$ subject to $\phi_1(q) = 0$, $\phi_2(q) \in C$. Q.E.D.

REFERENCES

- [1] A. BEN-TAL, *Second-order and related extremality conditions in nonlinear programming*, J. Optim. Theory Appl., 31 (1980), pp. 143–165.
- [2] DENNIS S. BERNSTEIN, *A systematic approach to higher-order necessary conditions in optimization theory*, this Journal, 22 (1984), pp. 211–238.
- [3] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [4] ———, *A second-order condition that strengthens Pontryagin's maximum principle*, J. Differential Equations, 28 (1978), pp. 284–307.

DIFFERENTIABILITY OF RELATIONS AND DIFFERENTIAL STABILITY OF PERTURBED OPTIMIZATION PROBLEMS*

JEAN-PAUL PENOT†

Abstract. Several new concepts of differentiability of multifunctions are introduced. Special attention is paid to relations given by perturbed inequalities. Applications are given to the study of the value (or marginal) function of a perturbed nonlinear program.

Key words. correspondence, relation, multifunction differentiability, Dini derivatives, perturbation, sensitivity, stability, value function, marginal function, variation

Introduction. Several concepts of differentiability of relations have been proposed with various aims [8], [18], [22], [23], [34], [38], [42], [43], [59]. Our purpose here is to study the marginal function (or value function)

$$m(w) = \inf \{f(w, x) | x \in A(w)\}$$

of a perturbed optimization problem. We wish to relate the (generalized) derivative of m at a given point (taken to be 0 for simplicity) to the derivatives of f and A .

Even for mappings, a whole spectrum of definitions of differentiability can be given (see [6], [45], [46], [66], [71], for instance) due to the different kinds of approximations one can choose (and, secondarily, to the choice of topologies). Here we focus our attention to the specific difficulties which appear with multivaluedness.

The derivatives we introduce are strongly related to the subdifferential calculus introduced in [47], [48], [50], [51] (see also [9], [60]). In the same way that continuity has to be split into upper and lower semi-continuity when applied to multifunctions, we get here two Dini derivatives. We define F to be directionally differentiable at a if these two derivatives coincide.

We show that under standard assumptions convex relations are directionally differentiable. In § 3 we especially focus our attention on relations given by parametrized inequalities or equalities using regularity conditions of [54], [55], [62], [69].

In § 5 we derive some estimates on the Dini derivatives of the marginal function and some inclusions on its generalized subdifferential ([47]) somewhat in the spirit of [24]–[28]. Equality is shown to occur in these estimates in two important cases; one of them involves a kind of first order sufficient optimality condition while the other one uses another concept of directional derivative of a relation A called the variation of A developed in § 4.

The Dini derivatives of the marginal function m or more precisely the radial Dini derivative of m , have been estimated by a number of authors [5], [20], [24]–[27], [31], [36], [37], [39], [41]; the novelty here lies in the interpretation in terms of the subdifferential of m and in the connection we establish with the derivative of the relation which defines the constraints. The first point is also considered in [24], [28] and in the papers [27], [58], [63], [64], [65] we received during the preparation of this paper. However, as we have a special interest in the initial problem

$$(P_0) \quad \text{minimize } \{f(0, x) | x \in A(0)\}$$

* Received by the editors February 19, 1982, and in revised form December 6, 1982.

† Département de Mathématiques, Université de Pau, France.

the subdifferential of m we introduced in [47] seems to be more adapted than the strict subdifferential (or peridifferential or Clarke's subdifferential) of [15] or its variants.

After the results of this paper were completed, the references [2], [10], [44] were pointed out to us. The definitions used in [10] and [44] are different from ours, although [10, Definition 1] seems to be close to the concept of variation of a relation. While sections [2, §§ 10–21] deal with applications to dynamical systems and [2, §§ 1–5] are contained in [47], [48], [50], part of our Definition 3.1 and the obvious Propositions 3.9, 3.11 of the present paper are given in [2, §§ 3, 6 and 7]. Let us observe that § 8 and [2, §§ 8, 9] are devoted to nice results, attributed to I. Ekeland, dealing with Newton method and an inverse function theorem; these results enhance the versatility and the simplicity of Dini derivatives and show they can be applied to other problems.

1. Notation and preliminaries. In all the sequel W , X and Y are Banach spaces (in most cases normed spaces would be enough); their norms are denoted by $|\cdot|$. The ball with center a and radius r in X is denoted by $B(a, r)$.

The set of continuous linear mappings from X into Y is denoted by $L(X, Y)$. For two subsets A and B of X we set

$$d(x, A) = \inf \{d(x, a) | a \in A\} \quad \text{and} \quad e(A, B) = \sup \{d(a, B) | a \in A\}.$$

$e(A, B)$ is the *excess* of A over B ; $d(A, B) = \max(e(A, B), e(B, A))$ is the *Hausdorff distance* of A and B . Let \bar{A} be the closure of A , $\overset{\circ}{A}$ or $\text{int } A$ be its interior and $A^c = X \setminus A$.

We identify a multifunction (or relation) $F: X \rightrightarrows Y$ with its graph $F = \{(x, y) \in X \times Y | y \in F(x)\}$; this is of common use in the theory of monotone operators and is quite convenient here too.

The following definition was introduced in [56] where we showed its wide applicability. In particular it can be used to give a lower semi-continuity result on the marginal function m ([57]).

DEFINITION 1.1. The relation $F: X \rightrightarrows Y$ is *compact* at some point a in the closure of its domain D if each sequence $((x_n, y_n))$ in F has a converging subsequence as soon as (x_n) converges to a .

The following definition is the simplest and the most widely used concept of tangent cone. For other notions and their connections see for instance [53] and its references.

DEFINITION 1.2. The *tangent cone* $T_a A$ at a point $a \in \bar{A}$ to a subset A of X is $\limsup_{t \rightarrow 0_+} t^{-1}(A - a)$. In other words, $v \in T_a A$ if and only if there exists a sequence (t_n, a_n) in $(0, +\infty) \times A$ such that

$$v = \lim t_n^{-1}(a_n - a).$$

We can characterize $T_a A$ as the set of v such that

$$\liminf_{t \rightarrow 0_+} t^{-1} d(a + tv, A) = 0.$$

The set of v such that this limit inferior can be replaced by a limit is denoted by $T_a^i A$.

If A is convex $T_a^i A = T_a A$ is the closure of $T_a' A = [0, +\infty)(A - a)$. The set of *interiorly tangent vectors* $I_a A$ to A at a is given by

$$I_a A = (T_a A^c)^c, \quad \text{where } A^c = X \setminus A.$$

2. Dini derivatives of a relation. Most definitions of the derivative of a relation F involve the behavior of F around all the image $F(a)$ of the considered point. This

is not the case with the following concept, which depends only on the behavior of F around one point c of F .

DEFINITION 2.1. The *Dini upper* and *lower derivatives* of a relation $F: X \rightrightarrows Y$ at $c = (a, b) \in X \times Y$ in the direction $x \in X$ are given respectively by

$$\bar{D}F(c)(x) = \limsup_{(t,v) \rightarrow (0_+, x)} t^{-1}(F(a+tv) - b), \quad \underline{D}F(c)(x) = \liminf_{(t,b) \rightarrow (0_+, x)} t^{-1}(F(a+tv) - b).$$

The following characterizations will be useful:

$$\begin{aligned} y \in \bar{D}F(c)(x) &\Leftrightarrow (\liminf_{(t,v) \rightarrow (0_+, x)} d(y, t^{-1}(F(a+tv) - b)) = 0) \\ &\Leftrightarrow (\liminf_{(t,v) \rightarrow (0_+, x)} t^{-1}d(b + ty, F(a+tv)) = 0) \\ &\Leftrightarrow (\exists(t_n, x_n, y_n) \rightarrow (0_+, x, y): \forall n \in \mathbb{N}, c + t_n(x_n, y_n) \in F), \\ y \in \underline{D}F(c)(x) &\Leftrightarrow (\lim_{(t,v) \rightarrow (0_+, x)} d(y, t^{-1}(F(a+tv) - b)) = 0) \\ &\Leftrightarrow (\forall(t_n, x_n) \rightarrow (0_+, x) \exists(y_n) \rightarrow y, \exists m \in \mathbb{N}: \forall n \geq m, c + t_n(x_n, y_n) \in F). \end{aligned}$$

Obviously, $\underline{D}F(c)(x) \subset \bar{D}F(c)(x)$ and these sets are closed. Moreover, identifying the relations $\underline{D}F(c)$ and $\bar{D}F(c)$ with their graphs, the equivalences above show that

$$I_c F \subset \underline{D}F(c) \subset \bar{D}F(c) = T_c F.$$

It is possible to give several variants of the preceding definition. In particular, if the convergence of v towards x is taken in the discrete topology, we speak of *radial Dini derivatives* and we write $\bar{D}_r F(c)$ and $\underline{D}_r F(c)$ respectively. Other convergences could be used. It is also useful to restrict the convergence of (t, v) towards $(0, x)$ in imposing that $a + tv \in D$, the domain of F ; we then write $(t, x) \xrightarrow{D} (0_+, x)$. The corresponding Dini derivatives are $D^+ F(c) = \bar{D}F(c)$ and $D_+ F(c) \supset \underline{D}F(c)$. Obviously, $\underline{D}F(c)$ is the restriction to $I_a D$ of $D_+ F(c)$, while the domain of $D_+ F(c)$ is contained in $T_a D$.

DEFINITION 2.2. The relation F is said to be *semi-differentiable at c* if $\bar{D}F(c) = \underline{D}F(c)$. If $\bar{D}F(c) = D_+ F(c)$, F is said to be *semi-differentiable on its domain*. Then $\bar{D}F(c)$ is simply denoted by $DF(c)$.

The preceding definitions are motivated by the following examples.

Examples 2.3. a) Suppose $f: X \rightarrow Y$ is semi-differentiable at a in Hadamard's sense (i.e., $s(x) = \lim_{(t,v) \rightarrow (0_+, x)} t^{-1}(f(a+tv) - f(a))$ exists for each $x \in X$). Then the relation F given by the graph of $f(F(x) = \{f(x)\})$ is semi-differentiable at $c = (a, f(a))$ and $DF(c) = s$ (identified with its graph). Conversely, if $\underline{D}F(c)$ is the graph of a mapping s , then f is semi-differentiable at a .

If Y is finite dimensional, and if $\bar{D}F(c)$ is the graph of a mapping s , then f is semi-differentiable at a with $f'(a) = s$ (cf. [49]; for a slightly weaker statement, see [21]).

b) Let $f: x \rightarrow \bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ be a function on X , finite at $a \in X$. We introduced in [47], [48], [50], [51] the Dini derivatives of f at a in the direction x , given by

$$\bar{d}f(a, x) = \limsup_{(t,v) \rightarrow (0_+, x)} t^{-1}[f(a+tv) - f(a)], \quad \underline{d}f(a, x) = \liminf_{(t,v) \rightarrow (0_+, x)} t^{-1}[f(a+tv) - f(a)].$$

Let $E(f)$ and $H(f)$ be the epigraph and hypograph relations associated with f :

$$E(f) = \{(x, y) \in X \times \mathbb{R}: y \geq f(x)\}, \quad H(f) = \{(x, y) \in X \times \mathbb{R}: y \leq f(x)\}.$$

Then, with $c = (a, f(a))$ we have

$$\begin{aligned}\bar{D}E(f)(c) &= E(\underline{d}f(a, \cdot)), & \bar{D}H(f)(c) &= H(\bar{d}f(a, \cdot)), \\ \underline{D}E(f)(c) &= E(\bar{d}f(a, \cdot)), & \underline{D}H(f)(c) &= H(\underline{d}f(a, \cdot)), \\ D_+E(f)(c) &= E(d^+f(a, \cdot)), & D_+H(f)(c) &= H(d_+f(a, \cdot)),\end{aligned}$$

where $d^+f(a, x) = +\infty$ if $x \notin T_aD$, with $D = \{x: f(x) < +\infty\}$,

$$d^+f(a, x) = \limsup_{(t,v) \in \mathcal{Q}_{(0_+,x)}} t^{-1}[f(a+tv) - f(a)] \quad \text{if } x \in T_aD,$$

and $d_+f(a, x) = -d^+(-f)(a, x)$. These modified Dini derivatives coincide with those introduced by R. Janin [37] and have proved to be quite useful. Let us observe that the use of the epigraph and the hypograph relations enables one to consider the case in which f takes its values in an ordered vector space without using any notion of inferior or superior limit, although these notions do exist [48], [56].

c) If F is constant on a neighborhood W of c (i.e., $F \cap W = (X \times B) \cap W$ for some $B \subset Y$), then F is semi-differentiable at c and $DF(c) = X \times T_bB$.

The following results give important examples which occur in optimization theory.

PROPOSITION 2.4. *Suppose F is a convex relation (i.e., that F is a convex subset of $X \times Y$) or, more generally that F is starlike at c . Then F is radially semidifferentiable at c .*

Proof. If $y \in \bar{D}_rF(c)(x)$ there exists (t_n, y_n) in $(0, +\infty) \times Y$ with limit $(0, y)$ such that $c + t_n(x, y_n) \in F$ for each n . Then, if (s_n) converges to 0_+ , and if $k(n) = \sup \{k | s_n \leq t_k\}$ we have $\lim_n k(n) = +\infty$. Then

$$c + s_n(x, y_{k(n)}) = (1 - s_n t_{k(n)}^{-1})(c + t_{k(n)}(x, y_{k(n)})) + s_n t_{k(n)}^{-1} c \in F$$

and as $(y_{k(n)})$ converges to y we get $y \in \underline{D}_rF(c)$. \square

A stronger result holds under openness assumptions.

PROPOSITION 2.5. *Let F be a convex relation with domain D and let $c = (a, b) \in F$. If (i) F has a nonvoid interior or (ii) F is closed, D has a nonvoid interior and X and Y are Banach spaces, then for any $x \in \text{int } T_aD$*

$$\underline{D}F(c)(x) = \bar{D}F(c)(x).$$

In particular, if (i) or (ii) holds and $a \in \text{int } D$, then F is semi-differentiable at c .

Proof. Without loss of generality we suppose $c = 0$. As $\text{int } D$ is nonvoid under both assumptions, we have $\text{int } T_aD = (0, +\infty) \text{int } D$ and, by positive homogeneity, we may suppose $x \in \text{int } D$. Let $y \in F(x)$ be such that $p: F \rightarrow D$ given by $p((u, v)) = u$ is open at y : if $\text{int } F \neq \emptyset$, we take $y \in Y$ such that $(x, y) \in \text{int } F$, while under the other assumption, any $y \in F(x)$ is suitable as a result due to P. C. Duong-H. Tuy, S. Robinson, C. Ursescu (see also [54, Thm. 3.5]) shows. Let $\bar{y} \in \bar{D}F(c)(x)$ and let (t_n, x_n) with limit $(0, x)$ in $(0, +\infty) \times X$. We can find a sequence (u_n, v_n) in F such that $(x, \bar{y}) = \lim t_n^{-1}(u_n, v_n)$. Let (s_n) be a sequence in $(0, 1)$ such that $(s_n) \rightarrow 0$, $(s_n^{-1}|x_n - x|) \rightarrow 0$, $(s_n^{-1}|t_n^{-1}u_n - x|) \rightarrow 0$. Then $w_n = x + s_n^{-1}(x_n - x) + s_n^{-1}(1 - s_n)(x - t_n^{-1}u_n)$ has limit x and

$$x_n = (1 - s_n)t_n^{-1}u_n + s_n w_n.$$

As p is open at y , we can find a sequence (y_n) in Y with limit y such that $(w_n, y_n) \in F$ for each n . Then

$$\bar{y}_n = (1 - s_n)t_n^{-1}v_n + s_n y_n$$

converges to \bar{y} and $(t_n x_n, t_n \bar{y}_n) = (1 - s_n)(u_n, v_n) + s_n(t_n w_n, t_n y_n) \in F$ by convexity, so that $\bar{y} \in \underline{D}F(c)(x)$. \square

COROLLARY 2.6. *If $f: X \rightarrow \mathbb{R}^* = \mathbb{R} \cup \{+\infty\}$ is a convex function which is finite and continuous at some point, then for each a in $D = \text{dom } f$ and each $x \in I_a D$ (in particular for each $x \in X$ if $a \in \text{int } D$)*

$$\underline{d}f(a, x) = \inf_{t \rightarrow 0} \frac{1}{t} [f(a + tx) - f(a)] = \bar{d}f(a, x).$$

This follows from Proposition 2.5 and Example 2.3b by introducing F given by $F(x) = f(x) + \mathbb{R}_+$ for $x \in D$.

DEFINITION 2.7. If F is a relation from X into Y , if $c \in F$, then

$$\bar{\partial}F(c) = \{u \in L(X, Y) | \forall x \in X \ u(x) \in \bar{D}F(c)(x)\},$$

$$\partial F(c) = \{u \in L(X, Y) | \forall x \in X \ u(x) \in \underline{D}F(c)(x)\}.$$

If F is a differentiable mapping (in Hadamard's sense) we find for $\bar{\partial}F(c)$ and $\partial F(c)$ the usual derivative at a , with $c = (a, F(a))$. If F is the hypograph relation associated with $f: X \rightarrow \mathbb{R}^*$ as in Example 2.3b, for $c = (a, f(a))$ we find for $\partial F(c)$ and $\bar{\partial}F(c)$ the subdifferentials $\partial f(a)$ and $\bar{\partial}f(a)$ defined in [9], [47], [48], [50], [51]. If F is any relation from $X = \mathbb{R}$ into Y , then $\bar{\partial}F(c)$ is nonvoid if and only if $T_c F \setminus \{0\} \times Y$ is a nonpointed cone, or in other words, if and only if $\bar{D}F(c)(1) \cap (-\bar{D}F(c)(-1)) \neq \emptyset$; a similar result holds for $\partial F(c)$. If Z is any set and $f: X \times Z \rightarrow Y$ is a mapping, differentiable at $(a, z_0) \in X \times Z$, with respect to x , then for F given by $F(x) = f(x, Z)$ and $c = (a, f(a, z_0))$ we have $D_1 f(a, z_0) \in \partial F(c)$.

The comparison of our definitions with known concepts of differentiability of relations will be given elsewhere [23]. Let us point out an easy connection with H. Methlouthi's definition [43] of a lower derivative of F at $c \in F$ (or rather an equivalent definition): $u \in L(X, Y)$ is a lower derivative of $F: X \rightrightarrows Y$ at $c = (a, b) \in F$ if there exists a neighborhood U of a in X and a selection $f: U \rightarrow Y$ of $F|U$, Fréchet differentiable at a , with $f(a) = b$, $f'(a) = u$. Even if Fréchet differentiability of f is replaced with Hadamard differentiability, this condition is more stringent than the conditions of the following proposition, whose proof is an immediate consequence of the equivalences following Definition 2.1.

PROPOSITION 2.8. *For $u \in L(X, Y)$ the following conditions are equivalent:*

- 1) $u \in \partial F(c)$;
- 2) *for each $x \in X$ $\lim_{(t,v) \rightarrow (0,+,x)} t^{-1} d(b + u(tv), F(a + tv)) = 0$.*

Although these conditions are stringent, H. Methlouthi's definition has proved to be quite useful (see [43] and [7]).

Demanding that the Dini derivatives of F be nonvoid is a much weaker requirement. For instance, the following lemma shows that Lipschitzian relations with finite dimensional image space have nonvoid Dini upper derivatives. The relation $F: X \rightrightarrows Y$ is said to be *B-tangentially compact* at $c = (a, b) \in F$ in the direction $x \in X$ if for any sequence (t_n, x_n) with limit $(0, x)$ in $(0, +\infty) \times X$ and any bounded sequence (y_n) in Y with $c + t_n(x_n, y_n) \in F$ for each n , the set of cluster points of (y_n) is nonvoid. This is obviously the case if Y is finite dimensional or if F is a finite dimensional submanifold or convex subset of $X \times Y$, or if F takes its values in such a subset of Y .

LEMMA 2.9. *Let $F: X \rightrightarrows Y$ be B-tangentially compact at $c = (a, b) \in F$ in the direction $x \in X$ and such that $\limsup_{t \rightarrow 0, +} t^{-1} d(b, F(a + tx)) < +\infty$. Then $\bar{D}F(c)(x) \neq \emptyset$.*

Proof. By assumption, there exist $k > 0$ and $\varepsilon > 0$ with $d(b, F(a + tx)) < kt$ for each $t \in (0, \varepsilon)$. Hence we can choose $z_t \in F(a + tx)$ such that $|b - z_t| \leq kt$. As $y_t = t^{-1}(z_t - b)$ is bounded, we can find a sequence (t_n) with limit 0 in $(0, \varepsilon)$ such that (y_{t_n}) has a limit y in Y . Then $y \in \bar{D}_r F(c)(x) \subset \bar{D}F(c)(x)$. \square

Under some further requirements we get that F is semi-differentiable at c .

LEMMA 2.10. *Let $F: X \rightrightarrows Y$ be B -tangentially compact at $c = (a, b)$ in the direction x and l.s.c. at (a, b) on its domain D . If $\bar{D}F(c)(x) \subset \{y\}$ and if $\bar{D}F(c)(0) = \{0\}$ then $D_+F(c)(x) = \bar{D}F(c)$, hence $\underline{D}F(c)(x) = \bar{D}F(c)(x)$ if $x \in I_c D$.*

Proof. Let $(t_n, x_n) \xrightarrow{D} (0, x)$. By assumption there exists a sequence $z_n \in F(a + t_n x_n)$ with limit b . Let $r_n = |z_n - b|$. If $(t_n^{-1} r_n)$ is not bounded, we can find an infinite subset K of \mathbb{N} with $r_k > 0$, $\lim_{k \in K} t_k r_k^{-1} = 0$, $\lim_{k \in K} t_k r_k^{-1} x_k = 0$. As F is B -tangentially compact at c , and as $(r_k^{-1}(z_k - b))$ is bounded, and $b + r_k r_k^{-1}(z_k - b) \in F(a + r_k t_k r_k^{-1} x_k)$ we have $(r_h^{-1}(z_h - b))_{h \in H} \rightarrow v$ for some infinite subset H of K and some $v \in Y$ with $|v| = 1$. Then $v \in \bar{D}F(c)(0)$ so that $v = 0$, a contradiction with $|v| = 1$.

As $(t_n^{-1} r_n)$ and $(r_n^{-1}(z_n - b))$ are bounded, for any infinite subset I of \mathbb{N} we can find an infinite subset J of I with $\lim_{j \in J} t_j^{-1} r_j = q$, $\lim_{j \in J} r_j^{-1}(z_j - b) = v$ for some $q \in [0, +\infty)$ and some $v \in Y$ with $|v| = 1$. Then $(t_j^{-1}(z_j - b)) \rightarrow qv$, so that $qv \in \bar{D}F(c)(x) = \{y\}$. This shows that the whole sequence $(t_n^{-1}(z_n - b))$ converges to y , thus $y \in D_+F(c)(x)$. \square

It is not our purpose here to develop a complete list of differential calculus rules for relations. Most of these rules are as trivial as the following ones in which $f^{-1} = (b, a)$ if $f = (a, b)$, $F^{-1} = \{f^{-1} | f \in F\}$, $F^c(x) = F(x)^c = (X \times Y \setminus F)(x)$:

$$\bar{D}F^{-1}(f^{-1}) = (\bar{D}F(f))^{-1},$$

$$\bar{D}F^c(z) \supset (\underline{D}F(z))^c, \quad \underline{D}F^c(z) \supset (\bar{D}F(z))^c,$$

$$\underline{D}F(z) \subset \bigcap_{i \in I} \underline{D}F_i(z), \quad \bar{D}F(z) \subset \bigcap_{i \in I} \bar{D}F_i(z) \quad \text{if } F = \bigcap_{i \in I} F_i,$$

$$\underline{D}F(z) \supset \bigcup_{i \in I} \underline{D}F_i(z), \quad \bar{D}F(z) \supset \bigcup_{i \in I} \bar{D}F_i(z) \quad \text{if } F = \bigcup_{i \in I} F_i$$

with equality in this last inclusion when I is finite.

Moreover, if $Y = Y_1 \times Y_2$ and $F = (F_1, F_2)$ is given by $F(x) = F_1(x) \times F_2(x)$ with $\text{dom } F_1 = \text{dom } F_2 = \text{dom } F$, for $c = (a, b_1, b_2) \in F$, $c_1 = (a, b_1) \in F_1$, $c_2 = (a, b_2) \in F$ we have

$$\underline{D}F(c) = (\underline{D}F_1(c_1), \underline{D}F_2(c_2)), \quad (\bar{D}F_1(c_1), \underline{D}F_2(c_2)) \subset \bar{D}F(c) \subset (\bar{D}F_1(c_1), \bar{D}F_2(c_2)).$$

The proofs of these assertions (and analogous ones with D_+F instead of $\underline{D}F$) are similar to the proof of Lemma 2.11.

LEMMA 2.11. *Let $F: X \rightrightarrows Y$, $G: Y \rightrightarrows Z$, $f = (a, b) \in F$, $g = (b, c) \in G$, $h = (a, c) \in H = G \circ F$.*

1) *If $x \in X$ is such that $D_+F(f)(x) \cap \text{dom } D_+G(g) \subset I_b \text{ dom } G$ then*

$$D_+G(g)(D_+F(f)(x)) \subset D_+H(h)(x).$$

2) *If $x \in X$ is such that $\bar{D}F(f)(x) \cap \text{dom } D_+G(g) \subset I_b \text{ dom } G$ then*

$$D_+G(g)(\bar{D}F(f)(x)) \subset \bar{D}H(h)(x).$$

In particular, if $\text{dom } G$ is a neighborhood of b we have

$$D_+G(g) \circ D_+F(f) \subset D_+H(h), \quad D_+G(g) \circ \bar{D}F(f) \subset \bar{D}H(h).$$

Proof. 1) Let $y \in D_+F(f)(x) \cap \text{dom } D_+G(g)$ and let $z \in D_+G(g)(y)$. Let $(t_n, x_n) \xrightarrow{D} (0_+, x)$, where $D = \text{dom } H$. Then $a + t_n x_n \in F^{-1}(\text{dom } G)$ for each $n \in \mathbb{N}$ and we can find $(y_n) \subset Y$ with $\lim (y_n) = y$, $b + t_n y_n \in F(a + t_n x_n)$ for each $n \in \mathbb{N}$. As $y \in$

$I_b \text{ dom } G$ we have $b + t_n y_n \in \text{dom } G$ for n large enough, hence there exists a sequence (z_n) in Z with limit z such that $c + t_n z_n \in G(b + t_n y_n)$ for n large enough. Thus $c + t_n z_n \in G(F(a + t_n x_n)) = H(a + t_n x_n)$ for n large enough and $z \in D_+ H(h)(x)$.

2) The proof is similar, starting from a particular sequence (t_n, x_n, y_n) in $(0, +\infty) \times X \times Y$ with limit $(0, x, y)$ and such that $f + t_n(x_n, y_n) \in F$ for each n , the assumption $y \in \bar{D}F(f)(x) \cap \text{dom } D_+ G(g) \subset I_b \text{ dom } G$ yielding $b + t_n y_n \in \text{dom } G$ for n large enough. \square

Example 2.12. The inclusion $\bar{D}G(g) \circ D_+ F(f) \subset \bar{D}H(h)$ is not true, even for single-valued relations with $X = Y = Z = \mathbb{R}$ and $\text{dom } G = \mathbb{R}$. To see that, let S be a subset of $[0, +\infty)$ such that 0 is an accumulation point of S and $[0, +\infty) \setminus S$, let $\text{dom } F = S$, with $F(x) = \{x\}$ for $x \in S$, let $G(y) = \{\sqrt{y}\}$, if $y \in S$, $G(y) = \{y\}$ if $y \in \mathbb{R} \setminus S$, so that $H(x) = \{\sqrt{x}\}$, for $x \in S$, $H(x) = \emptyset$ if $x \notin S$. Then $\bar{D}H(0)(1) = \emptyset$, but $D_+ F(0)(1) = \{1\}$, $\bar{D}G(0)(1) = \{1\}$.

PROPOSITION 2.13. *For $F: X \rightrightarrows Y$, $G: Y \rightrightarrows Z$ and $H = G \circ F$ as above we have*

$$\underline{D}G(g) \circ \underline{D}F(f) \subset \underline{D}H(h), \quad \underline{D}G(g) \circ \bar{D}F(f) \subset \bar{D}H(h).$$

Proof. This is a consequence of Lemma 2.11 and of the equality $\text{dom } \underline{D}G(g) = \text{dom } D_+ G(g) \cap I_b \text{ dom } G$. A direct proof is also quite simple. \square

PROPOSITION 2.14. *Let $F_1: X \rightrightarrows Y$, $F_2: X \rightrightarrows Y$ be two relations with domain D and let $F_3 = F_1 + F_2$ be given by $F_3(x) = \{y_1 + y_2 | y_1 \in F_1(x), y_2 \in F_2(x)\}$. Then, for $(a, b_1) \in F_1$, $(a, b_2) \in F_2$ and $b = b_1 + b_2$ we have, for any $x \in X$,*

$$\underline{D}F_1(a, b_1)(x) + \underline{D}F_2(a, b_2)(x) \subset \underline{D}F_3(a, b)(x),$$

$$\bar{D}F_1(a, b_1)(x) + \underline{D}F_2(a, b_2)(x) \subset \bar{D}F_3(a, b)(x).$$

Proof. A direct proof is easy. Let us observe that the first inclusion is also a consequence of the relation $\underline{D}(F_1, F_2)(a, b_1, b_2)(x) = \underline{D}F_1(a, b_1)(x) \times \underline{D}F_2(a, b_2)(x)$ and of Proposition 2.12 defining $G: Y \times Y \rightrightarrows Y$ by $G(y_1, y_2) = y_1 + y_2$ and $F: X \rightrightarrows Y \times Y$ by $F = (F_1, F_2)$. Both inclusions can also be deduced from Proposition 2.12 by considering the relations $F: x \rightrightarrows \{x\} \times F_1(x)$ and $G: (x, y) \rightrightarrows F_2(x) + y$. \square

3. Relations given by inequalities. In this section we consider the important case of a relation given by implicit and explicit constraints, for instance inequalities. This is of fundamental importance for mathematical programming. More precisely, we suppose W, X, Y are Banach spaces, $f: W \times X \rightarrow Y$ is a mapping, $G: W \rightrightarrows X$ and $H: W \rightrightarrows Y$ are relations and we set $F = G \cap \hat{f}^{-1}(H)$ with $\hat{f}(w, x) = (w, f(w, x))$:

$$F(w) = \{x \in X | x \in G(w), f(w, x) \in H(w)\}.$$

Of special interest to us will be the case where G and H are constant relations with values B and C respectively. Then we set

$$A(w) = \{x \in X | x \in B, f(w, x) \in C\}.$$

Here B represents a basic constraint while $f(w, x) \in C$ represents an implicit constraint.

Following S. Robinson [62] and A. Ioffe [35] we define F to be *metrically regular* at $(w_0, x_0) \in F$ if there exists $k > 0$, neighborhoods W_0, X_0 of w_0, x_0 such that

$$(R_m) \quad d(x, F(w)) \leq kd(f(w, x), H(w))$$

for every $w \in W_0, x \in X_0 \cap G(w)$. Given (\dot{w}, \dot{x}) in $W \times X$, if there exists a neighborhood $\dot{W} \times \dot{X}$ of (\dot{w}, \dot{x}) and $\varepsilon > 0$ such that (R_m) holds for $(w, x) \in [(w_0, x_0) + (0, \varepsilon) \dot{W} \times \dot{X}] \cap G$, we call F *metrically regular* at (w_0, x_0) in the direction (\dot{w}, \dot{x}) . In the sequel we suppose $w_0 = 0$ for simplicity.

THEOREM 3.1. *Let F be given as above and let $a \in F(0)$. Suppose f is Hadamard semi-differentiable at $(0, a)$; let $c = f(0, a)$. Then:*

1) $\bar{D}F(0, a)(w) \subset \{x \in \bar{D}G(0, a)(w) \mid f'(0, a)(w, x) \in \bar{D}H(0, c)\}$.

2) *If G and H are semi-differentiable at $(0, a)$ and $(0, c)$ respectively, and if F is metrically regular at $(0, a)$ in any direction, then F is semi-differentiable at $(0, a)$ and*

$$DF(0, a) = DG(0, a) \cap \tilde{f}'(0, a)^{-1}(DH(0, c)).$$

3) *In particular, if the relation A above is metrically regular at $(0, a)$ and $T_a B = T_a^i B$, $T_c C = T_c^i C$ (in particular if B and C are convex), then A is semi-differentiable at $(0, a)$ and*

$$DA(0, a)(\dot{w}) = \{x \in T_a B \mid f'(0, a)(w, x) \in T_c C\}.$$

Proof. 1) As $F = G \cap \tilde{f}^{-1}(H)$, $T_{(0,a)} F \subset T_{(0,a)} G \cap \tilde{f}'(0, a)^{-1}(T_{(0,a)} H)$.

2) Let $(\dot{w}, \dot{x}) \in \bar{D}G(0, a) \cap \tilde{q}^{-1}(\bar{D}H(0, c))$, with $q = f'(0, a)$; we have to show that $(\dot{w}, \dot{x}) \in \bar{D}F(0, a)$. Let (t_n, w_n) be a sequence in $(0, +\infty) \times W$ with limit $(0, \dot{w})$. As G and H are semi-differentiable at $(0, a)$ and $(0, c)$ respectively, there exist sequences (x_n) in X and (y_n) in Y with limits \dot{x} and $q(\dot{w}, \dot{x})$ respectively such that for n large enough $a + t_n x_n \in G(t_n w_n)$, $c + t_n y_n \in H(t_n w_n)$. For n still larger we have $t_n \in (0, \varepsilon)$ and $(w_n, x_n) \in \dot{W} \times \dot{X}$, where $\varepsilon, \dot{W}, \dot{X}$ are given as above, so that, setting

$$z_n = f(t_n w_n, a + t_n x_n) - f(0, a) - t_n f'(0, a)(\dot{w}_n, \dot{x}_n)$$

we have $(t_n^{-1} z_n) \rightarrow 0$ and

$$\begin{aligned} t_n^{-1} d(a + t_n x_n, F(t_n w_n)) &\leq t_n^{-1} k d(c + t_n q(\dot{w}_n, \dot{x}_n) + z_n, H(t_n w_n)) \\ &\leq t_n^{-1} k d(c + t_n y_n, H(t_n w_n)) + k |q(\dot{w}_n, \dot{x}_n) - y_n| + t_n^{-1} |z_n| \rightarrow 0, \end{aligned}$$

since the first term is 0. Hence $(\dot{w}, \dot{x}) \in \bar{D}F(0, a)$.

3) If $T_a^i B = T_a B$, $T_c C = T_c^i C$ then the constant relations G and H with values B and C respectively are semi-differentiable at $(0, a)$ and $(0, c)$ respectively with derivatives $W \times T_a B$ and $W \times T_c C$ respectively. \square

Remark 3.2. The equality $\bar{D}F(0, a) = \bar{D}G(0, a) \cap \tilde{f}'(0, a)^{-1}(\bar{D}H(0, a))$ holds if f is of class C^1 around $(0, a)$, if G and H are closed convex and

$$\tilde{f}'(0, a)(T'_{(0,a)} G) - T'_{(0,c)} H = W \times Y,$$

from [55, Corollary 5.8]. This relation, which is weaker than relation (R') below when $G = W \times B$, $H = W \times C$, does not imply metrical regularity of F .

Now we give a criterion for F to be metrically regular at $(0, a)$ in any direction. We suppose f satisfies the following differentiability assumption for some positively homogeneous mapping s :

(H) For each $\dot{w} \in W$,

$$\lim_{\substack{(t, w) \rightarrow (0, \dot{w}) \\ x, x' \rightarrow a, x' \neq x}} |x - x'|^{-1} |f(tw, x) - f(tw, x') - (s(x - a) - s(x' - a))| = 0.$$

This assumption is satisfied if the derivative $D_2 f$ exists and is continuous at $(0, a)$.

PROPOSITION 3.3. *With the preceding notations, suppose Y is ordered by some closed convex cone P , suppose B and C are closed convex subsets with $C - P = C$ (for instance $C = -P$), suppose (H) holds with some continuous sublinear map s . If the following regularity assumption holds:*

$$(R') \quad s(T_a^i B) - T_c^i C = Y$$

then A is metrically regular at $(0, a)$ in any direction $(\dot{w}, \dot{x}) \in W \times X$.

Proof. When $C = -P$, $s \in L(X, Y)$ and D_2f exists and is continuous, this is [62, Thm. 1]. The general case relies on [55, Thm. 5.5 and Corollary 5.8]. Let $(\dot{w}, \dot{x}) \in W \times X$. We choose the map g of this theorem to be given by $g(x) = s(x - a) + f(0, a)$. Then assumption a) of this theorem follows from (R') which can be written

$$(0, +\infty)(s(B - a) + f(0, a) - C) = Y,$$

while assumption b) is a consequence of the inclusion $C - P \subset C$ and the sublinearity of s . Taking γ as given by [55, Thm. 5.5], we choose $\lambda \in (0, \gamma^{-1})$ and then $\varepsilon > 0$ and neighborhoods B_0 of a in B , \dot{W} of \dot{w} small enough, for having d , given by

$$d(x) = f(tw, x) - f(tw, a) - s(x - a)$$

λ -Lipschitzian on B_0 for each $(t, w) \in (0, \varepsilon) \times \dot{W}$. Then [55, Thm. 5.5 and Corollary 5.8] give some $k > 0$ and some neighborhood X_0 of a in X such that

$$d(x, A(w)) \leq kd(f(tw, x), C)$$

for $(t, w) \in (0, \varepsilon) \times \dot{W}$ and $x \in X_0 \cap B$, a slightly stronger result than previously stated (take $\varepsilon > 0$ and \dot{X} small enough for having $a + (0, \varepsilon)\dot{X} \subset X_0$). \square

The following result was stated in [12] without any other proof than a reference to [17, Thm. 2.2, p. 539]. We recast it in our formulation and present a simple direct proof which does not suppose C is polyhedral.

PROPOSITION 3.4. *Let $F = \{(w, x) \in G \mid f(w, x) \in H(w)\}$ be as above. Suppose X is finite dimensional and F is l.s.c. at $(0, a) \in F$. Suppose $\bar{D}G(0, a)$ and $\bar{D}H(0, c)$ are convex, with $c = f(0, a)$. If for some $w \in W$ the inclusions*

$$x \in \bar{D}G(0, a)(w), \quad f'(0, a)(w, x) \in \bar{D}H(0, c)(w)$$

have exactly one solution x then F is semi-differentiable at $(0, a)$ in the direction w and $DF(0, a)(w) = \{x\}$.

For $G = W \times B$, $H = W \times C$ the inclusions above reduce to

$$x \in T_a B, \quad f'(0, a)(w, x) \in T_c C.$$

Furthermore, if $B = X$, $Y = \mathbb{R}^p$ and $C = \mathbb{R}_+^p$ the inclusions are equivalent to the following system, in which $I = \{i = 1, \dots, p \mid f_i(0, a) = 0\}$:

$$f'_i(0, a)(w, x) \leq 0, \quad i \in I.$$

Proof. F is B -tangentially compact at $(0, a)$ since X is finite dimensional. Moreover if $\dot{x} \in \bar{D}F(0, a)(\dot{w})$ there exists a sequence (t_n, w_n, x_n) in $(0, +\infty) \times W \times X$ with limit $(0, \dot{w}, \dot{x})$ such that $f(t_n w_n, a + t_n x_n) \in H(t_n w_n)$, $a + t_n x_n \in G(t_n w_n)$. Thus $\dot{x} \in \bar{D}G(0, a)(\dot{w})$, $f'(0, a)(\dot{w}, \dot{x}) \in \bar{D}H(0, c)(\dot{w})$. In particular $\bar{D}F(0, a)(w) \subset \{x\}$. Moreover if $\dot{x} \in \bar{D}F(0, a)(0)$ we have $\dot{x} \in \bar{D}G(0, a)(0)$, $f'(0, a)(0, \dot{x}) \in \bar{D}H(0, c)(0)$ hence

$$x + \dot{x} \in \bar{D}G(0, a)(w), \quad f'(0, a)(w, x + \dot{x}) \in \bar{D}H(0, c)(w)$$

hence $x + \dot{x} = x$ and $\dot{x} = 0$. The assumptions of Lemma 2.10 are satisfied and $0 \in \text{int dom } F$, hence the result follows. \square

Now we consider the case of a relation A defined by equalities: $B = X$ and $C = \{0\}$ above. The following example shows that even when $f'(0, a)$ is surjective A may not be semi-differentiable at $(0, a)$.

Example 3.5. $X = \mathbb{R}$, $W = \mathbb{R}^2$, $f(w_1, w_2, x) = w_1 - x^2(1 - x^2)$, $a = 0$. As $f'(0, a)$ is surjective, we have $\bar{D}A(0, a) = \text{Ker } f'(0, a) = \{0\} \times \mathbb{R} \times \mathbb{R}$. However

$(0, 1, 1) \notin \bar{D}A(0, a)$: if (t_n) is any sequence with limit 0 in $(0, +\infty)$ and if $w_n = (-t_n^2, t_n)$, then $t_n w_n \in \text{dom } A$ for n large enough but there is no $x_n \in A(t_n w_n)$ with $(x_n) \rightarrow 0$.

Of course, when f is C^1 and $D_2f(0, a)(X) = Y$, A is semi-differentiable at $(0, a)$ by Theorem 3.1 and Proposition 3.3. Here is another instance:

PROPOSITION 3.6. *Let $A = f^{-1}(0)$ where $f: W \times X \rightarrow Y$ is of class C^1 around $(0, a)$. Under one of the following assumptions A is semi-differentiable at $(0, a)$ and $DA(0, a) = \text{Ker } f'(0, a)$.*

1) $\text{Ker } D_2f(w, x)$ and $\text{Im } D_2f(w, x)$ vary continuously with (w, x) in the sets of complemented subspaces of X and Y respectively.

2) $\dim X < +\infty$, $\dim Y < +\infty$ and the rank of $D_2f(w, x)$ is constant around $(0, a)$.

Proof. Both assumptions ensure that there exist open subsets U and V in $W \times X$ and $W \times Y$ respectively with $(0, a) \in U$, $0 \in V$ and diffeomorphisms $\varphi: U \rightarrow \varphi(U)$, $\psi: V \rightarrow \psi(V)$ onto open subsets of $W \times X$ and $W \times Y$ with $\varphi(0, a) = 0$, $\psi(0) = 0$, $\varphi'(0, a) = I_{W \times X}$, $\psi'(0) = I_{W \times Y}$, $\varphi(w, x) = (w, \varphi_X(w, x))$ for $(w, x) \in U$, $\psi(w, y) = (w, \psi_Y(w, y))$, $\psi_Y \circ f \circ \varphi^{-1}(w, x) = D_2f(0, a)x$. Let $r \in L(Y, X)$ be a right inverse of $s = D_2f(0, a)$. Then

$$p(w, x) = (w, x - r(f'(0, a)(w, x)))$$

defines a continuous projector of $W \times X$ onto $N = \text{Ker } f'(0, a)$. Let $(\dot{w}, \dot{x}) \in N \supset \bar{D}A(0, a)$ and let (t_n, w_n) be a sequence in $(0, +\infty) \times W$ with limit $(0, \dot{w})$. Then $t_n p(w_n, \dot{x}) \in N$ so that $\varphi^{-1}(t_n p(w_n, \dot{x})) \in A$. As $t_n^{-1}(\varphi^{-1}(t_n p(w_n, \dot{x})) - \varphi^{-1}(0, 0)) \rightarrow p(\dot{w}, \dot{x}) = (\dot{w}, \dot{x})$, we can write $\varphi^{-1}(t_n p(w_n, \dot{x})) = (t_n w_n, a + t_n x_n)$ for some (x_n) satisfying $\lim (x_n) = \dot{x}$. We have shown that $(\dot{w}, \dot{x}) \in \bar{D}A(0, a)$. \square

Finally, we extract from Janin [36], [37] the following deep result obtained for the case of equalities and inequalities.

THEOREM 3.7 (R. Janin [37]). *Let $g: \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}^p$ be of class C^1 around a and let*

$$A(w) = \{x \in \mathbb{R}^n \mid g_i(w, x) \leq 0, i \leq q, g_i(w, x) = 0, i > q, i \leq p\}$$

with $q \leq p$. If the following condition is satisfied for some $a \in X$ such that $a \in \limsup_{(t,v) \rightarrow (0, \cdot, w)} A(tv)$ then $D_+A(0, a)(w) = \bar{D}A(0, a)(w)$.

(J) *For each subset I of $I(a) = \{i = 1, \dots, p \mid g_i(0, a) = 0\}$ there exists a neighborhood of $(0, a)$ on which the rank of the family $\{D_2g_i \mid i \in I\}$ is constant.*

4. Variation of a relation. Let us observe that the contribution to $\bar{D}F(c)$ by those vectors $(0, y)$ which are tangent to the image $F(a)$ of F at a does not reflect properly the way F varies around a . The following definition seems to capture more adequately this variation as an application to perturbed mathematical programming problems will show in the next section.

DEFINITION 4.1. A subset V of Y is said to be a *variation of the relation $F: X \rightrightarrows Y$* at $c = (a, b)$ in the direction $x \in X$ if for each sequence $((t_n, x_n, z_n))$ in $(0, +\infty) \times X \times Y$ with limit $(0, x, b)$ and $z_n \in F(a + t_n x_n)$ for each $n \in \mathbb{N}$ there exists a sequence (b_n) of $F(a)$ with limit b such that $d(t_n^{-1}(z_n - b_n), V)$ converges to 0. If one can take $b_n = b$, V is said to be a *b-variation* of F at c in the direction x . If one can take (b_n) so that $(t_n^{-1}(z_n - b_n))$ is bounded V is said to be a *strong variation* of F at c in the direction x .

If V satisfies this requirement whenever (x_n) is the constant sequence with value x , then V is called a *radial variation* of F at c in the direction x . If for each $b \in F(a)$ the set V is a variation of F at (a, b) in the direction x , then V is simply called a variation of F at a in the direction x .

As Y itself is a variation of F at any $c \in F$ in any direction, it is natural to look for a variation V as small as possible. On the other hand, it is easy to see that the closure of any b -variation V of F at c in the direction x contains $\bar{D}F(c)(x)$.

The following examples show it is useful to allow the sequence (b_n) to be a nonconstant sequence.

Examples 4.2. a) Suppose $f: X \rightarrow Y$ has a derivative at a in the direction x : $y := \lim_{(t,v) \rightarrow (0,+,x)} t^{-1}(f(a+tv) - f(a))$ exists. Then for any nonvoid subset C of Y and any $b \in f(a) + C$, the relation $F: X \rightrightarrows Y$ given by $F(x) = \{f(x)\} + C$ has $V = \{y\}$ as a variation at (a, b) in the direction x . In fact, one can take $b_n = f(a) + z_n - f(a + t_n x_n)$ in Definition 4.1.

b) In particular, if Y is an ordered topological vector space, with positive cone P , the epigraph relation $F = E_f$ obtained by taking $C = P$ in the preceding example has the derivative of f at a in the direction x as a variation at $(a, f(a))$ in this direction when this derivative exists. When this is not the case but f is continuous and for some $b \in F(a)$, $\bar{D}F(a, b)(x)$ is nonvoid, then for any $y \in \bar{D}F(a, b)(x)$, $V = \{y\} - P$ is a variation of F at (a, b) in the direction x . In fact, if (t_n, x_n, z_n) are as in Definition 4.1, and if $y = \lim t_n^{-1}[f(a + t_n x_n) + q_n - b]$ for some $q_n \in P$, we take $b_n = b + z_n - f(a + t_n x_n)$ so that $t_n^{-1}(z_n - b_n) - (y - t_n^{-1}q_n) \rightarrow 0$ and $y - t_n^{-1}q_n \in V$.

b') When Y is a topological vector lattice whose positive cone P has a nonvoid interior, and when $F(x) = [f(x), g(x)]$ with $g(a) - f(a) \in \bar{P}$, a variation V of F at (a, b) in the direction x can be given when f and g have a derivative at a in the direction x : for $b = f(a)$, take $V = \{f'(a)x\}$, for $b = g(a)$ take $V = \{g'(a)x\}$ and for $b \in \text{int } F(a)$ take $V = \{0\}$, as follows from the next example.

c) If F is constant on a neighborhood of c or if b belongs to $\text{int } F(a)$ then $V = \{0\}$ is a variation of F at c in any direction.

d) If F is pseudo-convex at c , i.e., if $F \subset c + T_c F$, in particular if F is starlike at c or convex, then for each $x \in X$, $\bar{D}F(c)(x)$ is a radial b -variation of F at c in the direction x . In fact, if $(t_n, z_n) \in (0, +\infty) \times Y$ has limit $(0, b)$ and is such that $z_n \in F(a + t_n x)$ for each n , then

$$(x, t_n^{-1}(z_n - b)) \in t_n^{-1}(F - c) \subset T_c F$$

so that $d(t_n^{-1}(z_n - b), \bar{D}F(c)(x)) = 0$.

e) If F is convex and if the assumptions of Proposition 2.5 are satisfied, then for each $x \in \text{int } T_a D$, with $D = \text{dom } F$, $V = \bar{D}_r F(c)(x)$ is a b -variation of F at c in the direction x provided it is bounded.

To prove this assertion, we may suppose $c = 0$, and, by positive homogeneity, we may suppose $x \in \text{int } D$. Let $y \in Y$ be such that $(x, y) \in F$ or $\text{int } F$. Then, using the openness of p as in the proof of Proposition 2.5, we can find $r, s > 0$ such that $B(x, r) \subset F^{-1}(B(y, s)) \subset D$. Let (t_n, x_n, z_n) be as in Definition 4.1 and let $r_n = |x_n - x|$. Then $\lim r_n = 0$, $\lim s_n = 0$, where $s_n := r_n(r_n + r)^{-1}$. Moreover we have $x = s_n x'_n + (1 - s_n)x_n$ where x'_n is given by $x'_n := s_n^{-1}(x - x_n) + x_n$, so that $|x'_n - x| = (s_n^{-1} - 1)|x_n - x| = r$. Let $y'_n \in B(y, s)$ with $(x'_n, y'_n) \in F$. Then, $s_n(t_n x'_n, t_n y'_n) + (1 - s_n)(t_n x_n, z_n) \in F$, so that $v_n := s_n y'_n + (1 - s_n)t_n^{-1}z_n \in t_n^{-1}F(t_n x)$ hence $(x, v_n) \in T'_c F$ and $v_n \in \bar{D}_r F(c)(x)$. As $\lim t_n^{-1}(z_n - b) - v_n = \lim (-s_n y'_n) = 0$, since (v_n) , hence $(t_n^{-1}z_n)$ is bounded $\bar{D}_r F(c)(x)$ is a b -variation of F at c in the direction x .

The following result gives a characterization of bounded variations and facilitates the comparison with other concepts.

PROPOSITION 4.3. *Each one of the following properties is necessary, and, when V is bounded is also sufficient, for a subset V of Y to be a variation of F at (a, b) in the direction x :*

- 1) $\lim_{(r,t,x') \rightarrow (0,+,0,+,x)} t^{-1}e(F(a + tx') \cap B(b, r), (1 - t)F(a) + t(V + b)) = 0$.
- 2) $\lim_{(r,t,x') \rightarrow (0,+,0,+,x)} t^{-1}e(F(a + tx') \cap B(b, r), F(a) + tV) = 0$.

Proof. Suppose V is a variation of F at (a, b) in the direction x and 1) does not hold: there exists some $\alpha > 0$, a sequence (r_n, t_n, x_n) with limit $(0, 0, x)$ and $z_n \in F(a + t_n x_n) \cap B(b, r_n)$ with

$$t_n^{-1} d(z_n, (1 - t_n)F(a) + t_n(V + b)) \geq \alpha.$$

Taking (b_n) as in Definition 4.1, we obtain

$$t_n^{-1} d(z_n - b_n, t_n V) \geq t_n^{-1} d(z_n, (1 - t_n)b_n + t_n V + t_n b_n) \geq \alpha - |b_n - b|,$$

a contradiction as $\lim b_n = b$.

Conversely, if 1) holds true, for any sequence (t_n, x_n, z_n) as in Definition 4.1 we have

$$\lim_n t_n^{-1} d(z_n, (1 - t_n)F(a) + t_n(V + b)) = 0$$

so that there exists $(b_n, v_n) \in F(a) + V$ with

$$\lim t_n^{-1} |(1 - t_n)b_n + t_n(v_n + b) - z_n| = 0.$$

As V is bounded $t_n(v_n + b) \rightarrow 0$ and we have $\lim b_n = \lim (1 - t_n)b_n = \lim z_n = b$. Moreover $\lim t_n^{-1} d(z_n - b_n, t_n V) \leq \lim t_n^{-1} |b_n - z_n + t_n v_n| = 0$ and V is a variation of F at (a, b) in the direction x . The proof of the other assertion is similar. \square

COROLLARY 4.4. *If $A: X \rightrightarrows Y$ is an upper derivative of F at a in the direction $x \in X$, [23], i.e., if $\text{dom } A \supset a + (0, +\infty)x$ and A satisfies*

$$\lim_{(t, x') \rightarrow (0, +, x)} t^{-1} e(F(a + tx'), (1 - t)A(a) + tA(a + x)) = 0,$$

if $A(a) \subset F(a)$ and if $A(a + x)$ is bounded, then $A(a + x) - b$ is a variation of F at (a, b) in the direction x .

COROLLARY 4.5. *If $L: X \rightrightarrows Y$ is an upper differential of F at a in the (generalized) sense of Lasota and Strauss [38], i.e., if L is u.s.c., homogeneous and there exists $r > 0$ such that*

$$F(a + x) \subset F(a) + L(x) \quad \text{if } |x| < r$$

then $L(x) - b$ is a variation of F at (a, b) in the direction x when $L(x)$ is bounded.

Proof. As L is u.s.c. in the metric sense of [18], for each $\varepsilon > 0$ there exists $\delta > 0$ with $L(x') \subset L(x) + B(0, \varepsilon)$ when $x' \in B(x, \delta)$ thus, for $0 < t < r(|x| + \varepsilon)^{-1}$, $x' \in B(x, \delta)$ we have

$$e(F(a + tx'), F(a) + tL(x)) \leq t\varepsilon$$

and Proposition 4.3 condition 2 holds. \square

The proof of the following result is similar.

COROLLARY 4.6. *If $F: X \rightrightarrows Y$ is differentiable at $a \in X$ in the De Blasi's sense [18] with differential B , i.e., if $\lim_{|h| \rightarrow 0} |h|^{-1} d(F(a + h), F(a) + B(h)) = 0$ for some u.s.c., homogeneous, bounded closed convex valued relation B then $B(x)$ is a variation of F at (a, b) in the direction x .*

Finally we present a connection with the upper Dini derivative.

PROPOSITION 4.7. *Suppose Y is finite dimensional and F has a bounded rate of expansion at a in the direction x in the following sense: there exist $\varepsilon > 0$, $k \geq 0$ such that $e(F(a + tx'), F(a)) \leq kt|x'|$ for $t \in (0, \varepsilon)$, $x' \in B(x, \varepsilon)$. Then for each $b \in F(a)$, V given as follows is a strong variation of F at (a, b) in the direction x :*

$$V = \limsup_{\substack{(t, x', b') \rightarrow (0, +, x, b) \\ b' \in F(a)}} t^{-1} (F(a + tx') - b').$$

In particular, if $F(a) = \{b\}$, $\bar{D}F(a, b)(x)$ is a variation of F at (a, b) in the direction x .

Proof. For any sequence (t_n, x_n, z_n) as in Definition 4.1 there exists $b_n \in F(a)$ such that $|z_n - b_n| \leq kt_n|x_n| + t_n^2$. The conclusion follows by a compactness argument, each subsequence of $(t_n^{-1}(z_n - b_n))$ having a cluster point. \square

Now we consider the case of multifunctions given by implicit constraints as in § 3 (we change f to g for the sake of clarity in the next section):

$$A(w) = \{x \in B | g(w, x) \in C\}.$$

Given $a \in A(0)$ and $w \in W$ we introduce the following condition, in which $c = g(0, a)$:

(C) There exist $\varepsilon > 0$, $k \geq 0$, a closed cone K in Y and neighborhoods W_0 of w , X_0 of a such that for each $(w', x') \in (0, \varepsilon)W_0 \times X_0$ with $x' \in A(w')$ there exists $a' \in A(0)$ satisfying $|x' - a'| \leq k|w'|$,

$$g(w', x') - g(0, a') \in K.$$

Typically K will be $T_c C$ or $\{0\}$.

THEOREM 4.8. *Let A be given as above. Suppose X is finite dimensional, g is strictly differentiable at $(0, a)$, condition (C) is satisfied for $w \in W$ with some closed cone K , and $T_a B = \limsup_{(t,x) \rightarrow (0+,a)} t^{-1}(B - x)$. Then $V = \{x \in T_a B | g'(0, a)(w, x) \in K\}$ is a strong variation of A at $(0, a)$ in the direction w .*

Proof. Suppose this is not true: there exists a sequence (t_n, w_n, x_n) in $(0, +\infty) \times W \times X$ with limit $(0, w, a)$ and $x_n \in A(t_n w_n)$ such that for any sequence (a_n) in $A(0)$ with limit a we have $l = \limsup d(t_n^{-1}(x_n - a_n), V) > 0$.

Let $n_0 \in \mathbb{N}$ be such that $t_n \in (0, \varepsilon)$, $x_n \in X_0$, $w_n \in W_0$ for $n \geq n_0$ and let $a_n \in A(0)$ be such that $|x_n - a_n| \leq k|t_n w_n|$, $g(t_n w_n, x_n) - g(0, a_n) \in K$. Let M be an infinite subset of \mathbb{N} such that $m \geq n_0$ for $m \in M$ and $l = \limsup_n d(t_n^{-1}(x_n - a_n), V) = \lim_m d(t_m^{-1}(x_m - a_m), V)$. Let N be an infinite subset of M such that the bounded sequence $(t_n^{-1}(x_n - a_n))_{n \in N}$ converges to some $v \in X$ with $|v| \leq k|w|$. As g is strictly differentiable at $(0, a)$ and K is a closed cone we have

$$g'(0, a)(w, v) = \lim_{n \in N} t_n^{-1}(g(t_n w_n, x_n) - g(0, a_n)) \in K.$$

Hence $v \in V$, a contradiction with $d(v, V) = l > 0$. \square

PROPOSITION 4.9. *Suppose g is strictly differentiable at $(0, a)$, $B = X$ and C is arbitrary. Then, if $D_2 g(0, a)$ is surjective, for any $w \in W$ condition (C) holds true with $K = \{0\}$.*

The proof relies on the following lemma which is an easy modification of [52, Thm. 1].

LEMMA 4.10. *Suppose $g: W \times X \rightarrow Y$ is strictly differentiable at $(0, a)$, with $D_2 g(0, a)$ surjective (or is differentiable at $(0, a)$ and satisfies the differentiability assumption (H) with $s = D_2 g(0, a)$ surjective). Then for each $w \in W$ there exist $\gamma \geq 0$, $\varepsilon > 0$, neighborhoods W_0 of w , X_0 of a and Y_0 of $c = g(0, a)$ and a continuous mapping $h: (0, \varepsilon)W_0 \times X_0 \times Y_0 \rightarrow X$ such that for each $(w', x, y) \in (0, \varepsilon)W_0 \times X_0 \times Y_0$*

- 1) $h(w', x, g(w', x)) = x$;
- 2) $g(w', h(w', x, y)) = y$;
- 3) $|h(w', x, y) - x| \leq \gamma|y - g(w', x)|$.

Proof of Proposition 4.9. We take $\gamma, \varepsilon, W_0, X_0, Y_0$ as in Lemma 4.10 and for $(w', x') \in (0, \varepsilon)W_0 \times X_0$ satisfying $x' \in A(w')$, we take $a' = h(0, x', g(w', x'))$ so that,

by 2) of Lemma 4.10 $g(0, a') = g(w', x') \in C$, hence $a' \in A(0)$ and by 3) of the same lemma

$$|a' - x'| \leq \gamma |g(w', x') - g(0, x')| \leq k |w'|$$

with $k > \gamma |g'(0, a)|$, provided ε, W_0, X_0 are small enough. \square

The following criteria motivated the introduction of condition (C).

THEOREM 4.11 (R. Janin [37, Props. 2 and 3]). *Suppose $W = \mathbb{R}^m$, $B = X = \mathbb{R}^n$, $Y = \mathbb{R}^p$, $C = -\mathbb{R}_+^q \times \{0\}$, f is of class C^1 around $(0, a)$, the restriction of A to its domain D is l.s.c. at $(0, a)$ and condition (J) holds. Then for each $w \in W$ condition (C) is satisfied with $K = T_c C$, $c = g(0, a)$.*

Hence, combining Theorems 3.7, 4.8 and 4.11 we get that

$$\bar{D}A(0, a)(w) = D_+A(0, a)(w) = \{x \in X | g'(0, a)(w, x) \in T_c C\}$$

is a variation of A at $(0, a)$ under the conditions of these theorems.

5. Application to perturbed optimization problems. We apply the preceding concepts and results to the study of the perturbed optimization problem

$$(P_w) \quad \text{minimize } f(w, x) \quad \text{as } x \in A(w)$$

with value $m(w)$. Here and in the sequel $m(0)$ is finite, W, X, Y are Banach spaces, $f: W \times X \rightarrow \mathbb{R} \cup \{+\infty\}$ is an arbitrary function (possibly with value $+\infty$ outside $W_0 \times X$, where W_0 is a neighborhood of 0 in W) and $A: W \rightrightarrows X$ is a relation with $0 \in \text{dom } A$ ($A(w)$ is the set of admissible solutions to (P_w)). Let

$$S(w) = \{x \in A(w) | f(w, x) = m(w)\}$$

be the set of optimal solutions. For $\varepsilon > 0$ we define the set of ε -approximate solutions to (P_w) as

$$S_\varepsilon(w) = \{x \in A(w) | f(w, x) < m(w) + \varepsilon\} \quad \text{if } m(w) > -\infty,$$

$$S_\varepsilon(w) = \{x \in A(w) | f(w, x) < -\varepsilon^{-1}\} \quad \text{if } m(w) = -\infty.$$

The following estimates (in which $\inf \emptyset = +\infty$) are trivial consequences of the definitions.

PROPOSITION 5.1. *The following estimate holds for each $w \in W$:*

$$\bar{d}m(0, w) \leq \inf \{\bar{d}f(0, a; w, v) | a \in S(0), v \in \bar{D}A(0, a)(w)\}.$$

Proof. It suffices to show that for each $w \in W$, each $a \in S(0)$, each $v \in \bar{D}A(0, a)(w)$ and each sequence (t_n, w_n) in $(0, \infty) \times W$ with limit $(0, w)$ we have

$$\limsup t_n^{-1} [m(t_n w_n) - m(0)] \leq \bar{d}f(0, a; w, v).$$

By definition of $\bar{D}A(0, a)(w)$, we can find a sequence (v_n) with limit v in X and $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$, $(0, a) + t_n(w_n, v_n) \in A$, so that $m(t_n w_n) \leq f(t_n w_n, a + t_n v_n)$, hence

$$\begin{aligned} \limsup t_n^{-1} [m(t_n w_n) - m(0)] &\leq \limsup t_n^{-1} [f(t_n w_n, a + t_n v_n) - f(0, a)] \\ &\leq \bar{d}f(0, a; w, v). \end{aligned}$$

\square

Remark 5.2. Similarly one has

$$d^+m(0, w) \leq \inf \{\bar{d}f(0, a; w, v) | a \in S(0), v \in D_+A(0, a)(w)\}.$$

PROPOSITION 5.3. *If f_A is the extension of $f|A$ by $+\infty$ on A^c , then*

$$\underline{d}m(0, w) \leq \inf \{ \underline{d}f_A(0, a; w, v) | a \in S(0), v \in X \}, \quad \partial m(0) \times \{0\} \subset \bigcap_{a \in S(0)} \partial f_A(0, a).$$

Proof. Let $a \in S(0)$, $w \in W$, $v \in X$. It suffices to show that for any sequence (t_n, w_n, v_n) in $(0, \infty) \times W \times X$ with limit $(0, w, v)$ we have

$$\liminf t_n^{-1} [m(t_n w_n) - m(0)] \leq \liminf t_n^{-1} [f_A(t_n w_n, a + t_n v_n) - f_A(0, a)].$$

As $f_A(0, a) = f(0, a) = m(0)$, and as $m(t_n w_n) \leq f_A(t_n w_n, a + t_n v_n)$ for each n , the result is obvious. \square

Let us note that the inequality $\underline{d}m(0, w) \leq \underline{d}f_A(0, a; w, v)$ is obvious and useless if $v \notin \bar{D}A(0, a)(w)$. Thus it is more convenient to have an estimate using f itself instead of f_A . As $f_A = f + \psi_A$, where ψ_A is the indicator function of A ($\psi_A|A = 0$, $\psi_A|A^c = +\infty$), we always have

$$\underline{d}f_A(0, a; w, v) \leq \bar{d}f(0, a; w, v) + \underline{d}\psi_A(0, a; w, v).$$

As $\underline{d}\psi_A(0, a; w, v) = 0$ if and only if $(w, v) \in T_{(0,a)}A$, we get the following estimate.

COROLLARY 5.4. *For each $w \in W$ we have*

$$\underline{d}m(0, w) \leq \inf \{ \bar{d}f(0, a; w, v) | a \in S(0), v \in \bar{D}A(0, a)(w) \}.$$

If for each $a \in S(0)$ the cone $T_{(0,a)}A$ is convex and $\bar{d}f(0, a; \cdot, \cdot)$ is a convex function, finite at some point of $T_{(0,a)}A$, then, if $N_{(0,a)}A = (T_{(0,a)}A)^0$ is the normal cone to A at $(0, a)$:

$$\partial m(0) \times \{0\} \subset \bigcap_{a \in S(0)} [\bar{d}f(0, a) + N_{(0,a)}A].$$

This last formula follows from $\underline{d}\psi_A(0, a; w, v) = \psi_{T_{(0,a)}A}(w, v)$, from $\partial m(0) \times \{0\} = \partial \tilde{m}(0, a)$ if $\tilde{m}(w, x) = m(w)$ and a subdifferentiation formula ([48, Thm. 3.16]).

Under some vertical Lipschitzian property on f , which is satisfied if f is strictly Hadamard differentiable at $(0, a)$, the upper Dini derivatives of f and A can be replaced by the lower Dini derivatives in the above estimate.

PROPOSITION 5.5. *If for some $a \in S(0)$, $w \in W$, $v \in \bar{D}A(0, a)(w)$ there exist $\varepsilon > 0$, $c > 0$ and neighborhoods W' of w and V' of v such that*

$$|f(w', x) - f(w', x')| \leq c|x - x'|$$

for any $w' \in (0, \varepsilon)W'$, $x, x' \in a + (0, \varepsilon)V'$, then

$$\underline{d}m(0, w) \leq \underline{d}f(0, a; w, v).$$

Moreover if f satisfies the above condition for every $a \in S(0)$, $(w, v) \in \bar{D}A(0, a)$, if for each $a \in S(0)$ the relation A is differentiable at $(0, a)$, the cone $T_{(0,a)}A$ is convex and $\underline{d}f(0, a; \cdot, \cdot)$ is convex, finite and continuous at some point of $T_{(0,a)}A$, then

$$\partial m(0) \times \{0\} \subset \bigcap_{a \in S(0)} [\underline{d}f(0, a) + N_{(0,a)}A].$$

Proof. Let $((t_n, w_n, v_n))$ be a sequence with limit $(0_+, w, v)$ such that

$$\underline{d}f(0, a; w, v) = \lim t_n^{-1} (f(t_n w_n, a + t_n v_n) - f(0, a)).$$

Since $v \in \bar{D}A(0, a)(w)$ there exists a sequence (v'_n) with limit v such that $(t_n w_n, a + t_n v'_n) \in A$ for n large enough. For n still larger $t_n \in (0, \varepsilon)$, $w_n \in W'$, $v_n, v'_n \in V'$ so that

$$|f(t_n w_n, a + t_n v_n) - f(t_n w_n, a + t_n v'_n)| \leq c t_n |v_n - v'_n|$$

and

$$\begin{aligned} \underline{df}(0, a; w, v) &= \lim t_n^{-1}(f(t_n w_n, a + t_n v'_n) - f(0, a)) \\ &\geq \liminf t_n^{-1}(m(t_n w_n) - m(0)) \geq \underline{dm}(0, w). \end{aligned}$$

The last inclusion follows as above. \square

Setting $g = -f$, $p = -m$ and observing that $\underline{dg} = -\bar{df}$, we get the following version of Propositions 5.1 and 5.4 in which

$$M(w) = \{x \in A(w) | g(w, x) = p(w)\}.$$

PROPOSITION 5.6. *If $p(w) = \sup \{g(w, x) | x \in A(w)\}$, and if $p(0)$ is finite then*

$$\underline{dp}(0, w) \geq \sup \{\underline{dg}(0, a; w, v) | a \in M(0), v \in \underline{DA}(0, a)(w)\},$$

$$\bar{dp}(0, w) \geq \sup \{\bar{dg}(0, a; w, v) | a \in M(0), v \in \bar{DA}(0, a)(w)\},$$

hence for each $(w^*, x^*) \in \partial g(0, a)$ and each $u \in \partial A(0, a)$ (resp. $\bar{\partial} A(0, a)$) we have

$$w^* + x^* \circ u \in \underline{\partial} p(0) \quad (\text{resp. } \bar{\partial} p(0)).$$

We shall show that under a first order sufficient optimality condition the preceding estimates give the exact value of the Dini derivative of m .

We define a subset A_0 of X to be *tangentially compact* at $a \in A_0$ if the relation $A = \{0\} \times A_0$ is B -tangentially compact at $(0, a)$. In other words any bounded sequence (v_n) of X such that $a + t_n v_n \in A_0$ for some sequence (t_n) of $(0, +\infty)$ with limit 0 has a cluster point.

PROPOSITION 5.7 (First order sufficient optimality condition). *Suppose $A_0 \subset X$ is tangentially compact at $a \in A_0$ and one of the following conditions holds on $f_0: X \rightarrow \mathbb{R}$, finite at a :*

- a) *For each $v \in T_a A_0$, $v \neq 0$, $\underline{df}_0(a, v) > 0$.*
- b) *There exists $c > 0$ with $\underline{df}_0(a, v) \geq c|v|$ for each $v \in T_a A_0$.*
- c) $0 \in \text{int}(\partial f_0(a) + N_a A_0)$.

Then a is a local strict minimum of $f_0|_{A_0}$.

Proof. First, let us show the conclusion under assumption a). Suppose on the contrary there exists a sequence (a_n) in A_0 with limit a such that $f_0(a_n) \leq f_0(a)$ for each $n \in \mathbb{N}$. Taking a subsequence if necessary, we may suppose (x_n) given by $x_n = t_n^{-1}(a_n - a)$, $t_n = |a_n - a|$ converges to some limit x with $|x| = 1$, and, obviously, $x \in T_a A_0$. But $\underline{df}_0(a, x) \leq \liminf t_n^{-1}[f_0(a + t_n x_n) - f_0(a)] \leq 0$ gives a contradiction.

As b) implies a), it remains to show that c) implies b). Let $r > 0$ be such that any x^* in the ball X_r^* of center 0 and radius r in X^* belongs to $\partial f_0(a) + N_a A_0$. Let $v \in T_a A_0$ and let $x^* \in X_r^*$. We can find $y^* \in N_a A_0$ and $z^* \in \partial f_0(a)$ such that $x^* = y^* + z^*$, so that

$$\underline{df}_0(a, v) \geq \langle z^*, v \rangle = \langle x^* - y^*, v \rangle \geq \langle x^*, v \rangle.$$

Hence

$$\underline{df}_0(a, v) \geq \sup_{|x^*| \leq r} \langle x^*, v \rangle = r|v|. \quad \square$$

Remark 5.8. In fact under our assumption on A_0 , a) and b) are equivalent since $T_a A_0$ is locally compact as one can easily show. If $T_a A_0$ and $\underline{df}_0(A, \cdot)$ are convex and if

$$\text{dom } \underline{df}_0(a, \cdot) - T_a A_0 = X$$

then b) and c) are equivalent. In fact, any $x^* \in X_c^*$ belongs to

$$\begin{aligned} \partial(\underline{d}f_0(a, \cdot) + \psi_{T_a A_0})(0) &= \partial(\underline{d}f_0(a, \cdot))(0) + \partial\psi_{T_a A_0}(0) \\ &= \underline{d}f_0(a) + N_a A_0 \end{aligned}$$

by [48, Thm. 3.6], ψ_S being the indicator function of the set S ($\psi_S(s) = 0$ for $s \in S$, $\psi_S(x) = +\infty$ for $x \in X \setminus S$).

Example 5.9. The condition c) of Proposition 5.7 occurs quite often. For instance, suppose $A_0 = \{x \in X \mid g_i(x) \leq 0, i \in I\}$, where I is a finite set and f_0 and g_i are differentiable at a . Let $K = \{i \in I \mid g_i(a) = 0\}$, and suppose the linear forms $\{g'_k(a) \mid k \in K\}$ are linearly independent and that there exists $y_k \in (0, +\infty)$, for $k \in K$, with

$$f'_0(a) + \sum_{k \in K} y_k g'_k(a) = 0.$$

Then if $\text{card } K = \dim X$, condition c) holds because $N_a A_0$ is the convex cone generated by $\{g'_k(a) \mid k \in K\}$ and $-f'_0(a)$ belongs to the interior of this cone as one can see by taking $\{g'_k(a) \mid k \in K\}$ as a basis of X^* .

We need the following definition in order to state our main result; it is variant of [56, §§ 3A and 3C] and [57, Definition 3.1].

DEFINITION 5.10. *The perturbation (f, A) of the problem (P_0) is well-set at 0 in the direction w , if for each sequence (ε_n) in $(0, +\infty)$ with limit 0 and each sequence (t_n, w_n) in $(0, +\infty) + W$ with limit $(0, w)$ and $t_n w_n \in \text{dom } A$ for each n , there exists an infinite subset K of \mathbb{N} and a sequence (x_k) in X with limit in $S(0)$, such that $x_k \in S_{\varepsilon_k}(t_k w_k)$ for each $k \in K$.*

THEOREM 5.11. *Under the following three conditions:*

- a) *A is B -tangentially compact in the direction w ;*
- b) *the perturbation (f, A) is well-set in the direction w at 0;*
- 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:

$$\underline{d}m(0, w) \geq \inf \{ \underline{d}f(0, a; w, x) \mid a \in S(0), x \in \bar{D}A(0, a)(w) \}.$$

Comparing with the estimate of Corollary 5.4 we get the following equality.

COROLLARY 5.12. *If f is directionally differentiable at each point of $\{0\} + S(0)$ and if the assumptions a), b), c), of Theorem 5.11 hold, then*

$$\underline{d}m(0, w) = \inf \{ \underline{d}f(0, a; w, x) \mid a \in S(0), x \in \bar{D}A(0, a)(w) \}.$$

COROLLARY 5.13. *If f is directionally differentiable at each point of $\{0\} \times S(0)$ with a sublinear derivative, if assumptions a), b), c) of Theorem 5.11 hold, if $T_{(0,a)}A$ is convex, and if $m(0)$ is finite, then*

$$w^* \in \underline{d}m(0) \Leftrightarrow (w^*, 0) \in \bigcap_{a \in S(0)} [\partial f(0, a) + N_{(0,a)}A].$$

In particular, if f is differentiable at each point of $\{0\} \times S(0)$ then

$$w^* \in \underline{d}m(0) \Leftrightarrow (w^*, 0) \in \bigcap_{a \in S(0)} [f'(0, a) + N_{(0,a)}A].$$

Proof. If $w^* \in \underline{d}m(0)$, for each $w \in W$, each $a \in S(0)$ and each $x \in \bar{D}A(0, a)(w)$, we have

$$\langle w^*, w \rangle \leq \underline{d}f_A(0, a; w, x) \leq f'(0, a; w, x) + \underline{d}\psi_A(0, a; w, x),$$

and this also holds true if $x \notin \bar{D}A(0, a)(w)$, i.e., if $(w, x) \notin T_{(0,a)}A$. Thus

$$(w^*, 0) \in \partial(f + \psi_A)(0, a) = \partial f(0, a) + \partial \psi_A(0, a) = \partial f(0, a) + N_{(0,a)}A$$

as $f'(0, a; \cdot, \cdot)$ and $d\psi_A(0, a; \cdot, \cdot) = \psi_{T_{(0,a)}A}$ are sublinear mappings and $\text{dom } f'(0, a; \cdot, \cdot) = W \times X$.

Conversely, if $w^* \in W^*$ is such that for any $a \in S(0)$ we have $(w^*, 0) \in \partial f(0, a) + N_{(0,a)}A$, then for each $w \in W$ and each $x \in \bar{D}A(0, a)(w)$ we get

$$\langle w^*, w \rangle \leq f'(0, a; w, x)$$

as $(w, x) \in T_{(0,a)}A$, hence $\langle w^*, w \rangle \leq dm(0, w)$ by the preceding corollary. \square

Remark 5.14. Suppose $(f_i)_{i \in I}$ is a finite set of mappings on W and $m = \inf_{i \in I} f_i$. We can embed I in $X = \mathbb{R}$ as a discrete subset and take $A = W \times I$, $f(w, x) = f_i(w)$ if $x = i$ and $f(w, x) = +\infty$ if $x \notin I$. Then assumptions a), b) and c) of Theorem 5.11 are satisfied so that

$$dm(0, w) \geq \inf \{df_i(0, w) | i \in S(0)\}$$

with $S(0) = \{i \in I | f_i(0) = m(0)\}$, as $\bar{D}A(0, i)(w) = \{0\}$ and $df(0, i; w, 0) = df_i(0, w)$ if $i \in S(0)$. This result was obtained more directly in [48].

Proof of Theorem 5.11. Let (t_n, w_n) be a sequence of $(0, +\infty) \times W$ with limit $(0, w)$ such that $dm(0, w) = \lim t_n^{-1}(m(t_n w_n) - m(0))$. We may suppose $m(t_n w_n) < +\infty$ for each n hence $t_n w_n \in \text{dom } A$.

First we consider the case $m(t_n w_n) > -\infty$ for each $n \in \mathbb{N}$. We can choose a sequence (x_n) with limit a in $S(0)$ such that $x_n \in A(t_n w_n)$, $f(t_n w_n, x_n) \leq m(t_n w_n) + t_n^2$ and such that the sequence $(v_n) = (r_n^{-1}(x_n - a))$, with $r_n = |x_n - a|$, converges to a vector v in X , and $(r_n t_n^{-1})$ converges to s in $[0, +\infty]$. Then

$$\begin{aligned} dm(0, w) &\geq \liminf t_n^{-1}[f(t_n w_n, x_n) - f(0, a) - t_n^2] \\ &\geq \liminf (r_n t_n^{-1}) r_n^{-1}[f(r_n t_n^{-1} w_n, a + r_n v_n) - f(0, a)]. \end{aligned}$$

If $s = +\infty$ we get $dm(0, w) \geq sdf(0, a; 0, v) = +\infty$ as $(0, v) = \lim r_n^{-1}((t_n w_n, a + r_n v_n) - (0, a)) \in T_{(0,a)}A$ and $df(0, a; 0, v) > 0$, so that the inequality is satisfied. If $s \in [0, +\infty)$, we have $(w, sv) = \lim t_n^{-1}((t_n w_n, a + r_n v_n) - (0, a)) = \lim t_n^{-1}((t_n w_n, x_n) - (0, a)) \in T_{(0,a)}A$ and

$$dm(0, w) \geq \liminf t_n^{-1}[f(t_n w_n, x_n) - f(0, a)] \geq df(0, a; w, sv),$$

and the inequality is again satisfied.

Now we consider the case in which $m(t_n w_n) = -\infty$ for n in an infinite subset N of \mathbb{N} . Then we can find an infinite subset K of N and a sequence $(x_k)_{k \in K}$ with limit a in $S(0)$ such that $x_k \in A(t_k w_k)$, $f(t_k w_k, x_k) \leq f(0, a) - \sqrt{t_k}$. Furthermore, we can suppose $(v_k) = (r_k^{-1}(x_k - a))$ has limit v in X and $(r_k t_k^{-1})$ has a limit s in $[0, +\infty]$, with $r_k = |x_k - a|$.

If $s = +\infty$, then as above $(0, v) \in T_{(0,a)}A$ and $dm(0, w) = +\infty$, a contradiction with $m(t_n w_n) = -\infty$ for $n \in N$. Thus $s \in [0, +\infty)$ and again (w, sv) is in $T_{(0,a)}A$ and

$$df(0, a; w, sv) \leq \liminf t_k^{-1}[f(t_k w_k, a + t_k r_k t_k^{-1} v_k) - f(0, a)] \leq \lim -t_k^{-1} \sqrt{t_k} = -\infty$$

and the inequality holds in each case. \square

Now we present another result giving a lower bound to $dm(0, w)$. It uses the concept of variation introduced in § 4 and the following differentiability assumptions on $f: W \times X \rightarrow Y$:

(H₁) For each $a \in S(0)$, $f(0, \cdot)$ has a strict semi-derivative s at $(0, a)$ which is sublinear and continuous:

$$\lim_{\substack{x, x' \rightarrow a \\ x \neq x'}} |x - x'|^{-1} [f(0, x) - f(0, x') - s(x - a) + s(x' - a)] = 0.$$

(H₂) For each $a \in S(0)$ the following limit exists:

$$y = \lim_{(t, w', x) \rightarrow (0, w, a)} \frac{1}{t} [f(tw', x) - f(0, x)].$$

These assumptions are satisfied if f is strictly differentiable at $(0, a)$, in particular if f is of class C^1 around $(0, a)$, for each $a \in S(0)$. On the other hand, (H₁) and (H₂) imply that for each $a \in S(0)$ and each $v \in X$, f has a directional derivative $df(0, a)(w, v) = y + s(v)$ in the direction (w, v) at $(0, a)$.

THEOREM 5.15. *Under the following assumptions:*

- a) *the perturbed problem (P) is well-set in the direction w at 0;*
- b) *f satisfies assumptions (H₁) and (H₂);*
- c) *for each $a \in S(0)$, A has a strong variation V_a at $(0, a)$ in the direction w ;*

the following estimate holds:

$$\underline{d}m(0, w) \geq \inf \{df(0, a)(w, 0) - df(0, a)(0, -v) | a \in S(0), v \in V_a\}.$$

If moreover $f(0, \cdot)$ is differentiable at a , for each $a \in S(0)$, then

$$\underline{d}m(0, w) \geq \inf \{f'(0, a)(w, v) | a \in S(0), v \in V_a\}.$$

Proof. Let $((t_n, w_n))$ be a sequence in $(0, +\infty) \times W$ with limit $(0, w)$ such that $\underline{d}m(0, w) = \lim t_n^{-1}(m(t_n w_n) - m(0))$. We may suppose $m(t_n w_n) < +\infty$, hence $t_n w_n \in \text{dom } A$ for each $n \in \mathbb{N}$. Assumption a) gives a point $a \in S(0)$ and a sequence $(x_k)_{k \in K}$ with limit a such that $x_k \in S_{t_k^2}(t_k w_k)$ for each k in the infinite subset K of \mathbb{N} . Using c) we can find sequences (a_k) and (v_k) in $A(0)$ and V_a respectively such that (a_k) converges to a and

$$\lim_k |t_k^{-1}(x_k - a_k) - v_k| = 0.$$

Let

$$q_k = t_k^{-1}[f(t_k w_k, x_k) - f(0, a_k)].$$

As $x_k \in S_{t_k^2}(t_k w_k)$ and as (H₂) holds we cannot have $m(t_k w_k) = -\infty$, so that

$$t_k^{-1}[m(t_k w_k) - m(0)] \geq q_k - t_k$$

hence $\underline{d}m(0, w) \geq \liminf q_k$. Now

$$q_k = t_k^{-1}[f(t_k w_k, x_k) - f(0, x_k)] + t_k^{-1}[f(0, x_k) - f(0, a_k)],$$

where $q_k'' := t_k^{-1}[f(t_k w_k, x_k) - f(0, x_k)]$ converges to $df(0, a; w, 0)$ and

$$q_k' := t_k^{-1}[f(0, x_k) - f(0, a_k)] = t_k^{-1}[s(x_k - a) - s(a_k - a)] + r_k$$

with $\lim r_k = 0$, as $\limsup t_k^{-1}|x_k - a_k| < +\infty$. Let us set $v_k' := t_k^{-1}(x_k - a_k)$, so that $\lim (v_k' - v_k) = 0$. Then as $s = df(0, a; 0, \cdot)$ is sublinear

$$s(x_k - a) - s(a_k - a) = s(a_k + t_k v_k' - a) - s(a_k - a) \geq -s(-t_k v_k').$$

Thus

$$\begin{aligned}\liminf q'_k &= \liminf t_k^{-1}[s(x_k - a) - s(a_k - a)] \\ &\geq \liminf -s(-v'_k) \\ &\geq \inf \{-s(-v) | v \in V_a\},\end{aligned}$$

since s is continuous. The result follows by addition:

$$dm(0, w) \geq \liminf q_k = \liminf q'_k + df(0, a; w, 0),$$

taking the infimum over $a \in S(0)$. \square

Combining Theorems 4.8, 5.15 and Corollary 5.4, we get the following result:

COROLLARY 5.16. *Suppose A is given by $A(w) = \{x \in B | g(w, x) \in C\}$, where B and C are subsets of X and Y respectively, $f: W \times X \rightarrow \mathbb{R}$ and $g: W \times X \rightarrow Y$ are of class C^1 , X is finite dimensional. Suppose the perturbed problem is well-set at 0 in the direction w , Condition (C) of Theorem 4.8 holds for each $a \in S(0)$ with $K = T_{g(0,a)}C$ and $T_a B = \limsup_{(t,x) \rightarrow (0+,a)} t^{-1}(B - x)$, $\bar{D}A(0, a)(w) = \{v \in T_a B | g'(0, a)(w, v) \in K\}$. Then*

$$dm(0, w) = \inf \{f'(0, a)(w, v) | a \in S(0), v \in T_a B, g'(0, a)(w, v) \in C\}.$$

Using Remark 5.2 and Theorem 4.11, we get the main result of [37].

COROLLARY 5.17. *Suppose A is given by $A(w) = \{x \in \mathbb{R}^n | g(w, x) \in C\}$, where $C = -\mathbb{R}_+^q \times \{0\}$, and $g: \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}^p$ is of class C^1 . Suppose f is of class C^1 and condition (J) holds at each $(0, a)$ with $a \in S(0)$. Then if the perturbed problem is well-set at 0 in some direction w , the marginal function m is differentiable on $D = \text{dom } A$ in the direction w and*

$$\begin{aligned}dm(0, w) &= \lim_{(t,w') \xrightarrow{D} (0,w)} t^{-1}[m(tw') - m(0)] \\ &= \inf \{f'(0, a)(w, v) | a \in S(0), g'(0, a)(w, v) \in T_{g(0,a)}C\}.\end{aligned}$$

Let us compare the preceding results with the completely convex case.

THEOREM 5.18 ([31, Thm. 3]). *Suppose A is given by $A(w) = \{x \in X | g(w, x) \in C\}$ where $X = \mathbb{R}^n$, $C = -\mathbb{R}_+^p$. Suppose f and g are convex on $W_0 \times X$, where W_0 is a neighborhood of 0 in W and $m(0)$ is finite. Then for each $a \in S(0)$*

$$d_r m(0, w) = \inf \{df(0, a)(w, x) | x \in \bar{D}_r A(0, a)(w)\}.$$

Here the subscript r indicates that radial derivatives are taken. We observe that $\bar{D}_r A(0, a)(w) = \{x | (w, x) \in T'_{(0,a)}A\}$ is a radial variation of A at $(0, a)$ in the direction w (Example 4.2.d). The fact that no infimum on $a \in S(0)$ has to be taken is explained in [31, Thm. 4].

Finally, we consider [26, Ex. 3.1]:

$$W = \mathbb{R}^3, \quad B = X = \mathbb{R}^2, \quad C = -\mathbb{R}_+^3$$

and for $x = (x_1, x_2)$, $w = (w_1, w_2, w_3)$

$$\begin{aligned}g(w, x) &= (x_2 + x_1^2 - w_1, x_2 - x_1^2 - w_2, -x_2 - 1 - w_3), \\ f(w, x) &= -x_2.\end{aligned}$$

Then it is easy to see that for $w = (1, 0, 0)$ the set $V = \mathbb{R} \times (-\infty, \frac{1}{2}]$ is a variation of A at $(0, 0)$ in the direction w , and

$$dm(0, w) = \inf \{f'(0, 0)(w, v) | v \in V\} = -\frac{1}{2},$$

although the estimates of [26] present gaps for this direction. Variations for other directions are easily found, yielding the value of the directional derivative of m .

REFERENCES

- [1] W. ALT, *Stabilität mengenwertiger Abbildungen mit Anwendungen auf nichtlineare Optimierungsprobleme*, Dissertation, Bayreuth Univ., 1979.
- [2] J. P. AUBIN, *Contingent derivatives of set-valued maps and existence of solutions to nonlinear inclusions and differential inclusion*, MRC Tech. Summary Rept. 2044, Univ. of Wisconsin, Madison, 1980.
- [3] ———, *Further properties of Lagrange multipliers*, Appl. Math. Optim. 6 (1980), pp. 79–90.
- [4] J. P. AUBIN AND F. H. CLARKE, *Multiplicateur de Lagrange en optimisation non convexe et applications*, C.R. Acad. Sci. Paris A, (1977), pp. 451–454.
- [5] A. AUSLENDER, *Differential stability in nonconvex and nondifferentiable programming*, Math. Progr. Study, 10 (1979), pp. 29–41.
- [6] V. I. AVERBUKH AND O. G. SMOLYANOV, *The various definitions of differentiation in linear topological spaces*, Russ. Math. Surveys, 23 (1968), pp. 67–113.
- [7] M. BAKLEH, *Contribution à l'étude des systèmes régis par des équations différentielles multivoques*, Thèse de 3ème cycle Université Paris VI, 1981.
- [8] H. T. BANKS AND M. Q. JACOBS, *A differential calculus for multifunctions*, J. Math. Anal. Appl., 29 (1970), pp. 246–272.
- [9] M. S. BAZARAA, J. J. GOODE AND M. Z. NASHED, *On the cones of tangents with application to mathematical programming*, J. Optim. Theory Appl., 13 (1974), pp. 389–426.
- [10] V. V. BERESNEV, *Minimization of solution functions for a parametric minimization problem*, Kibernetika, 5 (1976), pp. 100–109.
- [11] C. BERGE, *Espaces topologiques, fonctions multivoques*, Dunod, Paris, 1959; English translation, *Topological Spaces*, Macmillan, New York, 1963.
- [12] J. H. BIGELOW AND N. Z. SHAPIRO, *Implicit function theorems for mathematical programming and for systems of inequalities*, Math. Progr., 6 (1974), pp. 141–156.
- [13] J. BORWEIN, *Convex relations in analysis and optimization*, Res. Rep. 80-5, Carnegie-Melon Univ., Pittsburgh, PA.
- [14] CH. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Mathematics, 580, Springer-Verlag, Berlin, 1977.
- [15] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
- [16] ———, *A new approach to Lagrange multipliers*, Math. Oper. Research, 1 (1976), pp. 165–174.
- [17] G. B. DANTZIG, J. FOLKMAN AND N. SHAPIRO, *On the continuity of the minimum set of a continuous function*, J. Math. Anal. Appl., 17 (1967), pp. 519–548.
- [18] F. S. DE BLASI, *On the differentiability of multifunctions*, Pacific J. Math., 66 (1976), pp. 67–81.
- [19] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization techniques*, John Wiley, New York, 1968.
- [20] A. V. FIACCO AND W. P. HUTZLER, *Extension of the Gauvin-Tolle optimal value differential stability results to general mathematical programs*, Tech. Paper T-393, George Washington Univ., Washington, 1979.
- [21] T. M. FLETT, *On differentiation in normed vector spaces*, J. London Math. Soc., 42 (1967), pp. 523–533.
- [22] S. GAUTIER, *Différentiabilité de multiapplications*, Publications Math. de Pau, (1978) (5), pp. 1–17.
- [23] S. GAUTIER AND J. P. PENOT, *Differentiability of multifunctions and subdifferential calculus*, in preparation.
- [24] J. GAUVIN, *The generalized gradient of a marginal function in mathematical programming*, Math. Oper. Res., 4 (1979), pp. 458–463.
- [25] J. GAUVIN AND F. DUBEAU, *Differential Properties of the Marginal Function in Mathematical Programming*, Math. Prog. Study, M. Guignard, ed., forthcoming.
- [26] J. GAUVIN AND J. W. TOLLE, *Differential stability in nonlinear programming*, this Journal, 15 (1977), pp. 294–311.
- [27] B. GOLLAN, *Perturbation theory for abstract optimization problems*, J. Optim. Theory Appl., 35 (1981), pp. 417–441.
- [28] J. B. HIRIART-URRUTY, *Gradients généralisés de fonctions marginales*, this Journal, 16 (1978), pp. 301–316.
- [29] J. B. HIRIART-URRUTY, *Tangent cones, generalized gradients and mathematical programming in Banach spaces*, Math. Oper. Research, 4 (1979), pp. 79–97.

- [30] W. HOGAN, *Point to set maps in mathematical programming*, SIAM Rev., 15 (1973), pp. 591–603.
- [31] ———, *Directional derivatives for extremal value functions with applications to the completely convex case*, Oper. Res. 21 (1973), pp. 188–209.
- [32] L. HÖRMANDER, *Sur la fonction d'appui des ensembles convexes d'un espace localement convexe*, Arkiv für Math., 3 (1965), pp. 181–186.
- [33] P. HUARD, *Optimization algorithms and point to set maps*, Math. Progr., 8 (1975), pp. 308–331.
- [34] M. HUKUHARA, *Intégration des applications mesurables dont la valeur est un compact convexe*, Funkcial. Ekvac., 10 (1967), pp. 205–223.
- [35] A. IOFFE, *Regular points of Lipschitz functions*, Trans. Amer. Math. Soc., 251 (1979), pp. 61–69.
- [36] R. JANIN, *On sensitivity in nonconvex programming*, Proc. Conf. Murat Le Quaire, March 1976, Lecture Notes in Economics and Mathematical Systems, 144. Springer-Verlag, Berlin, 1977, pp. 115–119.
- [37] ———, *First order differential stability in non convex mathematical programming*, Preprint, Centre Univ. Antilles-Guyane, Pointe-à-Pitre, Guadeloupe.
- [38] A. LASOTA AND A. STRAUSS, *Asymptotic behavior for differential equations which cannot be locally linearized*, J. Differential Equations, 10 (1971), pp. 152–172.
- [39] F. LEMPIO AND H. MAURER, *Differential stability in infinite dimensional nonlinear programs*, Appl. Math. Optim., 6 (1980), pp. 139–152.
- [40] E. S. LEVITIN, *On differential properties of the optimum value of parametric problems of mathematical programming*, Dokl. Akad. Nauk SSSR 215 (1974); Soviet Math. Dokl., 15 (1974), pp. 603–608.
- [41] ———, *Differentiability with respect to a parameter of the optimal value in parametric problems of mathematical programming*, Kibernetika, (1976), pp. 44–59.
- [42] M. MARTELLI AND A. VIGNOLI, *On differentiability of multivalued maps*, Boll. Un. Mat. Ital., 10 (1974), pp. 701–712.
- [43] H. METHLOUTHI, *Calcul différentiel multivoque*, Cahiers Mathématiques de la Décision, no. 7702, Université Paris IX, 1977.
- [44] S. MIRICA, *The contingent and the paratingent as generalized derivatives for vector-valued and set-valued mappings*, Preprint no. 31, INCREST, Bucharest, 1981.
- [45] M. Z. NASHED, *Differentiability and related properties of nonlinear operators: some aspects of the role of differentials in nonlinear functional analysis*, in Nonlinear Functional Analysis and Applications, L. B. Rall, ed., Academic Press, New York, 1971, pp. 102–311.
- [46] J. P. PENOT, *Calcul différentiel dans les espaces vectoriels topologiques*, Studia Math., 47 (1973), pp. 1–23.
- [47] ———, *Sous-différentiels de fonctions numériques non convexes*, C.R. Acad. Sci. Paris A, 278 (1974), pp. 1553–1555.
- [48] ———, *Calcul sous-différentiel et optimisation*, J. Funct. Anal., 27 (1978), pp. 248–276.
- [49] ———, *Optimisation et extrémisation*, Lecture Notes, Univ. of Pau, 1973–1974.
- [50] ———, *The use of generalized subdifferential calculus in optimization theory*, Proc. Third Symp. on Operations Research, Mannheim, 1978, Meth. Oper. Res. 31, Athenäum, Berlin, pp. 495–511.
- [51] ———, *Utilisation des sous-différentiels généralisés en optimisation*, First meeting AFCET-SMF on Applied Mathematics, Palaiseau 1978, AFCET-SMF Paris, Vol. 3, 1978, pp. 69–85.
- [52] ———, *Inversion à droite d'applications non linéaires. Applications*, C.R. Acad. Sci. Paris, 290 (1980), pp. 997–1000.
- [53] ———, *A characterization of tangential regularity*, Nonlinear Analysis, Theory, Methods and Appl., 5(6), (1981), pp. 625–643.
- [54] ———, *On the existence of Lagrange multipliers in nonlinear programming in Banach spaces*, Symp. Oberwolfach, March 1980, on Optimization and Optimal Control, A. Auslender, W. Oettli, J. Stoer, ed., Lecture Notes in Control and Information Science 30, Springer-Verlag, Berlin (1981), pp. 89–104.
- [55] ———, *On regularity conditions in mathematical programming*, Math. Progr. Study, 19 (1982), pp. 167–199.
- [56] ———, *Compact nets, filters and relations*, J. Math. Anal. Appl., to appear.
- [57] ———, *Continuity properties of performance functions*, to appear.
- [58] J. CH. POMEROL, *The Lagrange multipliers set and the generalized gradients set of the marginal function of differentiable programs in Banach space*, Preprint Univ., Paris VI, 1981.
- [59] PHAM HUU SACH, *Les points réguliers des applications multivoques et la commandabilité dans les systèmes discrets*, Preprint, Univ. de Bordeaux.
- [60] B. N. PSZENICNYI, *Necessary Conditions for an Extremum*, Nauka, Moscow, 1969, English translation, Marcel Dekker, New York, 1971, M.R. 43 #2584, 2585.
- [61] H. RÄDSTRÖM, *An embedding theorem for spaces of convex sets*, Proc. Amer. Math. Soc., 3 (1952), pp. 165–169.

- [62] S. M. ROBINSON, *Stability theory for systems of inequalities, Part II: differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
- [63] R. T. ROCKAFELLAR, *Proximal subgradients, marginal values and augmented Lagrangians in non convex optimization*, Math. Oper. Res., to appear.
- [64] ———, *Lagrange multipliers and subderivatives of optimal value functions in nonlinear programming*, Math. Progr. Studies, R. Wets ed., forthcoming.
- [65] ———, *Augmented Lagrangians and marginal values in parametrized optimization problems*, Preprint, IASA, Laxenbourg, Austria (1981).
- [66] E. SACHS, *Differentiability in optimization theory*, Math. Operationsforsch. Statist. Ser. Optim., 9 (1978), pp. 497–513.
- [67] J. E. SPINGARN, *Generic conditions for optimality in constrained minimization problems*, Ph.D. Thesis, Univ. of Washington, Seattle, 1977.
- [68] ———, *Fixed and variable constraints in sensitivity analysis*, this Journal, 18 (1980), pp. 297–310.
- [69] H. TUY, *On the convex approximation of nonlinear inequalities*, Math. Operationsforsch. U. Statist., 5 (1974), pp. 451–466.
- [70] C. URSESCU, *Sur une généralisation de la notion de différentiabilité*, Accad. Naz. Lincei, 8 (1973), pp. 199–204.
- [71] S. YAMAMURO, *Differential Calculus in Topological Linear Spaces*, Lecture Notes in Mathematics, 374, Springer-Verlag, Berlin, 1974.

ON THE CONTROL CANONICAL STRUCTURE OF A CLASS OF SCALAR INPUT SYSTEMS*

RUSSELL G. TEGLAS†

Abstract. In this paper, we demonstrate the equivalence of a scalar input system $\dot{x} = \mathcal{A}x + \ell u$, for which the eigenvalues of the generator \mathcal{A} coincide with the roots of the entire function

$$p(\omega) = e^{\omega T} + a_1 e^{\omega(T-\theta_1)} + \cdots + a_m + \int_0^T a(\theta) e^{\omega(T-\theta)} d\theta,$$

with the controlled scalar functional equation

$$y(t) + a_1 y(t - \theta_1) + \cdots + a_m y(t - T) + \int_0^T a(\theta) y(t - \theta) d\theta = u(t).$$

The theory of nonharmonic Fourier series is then employed to investigate the placement of eigenvalues in the closed-loop system with continuous state feedback.

Key words. canonical form, control theory, feedback, nonharmonic Fourier series, pole assignment

0. Introduction. Let \mathcal{A} be a generator of a strongly continuous group of operators $\{\mathcal{G}(t): t \in \mathbb{R}\}$ on a Hilbert space \mathcal{H} , and suppose that the spectrum of \mathcal{A} consists of an infinite sequence of simple eigenvalues $\{\omega_k\}$ which forms the zero set of an entire function having the form

$$(0.1) \quad p(\omega) = e^{\omega T} + a_1 e^{\omega(T-\theta_1)} + \cdots + a_m + \int_0^T a(\theta) e^{\omega(T-\theta)} d\theta.$$

In this paper, we will analyze in detail the transformation which carries the scalar input system

$$(0.2) \quad \dot{x}(t) = \mathcal{A}x(t) + \ell u(t)$$

to the scalar functional equation

$$(0.3) \quad y(t) + a_1 y(t - \theta_1) + \cdots + a_m y(t - T) + \int_0^T a(\theta) y(t - \theta) d\theta = u(t).$$

The latter constitutes the causal control canonical form for the pair (\mathcal{A}, ℓ) . The adjective “causal” is used here because a feedback law of the form

$$(0.4) \quad u(t) = (\ell, x(t))_{\mathcal{H}}$$

will be shown to transform to a feedback law of the form

$$(0.5) \quad u(t) = \int_0^T g(\theta) y(t - \theta) d\theta,$$

i.e., the input u depends only upon past values of y . The expression “control canonical form” refers to the fact that the above-mentioned transformation has a structure quite similar to its well-known finite dimensional counterpart as well as to the fact that the

* Received by the editors March 9, 1982, and in revised form May 11, 1983. This research was sponsored in part by the U.S. Air Force Office of Scientific Research under grant AFOSR-79-0018 and in part by the National Aeronautics and Space Administration under contract NAS1-15810.

† Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, Virginia 23665. Current address: Department of Mathematics, University of Vermont, Burlington, Vermont 05405.

effect of linear feedback (0.4) in the original system (0.2) can be readily analyzed in terms of the system (0.3) with corresponding feedback (0.5).

Before proceeding, we present an example of a system which possesses the structure indicated above. Consider the linear hyperbolic system in characteristic normal form [3]:

$$\frac{\partial}{\partial t} \begin{bmatrix} w^- \\ w^+ \end{bmatrix} = \begin{bmatrix} \Lambda^- & 0 \\ 0 & \Lambda^+ \end{bmatrix} \frac{\partial}{\partial x} \begin{bmatrix} w^- \\ w^+ \end{bmatrix} + A(x) \begin{bmatrix} w^- \\ w^+ \end{bmatrix},$$

with boundary conditions

$$w^-(0, t) = D_0 w^+(0, t), \quad w^+(1, t) = D_1 w^-(1, t).$$

Here

$$\Lambda^-(x) = \text{diag}(\lambda_1(x), \dots, \lambda_n(x)), \quad \Lambda^+(x) = \text{diag}(\lambda_{n+1}(x), \dots, \lambda_{2n}(x)),$$

are diagonal $n \times n$ matrices with

$$\lambda_1(x) \leq \dots \leq \lambda_n(x) < 0 < \lambda_{n+1}(x) \leq \dots \leq \lambda_{2n}(x).$$

Also,

$$A(x) = \begin{bmatrix} A^-(x) & A_{12}(x) \\ A_{21}(x) & A^+(x) \end{bmatrix},$$

and D_0 and D_1 are both $n \times n$ matrices which determine the manner in which "information" is reflected at the boundaries $x = 0$ and $x = 1$. Such systems arise in the study of counterflow heat exchangers [9] and in the study of multiconductor transmission lines [6].

If $A_{12} \equiv A_{21} \equiv 0$ and the off-diagonal components of both A^- and A^+ vanish identically, then the characteristic frequencies are the roots of

$$\pi(\omega) = \det [\exp(\omega \Lambda_+ + M_+) - D_1 \exp(\omega \Lambda_- + M_-) D_0],$$

where

$$\Lambda_{\pm} \equiv \int_0^1 (\Lambda^{\pm}(x))^{-1} dx \quad \text{and} \quad M_{\pm} \equiv \int_0^1 (\Lambda_{\pm}(x))^{-1} A^{\pm}(x) dx.$$

The roots of π will coincide with those of a function p having the form displayed in (0.1):

$$\pi(\omega) = e^{\omega\alpha} (e^{\omega T} + a_1 e^{\omega(T-\theta_1)} + \dots + a_m) \equiv e^{\omega\alpha} \cdot p(\omega).$$

In the case where A has nonzero off-diagonal components, one can show [19], under the assumption that both D_0 and D_1 are invertible, that p remains the same except for the addition of an integral term with $a(\cdot) \in L^2(0, T)$, as in (0.1).

In the above example, one may consider control problems wherein one introduces a forcing term proportional to a scalar function $u(t)$ in the differential equations (distributed control)

$$(0.6) \quad \frac{\partial w}{\partial t} = \Lambda(x) \frac{\partial w}{\partial x} + A(x) w + b(x) u(t),$$

or in the boundary conditions (point control), e.g.,

$$(0.7) \quad w^+(1, T) = D_1 w^-(1, T) + bu(t).$$

Either way, one is led to consider a system of the form

$$\dot{x}(t) = \mathcal{A}x(t) + \ell u(t),$$

where the element ℓ , in the former case, lies in the state space \mathcal{H} but, in the latter case, must be interpreted as a distribution. We will discuss this point in greater detail in § 2.

Of central interest in this paper is the spectral synthesis problem: given a set of "desired" eigenvalues $\{\nu_k\}$, can one construct a feedback law of the form (0.4) such that the eigenvalues of the closed-loop system

$$\dot{x} = \mathcal{A}x + \ell(\ell', x) \equiv (\mathcal{A} + \ell \otimes \ell')x$$

coincide precisely with $\{\nu_k\}$? Russell [14] has carried out a study of this question in the case corresponding to $m = 1$ above for a class of linear hyperbolic systems consisting of a pair of equations. In his study, the theory of nonharmonic Fourier series is used to study the canonical equation (0.3) and to conclude that any sequence $\{\nu_k\}$ for which

$$(0.8) \quad \left\{ \frac{(\omega_k - \nu_k)}{b_k} \right\} \in l^2,$$

where $\{b_k\}$ is the sequence of expansion coefficients of the control distribution element ℓ with respect to the eigenfunctions of \mathcal{A} , may be synthesized by continuous linear state feedback. An alternative approach, which apparently avoids an appeal to the theory of nonharmonic Fourier series, has been offered by Clarke and Williamson [2]. Perturbation theorems for spectral operators have been employed by Sun [18] to show that (0.8) is both a necessary and sufficient condition for a class of systems that would include the one studied in this paper. The main contributions of this paper include a representation of the transformation carrying (0.2) into (0.3) more explicit than that given in [14] as well as a formula for the expansion coefficients of (0.8) in terms of the desired set of eigenvalues $\{\nu_k\}$.

1. A discrete finite dimensional system. For the sake of motivation, we will review briefly the reduction of the finite dimensional discrete system, where $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$,

$$(1.1) \quad x_k = Ax_{k-1} + bu_k,$$

to its causal canonical scalar equation

$$(1.2) \quad y_k + a_1 y_{k-1} + \cdots + a_n y_{k-n} = u_k,$$

where $\det(\lambda I - A) \equiv \lambda^n + a_1 \lambda^{n-1} + \cdots + a_n$.

For $r = 1, \cdots, n$, we have

$$(1.3) \quad x_{k+r} = A^r x_k + \sum_{l=1}^r A^{r-l} b u_{k+l}.$$

In particular, if $x_0 = 0$,

$$(1.4) \quad x_n = [A^{n-1}b \quad \cdots \quad Ab \quad b] \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \equiv Cu.$$

It is well known that the pair (A, b) is controllable only when the $n \times n$ matrix C is invertible.

With $a_0 \equiv 1$, we find

$$\begin{aligned}
 x_{k+n} + a_1 x_{k+n-1} + \cdots + a_n x_k &= \sum_{m=0}^{n-1} a_m x_{k+n-m} + a_n x_k \\
 (1.5) \qquad &= \sum_{m=0}^{n-1} a_m \left(A^{n-m} x_k + \sum_{l=1}^{n-m} A^{n-m-l} b u_{k+l} \right) + a_n x_k \\
 &= p(A) x_k + \sum_{l=1}^n \sum_{m=0}^{n-l} a_m A^{n-m-l} b u_{k+l},
 \end{aligned}$$

where $p(A) = A^n + a_1 A^{n-1} + \cdots + a_n I = 0$ by the Cayley–Hamilton theorem, and where we have interchanged the order of summation in obtaining the last line. The remaining double sum may be rewritten as

$$\sum_{l=1}^n \sum_{m=0}^{n-l} a_m A^{n-m-l} b u_{k+l} = CM \begin{bmatrix} u_{k+1} \\ \vdots \\ u_{k+n} \end{bmatrix},$$

where

$$(1.6) \qquad M = \begin{bmatrix} 1 & & & & 0 \\ & a_1 & & & \\ & \vdots & & & \\ & a_{n-1} & \cdots & a_1 & 1 \end{bmatrix}.$$

Equation (1.5) now reads

$$(1.7) \qquad x_{k+n} + a_1 x_{k+n-1} + \cdots + a_n x_k = CM \begin{bmatrix} u_{k+1} \\ \vdots \\ u_{k+n} \end{bmatrix}.$$

It is important to note that the coefficients appearing on the left-hand side of (1.7) are scalars. Assuming the pair (A, b) to be controllable, or equivalently, the matrix C to be invertible, we define the variable η by

$$(1.8) \qquad \begin{bmatrix} \eta_{k,1} \\ \vdots \\ \eta_{k,n} \end{bmatrix} = M^{-1} C^{-1} x_{k+n}.$$

It is also important that $\eta_{k,r}$ depends only on $k+r$. This may be demonstrated as follows. Let e_r denote the r th standard unit vector in \mathbb{R}^n ; thus

$$\eta_{k,r} = e_r^T M^{-1} C^{-1} x_{k+n}.$$

According to (1.8),

$$\begin{bmatrix} \eta_{k+p,1} \\ \vdots \\ \eta_{k+p,n} \end{bmatrix} = M^{-1} C^{-1} x_{k+p+n}.$$

Suppose $p+q=r$. We wish to show that

$$\eta_{k,r} = \eta_{k+p,q},$$

i.e., that

$$e_r^T M^{-1} C^{-1} x_{k+n} = e_q^T M^{-1} C^{-1} x_{k+p+n}.$$

Since

$$x_{k+p+n} = A^p x_{k+n} + \sum_{l=1}^p A^{p-l} b u_{k+l+n},$$

we must have:

$$(i) \quad e_r^T = e_q^T M^{-1} C^{-1} A^p C M \equiv e_q^T \hat{A}^p;$$

$$(ii) \quad e_q^T M^{-1} C^{-1} \sum_{l=1}^p A^{p-l} b u_{k+n-l} = 0.$$

That (i) is true follows from the well-known fact (see e.g., [10, pp. 199–201] or [16, p. 126]) that \hat{A} is the companion matrix of A :

$$\hat{A} = M^{-1} C^{-1} A C M = \begin{bmatrix} 0 & 1 & & & 0 \\ 0 & 0 & 1 & & \\ \vdots & & & \ddots & \\ 0 & 0 & \cdots & 0 & 1 \\ -a_n & & \cdots & -a_2 & -a_1 \end{bmatrix}.$$

That (ii) holds true follows from the observation that

$$C^{-1} A^{p-l} b = e_{n-p+l}$$

and the fact that $\text{span}\{e_{n-p+b}, \dots, e_n\}$ is invariant under M^{-1} .

We now define the sequence $\{y_k\} \subset \mathbb{R}$ by

$$y_{k+1} = \eta_{k,1}.$$

Equation (1.8) now reads

$$\begin{bmatrix} y_{k+1} \\ \vdots \\ y_{k+n} \end{bmatrix} = M^{-1} C^{-1} x_{k+n}.$$

Multiplying (1.7) on the left by $M^{-1} C^{-1}$, we find that $\{y_k\}$ satisfies (1.2).

Let us now see how linear feedback in (1.1) manifests itself in (1.2). If

$$u_k = f^T x_{k-1}, \quad f^T = \text{row vector},$$

then, by (1.8),

$$(1.9) \quad u_k = f^T C M \begin{bmatrix} y_{k-n} \\ \vdots \\ y_{k-1} \end{bmatrix} \equiv g^T \begin{bmatrix} y_{k-n} \\ \vdots \\ y_{k-1} \end{bmatrix},$$

where $g^T \equiv (g_n, \dots, g_1)$. Thus, the closed-loop system

$$x_k = (A + b f^T) x_{k-1},$$

goes over to the closed-loop system

$$(1.10) \quad y_k + (a_1 - g_1) y_{k-1} + \cdots + (a_n - g_n) y_{k-n} = 0.$$

Let $\lambda_1, \dots, \lambda_n$ denote the (not necessarily distinct) eigenvalues of A . We are assuming that the entries in A are all real; hence $\{\lambda_1, \dots, \lambda_n\}$ is symmetric with respect to the real axis in the complex plane. Given any likewise symmetric set of complex

numbers $\{\nu_1, \dots, \nu_n\}$, it is now an easy matter to construct a feedback gain $f \in \mathbb{R}^n$ for which the eigenvalues of $A + bf^T$ are precisely $\{\nu_1, \dots, \nu_n\}$. Indeed, if

$$\lambda^n + c_1\lambda^{n-1} + \dots + c_n \equiv \prod_{i=1}^n (\lambda - \nu_i),$$

then $c_i \in \mathbb{R}$ for $i = 1, \dots, n$ and hence, referring to (1.10), we must have

$$g_i = a_i - c_i, \quad i = 1, \dots, n$$

or

$$g^T = a^T - c^T, \quad \text{where } a^T = (a_n, \dots, a_1), \text{ etc.}$$

Referring to (1.9), we arrive at the Bass–Gura formula [10, pp. 199–201] for f :

$$(1.11) \quad f^T = (a - c)^T M^{-1} C^{-1}.$$

We will show that the same formula holds true for the class of infinite dimensional systems mentioned above, under appropriate assumptions.

2. Nonharmonic Fourier series and controllability. In this section, we discuss the nature of the transformation which corresponds to the matrix C of the previous section. Before doing so, it is first necessary to recall some definitions which concern the scalar input system

$$(2.1) \quad \dot{x}(t) = \mathcal{A}x(t) + \ell u(t).$$

Let \mathcal{A} generate a strongly-continuous group $\{\mathcal{G}(t): t \in \mathbb{R}\}$ on the Hilbert space \mathcal{H} , and denote the action of a conjugate linear functional x on $\psi \in \mathcal{H}$ by $\langle \psi, x \rangle$. As with the inner product (\cdot, \cdot) on \mathcal{H} , the bracket $\langle \cdot, \cdot \rangle$ is understood to be conjugate linear in the first argument and linear in the second. It is also understood that $\langle \cdot, \cdot \rangle$ has the property that $\langle \psi, x \rangle = (\psi, x)$ if $x \in \mathcal{H}$. Let us define the weak solution $x(\cdot)$ to (2.1) satisfying $x(0) = x_0 \in \mathcal{H}$ by requiring

$$(2.2) \quad \langle \psi, x(t) \rangle = \langle \mathcal{G}(t)^* \psi, x_0 \rangle + \int_0^t \langle \mathcal{G}(t-s)^* \psi, \ell \rangle u(s) ds,$$

for all $\psi \in \mathcal{D}(\mathcal{A}^*)$. Here, $\{\mathcal{G}(t)^*: t \in \mathbb{R}\}$ denotes the adjoint group generated by \mathcal{A}^* with domain $\mathcal{D}(\mathcal{A}^*) \subset \mathcal{H}$, and ℓ is understood to represent a conjugate linear functional on \mathcal{H} .

DEFINITION 2.1 [4]. The control distribution element ℓ is said to be *admissible with respect to \mathcal{A}* if ℓ can be identified with a conjugate linear functional, also denoted by ℓ , for which

(i) $\mathcal{D}(\ell) \supseteq \mathcal{D}(\mathcal{A}^*)$.

(ii) For any initial state $x(0) \in \mathcal{H}$ and any locally square integrable input u , the weak solution $x(t)$ of (2.1) lies in \mathcal{H} for each $t \geq 0$.

We note that if $\ell \in \mathcal{H}$ in (2.1), then the weak solution coincides with the mild solution

$$(2.3) \quad x(t) = \mathcal{G}(t)x_0 + \int_0^t \mathcal{G}(t-s)\ell u(s) ds.$$

Since $x(t) \in \mathcal{H}$ for each $t \geq 0$ [5, p. 159], ℓ in this case is admissible.

To make (2.2) somewhat more constructive, we introduce the sequence $\{\phi_k\}$ of eigenfunctions of \mathcal{A} with corresponding eigenvalues $\{\omega_k\}$. We will assume that each ω_k is simple and that $\{\phi_k\}$ forms a Riesz basis for \mathcal{H} .

DEFINITION 2.2 ([7], [11]). A sequence $\{\phi_k\}$ of elements of a Hilbert space is said to form a *Riesz basis* for \mathcal{H} if there exist constants d, D with $0 < d \leq D < \infty$ such that every $x \in \mathcal{H}$ may be expanded uniquely in a series

$$(2.4) \quad x = \sum_k x_k \phi_k \quad \text{with} \quad d^2 \sum_k |x_k|^2 \leq \|x\|_{\mathcal{H}}^2 \leq D^2 \sum_k |x_k|^2.$$

Equivalently (see, e.g., [7]), a Riesz basis may be thought of as the image of an orthonormal basis $\{e_k\}$ of \mathcal{H} under a bounded and boundedly invertible transformation Φ

$$\phi_k = \Phi e_k.$$

With every Riesz basis $\{\phi_k\}$, there is associated a unique “dual” Riesz basis $\{\psi_k\}$ defined by

$$\psi_k = (\Phi^*)^{-1} e_k,$$

with the biorthonormality property $(\psi_k, \phi_l)_{\mathcal{H}} = \delta_{kl}$. The expansion coefficients of $x \in \mathcal{H}$ with respect to $\{\phi_k\}, \{\psi_k\}$ are $(\psi_k, x), (x, \phi_k)$, respectively.

Let us denote by $\{\psi_k\}$ the sequence of eigenfunctions of \mathcal{A}^* ; i.e., $\mathcal{A}^* \psi_k = \bar{\omega}_k \psi_k$. Suitably normalized, $\{\psi_k\}$ is easily seen to be dual to $\{\phi_k\}$. Now let $\psi = \psi_k$ in (2.2) and set $x_k(t) = \langle \psi_k, x(t) \rangle$ and $b_k = \langle \psi_k, \ell \rangle$. Then

$$(2.5) \quad x_k(t) = e^{\omega_k t} x_k(0) + b_k \int_0^t e^{\omega_k(t-s)} u(s) ds,$$

and hence, weak solutions may be represented as

$$x(t) = \sum_k x_k(t) \phi_k,$$

with the x_k 's given by (2.5). If $\sup |\operatorname{Re} \omega_k| < \infty$, admissibility reduces, via (2.4), to whether or not, for each $t \geq 0$, the sequence

$$\left\{ b_k \int_0^t e^{\omega_k(t-s)} u(s) ds \right\} \in l^2,$$

for every $u \in L^2(0, t)$.

In the course of reducing (2.1) to canonical form, we will carry out manipulations similar to those employed in the previous section and which involve the expression

$$\int_0^T \mathcal{G}(T-s) \ell u(s) ds.$$

The whole point of the above discussion is that the latter expression may, even if $\ell \notin \mathcal{H}$, be interpreted rigorously if the eigenfunctions $\{\phi_k\}$ of \mathcal{A} form a Riesz basis for \mathcal{H} and ℓ is admissible with respect to \mathcal{A} :

$$\int_0^T \mathcal{G}(t-s) \ell u(s) ds \equiv \sum_k \left(b_k \int_0^t e^{\omega_k(t-s)} u(s) ds \right) \phi_k.$$

In particular, if $x_0 = 0$, we may write

$$(2.6) \quad x(T) = \int_0^T \mathcal{G}(T-s) \ell u(s) ds \equiv \mathcal{C}(T) u.$$

This is the analogue of (1.4). Conversely, if $x \in \mathcal{H}$, the problem of constructing a control

$u \in L^2(0, T)$, which takes the origin to x at time T , is characterized by a moment problem

$$\mathcal{C}(T)u = x,$$

or, with $x = \sum_k x_k \phi_k$,

$$(2.7) \quad b_k \int_0^T e^{\omega_k(T-s)} u(s) ds = x_k.$$

DEFINITION 2.3 [15]. The pair (\mathcal{A}, θ) is said to be *approximately (exactly) controllable* in time T if $\mathcal{C}(T): L^2(0, T) \rightarrow \mathcal{H}$ is densely (boundedly) invertible.

Referring to (2.7), it is obvious that approximate controllability requires $b_k \neq 0$ for all k . But controllability also depends upon properties of the sequence of exponentials $\{e^{\omega_k}\}$. This paper is concerned with a very special class of \mathcal{A} 's, namely those whose spectrum coincides with the zero set $\{\omega_k\}$ of a function having the form

$$(2.8) \quad p(\omega) = e^{\omega T} + a_1 e^{\omega(T-\theta_1)} + \cdots + a_m + \int_0^T a(\theta) e^{\omega(T-\theta)} d\theta.$$

It is assumed that $a_i \in \mathbb{R}$, $i = 1, \dots, m$ with $a_m \neq 0$, that $a(\cdot) \in L^2(0, T)$ is real-valued, and that

$$0 < \theta_1 < \cdots < \theta_m \equiv T.$$

Let us write

$$p(\omega) = p_0(\omega) + \int_0^T a(\theta) e^{\omega(T-\theta)} d\theta,$$

and let $\{\sigma_k\}$ denote the zero set of the exponential polynomial p_0 . It is easy to see that the set $\{\sigma_k\}$ as well as the set $\{\omega_k\}$ must lie in a vertical strip of finite width in the complex plane. It has been shown [1], using the argument principle, that an infinite number of σ_k 's do exist and that the number of σ_k 's in any horizontal strip of fixed height is bounded.

The prototype in this situation is $p_0(\omega) = e^{2\omega} - 1$ which has for its set of zeros $\{\sigma_k\} = \{k\pi i\}$. It can also be shown [17], [19], using the above properties of $\{\sigma_k\}$ together with the Fourier transform in the complex domain, that the sequence of exponentials $\{e^{\sigma_k}\}$ forms a Riesz basis for $L^2(0, T)$ if, in addition, $\inf_k |p'_0(\sigma_k)| > 0$. The latter condition ensures the bounded invertibility of the map Φ discussed after Definition 2.2. One can also demonstrate using Rouché's theorem that the sequence $\{\omega_k\}$ of zeros of p may be indexed in such a way that $\{\omega_k - \sigma_k\} \in l^2$ and that $\inf_k |p'(\omega_k)| > 0$ is sufficient to yield the Riesz basis property for the sequence $\{e^{\omega_k}\}$.

We omit the somewhat detailed proof of these statements [19] and return to their relationship with the notion of controllability.

Let us denote the sequence biorthonormal to $\{e^{\omega_k}\}$ by $\{q_l(\cdot)\}$:

$$\int_0^T \bar{q}_l(s) e^{\omega_k s} ds = \delta_{kl}.$$

The construction of the sequence $\{q_l(\cdot)\}$ goes back to the work of Paley and Wiener [12] and is carried out as follows. One forms the function

$$\hat{q}_k(\omega) = \frac{p(\omega)}{p'(\omega_k)(\omega - \omega_k)},$$

and notes that $\hat{q}_l(\omega_k) = \delta_{kl}$. The Paley-Wiener theorem ([12], [13, Chapt. 19]) then

asserts that \hat{q}_l is the Laplace transform of a square integrable function $\bar{q}_l(\cdot)$ whose support $\text{supp}(q_l) \subseteq [0, T]$:

$$\hat{q}_l(\omega) = \int_0^T e^{\omega s} \bar{q}_l(s) ds.$$

Thus, \bar{q}_l may be represented via the inverse Laplace transform

$$(2.9) \quad \bar{q}_l(s) = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} e^{-s\omega} \hat{q}_l(\omega) d\omega.$$

The sequence $\{q_l(\cdot)\}$ so constructed is biorthonormal to $\{e^{\omega_k \cdot}\}$, since

$$\int_0^T \bar{q}_l(s) e^{\omega_k s} ds = \hat{q}_l(\omega_k) = \delta_{kl}.$$

However, in the next section, we shall derive an explicit formula for each q_l without any mention of the Paley–Wiener theory. In effect, the inverse transform (2.9) can be computed explicitly in closed form.

Assuming $b_k \neq 0$ for any k , we may now invert relationship (2.7) with the aid of the sequence $\{q_k(\cdot)\}$. Indeed, the control

$$u(t) = \sum_k \frac{x_k}{b_k} \bar{q}_k(T-t),$$

satisfies (2.7). Here, T is understood to be the quantity appearing in the definition (2.8) of p . Thus $\mathcal{C} \equiv \mathcal{C}(T)$ has the properties

$$(2.10) \quad \mathcal{C}: \bar{q}_k(T-\cdot) \rightarrow b_k \phi_k,$$

$$(2.11) \quad \mathcal{C}^{-1}: \phi_k \rightarrow b_k^{-1} \bar{q}_k(T-\cdot),$$

which will be used throughout the remainder of the paper.

We conclude this section with a summary of the various assumptions made and a brief discussion of their significance.

ASSUMPTION 2.4.

- (i) \mathcal{A} generates a strongly-continuous group $\{\mathcal{G}(t): t \in \mathbb{R}\}$ on a Hilbert space \mathcal{H} .
- (ii) The sequence of eigenfunctions $\{\phi_k\}$ of \mathcal{A} forms a Riesz basis for \mathcal{H} .
- (iii) The eigenvalues $\{\omega_k\}$ of \mathcal{A} are the roots of $p(\omega)$ given by (2.8) and $p'(\omega_k) \neq 0$ for any k .
- (iv) The control distribution element ℓ is admissible with respect to \mathcal{A} and has coefficients $b_k = \langle \psi_k, \ell \rangle \neq 0$ for any k ($\mathcal{A}^* \psi_k = \bar{\omega}_k \psi_k$).

Since finite linear combinations of the ϕ_k 's are dense in \mathcal{H} , the pair (\mathcal{A}, ℓ) , under the above assumption, is approximately controllable in time T (cf. [15]). If $\inf_k |b_k| > 0$ and $\inf_k |p'(\omega_k)| > 0$, exact controllability in time T may be established as well [15].

Part (iii) of the above assumption essentially limits the scope of our results to the class of linear hyperbolic systems described in the introduction. For such systems, $p'(\omega_k) \neq 0$ for all k does not hold true in general. We make this hypothesis here only to simplify our presentation—notwithstanding a good dose of tedium, it can be shown by introducing generalized exponentials (i.e. $t^j e^{\omega_k t}$) in the case of multiple roots ω_k that the canonical form presented in this paper remains valid.

The case of boundary control (see (0.7)), where $b \in \mathbb{R}^n$, does correspond to an admissible distribution ℓ [15, Thm. 3.1]

$$b_k = \langle \psi_k, \ell \rangle = \psi_k^+(1) * \Lambda^+(1) b,$$

and, if no component of b vanishes, approximate controllability may be established [19]. We do not wish to emphasize the abstract character of the above assumptions. However, the development of a canonical form is carried out most transparently in this setting.

3. Reduction to canonical form. In this section, we will derive the transformation which carries the system (0.2) to its canonical form (0.3), and in order to pursue as closely as possible the analogy between the pair (A, b) of the discrete system above with the pair (\mathcal{A}, ℓ) at hand, we will proceed somewhat formally. Assumption 2.4 will be understood to be in force throughout. We begin with an analogue of the Cayley-Hamilton theory.

LEMMA 3.1. *Under (i), (ii) and (iii) of Assumption 2.4,*

$$\mathcal{G}(t) + a_1 \mathcal{G}(t - \theta_1) + \cdots + a_m \mathcal{G}(t - T) + \int_0^T a(\theta) \mathcal{G}(t - \theta) d\theta \equiv 0.$$

Proof. Given an arbitrary initial state $x = \sum_k x_k \phi_k \in \mathcal{H}$, the solution to $\dot{x}(t) = \mathcal{A}x(t)$ is given formally by $x(t) = \mathcal{G}(t)x$ and concretely by

$$x(t) = \sum_k e^{\omega_k t} x_k \phi_k.$$

Thus

$$\begin{aligned} & \left[\mathcal{G}(t) + a_1 \mathcal{G}(t - \theta_1) + \cdots + a_m \mathcal{G}(t - T) + \int_0^T a(\theta) \mathcal{G}(t - \theta) d\theta \right] x \\ &= \sum_k p(\omega_k) e^{\omega_k(t-T)} x_k \phi_k = 0. \end{aligned}$$

This completes the proof of the lemma.

Formally, solutions to $\dot{x}(t) = \mathcal{A}x(t) + \ell u(t)$ satisfy

$$(3.1) \quad x(t+r) = \mathcal{G}(r)x(t) + \int_0^r \mathcal{G}(r-s)\ell u(s) ds.$$

Thus,

$$\begin{aligned} & x(t+T) + a_1 x(t+T-\theta_1) + \cdots + a_m x(t) + \int_0^T a(\theta) x(t+T-\theta) d\theta \\ &= \left[\mathcal{G}(t) + a_1 \mathcal{G}(t-\theta_1) + \cdots + a_m \mathcal{G}(t-T) + \int_0^T a(\theta) \mathcal{G}(t-\theta) d\theta \right] x(T) \\ (3.2) \quad & + \int_0^T \mathcal{G}(T-s)\ell u(t+s) ds + a_1 \int_0^{T-\theta_1} \mathcal{G}(T-\theta_1-s)\ell u(t+s) ds \\ & + \cdots + a_{m-1} \int_0^{T-\theta_{m-1}} \mathcal{G}(T-\theta_{m-1}-s)\ell u(t+s) ds \\ & + \int_0^T a(\theta) \int_0^{T-\theta} \mathcal{G}(T-\theta-s)\ell u(t+s) ds d\theta. \end{aligned}$$

The first term is zero by Lemma 3.1. Changing the inner variable of integration from s to $s + \theta$ and then interchanging the order of integration, the last term may be written

as

$$\int_0^T \mathcal{G}(T-s) \ell \int_0^s a(\theta) u(t+s-\theta) d\theta ds.$$

Likewise,

$$\begin{aligned} \int_0^{T-\theta_i} \mathcal{G}(T-\theta_i-s) \ell u(t+s) ds &= \int_0^T \mathcal{G}(T-\theta_i-s) \ell \chi_{[0, T-\theta_i]}(s) u(t+s) ds \\ &= \int_0^T \mathcal{G}(T-s) \ell \chi_{[0, T-\theta_i]}(s-\theta_i) u(t+s-\theta_i) ds \\ &= \int_0^T \mathcal{G}(T-s) \ell \chi_{[\theta_i, T]}(s) u(t+s-\theta_i) ds, \end{aligned}$$

where $\chi_{[a,b]}(\cdot)$ is the characteristic function of the interval $[a, b]$. The right-hand side of (3.2) may now be written as

$$\int_0^T \mathcal{G}(T-s) \ell (\mathcal{M}u(t+\cdot))(s) ds \equiv \mathcal{C}\mathcal{M}u(t+\cdot),$$

where \mathcal{C} is as before and $\mathcal{M}: L^2(0, T) \rightarrow L^2(0, T)$ is the transformation defined by

$$\begin{aligned} (3.3) \quad (\mathcal{M}y)(s) &= y(s) + a_1 \chi_{[\theta_1, T]}(s) y(s-\theta_1) + \cdots \\ &\quad + a_{m-1} \chi_{[\theta_{m-1}, T]}(s) y(s-\theta_{m-1}) + \int_0^s a(\theta) y(s-\theta) d\theta. \end{aligned}$$

Equation (3.2) now reads

$$(3.4) \quad x(t+T) + a_1 x(t+T-\theta_1) + \cdots + a_m x(t) + \int_0^T a(\theta) x(t+T-\theta) d\theta = \mathcal{C}\mathcal{M}u(t+\cdot).$$

Equations (3.3) and (3.4) should be compared with (1.6) and (1.7), respectively. To proceed further, the nature of the transformation \mathcal{M} must be analyzed.

LEMMA 3.2. (i) \mathcal{M} is bounded and boundedly invertible.

(ii) If $g \in L^2(0, T)$ with $\text{supp}(g) \subset [T-\theta, T]$ for some $\theta \in (0, T)$, then $\text{supp}(\mathcal{M}^{-1}g) \subset [T-\theta, T]$ as well.

Proof. That \mathcal{M} is bounded is clear from (3.3). Referring to the latter, let us write

$$\mathcal{M} = \mathcal{J} - \mathcal{N}_1 - \mathcal{N}_2,$$

with $(\mathcal{N}_2 y)(s) \equiv -\int_0^s a(\theta) y(s-\theta) d\theta$. A standard theorem from integral equations [8, p. 33] yields the fact that $(\mathcal{J} - \mathcal{N}_2)^{-1}$ exists and is bounded with

$$(\mathcal{J} - \mathcal{N}_2)^{-1} = \mathcal{J} + \sum_{k=1}^{\infty} \mathcal{N}_2^k.$$

It is easily established that $\text{supp}(\mathcal{N}_2 g)$ and hence $((\mathcal{J} - \mathcal{N}_2)^{-1} g) \subseteq [T-\theta, T]$ if $\text{supp}(g) \subseteq [T-\theta, T]$. Moreover, \mathcal{N}_1 is nilpotent since $\text{supp}(\mathcal{N}_1 g) \subseteq [T-\theta+\theta_1, T]$ if $\text{supp}(g) \subseteq [T-\theta, T]$. Thus $(\mathcal{J} - \mathcal{N}_2)^{-1} \mathcal{N}_1$ is nilpotent and

$$\mathcal{M}^{-1} = (\mathcal{J} - \mathcal{N}_2)^{-1} (\mathcal{J} - (\mathcal{J} - \mathcal{N}_2)^{-1} \mathcal{N}_1)^{-1} = (\mathcal{J} - \mathcal{N}_2)^{-1} \left(\mathcal{J} + \sum_{k=1}^n ((\mathcal{J} - \mathcal{N}_2)^{-1} \mathcal{N}_1)^k \right),$$

for some $n < \infty$, which shows that \mathcal{M}^{-1} is bounded and satisfies (ii). This completes the proof.

We remark that Lemma 3.2(ii) is the analogue of the fact that $\text{span}\{e_{n-p}, \dots, e_n\}$ is invariant under the matrix M^{-1} discussed in § 1. We next Fourier analyze the action of \mathcal{M} on the exponentials e^{ω_k} .

LEMMA 3.3.

$$\int_0^T e^{\omega(t-s)} (\mathcal{M} e^{\omega_k})(s) ds = \frac{p(\omega)}{\omega - \omega_k}.$$

Proof. In forming the expression on the left, we find terms of the form

$$\int_0^T e^{\omega(T-s)} \chi_{[\theta_i, T]}(s) e^{\omega_k(s-\theta_i)} ds,$$

and the term

$$\int_0^T e^{\omega_k(T-s)} \int_0^s a(\theta) e^{\omega_k(s-\theta)} d\theta ds.$$

Standard manipulations reduce the former to

$$\frac{e^{\omega(T-\theta_i)} - e^{\omega_k(T-\theta_i)}}{\omega - \omega_k},$$

and the latter to

$$\int_0^T a(\theta) \frac{e^{\omega(T-\theta)} - e^{\omega_k(T-\theta)}}{\omega - \omega_k} d\theta.$$

It follows that

$$\begin{aligned} \int_0^T e^{\omega(T-s)} (\mathcal{M} e^{\omega_k})(s) ds &= \left[e^{\omega T} - e^{\omega_k T} + a_1(e^{\omega(T-\theta_1)} - e^{\omega_k(T-\theta_1)}) \right. \\ &\quad \left. + \dots + a_{m-1}(e^{\omega(T-\theta_{m-1})} - e^{\omega_k(T-\theta_{m-1})}) \right. \\ &\quad \left. + \int_0^T a(\theta)(e^{\omega(T-\theta)} - e^{\omega_k(T-\theta)}) d\theta \right] / (\omega - \omega_k) \\ &= \frac{(p(\omega) - p(\omega_k))}{(\omega - \omega_k)} = \frac{p(\omega)}{(\omega - \omega_k)}, \end{aligned}$$

since ω_k is a zero of p . This completes the proof.

Recall that the basis $\{e^{\omega_k}\} \subset L^2(0, T)$ has a unique biorthonormal sequence which we denote by $\{q_k(\cdot)\} \subset L^2(0, T)$. The key property which \mathcal{M} possesses is the following:

LEMMA 3.4.

$$\mathcal{M}: e^{\omega_k} \rightarrow p'(\omega_k) \bar{q}_k(T - \cdot).$$

Proof. According to Lemma 3.3,

$$\int_0^T e^{\omega_k(T-s)} (\mathcal{M} e^{\omega_k})(s) ds = p'(\omega_k) \delta_{jk}.$$

Since the sequence $\{q_j(T - \cdot)\}$ is biorthonormal to $\{e^{\omega_k(T-\cdot)}\}$ and unique, we conclude that

$$(\mathcal{M} e^{\omega_k})(s) = p'(\omega_k) \bar{q}_k(T - s),$$

which completes the proof.

Remark. The preceding lemma allows us to construct the biorthonormal sequence explicitly

$$\bar{q}_k(s) = \frac{1}{p'(\omega_k)} (\mathcal{M} e^{\omega_k \cdot})(T-s).$$

Under Assumption 2.4 (iv), the map \mathcal{C} possesses a densely defined inverse \mathcal{C}^{-1} (see (2.11)). If $x(\cdot)$ is a solution of $\dot{x} = \mathcal{A}x + \mathcal{B}u$ with sufficiently regular initial state x_0 and input $u(\cdot)$, we may define

$$(3.5) \quad \eta(t, \cdot) = (\mathcal{M}^{-1} \mathcal{C}^{-1} x(t+T))(\cdot) \in L^2(0, T).$$

Let us assume that η is defined for every t and s . Then

LEMMA 3.5. $\eta(t, s)$ is a function of $s+t$.

Proof. Suppose $p+s=r < T$. We wish to show that $\eta(t, r) = \eta(t+p, s)$. According to (3.5),

$$\eta(t, r) = (\mathcal{M}^{-1} \mathcal{C}^{-1} x(t+T))(r).$$

If we let $x(t+T) = \sum_k x_k(t+T) \phi_k$ and use (2.11) and Lemma 3.4, we have

$$\eta(t, r) = \sum_k p'(\omega_k)^{-1} b_k^{-1} x_k(t+T) e^{\omega_k r}.$$

Likewise,

$$\eta(t+p, s) = \sum_k p'(\omega_k)^{-1} b_k^{-1} x_k(t+p+T) e^{\omega_k s}.$$

Since

$$x_k(t+p+T) = e^{\omega_k p} x_k(t+T) + b_k \int_0^p e^{\omega_k(p-\theta)} u(t+\theta+T) d\theta,$$

we have

$$\eta(t+p, s) - \eta(t, r) = \sum_k \left(p'(\omega_k)^{-1} \int_0^p e^{\omega_k(p-\theta)} u(t+\theta+T) d\theta \right) e^{\omega_k s}.$$

If we use the change of variable $p-\theta = T-\tau$ and invoke Lemma 3.4, the right-hand side becomes

$$= \left(\mathcal{M}^{-1} \sum_k \int_{T-p}^T e^{\omega_k(T-\tau)} u(t+p+\tau) d\tau \bar{q}_k(T-\cdot) \right)(s).$$

But the function on which \mathcal{M}^{-1} operates is nothing but $\chi_{[T-p, T]}(\cdot) u(t+p+\cdot)$ expanded in terms of the sequence $\{\bar{q}_k(T-\cdot)\}$. Hence

$$= (\mathcal{M}^{-1} \chi_{[T-p, T]}(\cdot) u(t+p+\cdot))(s) \equiv (\mathcal{M}^{-1} g)(s).$$

Clearly, $\text{supp}(g) \subset [T-p, T]$, and since $0 \leq p+s=r < T$ implies $s \in [0, T-p]$, we conclude from Lemma 3.2 that $(\mathcal{M}^{-1} g)(s) = 0$. Thus, $\eta(t+p, s) = \eta(t, r)$. This completes the proof.

Lemma 3.5 may be shown to hold true in the case where $\eta(t, \cdot)$ is merely a square-integrable function for each t , by approximating everything with continuous functions and then passing to the limit. We may now define the scalar variable

$$(3.6) \quad y(t+\cdot) = \eta(t, \cdot) = (\mathcal{M}^{-1} \mathcal{C}^{-1} x(t+T))(\cdot).$$

We conclude this section with:

THEOREM 3.6. *Under Assumption 2.4, the variable y defined by (3.6) satisfies the functional equation*

$$(3.7) \quad y(t) + a_1 y(t - \theta_1) + \cdots + a_m y(t - T) + \int_0^T a(\theta) y(t - \theta) d\theta = u(t).$$

Proof. Apply $\mathcal{M}^{-1}\mathcal{C}^{-1}$ on the left to both sides of (3.4).

4. Spectral synthesis. Let us first examine how continuous state feedback

$$u(t) = (\not{f}, x(t))_{\mathcal{X}}$$

in the original system (2.1) manifests itself in the canonical system (3.7). Let \not{f} have the expansion

$$\not{f} = \sum_j f_j \psi_j,$$

where $\{\psi_j\}$ denotes the sequence of eigenfunctions of \mathcal{A}^* . Let $x(t)$ and $y(t + \cdot)$ have the expansions

$$(4.1) \quad x(t) = \sum_k x_k(t) \phi_k, \quad y(t + \cdot) = \sum_k y_k(t) e^{\omega_k \cdot}.$$

Referring to (3.6), we have

$$x_k(t + T) = b_k p'(\omega_k) y_k(t).$$

Thus,

$$u(t) = (\not{f}, x(t))_{\mathcal{X}} = \sum_k f_k b_k p'(\omega_k) y_k(t - T).$$

Again denoting the sequence biorthonormal to $\{e^{\omega_k \cdot}\}$ by $\{q_j(\cdot)\}$, the right-hand side of the last equation may be written as

$$\int_0^T \left(\sum_j f_j b_j p'(\omega_j) \bar{q}_j(s) \right) \left(\sum_k y_k(t - T) e^{\omega_k s} \right) ds.$$

By (4.1),

$$\sum_k y_k(t - T) e^{\omega_k s} = y(t - T + s).$$

If we change the variable of integration via $s = T - \theta$, and define

$$(4.2) \quad g(\theta) \equiv \sum_j f_j b_j p'(\omega_j) \bar{q}_j(T - \theta),$$

we obtain

$$u(t) = \int_0^T g(\theta) y(t - \theta) d\theta.$$

The resulting closed-loop canonical system is simply

$$(4.3) \quad y(t) + a_1 y(t - \theta_1) + \cdots + a_m y(t - T) + \int_0^T (a(\theta) - g(\theta)) y(t - \theta) d\theta = 0.$$

Thus, the eigenvalues of the closed-loop system

$$x(t) = (\mathcal{A} + \mathcal{C} \otimes \not{f}) x(t)$$

are roots of

$$(4.4) \quad p_g(\omega) = p_0(\omega) + \int_0^T (a(\theta) - g(\theta)) e^{\omega(T-\theta)} d\theta.$$

As we have already mentioned in § 2, one can show using Rouché's theorem that the sequence $\{\nu_k\}$ of roots of p_g differs from the sequence $\{\sigma_k\}$ of roots of p_0 in the sense that $\{\nu_k - \sigma_k\} \in l^2$. This suggests that any sequence $\{\nu_k\}$ with this property can be synthesized with a certain feedback gain $f \in \mathcal{H}$.

LEMMA 4.1. *If $\{\nu_k\}$ is a sequence of distinct complex numbers for which $\{\nu_k - \sigma_k\} \in l^2$, then there exists a unique $c(\cdot) \in L^2(0, T)$ for which the zero set of*

$$p_c(\omega) \equiv p_0(\omega) + \int_0^T c(\theta) e^{\omega(T-\theta)} d\theta$$

coincides with $\{\nu_k\}$.

Proof. c is characterized by the moment problem

$$p_c(\nu_k) = 0 \Rightarrow \int_0^T c(\theta) e^{\nu_k(T-\theta)} d\theta = -p_0(\nu_k).$$

Using the properties of the exponential polynomial p_0 discussed in § 2 together with the assumption that $\{\nu_k - \sigma_k\} \in l^2$, it is readily established that $\{p_0(\nu_k)\} \in l^2$. Moreover, it has been shown [17] that the sequence $\{e^{\nu_k \cdot}\}$ forms a basis for $L^2(0, T)$. Denoting the sequence biorthonormal to $\{e^{\nu_k \cdot}\}$ by $\{h_k(\cdot)\}$, the unique solution to the above moment problem is given by

$$c(\cdot) = -\sum_k p_0(\nu_k) \bar{h}_k(T - \cdot),$$

which completes the proof.

We now come to our main theorem.

THEOREM 4.2. *Let $\{\nu_k\}$ be any sequence of distinct complex numbers for which*

$$\left\{ \frac{\nu_k - \omega_k}{p'(\omega_k) b_k} \right\} \in l^2,$$

and let Assumption 2.4 hold true. Then there exists a unique $f \in \mathcal{H}$ for which the spectrum of $\mathcal{A} + \mathcal{B} \otimes f$ coincides with $\{\nu_k\}$.

Proof. By Lemma 4.1, we may express the desired closed-loop characteristic function uniquely as

$$p_c(\omega) = p_0(\omega) + \int_0^T c(\theta) e^{\omega(T-\theta)} d\theta.$$

Referring to (4.2) and (4.4), the desired feedback gain element

$$f = \sum_k f_k \psi_k$$

must be such that

$$p_g(\omega) \equiv p_0(\omega) + \int_0^T (a(\theta) - g(\theta)) e^{\omega(T-\theta)} d\theta = p_c(\omega),$$

or $g = a - c$. Let a and c have the expansions

$$a(\cdot) = \sum_k a_k \bar{q}_k(T - \cdot), \quad c(\cdot) = \sum_k c_k \bar{q}_k(T - \cdot).$$

Referring to (4.2), we must have

$$f_k b_k p'(\omega_k) = a_k - c_k,$$

or

$$(4.5) \quad f_k = (a_k - c_k) p'(\omega_k)^{-1} b_k^{-1}.$$

This should be compared with (1.11). It remains only to show that our hypotheses ensure that $\{f_k\} \in l^2$. The expansion coefficients a_k and c_k are given by

$$a_k = \int_0^T a(\theta) e^{\omega_k(T-\theta)} d\theta = p(\omega_k) - p_0(\omega_k) = -p_0(\omega_k),$$

$$c_k = \int_0^T c(\theta) e^{\omega_k(T-\theta)} d\theta = p_c(\omega_k) - p_0(\omega_k).$$

Thus, by the mean value theorem,

$$a_k - c_k = -p_c(\omega_k) = -p_c(\nu_k) - p'_c(\nu_k)(\omega_k - \nu_k) = p'_c(\hat{\nu}_k)(\nu_k - \omega_k),$$

for some $\hat{\nu}_k$ on the line segment connecting ν_k and ω_k . Since $\sup_k |p'_c(\hat{\nu}_k)| < \infty$, we conclude that

$$\left\{ \frac{a_k - c_k}{p'(\omega_k) b_k} \right\} \in l^2 \quad \text{just in the case} \quad \left\{ \frac{\nu_k - \omega_k}{p'(\omega_k) b_k} \right\} \in l^2.$$

This completes the proof.

For linear hyperbolic systems of the type described in the introduction, \mathcal{A} is the real operator

$$\mathcal{A} = \Lambda(x) \frac{d}{dx} + A(x)$$

with domain

$$\mathcal{D}(\mathcal{A}) = \{\phi \in H_{2n}^1[0, 1]: \phi^-(0) - D_0 \phi^+(0) = 0 = \phi^+(1) - D_1 \phi^-(1)\},$$

where H_{2n}^1 is the Sobolev space of square integrable \mathbb{R}^{2n} -valued functions possessing a square integrable derivative. As such the eigenfunctions of \mathcal{A} as well as those of \mathcal{A}^* will have the property that $\bar{\phi}_k = \phi_{-k}$. Likewise, if $\{\nu_k\}$ is a sequence of complex numbers with $\bar{\nu}_k = \nu_{-k}$, then any biorthogonal set of functions $\{h_k(\cdot)\}$ associated with the exponentials $\{e^{\nu_k \cdot}\}$ will have the property that $\bar{h}_k = h_{-k}$. Thus, for any such sequence $\{\nu_k\}$ satisfying the hypothesis of Lemma 4.1, the function $c(\cdot)$ will be real-valued with expansion coefficients satisfying $\bar{c}_k = c_{-k}$, and hence the feedback gain element

$$f = \sum_k f_k \psi_k = \sum_k \frac{a_k - c_k}{p'(\omega_k) b_k} \psi_k,$$

constructed above will have each of its $2n$ component functions real-valued.

We conclude this section with a brief discussion of how to use the above theory to construct a feedback gain which shifts a finite number of eigenvalues. Suppose we wish to shift $\omega_1, \omega_2, \dots, \omega_n$ to ν_1, \dots, ν_n . The desired closed-loop characteristic function is thus

$$p_c(\omega) = \frac{\alpha(\omega)}{\beta(\omega)} p(\omega),$$

where $\alpha(\omega) = \prod_{k=1}^n (\omega - \nu_k)$ and $\beta(\omega) = \prod_{k=1}^n (\omega - \omega_k)$. Thus

$$p_c(\omega) = p(\omega) + \frac{\alpha(\omega) - \beta(\omega)}{\beta(\omega)} p(\omega).$$

Since the degree of $\alpha - \beta$ is strictly less than n , $(\alpha - \beta)/\beta$ admits a partial fractions decomposition

$$(4.6) \quad \frac{\alpha(\omega) - \beta(\omega)}{\beta(\omega)} = \sum_{k=1}^n \frac{r_k}{\omega - \omega_k}.$$

According to Lemma 3.3,

$$\frac{p(\omega)}{\omega - \omega_k} = \int_0^T (\mathcal{M} e^{\omega_k \cdot})(\theta) e^{\omega(T-\theta)} d\theta.$$

Thus

$$p_c(\omega) = p(\omega) + \int_0^T \left(\mathcal{M} \sum_{k=1}^n r_k e^{\omega_k \cdot} \right)(\theta) e^{\omega(T-\theta)} d\theta.$$

By Lemma 3.4, this is

$$p_c(\omega) = p(\omega) + \int_0^T \sum_{k=1}^n r_k p'(\omega_k) \bar{q}_k(T-\theta) e^{\omega(T-\theta)} d\theta = p(\omega) + \int_0^T c(\theta) e^{\omega(T-\theta)} d\theta.$$

Thus, the desired expansion coefficients of c are given simply by $c_k = r_k p'(\omega_k)$, with the residues r_k being determined by (4.6).

5. Concluding remarks. In this paper, we have demonstrated precisely to what extent the eigenvalues associated with a controllable pair (\mathcal{A}, ℓ) of a certain type may be modified via continuous linear state feedback. Our results parallel those of Russell [14] for the case $m = 1$, as explained in the introduction. Our main contributions here include a more detailed analysis of the canonical transformation and a direct method for computing feedback gains in the case of shifting a finite number of eigenvalues. All of this requires, of course, a rather detailed knowledge of the spectral structure of \mathcal{A} . In practice, one can only obtain this information approximately by numerical computation. In this regard, our results might prove useful in providing exact solutions in special cases which could then be used to determine the accuracy of numerical computations in more general cases. It should be pointed out that it is unrealistic to assume the availability of state feedback in systems of the type discussed in this paper. Nevertheless, we feel that the results herein obtained should play a role in the development of an observer or asymptotic state estimator theory for such systems.

Acknowledgments. I wish to thank the referees for their valuable criticisms and comments concerning the first draft of this paper.

REFERENCES

- [1] R. BELLMAN AND K. COOKE, *Differential-Difference Equations*, Academic Press, New York, 1963.
- [2] B. M. C. CLARKE AND D. WILLIAMSON, *Control canonical forms and eigenvalue assignment by feedback for a class of linear hyperbolic systems*, this Journal, 19 (1981), pp. 711-729.
- [3] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. II, Interscience, New York, 1962.
- [4] R. F. CURTAIN AND A. J. PRITCHARD, *An abstract theory for unbounded control action for distributed parameter systems*, this Journal, 15 (1977), pp. 566-611.

- [5] R. F. CURTAIN AND A. J. PRITCHARD, *Functional Analysis in Modern Applied Mathematics*, Academic Press, London, 1977.
- [6] S. FRANKEL, *Multiconductor Transmission Line Analysis*, Artech House, San Francisco, 1977.
- [7] J. HIGGINS, *Completeness and Basis Properties of Sets of Special Functions*, Cambridge Univ. Press, New York, 1977.
- [8] H. HOCHSTADT, *Integral Equations*, Wiley-Interscience, New York, 1973.
- [9] IFAC Symposium on the Control and Distributed Parameter Systems, June 1971 Conference Proceedings, Banff, Canada, papers 6-1, 6-2 and 6-3.
- [10] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
- [11] M. G. KREIN AND I. C. GOHBERG, *Introduction to the Theory of Nonself-adjoint Operators*, Translations of Mathematical Monographs, Vol. 18, American Mathematical Society, Providence, RI, 1969.
- [12] R. E. A. C. PALEY AND N. WIENER, *Fourier Transforms in the Complex Domain*, AMS Colloquium Publications 19, American Mathematical Society, New York, 1934.
- [13] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
- [14] D. L. RUSSELL, *Canonical forms and spectral determination for a class of hyperbolic distributed parameter control systems*, J. Math. Anal., 62 (1978), pp. 186-225.
- [15] ———, *Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions*, SIAM Rev. 20 (1978), pp. 639-739.
- [16] ———, *Mathematics of Finite-Dimensional Control Systems*, Marcel Dekker, New York, 1979.
- [17] ———, *Nonharmonic Fourier series in the control theory of distributed parameter systems*, J. Math. Anal. Appl., 18 (1967), pp. 542-559.
- [18] S.-H. SUN, *On spectrum distribution of completely controllable linear systems*, this Journal, 19 (1981), pp. 730-744.
- [19] R. TEGLAS, *A control canonical form for a class of linear hyperbolic systems*, Doctoral thesis, Mathematics Dept., Univ. Wisconsin-Madison, 1981.

A GENERAL APPROACH TO LOWER SEMICONTINUITY AND LOWER CLOSURE IN OPTIMAL CONTROL THEORY*

E. J. BALDER†

Abstract. A self-contained approach to lower semicontinuity and lower closure evolves from an extension of relaxed control theory, which is based on a central relative weak compactness criterion (called tightness) and relaxation in all but one variable. Two lower closure results for outer integral functionals with variable abstract time domain are developed. The first of these has a convexity condition for the integrand and generalizes all similar results in the literature. The second lower closure result is of a new kind; among other things, it implies a quite general version of Fatou's lemma in several dimensions.

Key words. relaxed control theory, tightness, normal integrands, outer integral functionals, lower semicontinuity, lower closure, Fatou's lemma in several dimensions

1. Introduction. This paper presents a rather self-contained approach to lower closure—and lower semicontinuity—in optimal control theory. (An excellent description of the role of lower closure in the existence theory for optimal control has been given in [27a].) Quite similar approaches lead to existence results in other areas of the decision sciences, notably in economics (competitive equilibria, optimal growth theory) and statistics (statistical decision theory); cf. e.g. [3d, h, i, l]. Essentially, this approach is an extension of relaxed control theory [33], [23], [32], [6], [3]. Thus, in our approach the subjects of relaxed control theory and lower closure are brought together.

We shall obtain here essentially two quite general lower closure results. The first step in either proof is the same; it depends on relaxation in all but one variable (cf. [22]) and an associated relative weak compactness criterion, called *tightness*, for sets of measurable functions (considered as parametrized measures); cf. [3]. It turns out that tightness can hold naturally for the trajectories, time domains (variable), “derivative functions”, “singular component functions” and control functions of a control problem; cf. Examples 2.1–2.5. To facilitate the presentation of our results, central results of relaxed control theory have been concentrated in Theorem I, which can be regarded as an extension of the classical theorem by Yu. V. Prokhorov in topological measure theory [7]; cf. [3i].

After this common step the proofs are quickly finished by two quite different continuations. Theorem 3.1 follows by an application of Jensen's inequality, and Theorem 3.7 by applying Lyapunov's theorem. These results are then expanded by the consideration of variable time domains (as introduced in [3g]) and nonmeasurable integrands (by means of Lemma II, first formulated in [3k]). This leads to Theorems 3.3 and 3.8. The former result is a generalization of a well-known lower semicontinuity result (e.g. [16]); cf. Corollary 3.5. In § 4 it is shown to be equivalent to a lower closure result for abstract finite-dimensional orientor fields (Theorem 4.3). As such it generalizes all similar lower closure results in the literature (e.g. [10d, (6.i)], [30a, Thm. 3.1], [11b, Thm. 4.1], [11c, Thm. 3.1], [11d, Thm. 3.1], [3e, Thm. 5]). For a more concrete orientor field it leads to Theorem 4.6, a lower closure result of Lagrange type; this, too, subsumes all similar results in the literature, such as [10d, (6.ii)], [11b, Thm. 4.2], [3e, Thms. 7, 8, 10, Prop. 9]. The other main lower closure result, Theorem 3.8, is more general than Corollary 3.9, the multidimensional Fatou lemma of [3l], which in

* Received by the editors July 31, 1982, and in revised form April 25, 1983.

† Mathematical Institute, University of Utrecht, Utrecht, The Netherlands.

turn subsumes all previous versions of this lemma ([29], [11c], [1b]) as well as certain existence results for allocation problems ([2], [6, Prop. III.2.1], [1a], [3a]).

In our presentation of the major results on the subject of lower closure we have tried to keep the principal results as uncluttered as possible. In subsequent remarks we then present alternative formulations, “extension modules”, etc. We hope that this enables the reader to see the main lines of thought more clearly.

Let us now introduce some conventions and definitions concerning (outer) integration. Let (T, \mathcal{T}, μ) be a finite measure space. The set of all \mathcal{T} -measurable functions from T into $[-\infty, +\infty]$ will be denoted by $\mathcal{M}(T; [-\infty, +\infty])$. For any nonnegative $\phi \in \mathcal{M}(T; [-\infty, +\infty])$ the integral $\int_T \phi d\mu$ —possibly equal to $+\infty$ —is defined in the classical sense [26]. For any $\phi \in \mathcal{M}(T; [-\infty, +\infty])$ we define

$$\int_T \phi d\mu \equiv \int_T \phi^+ d\mu - \int_T \phi^- d\mu,$$

where $\phi^+ \equiv \max(\phi, 0)$, $\phi^- \equiv \max(-\phi, 0)$, with the convention $(+\infty) - (+\infty) = +\infty$. For any $\psi: T \rightarrow [-\infty, +\infty]$, possibly not \mathcal{T} -measurable, the outer integral of ψ over T with respect to μ is defined by

$$\int_T \tilde{\psi} d\mu \equiv \inf \left\{ \int_T \phi d\mu : \phi \in \mathcal{M}(T; [-\infty, +\infty]), \phi \geq \psi \text{ a.e. in } T \right\};$$

it is easy to see that this infimum is attained for some $\phi \in \mathcal{M}(T; [-\infty, +\infty])$, $\phi \geq \psi$ a.e. in T . Also, it is obvious that outer and ordinary integration coincide on $\mathcal{M}(T; [-\infty, +\infty])$.

2. Tightness. Let (T, \mathcal{T}, μ) be a finite measure space and S a standard Borel space (alias metrizable Lusin space [12]). Let $\mathcal{B}(S)$ stand for the Borel σ -algebra on S , and $M_1^+(S)$ for the set of all probability measures on $(S, \mathcal{B}(S))$; equipped with the usual weak (alias narrow) topology, $M_1^+(S)$ is also standard Borel [12, III.60].

The set of all $\mathcal{B}(S)$ -measurable functions from T into S will be denoted by $\mathcal{M}(T; S)$. Instead of $\mathcal{M}(T; M_1^+(S))$ we shall write $\mathcal{R}(T; S)$; the elements of this set are frequently referred to as “parametrized measures”, “relaxed controls”, etc. [23], [32], [6].

An *integrand* on $T \times S$ is a function from $T \times S$ into $(-\infty, +\infty]$. An integrand g on $T \times S$ is said to be *lower semicontinuous* if $g(t, \cdot): t \mapsto g(t, s)$ is lower semicontinuous on S for every $t \in T$ and it is said to be *normal* if it is lower semicontinuous and $\mathcal{T} \times \mathcal{B}(S)$ -measurable. Let $\mathcal{G}(T; S)$ denote the set of all normal integrands on $T \times S$; $\mathcal{G}^+(T; S)$ will then stand for the set of all nonnegative normal integrands on $T \times S$. The subset $\mathcal{H}(T; S)$ of $\mathcal{G}^+(T; S)$ is defined to consist of all $h \in \mathcal{G}^+(T; S)$ such that for every $t \in T$, $\gamma \in \mathbb{R}$

$$\{s \in S: h(t, s) \leq \gamma\} \text{ is compact.}$$

For $s_0 \in \mathcal{M}(T; S)$, $g \in \mathcal{G}(T; S)$ we shall frequently denote the function $t \mapsto g(t, s_0(t))$ by $g(\cdot, s_0)$. A sequence $\{s_k\}_1^\infty \subset \mathcal{M}(T; S)$ is defined to be *tight* if there exists $h \in \mathcal{H}(T; S)$ such that

$$\sup_k \int_T h(t, s_k(t)) \mu(dt) < +\infty.$$

When formulated in terms of $\mathcal{R}(T; S)$, this concept is a generalization of tightness in topological measure theory [7], [3i]; cf. Appendix A and Remark 2.6 below.

The following examples illustrate the various forms in which tightness can manifest itself in the existence theory for optimal control. Let X and V be standard Borel spaces and r, \bar{r} given dimensions.

Example 2.1. Let $\{x_k\}_0^\infty \subset \mathcal{M}(T; X)$ be such that

$$(2.1) \quad x_k(t) \rightarrow x_0(t) \text{ a.e. in } T.$$

Then $\{x_k\}_1^\infty$ is tight, as we can see as follows. Let N stand for the exceptional null set in (2.1). For $t \in T \setminus N$ we define

$$h(t, x) \equiv \begin{cases} 0 & \text{if } x \in \{x_k(t)\}_0^\infty, \\ +\infty & \text{else.} \end{cases}$$

For $t \in N$ we define

$$h(t, x) \equiv \begin{cases} 0 & \text{if } x = x_0(t), \\ +\infty & \text{else.} \end{cases}$$

Then $h \in \mathcal{H}(T; X)$, as is easy to see. Also,

$$\sup_k \int_T h(\cdot, x_k) d\mu = 0.$$

Example 2.2. Let $\{\xi_k\}_0^\infty \subset \mathcal{L}_1(T; \mathbb{R}^r)$ be such that

$$(2.2) \quad \{\xi_k\}_1^\infty \text{ converges weakly in } \sigma(\mathcal{L}_1^r, \mathcal{L}_\infty^r) \text{ to } \xi_0.$$

Then $\{\xi_k\}_1^\infty$ is tight, as can be seen, for instance, by applying de la Vallée-Poussin's theorem [12, II.22, 25]. By this result there exists a lower semicontinuous function $h': \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that $h'(\gamma)/\gamma \rightarrow +\infty$ as $\gamma \rightarrow +\infty$ and

$$\sup_k \int_T h'(|\xi_k(t)|) \mu(dt) < +\infty,$$

where $|\cdot|$ indicates the usual Euclidean norm. Now take $h(t, \xi) \equiv h'(|\xi|)$.

Example 2.3. Let $\{\eta_k\}_1^\infty \subset \mathcal{L}_1(T; \mathbb{R}^{\bar{r}})$ be such that

$$(2.3) \quad \sup_k \int_T |\eta_k| d\mu < +\infty.$$

Then $\{\eta_k\}_1^\infty$ is tight, as is evident from setting $h(t, \eta) \equiv |\eta|$.

Example 2.4. Let $\{v_k\}_1^\infty \subset \mathcal{M}(T; V)$. Then we shall have

$$(2.4) \quad \{v_k\}_1^\infty \text{ is tight,}$$

when, for instance,

$$\{v_k(t)\}_1^\infty \text{ is relatively compact a.e. in } T.$$

This is seen by introducing

$$h(t, v) \equiv \begin{cases} 0 & \text{if } v \in \text{cl} \{v_k(t): k \in \mathbb{N}\}, \\ +\infty & \text{else.} \end{cases}$$

The above examples will play roles further on. So as to illustrate the connection of tightness with topological measure theory, we consider one more example.

Example 2.5. Let Z be a Polish space and let $\{z_k\}_0^\infty \subset \mathcal{M}(T; Z)$ be such that

$$\{\mu_k\}_1^\infty \text{ converges weakly to } \mu_0,$$

where μ_k stands for the image of the measure μ under z_k . Then $\{z_k\}_1^\infty$ is tight, as seen by applying Prokhorov's theorem [7, Thm. 6.2]. By this result there exists for every $p \in \mathbb{N}$ a compact subset K_p of Z such that

$$\sup_k \mu_k(Z \setminus K_p) \leq 3^{-p}.$$

Setting

$$h(t, z) \equiv \begin{cases} 2^p & \text{if } z \in K_p \setminus \left(\bigcup_{j \leq p-1} K_j \right), \quad p \in \mathbb{N}, \\ +\infty & \text{if } z \in Z \setminus \left(\bigcup_{p=1}^\infty K_p \right), \end{cases}$$

we see that h belongs to $\mathcal{H}(T; Z)$ and

$$\sup_k \int_T h(\cdot, z_k) d\mu \leq 2\mu(T) + 4.$$

Remark 2.6. It is easy to see from the previous example that a subset Ω of $M_1^+(Z)$ is tight in the sense of topological measure theory [7, p. 37] if and only if there exists a function $h: Z \rightarrow [0, +\infty]$ such that $\{z \in Z: h(z) \leq \gamma\}$ is compact for every $\gamma \in \mathbb{R}$ and

$$\sup_{\nu \in \Omega} \int_Z h d\nu < +\infty.$$

This shows that our definition of tightness is a generalization of the classical one.

Good reasons for considering the tightness property will be produced now (and in later proofs). In Appendix A it is shown that tightness implies relative sequential compactness in some suitable topology on $\mathcal{R}(T; S)$; this actually generalizes one half of Prokhorov's theorem [7, Thms. 6.1, 6.2]. (Generalization of the other half is almost trivial; cf. Example 2.5.) The gist of this can be formulated as follows.

THEOREM I. *Suppose that $\{s_k\}_1^\infty \subset \mathcal{M}(T; S)$ is tight. Then there exist a subsequence $\{\ell\}$ of $\{k\}$ and a parametrized measure $\delta_* \in \mathcal{R}(T; S)$ such that for every $g \in \mathcal{G}(T; S)$*

$$(2.5) \quad \lim_{\ell} \int_T g(\cdot, s_\ell) d\mu \equiv \int_T g(\cdot, \delta_*) d\mu = \int_T \left[\int_S g(t, s) \delta_*(t)(ds) \right] \mu(dt),$$

provided that

$$(2.6) \quad \{g^-(\cdot, s_\ell)\} \text{ is uniformly integrable.}$$

Moreover, for a.e. $t \in T$ the measure $\delta_*(t)$ is carried by the set

$$\bigcap_{p=1}^\infty \text{cl} \{s_\ell(t): \ell \geq p\}$$

of all limit points of $\{s_\ell(t)\}$.

Remark 2.7. The following obvious addition can be made in Theorem I: the generalized limit δ_* of the subsequence $\{s_\ell\}$ satisfies

$$(2.7) \quad \int_T h(\cdot, \delta_*) d\mu \leq \sup_k \int_T h(\cdot, s_k) d\mu < +\infty,$$

where $h \in \mathcal{H}(T; S)$ is as in the definition of tightness for $\{s_k\}_1^\infty$.

An important property of tightness is the following. Let S' be another standard Borel space. Then marginal tightness implies joint tightness, as is expressed more formally below.

PROPOSITION 2.8. Suppose that $\{s_k\}_1^\infty \subset \mathcal{M}(T; S)$ and $\{s'_k\}_1^\infty \subset \mathcal{M}(T; S')$ are tight. then $\{(s_k, s'_k)\}_1^\infty$ is tight in $\mathcal{M}(T; S \times S')$.

Proof. It is enough to remark that for $h \in \mathcal{H}(T; S)$, $h' \in \mathcal{H}(T; S')$ an integrand $\tilde{h} \in \mathcal{H}(T; S \times S')$ is defined by

$$\tilde{h}(t, (s, s')) = h(t, s) + h'(t, s'). \quad \text{QED}$$

Let us see what more can be said about the “generalized limits” of a tight sequence by considering again the above examples.

Example 2.1 (continued). Every generalized limit δ_* of $\{x_\epsilon\}$ can be identified with x_0 , in that

$$(2.8) \quad \delta_*(t) \text{ is the Dirac (or point) measure at } x_0(t) \text{ a.e. in } T.$$

This follows from Theorem I by observing that for a.e. $t \in T$ the only limit point of $\{x_\epsilon(t)\}$ is $x_0(t)$.

Example 2.2 (continued). Every generalized limit δ_* of $\{\xi_\epsilon\}$ is such that

$$(2.9) \quad \text{bar } \delta_*(t) \equiv \int_{\mathbb{R}'} \xi \delta_*(t)(d\xi) \text{ exists and equals } \xi_0(t) \text{ a.e. in } T.$$

To see this connection, note first that for a.e. $t \in T$

$$\int_{\mathbb{R}'} |\xi| \delta_*(t)(d\xi) < +\infty$$

by (2.7) and the properties of h' . These same properties imply that for every $\tilde{\xi} \in \mathbb{R}'$, $B \in \mathcal{T}$ the sequences $\{g^-(\cdot, \xi_\epsilon)\}$ and $\{g^+(\cdot, \xi_\epsilon)\}$ are uniformly integrable, where

$$g(t, \xi) \equiv \begin{cases} \langle \tilde{\xi}, \xi \rangle & \text{if } t \in B, \\ 0 & \text{else.} \end{cases}$$

Here $\langle \cdot, \cdot \rangle$ stands for the usual inner product. Hence, we may invoke Theorem I for both g and $-g$. By (2.2) it follows that

$$\int_B \langle \tilde{\xi}, \xi_0(t) \rangle \mu(dt) = \int_B \langle \tilde{\xi}, \text{bar } \delta_*(t) \rangle \mu(dt).$$

Since $\tilde{\xi}$ and B were arbitrary, the result follows.

Example 2.3 (continued). Every generalized limit δ_* of $\{\eta_\epsilon\}$ is such that

$$(2.10) \quad \eta_*(t) \equiv \text{bar } \delta_*(t) \text{ exists a.e. in } T,$$

$$(2.11) \quad \eta_*(t) \in \bigcap_{p=1}^{\infty} \text{cl co } \{\eta_\epsilon(t) : \epsilon \geq p\} \text{ a.e. in } T,$$

$$(2.12) \quad \eta_* \in \mathcal{L}_1(T; \mathbb{R}^f).$$

To prove this, let us first note that by (2.7)

$$\int_T \left[\int_{\mathbb{R}'} |\eta| \delta_*(t)(d\eta) \right] \mu(dt) < +\infty;$$

therefore (2.10) and (2.12) hold. By Theorem I we also have that the probability measure $\delta_*(t)$ is carried by the set $\bigcap_{p=1}^{\infty} \text{cl } \{\eta_\epsilon(t) : \epsilon \geq p\}$ a.e. in T ; hence, the barycenter of $\delta_*(t)$ belongs to the closed convex hull of that same set. This proves (2.11).

Example 2.3 (variant). In addition to the usual suppositions in Example 2.3, suppose that T is the unit interval, \mathcal{T} the Lebesgue σ -algebra and μ the Lebesgue

measure on T . Suppose now also that there exists $F: T \rightarrow \mathbb{R}_+^{\bar{r}}$, componentwise non-decreasing and right-continuous on T , such that for every $t \in T$

$$(2.13) \quad \lim_k \int_0^t \eta_k^+(\tau) d\tau = F(t),$$

where $(\eta_k^+)^j \equiv \max(\eta_k^j, 0)$ defines η_k^+ in terms of its component functions, $j = 1, \dots, \bar{r}$. Then we have in addition to (2.10)–(2.12) that

$$(2.14) \quad \eta_*^+(t) \leq \frac{dF^{ac}}{dt}(t) \text{ a.e. in } T,$$

where F^{ac} stands for the absolutely continuous part (componentwise) of F with respect to the Lebesgue measure μ (Lebesgue decomposition). This is demonstrated by an application of Theorem I to

$$g(t, \eta) \equiv \begin{cases} \max(\eta^j, 0) & \text{if } \alpha \leq t \leq \beta, \\ 0 & \text{else,} \end{cases}$$

for arbitrary $\alpha, \beta \in T, j = 1, \dots, \bar{r}$. In view of (2.13) it follows then that

$$\nu_F((\alpha, \beta]) \equiv F^j(\beta) - F^j(\alpha) \geq \int_\alpha^\beta (\eta_*^+)^j(t) dt.$$

Since the collection of finite disjoint unions of intervals $(\alpha, \beta]$ forms an algebra generating the Borel σ -algebra on T , it follows by Carathéodory's extension theorem that for every Borel set B in T

$$\nu_F(B) \geq \int_B \eta_*^+(t) dt.$$

Augmenting B by a negligible set can only increase the left side of this inequality. Therefore the inequality also holds for $B \in \mathcal{T}$. Now (2.14) follows from a well-known property of Lebesgue decomposition [26, IV.1.3].

Example 2.4 (continued). There exists a subsequence of $\{v_k\}_1^\infty$ of which every generalized limit δ_* is such that

$$\delta_*(t) \text{ is a Dirac measure a.e. in } T^{\text{pa}},$$

where T^{pa} denotes the purely atomic part of T . (For the sake of clarity we remark that this statement is made under the mere assumption (2.4) of tightness.) We prove this by fixing a collection of atoms A_p of which T^{pa} is the union; of course, this collection can be taken so as to be at most countable. Let $h \in \mathcal{H}(T; V)$ be as in the definition of tightness; it follows from Lemma A.1 (Appendix A) that for every atom A_p there is $h_p: V \rightarrow [0, +\infty]$ such that $h(t, \cdot) = h_p$ a.e. on A_p . Also, since a standard Borel space is isomorphic to a Borel set in \mathbb{R} [12, III.20], every function v_k is equal to a constant a.e. on A_p ; this constant will be denoted by $v_{k,p}$. It is now easy to see from the definition of tightness that for every atom A_p the sequence $\{v_{k,p}\}$ is relatively compact. Hence, by a diagonal extraction argument we can find a subsequence $\{k'\}$ of $\{k\}$ such that for every p the sequence $\{v_{k',p}\}$ converges. We conclude that on T^{pa} , with $\{k\}$ replaced by $\{k'\}$, the situation of Example 2.1 prevails. The proof is now easily finished.

Example 2.5 (continued). Every generalized limit δ_* of $\{z_k\}$ is such that the marginal on Z of the product measure of μ and δ_* equals μ_0 . This is seen as follows. Let $c \in \mathcal{C}_b(Z)$ be arbitrary, where $\mathcal{C}_b(Z)$ stands for the set of all bounded continuous

functions on Z . We can apply Theorem I to c and $-c$. This gives

$$\int_Z c(z) \mu_0(dz) = \int_T \left[\int_Z c(z) \delta_*(t)(dz) \right] \mu(dt),$$

and since c is arbitrary the result has been proven.

3. Lower closure for outer integral functionals. The first in a series of lower closure results for integral functionals will now be formulated.

THEOREM 3.1. *Suppose that $\{x_k\}_0^\infty, \{\xi_k\}_0^\infty, \{\eta_k\}_1^\infty$ satisfy (2.1)–(2.3). Then there exist a subsequence $\{\ell\}$ of $\{k\}$ and $\eta_* \in \mathcal{L}_1(T; \mathbb{R}^f)$ such that*

$$(3.1) \quad \lim_{\ell} \int_T g(\cdot, x_\ell, \xi_\ell, \eta_\ell) d\mu \geq \int_T g(\cdot, x_0, \xi_0, \eta_*) d\mu$$

for every normal integrand g on $T \times (X \times \mathbb{R}^r \times \mathbb{R}^f)$ satisfying

$$(3.2) \quad g(t, x_0(t), \cdot, \cdot) \text{ is convex on } \mathbb{R}^r \times \mathbb{R}^f \text{ a.e. in } T,$$

$$(3.3) \quad \{g^-(\cdot, x_\ell, \xi_\ell, \eta_\ell)\} \text{ is uniformly integrable.}$$

Moreover, η_* is such that

$$(3.4) \quad \eta_*(t) \in \bigcap_{p=1}^{\infty} \text{cl co } \{\eta_\ell(t) : \ell \geq p\} \text{ a.e. in } T.$$

Proof. By what was proven for Examples 2.1–2.3 the sequence $\{(x_k, \xi_k, \eta_k)\}_1^\infty \subset \mathcal{M}(T; X \times \mathbb{R}^r \times \mathbb{R}^f)$ is tight (Proposition 2.8). It follows from Theorem I that there exist a subsequence $\{\ell\}$ of $\{k\}$ and $\delta_* \in \mathcal{R}(T; X \times \mathbb{R}^r \times \mathbb{R}^f)$ such that for every normal integrand g satisfying (3.3)

$$(3.5) \quad \alpha \geq \int_T g(\cdot, \delta_*) d\mu,$$

where α denotes the left side of (3.1). Moreover, we know that for a.e. $t \in T$ the measure $\delta_*(t)$ is carried by the set of limit points of $\{(x_\ell(t), \xi_\ell(t), \eta_\ell(t))\}$, i.e., by the Cartesian product of $\{x_0(t)\}$ and the set of limit points of $\{(\xi_\ell(t), \eta_\ell(t))\}$, in view of (2.1). Denote by $\delta^*(t)$ the marginal probability measure of $\delta_*(t)$ on $\mathbb{R}^r \times \mathbb{R}^f$ and the submarginals on \mathbb{R}^r and \mathbb{R}^f by $\delta_1^*(t)$ and $\delta_2^*(t)$ respectively. For g as in (3.5) it now follows that

$$(3.6) \quad \alpha \geq \int_T g(\cdot, x_0, \delta^*) d\mu.$$

A fortiori, we now have for every $g' \in \mathcal{G}(T; \mathbb{R}^r)$ for which $\{g'^-(\cdot, \xi_\ell)\}$ is uniformly integrable, that marginally

$$\lim_{\ell} \int_T g'(\cdot, \xi_\ell) d\mu \geq \int_T g'(\cdot, \delta_1^*) d\mu.$$

A similar situation is found for δ_2^* . Thus, marginally we find precisely the situations investigated in Examples 2.2–2.3. this means

$$\xi_0(t) = \text{bar } \delta_1^*(t) \text{ a.e. in } T,$$

$$\eta_*(t) = \text{bar } \delta_2^*(t) \text{ exists a.e. in } T,$$

with η_* satisfying (2.11)–(2.12). Hence, by definition of barycenter

$$(3.7) \quad \text{bar } \delta^*(t) = (\xi_0(t), \eta_*(t)) \text{ a.e. in } T.$$

We finish the proof by applying Jensen's inequality. Suppose for a moment that g is bounded from below by a constant. Then for a.e. $t \in T$ the function $g(t, x_0(t), \cdot, \cdot)$ is proper convex, unless it is identically equal to $+\infty$. The latter possibility is trivial to deal with, so let us only look at the former. By applying a well-known result about proper convex functions and their affine minorants [9, I.4] and using (3.7) it follows that

$$g(t, x_0(t), \delta^*(t)) \geq g(t, x_0(t), \xi_0(t), \eta_*(t)) \text{ a.e. in } T.$$

In view of (3.6) the desired inequality (3.1) then follows. Thus far, we worked under the extra assumption that g is bounded from below. For general g it follows easily from (3.3)—cf. [16]—that for every $\varepsilon > 0$ there exists $\gamma > 0$ such that

$$\int_T g(\cdot, x_\ell, \xi_\ell, \eta_\ell) d\mu \geq \int_T \max(-\gamma, g(\cdot, x_\ell, \xi_\ell, \eta_\ell)) d\mu - \varepsilon$$

for all ℓ . By the inequality (3.1) for the normal integrand $\max(-\gamma, g)$ established above, by the inequality $\max(-\gamma, g) \geq g$ and the arbitrary choice of ε , the inequality (3.1) must now also hold for g . QED

This is essentially our main lower closure result “with convexity”. Its relation to other results in the literature will be discussed after Theorem 4.3. Next we shall derive more general and useful versions of this result; these apply to general integrands, whether measurable or not. The following lemma is instrumental [3k]; its proof can be found in Appendix A.

LEMMA II. Suppose that \mathcal{g} is a lower semicontinuous integrand on $T \times (X \times \mathbb{R}^r \times \mathbb{R}^{\bar{r}})$ with

$$(3.8) \quad \mathcal{g}(t, x_0(t), \cdot, \cdot) \text{ is convex on } \mathbb{R}^r \times \mathbb{R}^{\bar{r}} \text{ a.e. in } T.$$

Then there exists a normal integrand g on $T \times (X \times \mathbb{R}^r \times \mathbb{R}^{\bar{r}})$ such that

$$(3.9) \quad g \geq \mathcal{g},$$

$$(3.10) \quad g(t, x_0(t), \cdot, \cdot) \text{ is convex on } \mathbb{R}^r \times \mathbb{R}^{\bar{r}} \text{ a.e. in } T$$

and for every $x \in \mathcal{M}(T; X)$, $\xi \in \mathcal{M}(T; \mathbb{R}^r)$, $\eta \in \mathcal{M}(T; \mathbb{R}^{\bar{r}})$

$$(3.11) \quad \int_T \mathcal{g}(t, x(t), \xi(t), \eta(t)) \mu(dt) = \int_T g(t, x(t), \xi(t), \eta(t)) \mu(dt),$$

where outer integration and integration are subject to conventions introduced in § 1.

THEOREM 3.2. Suppose that $\{x_k\}_0^\infty$, $\{\xi_k\}_0^\infty$, $\{\eta_k\}_1^\infty$ satisfy (2.1)–(2.3). Then there exist a subsequence $\{\ell\}$ of $\{k\}$ and $\eta_* \in \mathcal{L}_1(T; \mathbb{R}^{\bar{r}})$ such that

$$(3.12) \quad \lim_{\ell} \int_T l(\cdot, x_\ell, \xi_\ell, \eta_\ell) d\mu \geq \int_T l(\cdot, x_0, \xi_0, \eta_*) d\mu$$

for every integrand l on $T \times X \times \mathbb{R}^r \times \mathbb{R}^{\bar{r}}$ satisfying

$$(3.13) \quad l(t, \cdot, \cdot, \cdot) \text{ is lower semicontinuous at every point in } \{x_0(t)\} \times \mathbb{R}^r \times \mathbb{R}^{\bar{r}} \text{ a.e. in } T,$$

$$(3.14) \quad l(t, x_0(t), \cdot, \cdot) \text{ is convex on } \mathbb{R}^r \times \mathbb{R}^{\bar{r}} \text{ a.e. in } T,$$

$$(3.15) \quad \text{there is a uniformly integrable sequence } \{\lambda_\ell\} \subset \mathcal{L}_1(T; \mathbb{R}) \text{ with} \\ l(\cdot, x_\ell, \xi_\ell, \eta_\ell) \geq \lambda_\ell \text{ for all } \ell.$$

Moreover, η_* satisfies (3.4).

Proof. Suppose first that l is bounded from below by a constant. The lower semicontinuous integrand \bar{l} on $T \times (X \times \mathbb{R}^r \times \mathbb{R}^{\bar{r}})$ is defined by

$$(3.16) \quad \bar{l}(t, x, \xi, \eta) \equiv \lim_{x' \rightarrow x, \xi' \rightarrow \xi, \eta' \rightarrow \eta} l(t, x', \xi', \eta').$$

By (3.13), (3.16) we have

$$(3.17) \quad l(t, x_0(t), \cdot, \cdot) = \bar{l}(t, x_0(t), \cdot, \cdot) \text{ a.e. in } T.$$

Clearly, (3.8) now holds for $g \equiv \bar{l}$ in view of (3.14). Applying Lemma II to $g \equiv \bar{l}$ gives that there exists a normal integrand g on $T \times (X \times \mathbb{R}^r \times \mathbb{R}^{\bar{r}})$ such that (3.9)–(3.11) hold. We can now apply Theorem 3.1 to g , since (3.2) holds by (3.10) and (3.3) by (3.9), (3.16) and the extra supposition. Thus, (3.1) holds for g . By using successively the inequality $l \geq g$, (3.11), (3.1) and (3.17), the inequality (3.12) follows. Secondly, consider the general case. By elementary properties of outer integration it follows easily from (3.15) that for every $\varepsilon > 0$ there exists $\gamma > 0$ such that for all k

$$\int_T l(\cdot, x_k, \xi_k, \eta_k) d\mu \geq \int_T \max(-\gamma, l(\cdot, x_k, \xi_k, \eta_k)) d\mu - \varepsilon.$$

(Let $\phi_k \in \mathcal{M}(T; (-\infty, +\infty])$) correspond to $l(\cdot, x_k, \xi_k, \eta_k)$ as in the attainment property for outer integrals mentioned in § 1, $\phi_k \geq l(\cdot, x_k, \xi_k, \eta_k)$ a.e. in T , and consider the identity

$$\int_T l(\cdot, x_k, \xi_k, \eta_k) d\mu = \int_{\{\phi_k \geq -\gamma\}} \phi_k d\mu + \int_{\{\phi_k < -\gamma\}} \phi_k d\mu.$$

From this the above inequality follows quickly.) Just as in the proof of Theorem 3.1, the inequality (3.12) now follows by the previous step. QED

It is easy to convert Theorem 3.2 into a more useful result by “recombination of variables” [3e, g]. In particular, this will lead to closure results for models with variable time domain.

Let $\{T_k\}_0^\infty \subset \mathcal{T}$ be such that

$$(3.18) \quad 1_{T_k}(t) \rightarrow 1_{T_0}(t) \text{ a.e. in } T,$$

where 1_{T_k} stands for the characteristic function of the set T_k . Note that this is equivalent to saying that the set $T_0 \equiv \lim_k T_k$ exists modulo a null set; cf. [26, I.4]. Further, $\{x_k\}_0^\infty \subset \mathcal{M}(T; X)$, $\{\xi_k\}_0^\infty \subset \mathcal{L}_1(T; \mathbb{R}^r)$ and $\{\eta_k\}_1^\infty \subset \mathcal{L}_1(T; \mathbb{R}^{\bar{r}})$ are now also allowed to satisfy

$$(3.19) \quad x_k(t) \rightarrow x_0(t) \text{ a.e. in } T_0,$$

$$(3.20) \quad \{1_{T_k} \xi_k\}_1^\infty \text{ converges weakly in } \sigma(\mathcal{L}_1^r, \mathcal{L}_\infty^r) \text{ to } 1_{T_0} \xi_0,$$

$$(3.21) \quad \sup_k \int_{T_k} |\eta_k| d\mu < +\infty.$$

Note that (3.20) is already implied by (2.2) and (3.18); this follows by a simple imitation of the proof of Theorem 3.3 below. Further, let $\{d_k\}_1^\infty \subset \mathcal{M}(T; \mathbb{R}^r)$, $\{\bar{d}_k\}_1^\infty \subset \mathcal{M}(T; \mathbb{R}^{\bar{r}})$ and $\{\bar{e}_k\}_1^\infty \subset \mathcal{L}_1(T; \mathbb{R}^{\bar{r}})$ be such that

$$(3.22) \quad d_k(t) \rightarrow 0 \text{ a.e. in } T_0,$$

$$(3.23) \quad \bar{d}_k(t) \rightarrow 0 \text{ a.e. in } T_0,$$

$$(3.24) \quad \{1_{T_k} \bar{e}_k\}_1^\infty \text{ converges weakly in } \sigma(\mathcal{L}_1^{\bar{r}}, \mathcal{L}_\infty^{\bar{r}}) \text{ to } 0.$$

THEOREM 3.3. *Suppose that $\{T_k\}_0^\infty, \{x_k\}_0^\infty, \{\xi_k\}_0^\infty, \{\eta_k\}_1^\infty$ satisfy (3.18)–(3.21) and that $\{d_k\}_1^\infty, \{\bar{d}_k\}_1^\infty, \{\bar{e}_k\}_1^\infty$ satisfy (3.22)–(3.24). Then there exist a subsequence $\{\ell\}$ of $\{k\}$ and $\eta_* \in \mathcal{L}_1(T; \mathbb{R}^f)$ such that*

$$(3.25) \quad \lim_{\ell} \int_{T_\ell} l(\cdot, x_\ell, \xi_\ell + d_\ell, \eta_\ell + \bar{d}_\ell + \bar{e}_\ell) d\mu \geq \int_{T_0} l(\cdot, x_0, \xi_0, \eta_*) d\mu,$$

for every integrand l on $T \times X \times \mathbb{R}^r \times \mathbb{R}^f$ satisfying

$$(3.26) \quad l(t, \cdot, \cdot, \cdot) \text{ is lower semicontinuous at every point in } \{x_0(t)\} \times \mathbb{R}^r \times \mathbb{R}^f \text{ a.e. in } T_0,$$

$$(3.27) \quad l(t, x_0(t), \cdot, \cdot) \text{ is convex on } \mathbb{R}^r \times \mathbb{R}^f \text{ a.e. in } T_0,$$

$$(3.28) \quad \text{there is a uniformly integrable sequence } \{\lambda_\ell\} \subset \mathcal{L}_1(T; \mathbb{R}) \text{ with } 1_{T_\ell} l(\cdot, x_\ell, \xi_\ell + d_\ell, \eta_\ell + \bar{d}_\ell + \bar{e}_\ell) \geq \lambda_\ell \text{ for all } \ell.$$

Moreover, η_* satisfies

$$(3.29) \quad \eta_*(t) \in \bigcap_{p=1}^{\infty} \text{cl co } \{\eta_\ell(t) : \ell \geq p\} \text{ a.e. in } T_0.$$

Proof. Let $\bar{x} \in X$ be arbitrary but fixed. Let $\bar{x}_k \in \mathcal{M}(T; X)$ be such that it coincides with x_k on T_k and has the constant value \bar{x} on $T \setminus T_k$; then $\bar{x}_k(t) \rightarrow \bar{x}_0(t)$ a.e. in T by (3.18)–(3.19). Let us now define $\tilde{x}_k \in \mathcal{M}(T; X \times \{0, 1\} \times \mathbb{R}^r \times \mathbb{R}^f)$ by $\tilde{x}_k \equiv (\bar{x}_k, 1_{T_k}, 1_{T_k} d_k, 1_{T_k} \bar{d}_k)$ for $k \in \mathbb{N}$ and $\tilde{x}_0 \equiv (\bar{x}_0, 1_{T_0}, 0, 0)$ for $k = 0$. Also, we define $\{\tilde{\xi}_k\}_0^\infty \subset \mathcal{L}_1(T; \mathbb{R}^{r+f})$ by $\tilde{\xi}_k \equiv (1_{T_k} \xi_k, 1_{T_k} \bar{e}_k)$ and $\tilde{\xi}_0 \equiv (1_{T_0} \xi_0, 0)$, and finally $\{\tilde{\eta}_k\}_1^\infty$ by $\tilde{\eta}_k \equiv 1_{T_k} \eta_k$. By (3.18)–(3.24) and the above we have

$$\begin{aligned} \tilde{x}_k(t) &\rightarrow \tilde{x}_0(t) \text{ a.e. in } T, \\ \{\tilde{\xi}_k\}_1^\infty &\text{ converges weakly in } \sigma(\mathcal{L}_1^{r+f}, \mathcal{L}_\infty^{r+f}) \text{ to } \tilde{\xi}_0, \\ \sup_k \int_T |\tilde{\eta}_k| d\mu &< +\infty. \end{aligned}$$

Given l with (3.26)–(3.28), we define the integrand \tilde{l} by

$$\tilde{l}(t, \tilde{x}, \tilde{\xi}, \eta) \equiv \gamma l(t, x, \xi + d, \eta + \bar{d} + \bar{e})$$

for $\tilde{x} \equiv (x, \gamma, d, \bar{d}) \in X \times \{0, 1\} \times \mathbb{R}^r \times \mathbb{R}^f$, $\tilde{\xi} \equiv (\xi, \bar{e}) \in \mathbb{R}^r \times \mathbb{R}^f$. By (3.26)

$$\tilde{l}(t, \cdot, \cdot, \cdot) \text{ is lower semicontinuous at every point in } \{\tilde{x}_0(t)\} \times \mathbb{R}^{r+f} \times \mathbb{R}^f \text{ a.e. in } T.$$

Also, in the same notation, we have

$$\tilde{l}(t, \tilde{x}_0(t), \tilde{\xi}, \eta) = \begin{cases} l(t, x_0(t), \xi, \eta + \bar{e}) & \text{if } t \in T_0, \\ 0 & \text{else.} \end{cases}$$

This shows that by (3.27)

$$\tilde{l}(t, \tilde{x}_0(t), \cdot, \cdot) \text{ is convex on } \mathbb{R}^{r+f} \times \mathbb{R}^f \text{ a.e. in } T.$$

Finally, it follows from (3.28) by definition of \tilde{l} that for all ℓ

$$\tilde{l}(\cdot, \tilde{x}_\ell, \tilde{\xi}_\ell, \tilde{\eta}_\ell) \geq \lambda_\ell.$$

Hence the situation found in the statement of this theorem has been reduced completely to that of Theorem 3.2. It remains to invoke this result. \square

Remark 3.4. Suppose that instead of the suppositions (2.1), (3.18)–(3.19) regarding convergence a.e. we make the following weaker suppositions about convergence in measure:

$$(2.1') \quad \mu(\{t \in T: \text{dist}(x_k(t), x_0(t)) > \varepsilon\}) \rightarrow 0 \text{ for every } \varepsilon > 0,$$

$$(3.18') \quad \mu((T_k \setminus T_0) \cup (T_0 \setminus T_k)) \rightarrow 0 \text{ for every } \varepsilon > 0,$$

$$(3.19') \quad \mu(\{t \in T_0: \text{dist}(x_k(t), x_0(t)) > \varepsilon\}) \rightarrow 0 \text{ for every } \varepsilon > 0,$$

to be used instead of (2.1), (3.18) and (3.19) respectively. Then our previous results will remain valid, since we will now have (2.1), (3.18) and (3.19) respectively for suitable subsequences of $\{x_k\}_0^\infty$ and $\{T_k\}_0^\infty$. The same can be said if instead of (2.2), (3.20) we suppose

$$(2.2') \quad \{\xi_k\}_1^\infty \text{ is uniformly integrable,}$$

$$(3.20') \quad \{1_{T_k} \xi_k\}_1^\infty \text{ is uniformly integrable,}$$

provided that we denote any weak limit point in either case by ξ_0 ; here we use the Dunford–Pettis theorem [12, II.25] (or, alternatively, Theorem I and Example 2.2).

COROLLARY 3.5. *Suppose that $\{T_k\}_0^\infty$, $\{x_k\}_0^\infty$, $\{\xi_k\}_0^\infty$ satisfy (3.18)–(3.20). Then*

$$\lim_k \int_T l(\cdot, x_k, \xi_k) d\mu \geq \int_{T_0} l(\cdot, x_0, \xi_0) d\mu$$

for every integrand l on $T \times X \times \mathbb{R}^r$ satisfying

$l(t, \cdot, \cdot)$ is lower semicontinuous at every point in $\{x_0(t)\} \times X \times \mathbb{R}^r$ a.e. in T_0 ,

$l(t, x_0(t), \cdot)$ is convex on \mathbb{R}^r a.e. T_0 ,

there is a uniformly integrable sequence $\{\lambda_k\}_1^\infty \subset \mathcal{L}_1(T; \mathbb{R})$ with $1_{T_k} l(\cdot, x_k, \xi_k) \geq \lambda_k$ for all $k \in \mathbb{N}$.

This generalizes a classical lower semicontinuity result for integral functionals to outer integral functionals with variable time domain [13, VIII.2.2], [17, 9.1.4], [10d, (1.iii)], [27c, d], [16], [3c, f].

Necessary conditions for lower semicontinuity in similar setups have been obtained in e.g. [16], [27d]; they indicate that the conditions of Corollary 3.5—and Theorem 3.3 by implication—are quite sharp.

COROLLARY 3.6. *Suppose that (T, \mathcal{T}, μ) is as in the variant of Example 2.3, that $\{\alpha_k\}_0^\infty$, $\{\beta_k\}_0^\infty \subset T$ are such that*

$$\alpha_k \rightarrow \alpha_0, \quad \beta_k \rightarrow \beta_0$$

and that for $T_k \equiv [\alpha_k, \beta_k]$ the sequences $\{x_k\}_0^\infty$, $\{\xi_k\}_0^\infty$, $\{\eta_k\}_1^\infty$ satisfy (3.18)–(3.21) and (2.13). Then

$$\lim_k \int_{\alpha_k}^{\beta_k} g(\cdot, x_k, \xi_k, \eta_k) d\mu \geq \int_{\alpha_0}^{\beta_0} g\left(t, x_0(t), \xi_0(t), \frac{dF^{ac}}{dt}(t)\right) dt$$

for every normal integrand g on $T \times (X \times \mathbb{R}^r \times \mathbb{R}^{\bar{r}})$ satisfying

$g(t, x_0(t), \cdot, \cdot)$ is convex on $\mathbb{R}^r \times \mathbb{R}^{\bar{r}}$ a.e. in T_0 ,

$g(t, x_0(t), \xi_0(t), \cdot)$ is nonincreasing on $\mathbb{R}^{\bar{r}}$ a.e. in T_0 ,

$\{g^-(\cdot, x_k, \xi_k, \eta_k)\}_1^\infty$ is uniformly integrable.

It would seem that this corollary is a new result; it can also be derived from [3c, Thm. 5], as was already argued in [3f, Thm. 3.2] for a slightly less general version.

Let us return to the proof of Theorem 3.1, which has been essential for the developments thus far. We can see that in deriving the inequality (3.6) the convexity property (3.2) was not used. Therefore (3.6) suggests a radically different way to obtain a lower closure result for integral functionals; namely, we can try to trade the convexity condition (3.2) for “extreme point considerations and Lyapunov’s theorem”. Thus, we enter the domain of “existence without convexity”, explored for the first time in optimal control theory by L. W. Neustadt [25], from a completely new angle. The following lemma, proven in Appendix A by extreme point considerations and Lyapunov’s theorem, will bring us the desired results.

LEMMA III. Suppose that (T, \mathcal{T}, μ) is nonatomic and that $\delta^* \in \mathcal{R}(T; V)$ is tight with respect to $h \in \mathcal{H}(T; V)$, i.e.,

$$\int_T h(\cdot, \delta^*) d\mu < +\infty.$$

Suppose that the normal integrands g_1, \dots, g_m on $T \times V$ satisfy the following growth condition with respect to h : for every $\varepsilon > 0$ there is $\phi_\varepsilon \in \mathcal{L}_1(T; \mathbb{R})$ with

$$(3.30) \quad g_j^-(t, v) \leq \varepsilon h(t, v) + \phi_\varepsilon(t), \quad j = 1, \dots, m.$$

Then there exists $v^* \in \mathcal{M}(T; V)$ such that

$$\int_T g_j(\cdot, v^*) d\mu \leq \int_T g_j(\cdot, \delta^*) d\mu, \quad j = 1, \dots, m.$$

THEOREM 3.7. Suppose that $\{x_k\}_0^\infty, \{v_k\}_1^\infty$ satisfy (2.1), (2.4). Then there exists a subsequence $\{\ell\}$ of $\{k\}$ such that to every finite collection g_1, \dots, g_m of normal integrands on $T \times (X \times V)$ there corresponds $v_* \in \mathcal{M}(T; V)$ with

$$(3.31) \quad \lim_{\ell} \int_T g_j(\cdot, x_\ell, v_\ell) d\mu \geq \int_T g_j(\cdot, x_0, v_*) d\mu, \quad j = 1, \dots, m,$$

provided that

$$(3.32) \quad \{g_j^-(\cdot, x_\ell, v_\ell)\} \text{ is uniformly integrable,} \quad j = 1, \dots, m.$$

Moreover, v_* satisfies

$$(3.33) \quad v_*(t) \in \bigcap_{p=1}^{\infty} \text{cl} \{v_\ell(t) : \ell \geq p\} \text{ a.e. in } T.$$

Proof. We start by considering a preliminary subsequence $\{k'\}$ of $\{k\}$ which is such that $\{v_{k'}(t)\}$ converges a.e. on T^{pa} (Example 2.4). It is left to the reader to see that the proof of Theorem 3.1 can now be imitated up to formula (3.6). Here $\mathbb{R}^r \times \mathbb{R}^r$ is replaced by V and $\{\xi_k, \eta_k\}$ by $v_{k'}$; also we use tightness of $\{x_{k'}\}$, $\{v_{k'}\}$ and condition (3.32). Hence, there exist a subsequence $\{\ell\}$ of $\{k'\}$ and $v^* \in \mathcal{M}(T^{\text{pa}}; V)$, $\delta^* \in \mathcal{R}(T^{\text{na}}; V)$ such that for all j

$$(3.34) \quad \lim_{\ell} \int_T g_j(\cdot, x_\ell, v_\ell) d\mu \geq \int_{T^{\text{pa}}} g_j(\cdot, x_0, v^*) d\mu + \int_{T^{\text{na}}} g_j(\cdot, x_0, \delta^*) d\mu,$$

where $T^{\text{na}} \equiv T \setminus T^{\text{pa}}$ denotes the nonatomic part of T and where the pointwise convergence on T^{pa} has been taken into account (cf. Example 2.1). Moreover, we have that

$$(3.35) \quad v^*(t) \in V_1(t) \equiv \bigcap_{p=1}^{\infty} \text{cl} \{v_k(t) : k \geq p\} \text{ a.e. in } T^{\text{pa}},$$

$$(3.36) \quad \delta^*(t) \text{ is carried by } V_1(t) \text{ a.e. in } T^{\text{na}},$$

$$\int_{T^{\text{na}}} h(\cdot, \delta^*) d\mu < +\infty,$$

where h is as in Example 2.4. Momentarily we shall make an extra assumption: we assume that for every $\varepsilon > 0$ there exists $\phi_\varepsilon \in \mathcal{L}_1(T; \mathbb{R})$ with

$$(3.37) \quad g_j^-(t, x_0(t), v) \leq \varepsilon h(t, v) T\phi_\varepsilon(t), \quad j = 1, \dots, m.$$

This allows us to invoke Lemma III: there exists $v^{**} \in \mathcal{M}(T^{\text{na}}; V)$ such that

$$(3.38) \quad \int_{T^{\text{na}}} g_j(\cdot, x_0, v^{**}) d\mu \leq \int_{T^{\text{na}}} g_j(\cdot, x_0, \delta^*) d\mu, \quad j = 1, \dots, m,$$

$$(3.39) \quad \int_{T^{\text{na}}} g_{m+1}(\cdot, v^{**}) d\mu \leq 0,$$

where $g_{m+1} \in \mathcal{G}^+(T; V)$ is defined by

$$g_{m+1}(t, v) \equiv \begin{cases} 0 & \text{if } v \in V_1(t), \\ +\infty & \text{else.} \end{cases}$$

Defining $v_* = v^*$ on T^{pa} , $v_* = v^{**}$ on T^{na} , we see that (3.34) and (3.38) imply (3.31). Also, (3.35), (3.36) and (3.39) imply (3.33). Let us now see how the extra assumption (3.37) can be revoked. We shall apply the result established under (3.37) to the normal integrands $\tilde{g}_1, \dots, \tilde{g}_m$ on $T \times (X \times V \times \mathbb{R})$ defined by

$$(3.40) \quad \tilde{g}_j(t, x, v, \lambda) \equiv \max(g_j(t, x, v), \lambda).$$

We define also

$$\lambda_k(t) \equiv -|g^-(t, x_k(t), v_k(t))|,$$

where $(g^-)^j \equiv (g^j)^-$. By de la Vallée-Poussin's theorem there exists $h' : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ with $h'(\gamma)/\gamma \rightarrow +\infty$ as $\gamma \rightarrow +\infty$ and

$$\sup_k \int_T h'(|\lambda_k(t)|) \mu(dt) < +\infty,$$

in view of (3.32); cf. Example 2.2. Hence, by (2.4) the sequence $\{(v_k, \lambda_k)\}_1^\infty$ satisfies

$$(3.41) \quad \sup_k \int_T \tilde{h}(\cdot, v_k, \lambda_k) d\mu < +\infty,$$

where $\tilde{h}(t, v, \lambda) \equiv h(t, v) + h'(|\lambda|)$, and

$$(3.42) \quad \tilde{g}_j(\cdot, x_k, v_k, \lambda_k) = g_j(\cdot, x_k, v_k) \quad \text{for all } k \text{ and } j.$$

We may now apply the result established above. Note that (3.41) replaces (2.4) and that (3.37) obviously holds for \tilde{g}_j^- with respect to \tilde{h} : in view of (3.40) and the properties of h' , we have for every $\varepsilon > 0$ that there exists $\gamma_\varepsilon > 0$ with

$$\tilde{g}_j^-(t, x, v, \lambda) \leq \max(-\lambda, 0) \leq \varepsilon h'(|\lambda|) + \gamma_\varepsilon \leq \varepsilon \tilde{h}(t, v, \lambda) + \gamma_\varepsilon.$$

Thus, we find that there exists $(v_*, \lambda_*) \in \mathcal{M}(T; V \times \mathbb{R})$ with

$$\lim_{\mathcal{K}} \int_T \tilde{g}_j(\cdot, x_\ell, v_\ell, \lambda_\ell) d\mu \cong \int_T \tilde{g}_j(\cdot, x_0, v_*, \lambda_*) d\mu, \quad j = 1, \dots, m.$$

In view of (3.40), (3.42), this amounts to (3.31). Moreover, we have

$$(v_*(t), \lambda_*(t)) \text{ is a limit point of } \{(v_\ell(t), \lambda_\ell(t))\} \text{ a.e. in } T,$$

and this implies (3.33). QED

Theorem 3.7 is a quite new result. Just as Theorem 3.1 was upgraded by using Lemma II and recombination of variables, so can this be done with Theorem 3.7.

Let $\{T_k\}_0^\infty$ be as in (3.18) ff. We shall now also consider a sequence $\{v_k\}_1^\infty$ of measurable functions, $v_k \in \mathcal{M}(T_k; V)$, such that there exists $h \in \mathcal{H}(T; V)$ with

$$(3.43) \quad \sup_k \int_{T_k} h(\cdot, v_k) d\mu < +\infty.$$

THEOREM 3.8. *Suppose that $\{T_k\}_0^\infty$, $\{x_k\}_0^\infty$, $\{v_k\}_1^\infty$ satisfy (3.18), (3.19), (3.43). Then there exists a subsequence $\{\ell\}$ of $\{k\}$ such that to every finite collection l_1, \dots, l_m of integrands on $T \times (X \times V)$ there corresponds $v_* \in \mathcal{M}(T_0; V)$ with*

$$\lim_{\mathcal{K}} \int_{T_\ell} l_j(\cdot, x_\ell, v_\ell) d\mu \cong \int_{T_0} l_j(\cdot, x_0, v_*) d\mu, \quad j = 1, \dots, m,$$

provided that

there is a uniformly integrable sequence $\{\lambda_\ell\} \subset \mathcal{L}_1(T; \mathbb{R})$ with $\bar{l}_j(\cdot, x_\ell, v_\ell) \cong \lambda_\ell$ for all ℓ , $j = 1, \dots, m$,

where \bar{l}_j is defined by

$$\bar{l}_j(t, x, v) \equiv \lim_{x' \rightarrow x, v' \rightarrow v} l_j(t, x', v').$$

Moreover, v_* satisfies

$$v_*(t) \in \bigcap_{p=1}^\infty \text{cl} \{v_k(t): k \geq p\} \text{ a.e. in } T_0.$$

Proof. The proof is quite similar to the argument by which Theorem 3.1 was transformed—via Theorem 3.2—into Theorem 3.3. It will be left to the reader, except for the following point. Define $\bar{h}: T \rightarrow [0, +\infty]$ by $\bar{h}(t) \equiv \inf \{h(t, v): v \in V\}$. Then \bar{h} is measurable with respect to the completion $\bar{\mathcal{T}}$ of \mathcal{T} [9, III.39]. By Fatou's lemma it follows from (3.18) that

$$(3.44) \quad \int_{T_0} \bar{h} d\mu \leq \liminf_k \int_{T_k} h(\cdot, v_k) d\mu.$$

By inf-compactness of $h(t, \cdot)$ there exists for every $t \in T$ an element $v_t \in V$ with $h(t, v_t) = \bar{h}(t)$. Since the set of all $(t, v) \in T \times V$ for which $h(t, v) = \bar{h}(t)$ is $\bar{\mathcal{T}} \times \mathcal{B}(V)$ -measurable, it follows from Aumann's theorem [14] that there exists $\bar{v}: T \rightarrow V$, $\bar{\mathcal{T}}$ -measurable, such that $h(t, \bar{v}(t)) = \bar{h}(t)$ a.e. in T . Since V is isomorphic to a Borel subset of \mathbb{R} [12], it follows that there exists a \mathcal{T} -measurable modification $\bar{v} \in \mathcal{M}(T; V)$ of \bar{v} . In view of (3.44) we conclude that

$$\int_{T_0} h(\cdot, \bar{v}) d\mu \leq \sup_k \int_{T_k} h(\cdot, v_k) d\mu.$$

Hence we obtain a tight sequence $\{\bar{v}_k\}_1^\infty \subset \mathcal{M}(T_0; V)$ by defining $\bar{v}_k \equiv v_k$ on $T_0 \cap T_k$ and $\bar{v}_k \equiv \bar{v}$ on $T_0 \setminus T_k$. As explained above, the rest of the proof is quite simple. QED

COROLLARY 3.9 (Fatou's lemma in several dimensions). *Suppose that $\{\phi_k\}_1^\infty \subset \mathcal{L}_1(T; \mathbb{R}^m)$ is such that*

$$(3.45) \quad \lim_k \int_T \phi_k d\mu \text{ exists (in } \mathbb{R}^m),$$

$$(3.46) \quad \{\phi_k\}_1^\infty \text{ is uniformly integrable.}$$

Then there exists $\phi_ \in \mathcal{L}_1(T; \mathbb{R}^m)$ such that*

$$\int_T \phi_* d\mu \leq \lim_k \int_T \phi_k d\mu,$$

$\phi_(t)$ is a limit point of $\{\phi_k(t)\}_1^\infty$ a.e. in T .*

Proof. We define the normal integrands g_1, \dots, g_{3m} on $T \times V$ by $g_j(t, v) \equiv \max(v^j, 0)$, $g_{m+j}(t, v) \equiv \max(-v^j, 0)$ and $g_{2m+j}(t, v) \equiv \min(v^j, 0)$, $j = 1, \dots, m$. Also, we set $v_k \equiv \phi_k$. Note that (3.45)–(3.46) imply

$$\sup_k \int_T |\phi_k| d\mu < +\infty.$$

Hence (2.4) holds. Since (3.46) implies that (3.32) is fulfilled, we have by Theorem 3.7 the desired result, as is seen at once. QED

The above corollary was given in [31], where it was shown to be equivalent to slightly weaker form of Theorem 3.7. It generalizes the multidimensional Fatou lemmas of [29], [11c] and [1b], as well as a number of existence results for allocation problems arising in economics ([2], [6, Prop. III.2.1], [1a], [3a]). Corollary 3.9 can also be used directly to obtain existence results “without convexity conditions” for the optimal control of certain linear dynamical systems. In this way an existence result has been derived in [3m] for the optimal control of a linear integral equation having singular components; this generalizes the existence results of [25], [19], [4], [10b] and essentially also that of [30b].

4. Lower closure for orientor fields. It turns out that each of the main lower closure results of the previous section can be expressed in an alternative form, involving multifunctions. Here we shall only work out such a procedure for the lower closure result “with convexity”, i.e. Theorem 3.3. It will lead us to the so-called lower closure results for orientor fields.

For a multifunction $Q: T \times X \rightrightarrows \mathbb{R}^r \times \mathbb{R}^f$ we define $\text{dom } Q$ to be the set of those $(t, x) \in T \times X$ for which the set $Q(t, x)$ is nonempty. We shall say that Q has *property (K)* at a point $(t, x) \in T \times X$ if

$$(4.1) \quad Q(t, x) = \bigcap_{\gamma > 0} \text{cl} \bigcup \{Q(t, x') : x' \in X, \text{dist}(x', x) < \gamma\},$$

where “dist” refers to any compatible metric on X . Note that x' runs effectively in the section at t of $\text{dom } Q$.

LEMMA 4.1. *Suppose $Q: T \times X \rightrightarrows \mathbb{R}^r \times \mathbb{R}^f$ and $(t, x^0) \in T \times X$ are given. Let l_i^Q, \dots, l_f^Q be integrands on $T \times X \times \mathbb{R}^r \times \mathbb{R}^f$, defined by*

$$(4.2) \quad l_j^Q(t, x, \xi, \eta) \equiv \begin{cases} \eta^j & \text{if } (\xi, \eta) \in Q(t, x), \\ +\infty & \text{else.} \end{cases}$$

(a) *The following are equivalent:*

(4.3) *Q has property (K) at (t, x^0) ,*

(4.4) *$l_j^Q(t, \cdot, \cdot, \cdot)$ is lower semicontinuous at every point of $\{x^0\} \times \mathbb{R}^r \times \mathbb{R}^{\bar{r}}$, $j = 1, \dots, \bar{r}$.*

(b) *The following are also equivalent:*

$Q(t, x^0)$ is convex,

$l_j^Q(t, x^0, \cdot, \cdot)$ is convex on $\mathbb{R}^r \times \mathbb{R}^{\bar{r}}$, $j = 1, \dots, \bar{r}$.

Proof. Suppose first that (4.3) holds. Let $\{(x^k, \xi^k, \eta^k)\}_1^\infty$ be arbitrary and such that $x^k \rightarrow x^0$, $\xi^k \rightarrow \xi^0$, $\eta^k \rightarrow \eta^0$ for certain ξ^0, η^0 . Without loss of generality we may assume that $\zeta \equiv \lim_k l_j^k$ is finite, where $l_j^k \equiv l_j^Q(t, x^k, \xi^k, \eta^k)$. Further, we can assume that $\zeta = \lim_k l_j^k$, instead of restricting ourselves to a suitable subsequence. It now follows that eventually l_j^k is finite, so without loss of generality we can suppose that $(\xi^k, \eta^k) \in Q(t, x^k)$ and $(\eta^k)^j = l_j^k$ for all $k \in \mathbb{N}$. Also, it follows that $\zeta = (\eta^0)^j$. For every $\gamma > 0$ we have now evidently

$$(\xi^0, \eta^0) \in \text{cl} \cup \{Q(t, x): x \in X, \text{dist}(x, x^0) < \gamma\},$$

so it follows from (4.3) that $(\xi^0, \eta^0) \in Q(t, x^0)$. By (4.2) we find $l_j^Q(t, x^0, \xi^0, \eta^0) = (\eta^0)^j = \zeta$. This shows that (4.4) holds.

Conversely, suppose that (4.4) holds. One inclusion in (4.1) is always trivial (take $x' = x$). To prove the other inclusion, let (ξ^0, η^0) belong to the right side in (4.1) (with $x = x^0$). This is easily seen to be equivalent to the following: for every $k \in \mathbb{N}$ there exist $x^k \in X$, $(\xi^k, \eta^k) \in Q(t, x^k)$ such that $\text{dist}(x^k, x^0)$, $|\xi^k - \xi^0|$ and $|\eta^k - \eta^0|$ are all smaller than k^{-1} . Hence $x^k \rightarrow x^0$, $\xi^k \rightarrow \xi^0$ and $\eta^k \rightarrow \eta^0$. By (4.4) we have for any j

$$(\eta^0)^j = \lim_k l_j^Q(t, x^k, \xi^k, \eta^k) \geq l_j^Q(t, x^0, \xi^0, \eta^0).$$

Since the left side is finite, (4.2) gives that $(\xi^0, \eta^0) \in Q(t, x^0)$.

(b) The demonstration of this part is trivial. QED

Remark 4.2. From the final step in the proof of Lemma 4.1(a) it appears clearly that (4.3)–(4.4) are also equivalent to

(4.4') *$l_j^Q(t, \cdot, \cdot, \cdot)$ is lower semicontinuous at every point of $\{x^0\} \times \mathbb{R}^r \times \mathbb{R}^{\bar{r}}$ for some j , $1 \leq j \leq \bar{r}$.*

A similar remark holds with regard to part (b) of Lemma 4.1.

We shall now state our main lower closure result for orientor fields and show that it is equivalent to Theorem 3.3.

THEOREM 4.3. *Suppose that $\{T_k\}_0^\infty, \{x_k\}_0^\infty, \{\xi_k\}_0^\infty, \{\eta_k\}_1^\infty$ satisfy (3.18)–(3.21) and that $\{d_k\}_1^\infty, \{\bar{d}_k\}_1^\infty, \{\bar{e}_k\}_1^\infty$ satisfy (3.22)–(3.24). Then there exist a subsequence $\{\ell\}$ of $\{k\}$ and $\eta_* \in \mathcal{L}_1(T, \mathbb{R}^{\bar{r}})$ such that*

$$(4.5) \quad (\xi_0(t), \eta_*(t)) \in Q(t, x_0(t)) \text{ a.e. in } T_0,$$

$$(4.6) \quad \lim_{\ell} \int_{T_\ell} \eta_\ell^j d\mu \geq \int_{T_0} \eta_*^j d\mu,$$

for every multifunction $Q: T \times X \rightrightarrows \mathbb{R}^r \times \mathbb{R}^{\bar{r}}$ and every j , $1 \leq j \leq \bar{r}$, such that

$$(4.7) \quad Q \text{ has property (K) at } (t, x_0(t)) \text{ a.e. in } T_0,$$

(4.8) $Q(t, x_0(t))$ is a convex subset of $\mathbb{R}^r \times \mathbb{R}^{\bar{r}}$ a.e. in T_0 ,

(4.9) $(\xi_\epsilon(t) + d_\epsilon(t), \eta_\epsilon(t) + \bar{d}_\epsilon(t) + \bar{e}_\epsilon(t)) \in Q(t, x_\epsilon(t))$ a.e. in T_ϵ ,

(4.10) $\{(1_{T_\epsilon} \eta_\epsilon^j)^-\}$ is uniformly integrable.

Moreover, η_* satisfies (3.29).

Proof. Suppose Q and j satisfy (4.7)–(4.10). Define $\tilde{x}_k \equiv (x_k, \bar{d}_k)$ for $k \in \mathbb{N}$ and $\tilde{x}_0 \equiv (x_0, 0)$. Then by (3.19), (3.22)

$$\tilde{x}_k(t) \rightarrow \tilde{x}_0(t) \text{ a.e. in } T_0.$$

We shall apply Theorem 3.3 to l_j defined by

$$l_j(t, \tilde{x}, \xi, \eta) \equiv l_j^Q(t, x, \xi, \eta + \bar{d}) - \bar{d}^j$$

for $\tilde{x} \equiv (x, \bar{d})$. Hence, in view of (4.2), (4.9),

$$1_{T_\epsilon} l_j(\cdot, \tilde{x}_\epsilon, \xi_\epsilon, \eta_\epsilon + \bar{e}_\epsilon) = 1_{T_\epsilon}(\eta_\epsilon^j + \bar{e}_\epsilon^j).$$

Further, by Lemma 4.1 it follows from (4.7)–(4.8) that

$l_j(t, \cdot, \cdot, \cdot)$ is lower semicontinuous at every point in $\{\tilde{x}_0(t)\} \times \mathbb{R}^r \times \mathbb{R}^{\bar{r}}$,

$l_j(t, \tilde{x}_0(t), \cdot, \cdot)$ is convex on $\mathbb{R}^r \times \mathbb{R}^{\bar{r}}$.

Hence we conclude that the conditions of Theorem 3.3 are fulfilled. We find therefore

$$(4.11) \quad \lim_{\epsilon} \int_{T_\epsilon} (\eta_\epsilon^j + \bar{e}_\epsilon^j) d\mu \geq \int_{T_0} l_j(\cdot, \tilde{x}_0, \xi_0, \eta_*) d\mu.$$

In view of (3.21), (3.24) and elementary properties of the outer integral it follows that

$$l_j(t, \tilde{x}_0(t), \xi_0(t), \eta_*(t)) < +\infty \text{ a.e. in } T_0.$$

This gives (4.5) by definition of l_j . Finally, (4.11) implies now (4.6). QED

Remark 4.4. Evidently, it follows from (4.5) that

$$(4.12) \quad (t, x_0(t)) \in \text{dom } Q \text{ a.e. in } T_0.$$

In the literature one usually considers only the restriction Q' of Q to $\text{dom } Q$. Let $A(t)$ denote the set of all $x' \in X$ with $(t, x') \in \text{dom } Q$. The multifunction $Q_D: \text{dom } Q \rightrightarrows \mathbb{R}^r \times \mathbb{R}^{\bar{r}}$ is said to have *property (K) with respect to $A(t)$* at $(t, x) \in \text{dom } Q$ if

$$Q_D(t, x) = \bigcap_{\gamma > 0} \text{cl} \bigcup \{Q_D(t, x'): x' \in A(t), \text{dist}(x', x) < \gamma\}.$$

To connect the formulation of results in the literature with that employed here, it is enough to observe that Q has property (K) at a point $(t, x) \in T \times X \setminus \text{dom } Q$ if (but not only if!) $A(t)$ is closed, whereas Q has property (K) at $(t, x) \in \text{dom } Q$ if and only if Q_D has property (K) with respect to $A(t)$ at (t, x) . This explains also why (4.12) is a final conclusion in our somewhat more general approach, while it is a necessary preliminary step for the usual approach in the literature.

Before discussing Theorem 4.3, we show that Theorems 3.2, 3.3 are in fact equivalent to it.

PROPOSITION 4.5. *The lower closure results obtained in Theorems 3.2, 3.3 and 4.3 are equivalent.*

Proof. Theorem 3.3 was shown to follow from Theorem 3.2 by “recombination of variables”. Theorem 4.3 was derived from Theorem 3.3. Hence it is enough to show that Theorem 4.3 implies Theorem 3.2.

Suppose that l satisfies (3.13)–(3.15). If the left side in (3.12) equals $+\infty$, there is nothing left to prove. Thus, we shall assume that this is not the case. We shall apply Theorem 4.3 to $Q: T \times X \rightrightarrows \mathbb{R}^r \times \mathbb{R}^{\bar{r}+1}$, defined by

$$(4.13) \quad Q(t, x) = \{(\xi, \eta, \gamma) \in \mathbb{R}^r \times \mathbb{R}^{\bar{r}} \times \mathbb{R}: \gamma \geq l(t, x, \xi, \eta)\};$$

in particular, we shall look at the coordinate $j = \bar{r} + 1$. In the present case (4.2) gives

$$l_{\bar{r}+1}^Q(t, x, \xi, \eta, \gamma) = \begin{cases} \gamma & \text{if } \gamma \geq l(t, x, \xi, \eta), \\ +\infty & \text{else.} \end{cases}$$

Hence, (3.13)–(3.14) imply (4.7)–(4.8), as follows by Lemma 4.1 and Remark 4.2. By the elementary properties of outer integrals, established in § 1, there exists for every $k \in \mathbb{N}$ a function $\gamma_k \in \mathcal{M}(T; (-\infty, +\infty])$ with

$$(4.14) \quad \gamma_k(t) \geq l(t, x_k(t), \xi_k(t), \eta_k(t)) \text{ a.e. in } T,$$

$$(4.15) \quad \int_T l(\cdot, x_k, \xi_k, \eta_k) d\mu = \int_T \gamma_k d\mu.$$

Since we work under the assumption that the left side in (3.12) is not equal to $+\infty$, we can suppose without loss of generality that

$$\sup_k \int_T |\gamma_k| d\mu < +\infty,$$

in view of (3.15). A fortiori, we have $\gamma_k(t) < +\infty$ a.e. in T for all $k \in \mathbb{N}$. Thus, (4.13)–(4.14) entail that for all $k \in \mathbb{N}$

$$(\xi_k(t), \eta_k(t), \gamma_k(t)) \in Q(t, x_k(t)) \text{ a.e. in } T.$$

Hence, condition (4.9)—with $T_k = T$, $d_k = 0$, $\bar{d}_k = 0$, $\bar{e}_k = 0$ —is also satisfied. Further, (3.15) and (4.14) imply that (4.10) holds. Application of Theorem 4.3 gives the existence of a subsequence $\{\ell\}$ of $\{k\}$ and of $(\eta_*, \gamma_*) \in \mathcal{L}_1(T; \mathbb{R}^{\bar{r}+1})$ such that

$$\gamma_*(t) \geq l(t, x_0(t), \xi_0(t), \eta_*(t)) \text{ a.e. in } T,$$

$$\varliminf_{\ell} \int_T \gamma_{\ell} d\mu \geq \int_T \gamma_* d\mu.$$

In view of (4.15) this establishes the inequality (3.12). Note that the subsequence $\{\ell\}$ and the function η_* would seem to depend upon the choice of the integrand l . Although this certainly applies to γ_* , it is easy to see from Example 2.3 and formula (3.7) that $\{\ell\}$ and η_* can indeed be chosen independently from l . QED

A result which is very closely related to Theorem 4.3 is due to Cesari and Suryanarayana [11c, Thm. 3.1] (rather similar results already figure in [10a]). In several respects this result is generalized by our present result. In [11c] the orientor field Q has to have the following property:

$$\eta' \geq \eta \text{ and } (\xi, \eta) \in Q(t, x) \text{ imply } (\xi, \eta') \in Q(t, x).$$

Also, it is assumed there that for all k $T_k = T$, $d_k = 0$, $\bar{d}_k = 0$, $\bar{e}_k = 0$. Other restrictions are that (T, \mathcal{T}, μ) must be nonatomic, complete and that X must be finite-dimensional. There are other, less significant differences; with respect to each of these Theorem 4.3 is the more general result. (Apart from this, it should be pointed out that the argument in [11c] is incomplete: the proof of [11c, 2.2] is not given, even though this concerns a quite nontrivial extension of Fatou's lemma in several dimensions.)

Further, Theorem 4.3 generalizes [3e, Thm. 5], which has $\bar{r} = 1$. Therefore we can refer the reader to a number of comparisons with other results in the literature made in [3e]. (Note that the measurability conditions for the orientor field there are superfluous in the light of Theorem 4.3.)

The main point made in [3e] is that when $\bar{r} = 1$ a large number of lower closure results follows from the lower semicontinuity result in Corollary 3.5. Conversely, it is well known that such lower semicontinuity results follow from lower closure results. Needless to say, the equivalence result of Proposition 4.5 advances such insights. Interestingly enough, by making use of R. V. Chacon's "biting lemma" [8] one can also obtain Theorem 4.3 from Corollary 3.5 (this has been observed independently by the referee and the author). From the above it will be clear that the necessary conditions for lower semicontinuity of e.g. [16], [27d] can also be converted into necessary conditions for lower closure in certain problems.

Let us now look at more concrete orientor fields. Similar fields figure in many existence problems of optimal control theory.

Let $q: T \times X \times V \rightarrow \mathbb{R}^r$ and $\bar{q}: T \times X \times V \rightarrow (-\infty, +\infty]^r$ be $\mathcal{T} \times \mathcal{B}(X \times V)$ -measurable functions. We shall consider the multifunction $\tilde{Q}: T \times X \rightrightarrows \mathbb{R}^r \times \mathbb{R}^r$ defined by

$$\tilde{Q}(t, x) = \{(q(t, x, v), \eta) \in \mathbb{R}^r \times \mathbb{R}^r : v \in V, \eta \geq \bar{q}(t, x, v)\}.$$

In what follows we shall consider the sequence $\{v_k\}_1^\infty$, $v_k \in \mathcal{M}(T_k; V)$, of the previous section. Let us agree to set $q(\cdot, x_k, v_k) \equiv 0$, $\bar{q}(\cdot, x_k, v_k) \equiv 0$ on $T \setminus T_k$.

THEOREM 4.6. *Suppose that $\{T_k\}_0^\infty$, $\{x_k\}_0^\infty$ satisfy (3.18)–(3.19), that*

$$(4.16) \quad \{q(\cdot, x_k, v_k)\}_1^\infty \text{ converges weakly in } \sigma(\mathcal{L}_1^r, \mathcal{L}_\infty^r) \text{ to } \xi_0 \in \mathcal{L}_1(T; \mathbb{R}^r),$$

$$(4.17) \quad \{\bar{q}(\cdot, x_k, v_k)\}_1^\infty \text{ is uniformly integrable,}$$

$$(4.18) \quad \lim_k \int_{T_k} \bar{q}(\cdot, x_k, v_k) d\mu \text{ exists (in } \mathbb{R}^r),$$

and that

$$(4.19) \quad \tilde{Q} \text{ has property (K) at } (t, x_0(t)) \text{ a.e. in } T_0,$$

$$(4.20) \quad \tilde{Q}(t, x_0(t)) \text{ is a convex subset of } \mathbb{R}^r \times \mathbb{R}^r \text{ a.e. in } T_0.$$

Then there exists $v_* \in \mathcal{M}(T_0; V)$ such that

$$(4.21) \quad \xi_0 = q(\cdot, x_0, v_*) \text{ a.e. in } T_0,$$

$$(4.22) \quad \lim_k \int_{T_k} \bar{q}(\cdot, x_k, v_k) d\mu \geq \int_{T_0} \bar{q}(\cdot, x_0, v_*) d\mu.$$

Moreover, condition (4.19) can be lifted altogether either: if (3.43) holds and a.e. in T_0

$$(4.23) \quad q(t, \cdot, \cdot) \text{ is continuous at every point of } \{x_0(t)\} \times V,$$

$$(4.24) \quad \bar{q}(t, \cdot, \cdot) \text{ is lower semicontinuous at every point of } \{x_0(t)\} \times V,$$

or under the following set of conditions

$$(4.25) \quad q(t, x_k(t), v_k(t)) - q(t, x_0(t), v_k(t)) \rightarrow 0 \text{ a.e. in } T_0,$$

$$(4.26) \quad \bar{q}(t, x_k(t), v_k(t)) - \bar{q}(t, x_0(t), v_k(t)) \rightarrow 0 \text{ a.e. in } T_0.$$

Proof. It follows from (4.17)–(4.18) that

$$\sup_k \int_{T_k} |\bar{q}(\cdot, x_k, v_k)| d\mu < +\infty.$$

Also, by definition of \tilde{Q}

$$(q(t, x_k(t), v_k(t)), \bar{q}(t, x_k(t), v_k(t))) \in \tilde{Q}(t, x_k(t)) \text{ a.e. in } T_k.$$

Applying Theorem 4.3 (with $d_k=0$, $\bar{d}_k=0$, $\bar{e}_k=0$) we find that there exists $\eta_* \in \mathcal{L}_1(T; \mathbb{R}^r)$ such that for a.e. $t \in T_0$ there is $v \in V$ with

$$(4.27) \quad \xi_0(t) = q(t, x_0(t), v), \quad \eta_*(t) \geq \bar{q}(t, x_0(t), v),$$

$$(4.28) \quad \lim_k \int_{T_k} \bar{q}(\cdot, x_k, v_k) d\mu \geq \int_{T_0} \eta_* d\mu.$$

The set of all $(t, v) \in T_0 \times V$ for which (4.27) holds is $\mathcal{T} \times \mathcal{B}(V)$ -measurable. Hence, by Aumann's measurable selection theorem [14] there exists $v_* \in \mathcal{M}(T_0; V)$ such that

$$\xi_0 = q(\cdot, x_0, v_*), \quad \eta_* \geq \bar{q}(\cdot, x_0, v_*) \text{ a.e. in } T_0.$$

Together with (4.28) this shows that (4.21)–(4.22) hold.

Next, we show that in the specified special cases condition (4.19) can be omitted. In the first case we define $\tilde{q} \equiv (\bar{q}, h)$, where h is as in (3.43). We then consider the multifunction $\tilde{Q}' : T \times X \rightrightarrows \mathbb{R}^r \times \mathbb{R}^{\bar{r}+1}$ defined as follows

$$\tilde{Q}'(t, x) \equiv \begin{cases} \tilde{Q}(t, x_0(t)) \times \mathbb{R} & \text{if } x = x_0(t), \\ \{(q(t, x, v), \eta) \in \mathbb{R}^r \times \mathbb{R}^{\bar{r}+1} : v \in V, \eta \geq \bar{q}(t, x, v)\} & \text{else.} \end{cases}$$

From the inf-compactness property of $h(t, \cdot)$ and (4.23)–(4.24) it follows by elementary reasoning that \tilde{Q}' has property (K) at $(t, x_0(t))$, irrespective of condition (4.19). We now have

$$(q(t, x_k(t), v_k(t)), \bar{q}(t, x_k(t), v_k(t))) \in \tilde{Q}'(t, x_k(t)) \text{ a.e. in } T_k,$$

and the remaining conditions of the previously established part of the theorem are easily seen to hold (with \tilde{Q}' , \bar{q} instead of \tilde{Q} , \bar{q} ; note that in view of (3.43), condition (4.18) is fulfilled without loss of generality). Hence, there exists $v_* \in \mathcal{M}(T_0; V)$ such that (4.21) holds and

$$\lim_k \int_{T_k} \bar{q}(\cdot, x_k, v_k) d\mu \geq \int_{T_0} \bar{q}(\cdot, x_0, v_*) d\mu.$$

By definition of \bar{q} this entails (4.22).

For the second case we define

$$d_k \equiv q(\cdot, x_0, v_k) - q(\cdot, x_k, v_k),$$

$$\bar{d}_k \equiv \bar{q}(\cdot, x_0, v_k) - \bar{q}(\cdot, x_k, v_k),$$

$$\tilde{Q}''(t, x) \equiv \tilde{Q}(t, x_0(t)).$$

Then (3.22)–(3.23) hold by (4.25)–(4.26). Evidently

\tilde{Q}'' has property (K) at every $(t, x) \in T \times X$,

$$(q(t, x_k(t), v_k(t)) + d_k(t), \bar{q}(t, x_k(t), v_k(t)) + \bar{d}_k(t)) \in \tilde{Q}''(t, x_k(t)) \text{ a.e. in } T_k.$$

It is now easy to verify that we may invoke Theorem 4.3 and Aumann's theorem as before to arrive at (4.21)–(4.22). QED

Remark 4.7. Define $\text{dom } \bar{q}^j$ to be the set of all $(t, x, v) \in T \times X \times V$ with $\bar{q}^j(t, x, v) < +\infty$ ($\text{dom } \tilde{Q}$ is precisely the projection of $\bigcap_{j=1}^{\bar{r}} \text{dom } \bar{q}^j$ on $T \times X$; cf. Remark 4.4). It is easy to verify that (4.23) can be replaced by the weaker condition

$$(4.23') \quad q(t, \cdot, \cdot) \text{ is continuous at every point of } \{x_0(t)\} \times V \text{ relative to the section at } t \text{ of } \bigcap_{j=1}^{\bar{r}} \text{dom } \bar{q}^j \text{ a.e. in } T_0.$$

Remark 4.8. In the literature one usually restricts the considerations from the beginning to a $\mathcal{T} \times \mathcal{B}(X \times V)$ -measurable subset D of $T \times X \times V$. One introduces functions $q_D: D \rightarrow \mathbb{R}^r$, $\bar{q}_D: D \rightarrow \mathbb{R}^{\bar{r}}$ and the multifunction $\tilde{Q}_D: D_0 \rightrightarrows \mathbb{R}^r \times \mathbb{R}^{\bar{r}}$ given by

$$\tilde{Q}_D(t, x) \equiv \{(q_D(t, x, v), \eta) \in \mathbb{R}^r \times \mathbb{R}^{\bar{r}}: (t, x, v) \in D, \eta \geq \bar{q}_D(t, x, v)\},$$

where D_0 stands for the projection of D on $T \times X$. The present setup is regained by introducing the integrand $\bar{q}^{\bar{r}+1}$ on $T \times X \times V$, given by

$$\bar{q}^{\bar{r}+1}(t, x, v) \equiv \begin{cases} 0 & \text{if } (t, x, v) \in D, \\ +\infty & \text{else,} \end{cases}$$

by letting $q: T \times X \times V \rightarrow \mathbb{R}^r$ be the extension of q_D , obtained by setting $q \equiv 0$ on $(T \times X \times V) \setminus D$, and by letting \bar{q}^j be the extension of \bar{q}_D^j with $\bar{q}^j \equiv +\infty$ on $(T \times X \times V) \setminus D$. As for (4.19), Remark 4.4 holds. Concerning the use of (4.23')–(4.24), we note that these are satisfied if a.e. in T_0

$q_D(t, \cdot, \cdot)$ is continuous at every point of $(\{x_0(t)\} \times V) \cap D_t$,

$\bar{q}_D^j(t, \cdot, \cdot)$ is lower semicontinuous at every point of $(\{x_0(t)\} \times V) \cap D_t$,
 $j = 1, \dots, \bar{r}$,

$\bar{q}^{\bar{r}+1}(t, \cdot, \cdot)$ is lower semicontinuous at every point of $\{x_0(t)\} \times V$;

here D_t stands for the section at t of D . As for (4.25)–(4.26), note that they are equivalent to having a.e. in T_0

$$q_D(t, x_k(t), v_k(t)) - q_D(t, x_0(t), v_k(t)) \rightarrow 0,$$

$$\bar{q}_D(t, x_k(t), v_k(t)) - q_D(t, x_0(t), v_k(t)) \rightarrow 0,$$

$$(t, x_k(t), v_k(t)) \in D, (t, x_0(t), v_k(t)) \in D \text{ for large enough } k.$$

Remark 4.9. Conditions (3.43), (4.23)–(4.24) suggest that this special case of Theorem 4.6 can also be proven directly, in the style of § 3. Indeed, this is true. We shall leave it to the reader to work out the details for the integrands g_j on $T \times X \times \mathbb{R}^r \times V$, defined by

$$g_j(t, x, \xi, v) \equiv \begin{cases} \bar{q}^j(t, x, v) & \text{if } q(t, x, v) = \xi, \\ +\infty & \text{else.} \end{cases}$$

Remark 4.10. An obvious extension of (3.43), (4.23)–(4.24) is obtained by letting h also depend on the x -variable: Suppose instead of (3.43) that there exists a nonnegative $\mathcal{T} \times \mathcal{B}(X \times V)$ -measurable integrand h on $T \times X \times V$ such that a.e. in T_0

$$(3.43') \quad h(t, \cdot, \cdot) \text{ is lower semicontinuous at every point of } \{x_0(t)\} \times V,$$

$$(3.43'') \quad \sup_k h(t, x^k, v^k) < +\infty \text{ implies that } \{v^k\}_1^\infty \text{ has a limit point for every } \{x^k\}_1^\infty \subset X, \{v^k\}_1^\infty \subset V \text{ with } x^k \rightarrow x_0(t).$$

Suppose further that

$$(3.43''') \quad \sup_k \int_{T_k} h(\cdot, x_k, v_k) d\mu < +\infty.$$

Then the conclusions of Theorem 4.6 regarding (3.43), (4.23)–(4.24) remain valid. This is seen by noting that the multifunction \tilde{Q}' in the proof of Theorem 4.6 then still has property (K) at $(t, x_0(t))$ a.e. in T_0 .

Remark 4.11. Note that the graph of the multifunction \tilde{Q} is the projection on $T \times X \times \mathbb{R}^r \times \mathbb{R}^{\bar{r}}$ of a $\mathcal{T} \times \mathcal{B}(X \times \mathbb{R}^r \times \mathbb{R}^{\bar{r}} \times V)$ -measurable set. In itself this does not bestow any useful sort of measurability on \tilde{Q} . Hence, our consideration of nonmeasurable integrands in § 3 is justified. Note also that conditions like (4.25)–(4.26) warrant the use of the perturbations d_k, \bar{d}_k in §§ 3–4.

By taking into account our Remarks 4.4, 4.8 it is easy to see that Theorem 4.6 generalizes [3e, Thms. 7, 10] (where $\bar{r} = 1$ among other things). This also means that a large number of lower closure results in the literature (e.g. [5], [10c, d], [11a, b, d]) follow from our result.

In conclusion, we wish to remark that a large number of existence results can be obtained as follows (applications include the optimal control of ordinary differential equations, functional-differential equations, nonlinear integral equations and elliptic boundary value problems): One applies Corollary 3.5 to the dynamical system and Theorem 4.6 to the orientor field \tilde{Q} , where \tilde{q}^1 may stand for the usual cost functional (for instance). Details can be found in [3g] and in forthcoming work by the author.

Appendix A. In this appendix we shall gather some facts about relaxed control theory that were established in [3]. In particular, we shall prove Theorem I and Lemmas II, III.

Since S is a standard Borel space, we may identify it with a Borel subset of a compact metric space \hat{S} , the metric of which will be denoted by ρ [12, III]. Hence $\mathcal{M}(T; S)$ and $\mathcal{R}(T; S)$ are subsets of $\mathcal{M}(T; \hat{S})$ and $\mathcal{R}(T; \hat{S})$ respectively.

Define $\mathcal{C}_e(\hat{S}) \subset \mathcal{C}(\hat{S})$ as follows:

$$\mathcal{C}_e(\hat{S}) \equiv \{-n\rho(\cdot, s) + \eta; n \in \mathbb{N}, s \in \hat{S}, \eta \in \mathbb{R}\}.$$

LEMMA A.1. *For every $g \in \mathcal{G}^+(T; S)$ there exist a null set $N \subset T$ and sequences $\{T_p\}_1^\infty \subset \mathcal{T}$, $\{c_p\}_1^\infty \subset \mathcal{C}_e(\hat{S})$ such that*

$$(a.1) \quad g(t, s) = \sup_p 1_{T_p}(t) c_p(s) \text{ on } (T \setminus N) \times S.$$

Proof. Let $\{s^j\}_1^\infty$ be a countable dense sequence in S and let $\{\gamma^k\}_1^\infty$ be an enumeration of the rationals. For $j, k, m \in \mathbb{N}$ we define $c_{jkm} \in \mathcal{C}_e(\hat{S})$ by $c_{jkm} \equiv \gamma^k - m\rho(s^j, \cdot)$ and $B_{jkm} \equiv \{t \in T: c_{jkm}(s) \leq g(t, s) \text{ on } S\}$. Then B_{jkm} is the projection of the set of all $(t, s) \in T \times S$ such that $c_{jkm}(s) > g(t, s)$ onto T . Hence, B_{jkm} belongs to the completion of \mathcal{T} with respect to μ [9, III.23]; this implies that there exists $T_{jkm} \in \mathcal{T}$, $T_{jkm} \subset B_{jkm}$, such that $B_{jkm} \setminus T_{jkm}$ is contained in a null set N_{jkm} . Using nonnegativity and lower semicontinuity of $g(t, \cdot)$ it is entirely elementary to prove that $\sup_{j,k,m} 1_{B_{jkm}}(t) c_{jkm}(s) = g(t, s)$ on $T \times S$. By taking N to be the union of all N_{jkm} the result now follows. QED

A $\mathcal{T} \times \mathcal{B}(S)$ -measurable integrand g on $T \times S$ is said to be a *Carathéodory integrand* if $g(t, \cdot)$ is continuous on S for every $t \in T$ and there exists $\phi \in \mathcal{L}_1(T; \mathbb{R})$ such that

$$|g(t, s)| \leq \phi(t) \text{ on } T \times S.$$

The set of all Carathéodory integrands on $T \times S$ will be denoted by $\mathcal{G}_C(T; S)$.

We shall equip $\mathcal{R}(T; \hat{S})$ with the coarsest topology $\hat{\mathcal{O}}$ for which all functions $\delta \mapsto \int g(\cdot, \delta) d\mu$ are continuous on $\mathcal{R}(T; \hat{S})$, $g \in \mathcal{G}_C(T; \hat{S})$. Its relative topology on $\mathcal{R}(T; S)$ will be denoted \mathcal{O} .

LEMMA A.2. *The topology \mathcal{O} is the coarsest topology for which all functions $\delta \mapsto \int_T g(\cdot, \delta) d\mu$ are lower semicontinuous on $\mathcal{R}(T; S)$, $g \in \mathcal{G}^+(T; S)$.*

Proof. Call the topology that figures in the statement above \mathcal{O}' . Given $g \in \mathcal{G}_C(T; \hat{S})$, its restriction to $T \times S$ is easily seen to belong to $\mathcal{G}_C(T; S)$. It follows easily that $\delta \mapsto \int_T g(\cdot, \delta) d\mu$ is \mathcal{O}' -continuous on $\mathcal{R}(T; S)$. Conversely, given $g \in \mathcal{G}^+(T; S)$, let $\{T_p\}_1^\infty$ and $\{c_p\}_1^\infty$ be as in Lemma A.1. Define $\kappa_p: \mathbb{R} \rightarrow [0, p]$ by $\kappa_p(\gamma) = \max(\min(\gamma, p), 0)$ and set

$$g_p(t, s) \equiv \kappa_p(\sup_{j \leq p} 1_{T_j}(t) c_j(s)).$$

Then it follows by the monotone convergence theorem and (a.1) that for every $\delta \in \mathcal{R}(T; S)$

$$\int_T g(\cdot, \delta) d\mu = \sup_p \int_T g_p(\cdot, \delta) d\mu.$$

Hence, $\delta \mapsto \int_T g(\cdot, \delta) d\mu$ is \mathcal{O} -lower semicontinuous on $\mathcal{R}(T; S)$ for every $g \in \mathcal{G}^+(T; S)$. This finishes the proof. QED

Let $M^+(\hat{S})$ denote the set of all bounded nonnegative measures on \hat{S} ; set $M(\hat{S}) \equiv M^+(\hat{S}) - M^+(\hat{S})$. It is well known that the usual L_∞ -space $L_\infty(T, \mathcal{T}, \mu; M(\hat{S}))$ of essentially bounded \mathcal{T} -measurable functions from T into $M(\hat{S})$ has as its dual the usual L_1 -space $L_1(T, \mathcal{T}, \mu; \mathcal{C}(\hat{S}))$ of integrable functions from T into $\mathcal{C}(\hat{S})$ [18, VII.7]; cf. [24, p. 301] for a short proof. (For a good understanding we note that $L_\infty(T, \mathcal{T}, \mu; M(\hat{S}))$ consists of (equivalence classes of) functions $\sigma: T \rightarrow M(\hat{S})$ that are Borel measurable with respect to the usual weak topology on $M(\hat{S})$ and have $\text{ess sup}_t |\sigma(t)|_v < +\infty$, where $|\cdot|_v$ stands for the total variation norm.) Let Σ be the set of all $\sigma \in L_\infty(T, \mathcal{T}, \mu; M(\hat{S}))$ for which $\sigma(t) \in M_1^+(\hat{S})$ a.e. in T . It will be equipped with the relative $\sigma(L_\infty, L_1)$ -topology.

LEMMA A.3. Σ is compact and sequentially compact for the topology $\sigma(L_\infty, L_1)$.

Proof. Compactness of Σ is well known; it follows from the above by a simple application of the Alaoglu-Bourbaki theorem [9, V.2], [32, IV]. Further, it is well known that Σ is metrizable if the σ -algebra \mathcal{T} is countably generated, since in that case $L_1(T, \mathcal{T}, \mu; \mathcal{C}(\hat{S}))$ is separable [15, 12.F]; cf. [32, IV]. Sequential compactness is proven next (cf. [20]). Let $\{\sigma_k\}_1^\infty \subset \Sigma$ be arbitrary and let \mathcal{T}_0 stand for the σ -algebra generated by this sequence; it is countably generated, since $M_1^+(\hat{S})$ is metrizable and separable for the weak topology. Hence, it follows from the above that there exist a subsequence $\{\ell\}$ of $\{k\}$ and a \mathcal{T}_0 -measurable $\sigma_* \in \Sigma$ such that $\{\sigma_\ell\}$ converges to σ_* in the topology $\sigma(L_\infty(T, \mathcal{T}_0, \mu; M(\hat{S})), L_1(T, \mathcal{T}_0, \mu; \mathcal{C}(\hat{S})))$. Since each element of $L_1(T, \mathcal{T}, \mu; \mathcal{C}(\hat{S}))$ has a conditional expectation in $L_1(T, \mathcal{T}_0, \mu; \mathcal{C}(\hat{S}))$ [9, VIII.32], it follows directly that $\{\sigma_\ell\}$ also converges to σ_* in $\sigma(L_\infty, L_1)$. QED

LEMMA A.4. $\mathcal{R}(T; \hat{S})$ is compact and sequentially compact for the topology $\hat{\mathcal{O}}$.

Proof. Denote by χ the usual quotient mapping from the set of all $\delta \in \mathcal{M}(T; M(\hat{S}))$ such that $\sup_{t \in T} |\delta(t)|_v < +\infty$ into $L_\infty(T, \mathcal{T}, \mu; M(\hat{S}))$. It is easy to see that $\hat{\mathcal{O}}$ is the coarsest topology for which χ is continuous with respect to $\sigma(L_\infty, L_1)$. Since χ is a surjection from $\mathcal{R}(T; \hat{S})$ into Σ , the desired result now follows directly from Lemma A.3. QED

Proof of Theorem I. By supposition there exists $h \in \mathcal{H}(T; S)$ such that $\sup_k \int_T h(\cdot, s_k) d\mu < +\infty$. For every k a parametrized measure $\delta_k \in \mathcal{R}(T; \hat{S})$ is defined by taking $\delta_k(t)$ to be the Dirac measure (point mass) at $s_k(t)$. By Lemma A.4 there

exist a subsequence $\{\ell\}$ of $\{k\}$ and $\delta_* \in \mathcal{R}(T; \hat{S})$ such that $\{\delta_\ell\}$ converges to δ_* in the topology $\hat{\mathcal{O}}$. We shall show that in fact $\delta_* \in \mathcal{R}(T; S)$, which would mean that $\{\delta_\ell\}$ converges to δ_* in the relative topology \mathcal{O} . To see this, we define $\hat{h}: T \times \hat{S} \rightarrow [0, +\infty]$ by

$$\hat{h}(t, s) \equiv \begin{cases} h(t, s) & \text{on } T \times S, \\ +\infty & \text{on } T \times (\hat{S} \setminus S). \end{cases}$$

Since S is Borel in \hat{S} , \hat{h} is $\mathcal{T} \times \mathcal{B}(\hat{S})$ -measurable. Also, the topological homeomorphism that makes S into a subset of \hat{S} turns compact subsets of S into compact subsets of \hat{S} . Hence, for every $t \in T$, $\gamma \in \mathbb{R}$ the set $\{s \in \hat{S}: \hat{h}(t, s) \leq \gamma\}$ is compact. We conclude that $\hat{h} \in \mathcal{H}(T; \hat{S})$. By compactness of \hat{S} the latter set is precisely $\mathcal{G}^+(T; \hat{S})$, so it follows from the above that

$$\int_T \hat{h}(\cdot, \delta_*) d\mu \leq \varliminf_{\ell} \int_T \hat{h}(\cdot, \delta_\ell) d\mu = \varliminf_{\ell} \int_T h(\cdot, \delta_\ell) d\mu < +\infty.$$

By definition of \hat{h} this implies that for a.e. $t \in T$ the probability measure $\delta_*(t)$ is carried by S . Of course, we can modify $\delta_*(t)$ on the exceptional set, and this does not affect the values of integrals. Thus we may conclude that $\delta_* \in \mathcal{R}(T; S)$ without loss of generality. By definition of \mathcal{O} we have now for every $g \in \mathcal{G}^+(T; S)$ that (2.5) holds. For $g \in \mathcal{G}(T; S)$ such that there exists $\phi \in \mathcal{L}_1(T; \mathbb{R})$ with

$$g(t, s) \geq \phi(t) \quad \text{on } T \times S,$$

(2.5) is also valid, as is easily seen by working with $g - \phi \in \mathcal{G}^+(T; S)$. Further, for $g \in \mathcal{G}(T; S)$ with (2.6) there exists for every $\varepsilon > 0$ a constant $\gamma > 0$ such that

$$\int_T g(\cdot, s_\ell) d\mu \geq \int_T \max(-\gamma, g(\cdot, s_\ell)) d\mu - \varepsilon \quad \text{for all } \ell.$$

In view of the above, (2.5) is now easy to derive.

Finally, we shall demonstrate that for a.e. $t \in T$ the measure $\delta_*(t)$ is carried by the limit points of $\{s_\ell(t)\}$. Let $\hat{\mathbb{N}} \equiv \mathbb{N} \cup \{\infty\}$ be the usual Alexandrov compactification of the natural numbers; this is a metrizable compact space. We define $\tilde{g}: T \times \hat{\mathbb{N}} \times S \rightarrow [0, +\infty]$ as follows:

$$\tilde{g}(t, p, s) \equiv \begin{cases} 0 & \text{if } s \in \text{cl} \{s_\ell(t): \ell \geq p\}, \\ +\infty & \text{else,} \end{cases}$$

for $p \in \mathbb{N}$. For $p = \infty$ we define

$$\tilde{g}(t, \infty, s) \equiv \begin{cases} 0 & \text{if } s \in \bigcap_{p=1}^{\infty} \text{cl} \{s_\ell(t): \ell \geq p\}, \\ +\infty & \text{else.} \end{cases}$$

It is easy to check that \tilde{g} belongs to $\mathcal{G}^+(T; \hat{\mathbb{N}} \times S)$. Let ρ' be a compatible metric on $\hat{\mathbb{N}}$; we shall equip $\hat{\mathbb{N}} \times \hat{S}$ with the metric $\rho' + \rho$. Let $\{T_m\}_1^\infty \subset \mathcal{T}$ and $\{c_m\}_1^\infty \subset \mathcal{C}_e(\hat{\mathbb{N}} \times \hat{S})$ correspond to \tilde{g} as asserted in Lemma A.1. Define $\tilde{g}_m \equiv \kappa_m(\sup_{j \leq m} 1_{T_j c_j})$, where κ_m is as in the proof of Lemma A.2. Note that for every m there exists a Lipschitz constant γ_m such that

$$|\tilde{g}_m(t, p, s) - \tilde{g}_m(t, p', s')| \leq \gamma_m(\rho'(p, p') + \rho(s, s')).$$

It follows that for every m

$$\lim_{\ell} \left(\int_T \tilde{g}_m(\cdot, \ell, s_\ell) d\mu - \int_T \tilde{g}_m(\cdot, \infty, s_\ell) d\mu \right) = 0.$$

Hence we find for every m

$$\lim_{\ell} \int_T \tilde{g}_m(\cdot, \ell, s_\ell) d\mu = \lim_{\ell} \int_T \tilde{g}_m(\cdot, \infty, s_\ell) d\mu \cong \int_T \tilde{g}_m(\cdot, \infty, \delta_*) d\mu,$$

since $\{\delta_\ell\}$ converges in $\hat{\mathcal{O}}$ to δ_* . Also, we have by the monotone convergence theorem for every $\delta \in \mathcal{R}(T; \mathbb{N} \times S)$

$$\int_T \tilde{g}(\cdot, \delta) d\mu = \sup_m \int_T \tilde{g}_m(\cdot, \delta) d\mu.$$

Combined with the above this gives

$$\int_T \tilde{g}(\cdot, \infty, \delta_*) d\mu \leq \lim_{\ell} \int_T \tilde{g}(\cdot, \ell, s_\ell) d\mu = 0,$$

where the latter equality follows by definition of \tilde{g} . The desired conclusion now also follows from the definition of \tilde{g} . QED

LEMMA A.5. *For every lower semicontinuous integrand l on $T \times S$ there exists a normal integrand $g \in \mathcal{G}(T; S)$, $g \geq l$, such that for every $u \in \mathcal{M}(T; S)$, $\phi \in \mathcal{M}(T; [-\infty, +\infty])$*

$$l(t, u(t)) \leq \phi(t) \text{ a.e. in } T \text{ implies that } g(t, u(t)) \leq \phi(t) \text{ a.e. in } T.$$

Proof. Suppose first that $l \geq 0$. Although l need not be $\mathcal{T} \times \mathcal{B}(S)$ -measurable, the proof of Lemma A.1 shows that there exist a sequence $\{B_p\}_1^\infty$ of subsets of T and $\{c_p\}_1^\infty \subset \mathcal{C}_e(\hat{S})$ such that

$$(a.2) \quad l(t, s) = \sup_p 1_{B_p}(t) c_p(s) \text{ on } T \times S.$$

For every p there exists $T_p \in \mathcal{T}$, $T_p \supset B_p$, such that $\mu(T_p)$ equals the outer measure of B_p [26, I.4]. Define on $T \times S$

$$g(t, s) \equiv \sup_p 1_{T_p}(t) c_p(s);$$

then $g \geq l$ and $g \in \mathcal{G}^+(T; S)$. Let $u \in \mathcal{M}(T; S)$, $\phi \in \mathcal{M}(T; [0, +\infty])$ be arbitrary with

$$l(t, u(t)) \leq \phi(t) \text{ for all } t \in T.$$

(Evidently, it is enough to prove the desired implication in this case.) We now have by (a.2) that for every p the set B_p is contained in $A_p \equiv \{t \in T: c_p(u(t)) \leq \phi(t)\}$. By \mathcal{T} -measurability of u , A_p is \mathcal{T} -measurable. By definition of T_p it follows that $T_p \setminus A_p$ is a null set; in other words, we must have

$$1_{T_p}(t) c_p(u(t)) \leq \phi(t) \text{ a.e. in } T.$$

Thus, the desired implication holds if $l \geq 0$. For general l we define $l' \equiv \exp(l)$. From the previous step the desired conclusion then follows easily by monotonicity and continuity of the transformation involved. QED

LEMMA A.6. *For every $\bar{\mathcal{T}} \times \mathcal{B}(S)$ -measurable lower semicontinuous integrand g on $T \times S$ there exist a null set $N \subset T$ and $g' \in \mathcal{G}(T; S)$ such that*

$$g'(t, s) = g(t, s) \text{ on } (T \setminus N) \times S;$$

here $\bar{\mathcal{T}}$ stands for the completion of \mathcal{T} with respect to μ .

Proof. Suppose first that $g \geq 0$. From the proof given for Lemma A.1 it follows that there exist sequences $\{B_p\}_1^\infty \subset \bar{\mathcal{T}}$, $\{c_p\}_1^\infty \subset \mathcal{C}_e(\hat{S})$ with

$$g(t, s) = \sup_p 1_{B_p}(t) c_p(s) \text{ on } T \times S.$$

For every $p \in \mathbb{N}$ there exists $T_p \in \mathcal{T}$, $B_p \subset T_p$, such that $T_p \setminus B_p$ is a null set. Defining on $T \times S$

$$g'(t, s) \equiv \sup_p 1_{T_p}(t) c_p(s),$$

we see that $g' \geq g \geq 0$ and that g' has the required properties. For general g we apply the previous step to $\exp(g)$. QED

Proof of Lemma II. Let \mathcal{g} be as given. By Lemma A.5 there exists $\tilde{g} \in \mathcal{G}(T; X \times \mathbb{R}^r \times \mathbb{R}^f)$, $\tilde{g} \geq \mathcal{g}$, such that for every $x \in \mathcal{M}(T; X)$, $\xi \in \mathcal{M}(T; \mathbb{R}^r)$, $\eta \in \mathcal{M}(T; \mathbb{R}^f)$, $\phi \in \mathcal{M}(T; [-\infty, +\infty])$

$$(a.3) \quad \mathcal{g}(\cdot, x, \xi, \eta) \leq \phi \text{ a.e. in } T \text{ implies } \tilde{g}(\cdot, x, \xi, \eta) \leq \phi \text{ a.e. in } T.$$

We define the following Fenchel conjugate functions:

$$\tilde{g}_0^*(t, \xi, \eta) \equiv \sup \{ \langle \xi, \xi' \rangle + \langle \eta, \eta' \rangle - \tilde{g}(t, x_0(t), \xi', \eta') : \xi' \in \mathbb{R}^r, \eta' \in \mathbb{R}^f \},$$

$$\tilde{g}_0^{**}(t, \xi, \eta) \equiv \sup \{ \langle \xi, \xi' \rangle + \langle \eta, \eta' \rangle - \tilde{g}_0^*(t, \xi', \eta') : \xi' \in \mathbb{R}^r, \eta' \in \mathbb{R}^f \}.$$

Since $(t, \xi, \eta) \mapsto \tilde{g}(t, x_0(t), \xi, \eta)$ is certainly $\tilde{\mathcal{T}} \times \mathcal{B}(\mathbb{R}^r \times \mathbb{R}^f)$ -measurable, it follows by [9, III.39] that \tilde{g}_0^* , \tilde{g}_0^{**} are also $\tilde{\mathcal{T}} \times \mathcal{B}(\mathbb{R}^r \times \mathbb{R}^f)$ -measurable. It is well known that for every $t \in T$, $\tilde{g}_0^{**}(t, \cdot, \cdot)$ is the lower semicontinuous convex hull of $\tilde{g}(t, x_0(t), \cdot, \cdot)$. By (3.8) and $\tilde{g} \geq \mathcal{g}$ this implies that for every $t \in T$

$$(a.4) \quad \tilde{g}_0(t, x_0(t), \cdot, \cdot) \geq \tilde{g}_0^{**}(t, \cdot, \cdot) \geq \mathcal{g}(t, x_0(t), \cdot, \cdot).$$

Also, by Lemma A.6 the above imply that there exist a null set $N \subset T$ and $\tilde{g}_0 \in \mathcal{G}(T; \mathbb{R}^r \times \mathbb{R}^f)$ such that for every $t \in T \setminus N$

$$(a.5) \quad \tilde{g}_0(t, \cdot, \cdot) = \tilde{g}_0^{**}(t, \cdot, \cdot).$$

For $t \in T \setminus N$ we now define

$$g(t, x, \xi, \eta) \equiv \begin{cases} \tilde{g}(t, x, \xi, \eta) & \text{if } x \neq x_0(t), \\ \tilde{g}_0(t, \xi, \eta) & \text{if } x = x_0(t), \end{cases}$$

and for $t \in N$ we set $g(t, x, \xi, \eta) \equiv +\infty$. Then $g \geq \mathcal{g}$ and g is $\mathcal{T} \times \mathcal{B}(X \times \mathbb{R}^r \times \mathbb{R}^f)$ -measurable. From the first inequality in (a.4), (a.5) and lower semicontinuity of $\tilde{g}(t, \cdot, \cdot, \cdot)$ it follows now by elementary reasoning that $g(t, \cdot, \cdot, \cdot)$ is lower semicontinuous. Hence $g \in \mathcal{G}(T; X \times \mathbb{R}^r \times \mathbb{R}^f)$ and (3.9)–(3.10) hold. To prove (3.11), let x, ξ, η be as in (a.3) and arbitrary. As remarked in § 1, there exists $\phi \in \mathcal{M}(T; [-\infty, +\infty])$, $\phi \geq \mathcal{g}(\cdot, x, \xi, \eta)$ a.e. in T , such that

$$\int_T \phi \, d\mu = \int_T \tilde{\mathcal{g}}(\cdot, x, \xi, \eta) \, d\mu.$$

Note that $\tilde{g}(t, \cdot, \cdot, \cdot) \geq g(t, \cdot, \cdot, \cdot)$ for $t \in T \setminus N$, by definition of g . Hence, it follows from (a.3) that $\phi \geq g(\cdot, x, \xi, \eta)$ a.e. in T and so

$$\int_T \tilde{\mathcal{g}}(\cdot, x, \xi, \eta) \, d\mu \geq \int_T g(\cdot, x, \xi, \eta) \, d\mu.$$

The converse inequality holds trivially. QED

Proof of Lemma III. We may work with $S \equiv V$, so that we can use the results obtained in this appendix. We shall have to work with Σ rather than $\mathcal{R}(T; \hat{S})$, since

Σ lies in the locally convex Hausdorff space $L_\infty(T, \mathcal{T}, \mu; M(\hat{S}))$. For this purpose we define, by abuse of notation, for any $\sigma \in \Sigma$, $g \in \mathcal{G}(T; \hat{S})$

$$\int_T g(\cdot, \sigma) d\mu \equiv \int_T g(\cdot, \delta) d\mu,$$

where δ stands for any $\delta \in \mathcal{R}(T; \hat{S})$ with $\chi(\delta) = \sigma$; cf. the proof of Lemma A.4. This definition makes sense and from what was said in proving Lemma A.4 it follows that lower semicontinuity of $\sigma \mapsto \int_T g(\cdot, \sigma) d\mu$ with respect to $\sigma(L_\infty, L_1)$ follows from lower semicontinuity of $\delta \mapsto \int_T g(\cdot, \delta) d\mu$ with respect to $\hat{\mathcal{O}}$ (and conversely). Let $h \in \mathcal{H}(T; S)$ be as given; we define $\hat{h} \in \mathcal{H}(T; \hat{S})$ to be its extension defined in the proof of Theorem I. Define $\Sigma(\hat{h})$ to be the (compact) set of all $\sigma \in \Sigma$ with $\int_T \hat{h}(\cdot, \sigma) d\mu \leq \int_T h(\cdot, \delta^*) d\mu$. Note that by definition of \hat{h} every $\sigma \in \Sigma(\hat{h})$ has $\sigma(t)$ carried by S a.e. in T . As follows from Theorem I (or at least its obvious nonsequential analogue), the functions $\sigma \mapsto \int_T g(\cdot, \sigma) d\mu$ are lower semicontinuous on $\Sigma(\hat{h})$; note that uniform integrability—as in (2.6)—is guaranteed by (3.30). Hence, the set of all $\sigma \in \Sigma(\hat{h})$ with

$$\int_T g_j(\cdot, \sigma) d\mu \leq \int_T g_j(\cdot, \delta^*) d\mu, \quad j = 1, \dots, m,$$

is compact; therefore it has an extreme point σ_* by the Krein–Milman theorem. By [6, Prop. II.2]—a consequence of Carathéodory's theorem; cf. [21]—it follows that σ_* is a convex combination of at most $m+1$ extreme points of $\Sigma(\hat{h})$. By the same result every extreme point of $\Sigma(\hat{h})$ is a convex combination of at most two extreme points of Σ . We thus conclude that σ_* is a convex combination of at most $2m+2$ extreme points $\sigma_1, \dots, \sigma_{2m+2}$ of Σ . By [14, Thms. 5.2, 9.3] there corresponds an $s_i \in \mathcal{M}(T; \hat{S})$ to each σ_i such that $\chi(\varepsilon_i) = \sigma_i$, where $\varepsilon_i(t)$ is the Dirac measure at $s_i(t)$; in fact, we have that $s_i \in \mathcal{M}(T; S)$, as follows easily by $\sigma_* \in \Sigma(\hat{h})$. We now find that for certain $\alpha_1, \dots, \alpha_{2m+2} \geq 0$, $\sum_{i=1}^{2m+2} \alpha_i = 1$,

$$\beta_j \equiv \sum_{i=1}^{2m+2} \alpha_i \int_T g_j(\cdot, s_i) d\mu \leq \int_T g_j(\cdot, \delta^*) d\mu, \quad j = 1, \dots, m.$$

By a well-known extension of Lyapunov's theorem [9, IV.17] there exists, in view of the nonatomicity supposition, a function $s^* \in \mathcal{M}(T; S)$ with

$$\beta_j = \int_T g_j(\cdot, s^*) d\mu, \quad j = 1, \dots, m,$$

and the proof is thereby finished. QED

A different proof of Lemma II, based on [31, Prop. 14], has been given in [3k]. (In turn, the above result from [31] has been generalized in [3j].)

Acknowledgments. Major sources of inspiration for the author have been the important contributions to relaxed control theory made by L. C. Young, E. J. McShane, J. Warga, H. Berliocchi and J.-M. Lasry. In another direction, the author has been inspired by the work of L. Cesari on lower closure and existence. Finally, our treatment of constraints—made implicit by working with extended real-valued functions and orientor fields which may have empty values—shows the influence of the work by R. T. Rockafellar on the subject of deparametrization.

REFERENCES

- [1a] Z. ARTSTEIN, *On a variational problem*, J. Math. Anal. Appl., 45 (1974), pp. 404–415.
- [1b] ———, *A note on Fatou's lemma in several dimensions*, J. Math. Econom., 6 (1979), pp. 277–282.
- [2] R. J. AUMANN AND M. PERLES, *A variational problem arising in economics*, J. Math. Anal. Appl., 11 (1965), pp. 488–503.
- [3a] E. J. BALDER, *On a useful compactification for optimal control problems*, J. Math. Anal. Appl., 72 (1979), pp. 391–398.
- [3b] ———, *Relaxed inf-compactness for variational problems by Hilbert cube compactification*, J. Math. Anal. Appl., 79 (1981), pp. 1–12.
- [3c] ———, *Lower semicontinuity of integral functionals with nonconvex integrands by relaxation-compactification*, this Journal, 19 (1981), pp. 533–542.
- [3d] ———, *A new look at the existence of p -optimal policies in dynamic programming*, Math. Oper. Res., 6 (1981), pp. 513–517.
- [3e] ———, *Lower closure problems with weak convergence conditions in a new perspective*, this Journal, 20 (1982), pp. 198–210.
- [3f] ———, *On lower closure and lower semicontinuity in the existence theory for optimal control*, in System Modelling and Optimization, R. F. Drenick and F. Kozin, eds., Lecture Notes in Control and Information Sciences 38, Springer-Verlag, Berlin, 1982, pp. 158–164.
- [3g] ———, *On existence problems for the optimal control of certain nonlinear integral equations of Urysohn type*, J. Optim. Theory Appl., 42 (1984), pp. 447–465.
- [3h] ———, *An existence result for optimal economic growth problems*, J. Math. Anal. Appl., 95 (1983), pp. 195–213.
- [3i] ———, *Mathematical foundations of statistical decision theory: A modern viewpoint*, Statistics and Decisions, to appear.
- [3j] ———, *An extension of the essential supremum concept with applications to normal integrands and multifunctions*, Bull. Austral. Math. Soc., 27 (1983), pp. 407–418.
- [3k] ———, *Lower closure for orientor fields by lower semicontinuity of outer integral functionals*, submitted for publication.
- [3l] ———, *A unifying note on Fatou's lemma in several dimensions*, Math. Oper. Res., 9 (1984), to appear.
- [3m] ———, *Existence results without convexity conditions for general problems of optimal control with singular components*, J. Math. Anal. Appl., 101 (1984), to appear.
- [4] H. T. BANKS AND M. Q. JACOBS, *The optimization of trajectories of linear functional differential equations*, this Journal, 8 (1970), pp. 461–488.
- [5a] L. D. BERKOVITZ, *Existence and lower closure theorems for abstract control problems*, this Journal, 12 (1974), pp. 27–42.
- [5b] ———, *A lower closure theorem for abstract control problems with L_p -bounded controls*, J. Optim. Theory Appl., 14 (1974), pp. 521–528.
- [6] H. BERLIOCCI AND J. M. LASRY, *Intégrales normales et mesures paramétrées en calcul des variations*, Bull. Soc. Math. France, 101 (1973), pp. 129–184.
- [7] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
- [8] J. K. BROOKS AND R. V. CHACON, *Continuity and compactness of measures*, Adv. in Math., 37 (1980), pp. 16–26.
- [9] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Mathematics 580, Springer-Verlag, Berlin, 1977.
- [10a] L. CESARI, *Existence theorems for optimal control problems of the Mayer type*, this Journal, 6 (1968), pp. 517–552.
- [10b] ———, *An existence theorem without convexity conditions*, this Journal, 12 (1974), pp. 319–331.
- [10c] ———, *Closure theorems for orientor fields and weak convergence*, Arch. Rational Mech. Anal., 55 (1974), pp. 332–356.
- [10d] ———, *Lower semicontinuity and lower closure theorems without seminormality conditions*, Ann. Mat. Pura Appl., 98 (1974), pp. 381–397.
- [11a] L. CESARI AND M. B. SURYANARAYANA, *Closure theorems without seminormality conditions*, J. Optim. Theory Appl., 15 (1975), pp. 441–465.
- [11b] ———, *Nemitsky's operators and lower closure theorems*, J. Optim. Theory Appl., 19 (1976), pp. 165–183.
- [11c] ———, *An existence theorem for Pareto problems*, Nonlinear Anal., 2 (1978), pp. 225–233.
- [11d] ———, *On recent existence theorems in the theory of optimization*, J. Optim. Theory Appl., 31 (1980), pp. 379–415.
- [12] C. DELLACHERIE AND P.-A. MEYER, *Probabilities and Potential*, Hermann, Paris, 1975; English transl., North-Holland, Amsterdam, 1978.

- [13] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, Dunod, Paris, 1972; English transl., North-Holland, Amsterdam, 1976.
- [14] C. J. Himmelberg, *Measurable relations*, Fund. Math., 87 (1975), pp. 53–72.
- [15] R. B. Holmes, *Geometric Functional Analysis and Its Applications*, Springer-Verlag, Berlin, 1975.
- [16] A. D. Ioffe, *On lower semicontinuity of integral functionals*, I, II, this Journal, 15 (1977), pp. 521–538, pp. 991–1000.
- [17] A. D. Ioffe and V. M. Tikhomirov, *Theory of Extremal Functions*, Nauka, Moscow, 1974; German transl., Deutscher Verlag der Wissenschaften, Berlin, 1979.
- [18] A. and C. Ionescu-Tulcea, *Topics in the Theory of Lifting*, Springer-Verlag, Berlin, 1969.
- [19] M. Q. Jacobs, *Attainable sets in linear systems with unbounded controls*, in Mathematical Theory of Control, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 46–53.
- [20] H.-P. Kirschner, *On the risk-equivalence of two methods of randomization in statistics*, J. Multivariate Anal., 6 (1976), pp. 159–166.
- [21] V. Klee, *On a theorem of Dubins*, J. Math. Anal. Appl., 7 (1963), pp. 425–427.
- [22] R. M. Lewis and R. B. Vinter, *Relaxation of optimal control problems to equivalent convex programs*, J. Math. Anal. Appl., 74 (1980), pp. 475–493.
- [23] E. J. McShane, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438–485.
- [24] P.-A. Meyer, *Probability and Potentials*, Hermann, Paris, 1966; English transl., Blaisdell, Waltham, MA, 1966.
- [25] L. W. Neustadt, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.
- [26] J. Neveu, *Mathematical Foundations of the Calculus of Probability*, Masson, Paris, 1964; English transl., Holden-Day, San Francisco, 1965.
- [27a] C. Olech, *Existence theory in optimal control problems—the underlying ideas*, in International Conference on Differential Equations, H. A. Antosiewicz, ed., Academic Press, New York, 1975, pp. 612–629.
- [27b] ———, *Existence theory in optimal control*, in Control Theory and Topics in Functional Analysis I, I.A.E.A., Vienna, 1976, pp. 291–328.
- [27c] ———, *Weak lower semicontinuity of integral functionals*, J. Optim. Theory Appl., 19 (1976), pp. 3–16.
- [27d] ———, *A characterization of L_1 -weak lower semicontinuity of integral functionals*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 25 (1977), pp. 135–142.
- [28] R. T. Rockafellar, *Integral functionals, normal integrands and measurable selections*, in Nonlinear Operators and the Calculus of Variations, J. P. Gossez et al., eds., Lecture Notes in Mathematics 543, Springer-Verlag, Berlin, 1975, pp. 157–207.
- [29] D. Schmeidler, *Fatou's lemma in several dimensions*, Proc. Amer. Math. Soc., 24 (1970), pp. 300–306.
- [30a] M. B. Suryanarayana, *Remarks on lower semicontinuity and lower closure*, J. Optim. Theory Appl., 19 (1976), pp. 125–140.
- [30b] ———, *Existence theorems for optimization problems concerning linear, hyperbolic partial differential equations without convexity conditions*, J. Optim. Theory Appl., 19 (1976), pp. 47–61.
- [31] M. Valadier, *Multi-applications mesurables à valeurs convexes compactes*, J. Math. Pures Appl., 50 (1971), pp. 265–297.
- [32] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [33] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, Saunders, Philadelphia, 1969.

A STUDY OF MINIMIZING SEQUENCES*

E. POLAK† AND Y. Y. WARDI‡

Abstract. Differentiable and nondifferentiable optimization problems in normed spaces may fail to have solutions. Even when they have solutions, optimization algorithms may produce minimizing sequences that have no accumulation points. To deal with this difficulty, this paper examines optimization problems as problems on sequences, in an extended normed space, and derives first and second order optimality conditions for them.

Key words. optimality conditions, extension of spaces, nondifferentiable optimization, nonlinear programming, optimal control

1. Introduction. Engineering design periodically produces new classes of optimization problems in normed spaces. Among the earliest of such problems were those of optimal control. More recently design centering and tuning [1, 2] has produced problems with maximinmax functions in the constraints, and the design of wings, turbine blades and bridges has produced problems with eigenvalue inequalities [3], [4], [5]. Thus, in designing the profile of a seismically resistant bridge, one may wish to minimize the weight of the structure, while considering both its low amplitude linear behavior and large amplitude nonlinear behavior. As a linear structure, the lowest natural frequency of the bridge must lie above a specified value; while as a nonlinear structure, its excursions, with respect to time, produced by a set of earthquakes, must be maintained within prescribed limits so as to avoid destruction. Of course, there are also constraints on the profile of the bridge itself.

Abstractly, such problems can be viewed as being of the form

$$(1.1) \quad P: \min \{f(x) | x \in X\}$$

where X is a subset of \mathcal{X} , a “convenient” topological space, in which P may or may not have a solution, and $f: \mathcal{X} \rightarrow \mathbb{R}$ is continuous and bounded on X . \mathcal{X} is convenient in the sense that it is reasonably easy to construct and analyse an optimization algorithm in its topology. If X is a finite dimensional space, then the algorithm is usually accompanied by a convergence theorem which states that if a sequence $\{x_i\}_{i=0}^{\infty}$ constructed by the algorithm has accumulation points, then all these accumulation points are in X and satisfy some condition of optimality. Now, even when X is closed and bounded, a sequence $\{x_i\}_{i=0}^{\infty}$, constructed by an optimization algorithm, may fail to have accumulation points either because (1.1) has no solution in the topology of \mathcal{X} or because of the particular process used by the algorithm in constructing the x_i . As an example of the latter, consider the case where $\mathcal{X} = L_{\infty}[0, 1]$, the x_i are all continuous functions, but all the local minima x^* are only piecewise continuous. Obviously, such phenomena are disturbing since they lead to the conclusion that the convergence theorems in question are vacuous.

A standard approach in dealing with the above described difficulties is to replace the space \mathcal{X} with a suitable extension, e.g., as in the case when ordinary controls are

* Received by the editors April 1, 1982, and in revised form June 17, 1983. This research was sponsored by the National Science Foundation under grants ECS-79-13148 and CEE-8105790, the Joint Services Electronics Program under contract F49620-79-C-0178, and the UK Science Research Council.

† Department of Electrical Engineering and Computer Sciences, and Electronics Research Laboratory, University of California, Berkeley, California, 94720.

‡ Bell Laboratories, Holmdel, New Jersey, 07733.

replaced by relaxed controls in optimal control problems [17]. Unfortunately, by and large, the construction of minimal extensions can be quite difficult, as was the case with relaxed controls, for example. Consequently, the question arises whether there may not be an easier approach, one based on the concept of minimizing sequences in the original space \mathcal{X} and of utilizing elements of optimization algorithm theory. This paper explores this question.

In constructing a theory of optimality conditions for minimizing sequences, we felt it was important to take into account the following facts and goals. Firstly, optimization algorithms construct sequences $\{x_i\}$ which may not be Cauchy and along which the cost $f(x_i)$ may not be monotonically decreasing. Secondly, by and large, any subsequence of a locally minimizing sequence, constructed by an optimization algorithm, is a locally minimizing sequence. Thirdly, in computing a search direction, many algorithms solve a relatively simple program whose value can be viewed as an *optimality function* $\theta(x)$. These optimality functions vanish only at stationary points. When an optimality function is continuous, it can be used loosely as an ε -solution detector.

We show that a natural extension of the space \mathcal{X} in (1.1) is to an extended norm space \mathcal{X}^s of sequences in \mathcal{X} and that many of the well known optimality functions can be used to provide both first and second order optimality conditions for characterizing minimizing sequences of P .

2. The space of minimizing sequences. Consider the problem

$$(2.1) \quad P: \inf \{f(x) | x \in X\}$$

where X is a subset of a normed space \mathcal{X} and $f: \mathcal{X} \rightarrow \mathbb{R}$ is continuous and bounded from below on X .

Our definitions of global and local minimizing sequences reflect the fact that when an optimization algorithm is applied to P , it may produce a sequence $\{x_i\}_{i=0}^\infty$ which is *not* Cauchy and for which $\{f(x_i)\}_{i=0}^\infty$ is *not* monotonically decreasing. However, any accumulation point x^* , that such a sequence $\{x_i\}_{i=0}^\infty$ may have will be a local minimum, or at least a stationary point.

DEFINITION 2.1. $\{x_i\}_{i=0}^\infty$, $x_i \in \mathcal{X}$, $i = 0, 1, 2, \dots$, is an *eventually feasible sequence* (for P) if

$$(2.2) \quad \liminf_{i \rightarrow \infty} \{\|x_i - x\| | x \in X\} = 0.$$

DEFINITION 2.2. A bounded eventually feasible sequence $\{\hat{x}_i\}_{i=0}^\infty$ is a *globally minimizing sequence* (for P) if for all bounded eventually feasible sequences $\{x_i\}_{i=0}^\infty$

$$(2.3) \quad \overline{\lim}_{i \rightarrow \infty} f(\hat{x}_i) \leq \overline{\lim}_{i \rightarrow \infty} f(x_i).$$

The following result is obvious.

PROPOSITION 2.1. Suppose that $\{\hat{x}_i\}_{i=0}^\infty$ is a globally minimizing sequence. Then

- a) $\lim_{i \rightarrow \infty} f(\hat{x}_i)$ exists;
- b) every infinite subsequence of $\{\hat{x}_i\}_{i=0}^\infty$ is a globally minimizing sequence.

Our definition of a locally minimizing sequence, below, ensures the property that every subsequence of a locally minimizing sequence, constructed by an algorithm, is also a locally minimizing sequence.

DEFINITION 2.3. A bounded eventually feasible sequence $\{\hat{x}_i\}_{i=0}^\infty$ is a *locally minimizing sequence* (for P) if there exists a $\rho > 0$ such that for all eventually feasible

sequences $\{x_i\}_{i=0}^\infty$ satisfying

$$(2.4) \quad \overline{\lim}_{i \rightarrow \infty} \|\hat{x}_i - x_i\| \leq \rho.$$

We have

$$(2.5) \quad \overline{\lim}_{\substack{K \\ i \rightarrow \infty}} f(\hat{x}_i) \leq \overline{\lim}_{\substack{K \\ i \rightarrow \infty}} f(x_i)$$

for all infinite subsets $K \subset \mathbb{N}_+ \triangleq \{0, 1, 2, \dots\}$, with $i \xrightarrow{K} \infty$ denoting: $i \in K, i \rightarrow \infty$.

PROPOSITION 2.2. Suppose that $\{\hat{x}_i\}_{i=0}^\infty$ is a locally minimizing sequence. Then for every infinite subset $K \subset \mathbb{N}_+$, $\{\hat{x}_i\}_{i \in K}$ is a locally minimizing sequence.

Proof. Let $\rho > 0$ be as specified in Definition (2.3) and suppose that there is a subsequence $\{\hat{x}_i\}_{i \in K'}$ for some infinite $K' \subset \mathbb{N}_+$ which is not a minimizing sequence. Then there must exist an infinite $K'' \subset K'$ and a sequence $\{\bar{x}_i\}_{i \in K''}$ such that

$$(2.6) \quad \overline{\lim}_{\substack{K'' \\ i \rightarrow \infty}} \|\bar{x}_i - \hat{x}_i\| \leq \rho$$

and

$$(2.7) \quad \overline{\lim}_{\substack{K'' \\ i \rightarrow \infty}} f(\bar{x}_i) < \overline{\lim}_{\substack{K'' \\ i \rightarrow \infty}} f(\hat{x}_i).$$

Let $x_i = \bar{x}_i$ for all $i \in K''$ and let $x_i = \hat{x}_i$ otherwise; then $\{x_i\}_{i=0}^\infty$ satisfies (2.4) but fails to satisfy (2.5) for $K = K''$, contradicting the assumption that $\{\hat{x}_i\}_{i=0}^\infty$ is a locally minimizing sequence. This completes our proof. \square

Remark. Note that for a locally minimizing sequence $\{x_i\}_{i=0}^\infty$, the bounded sequence $\{f(x_i)\}_{i=0}^\infty$ may have more than one accumulation point. This can be seen from the following example: Let \hat{x} and \bar{x} be two local minimizers for f (with $\Omega = \mathcal{X}$), such that $f(\bar{x}) \neq f(\hat{x})$. Let $x_{2i} = \bar{x}$ and $x_{2i+1} = \hat{x}$, $i = 0, 1, 2, \dots$.

We are now ready to put the problem P into one-to-one correspondence with a problem P^s defined in a space of sequences and thus remove the need for determining whether minimizing sequences do or do not have accumulation points in \mathcal{X} .

DEFINITION 2.4. a) We define $\tilde{\mathcal{X}}$ to be the class of infinite sequences $\{x_i\}_{i=0}^\infty$, with $x_i \in \mathcal{X}$, $i \in \mathbb{N}_+$.

b) We define $\{x_i\}_{i=0}^\infty, \{y_i\}_{i=0}^\infty$ in $\tilde{\mathcal{X}}$ to be *equivalent* if $\lim_{i \rightarrow \infty} \|x_i - y_i\| = 0$. We shall denote this equivalence relation by the symbol \sim .

c) We define the vector space \mathcal{X}^s to be $\tilde{\mathcal{X}}/\sim$, with addition and scalar (real) multiplication defined as follows:

$$(2.8) \quad \{x_i\}_{i=0}^\infty + \{y_i\}_{i=0}^\infty = \{x_i + y_i\}_{i=0}^\infty,$$

$$(2.9) \quad \alpha \{x_i\}_{i=0}^\infty = \{\alpha x_i\}_{i=0}^\infty.$$

PROPOSITION 2.3. The operations of addition and scalar multiplication in \mathcal{X}^s , as given by (2.8) and (2.9), are well defined, i.e. if $s_x, s'_x, s_y, s'_y \in \tilde{\mathcal{X}}$ are such that $s_x \sim s'_x$ and $s_y \sim s'_y$, then $(s_x + s_y) \sim (s'_x + s'_y)$ and for any $\alpha \in \mathbb{R}$, $\alpha s_x \sim \alpha s'_x$.

Next, we define the concepts of an extended norm and of an extended normed vector space.

DEFINITION 2.5. a) Let \mathcal{X} be a real vector space and let $\|\cdot\|$ be a functional on \mathcal{X} which can take on the value ∞ . We say that $\|\cdot\|$ is an *extended norm* if it has all

the properties of a norm on any subset $B \subset \mathcal{Z}$ with the property that B is closed under addition and scalar multiplication, on which $\|\cdot\|$ is finite, viz:

- i) $\|z\| \geq 0, \forall z \in \mathcal{Z}$;
- ii) $\|z\| = 0 \Leftrightarrow z = 0$;
- iii) $\|\alpha z\| = |\alpha| \|z\|, \forall z \in B, \forall \alpha \in \mathbb{R}$;
- iv) $\|z_1 + z_2\| \leq \|z_1\| + \|z_2\|, \forall z_1, z_2 \in B$.

b) With $\|\cdot\|$ an extended norm, we say that $(\mathcal{Z}, \|\cdot\|)$ is an *extended normed space* if the vector addition operation is bicontinuous (i.e. it is continuous in both operands) and the scalar multiplication is continuous on $B \times \mathbb{R}$ (with respect to $\|\cdot\|$), where B is any set in \mathcal{Z} , such that for all $b \in B, \|b\| < \infty$, and B is closed under addition and scalar multiplication.

We note that the scalar multiplication operator cannot be continuous at any $z \in \mathcal{Z}$, \mathcal{Z} an extended normed space, at which $\|z\| = \infty$, since for any $\alpha_i \rightarrow 0$ as $i \rightarrow \infty$, $\alpha_i > 0$, $\|\alpha_i z\| = \infty$ for all $i \in \mathbb{N}_+$, while $\|0 \cdot z\| = 0$.

To make \mathcal{X}^s an extended normed space, we define $\|\cdot\|$ on \mathcal{X}^s by

$$(2.10) \quad \|z\| \triangleq \overline{\lim}_{i \rightarrow \infty} \|x_i\|$$

where $\{x_i\}_{i=0}^\infty$ is any sequence in the equivalence class z . We shall use the notation $\{x_i\}_{i=0}^\infty$ to refer to an element in \mathcal{X}^s .

The following result is obvious in view of the definition of the equivalence relation \sim and the properties of the norm $\|\cdot\|$ on \mathcal{Z} .

PROPOSITION 2.4. *The functional $\|\cdot\|$ defined by (2.10) is a well defined extended norm on \mathcal{X}^s , and $(\mathcal{X}^s, \|\cdot\|)$ is an extended normed space.*

For the problem P to make sense in the context of \mathcal{X}^s , it is necessary that the asymptotic behavior of $f(\cdot)$ on $z \in \mathcal{X}^s$ be independent of the particular choice of a sequence $\{x_i\}_{i=0}^\infty$ in the equivalence class defined by z . Consequently we postulate as follows.

Assumption 2.1. *The function $f(\cdot)$ is uniformly continuous on bounded sets. We obtain immediately*

PROPOSITION 2.5. *If $\{x_i\}_{i=0}^\infty \sim \{x'_i\}_{i=0}^\infty$ and $\|\{x_i\}_{i=0}^\infty\| < \infty$, then $\overline{\lim}_{i \rightarrow \infty} f(x_i) = \overline{\lim}_{i \rightarrow \infty} f(x'_i)$. Furthermore, if $\{x_i\}_{i=0}^\infty$ is eventually feasible, then so is $\{x'_i\}_{i=0}^\infty$.*

Let $X^s \subset \mathcal{X}^s$ be defined by

$$(2.11) \quad X^s \triangleq \{z \in \mathcal{X}^s \mid \|z\| < \infty \text{ and } z \text{ is eventually feasible}\}$$

and let the extended function $f^s: \mathcal{X}^s \rightarrow \mathbb{R}$ be defined by

$$(2.12) \quad f^s(z) = \overline{\lim}_{i \rightarrow \infty} f(x_i)$$

where $z = \{x_i\}_{i=0}^\infty$. We now define the problem P^s as follows:

$$(2.13) \quad P^s: \min \{f^s(z) \mid z \in X^s\}$$

PROPOSITION 2.6. *Problem P^s has a solution if and only if P admits a bounded globally minimizing sequence. Furthermore, the values of P and P^s are the same.*

We note that if \hat{z} solves P^s and $\{\hat{x}_i\}_{i=0}^\infty$, a sequence in the equivalence class \hat{z} , has an accumulation point \hat{x} , then \hat{x} is a feasible minimizer for P . Similarly, if \hat{x} is a global minimizer for P , then $\hat{z} = \{\hat{x}_i\}_{i=0}^\infty, \hat{x}_i \equiv \hat{x}$, is a solution to P^s .

The introduction of the function f^s makes the formalism (2.13) very appealing with respect to global solutions. However, it fails with respect to local solutions. Thus, if we define, as is customary, $\hat{z} \in X^s$, to be a local minimizer of P^s if for some $\rho > 0$,

$$(2.14) \quad f^s(\hat{z}) \leq f^s(z) \quad \forall z \in B(\hat{z}, \rho)$$

with $B(\hat{z}, \rho) \triangleq \{z \mid \|z - \hat{z}\| \leq \rho\}$, we find that \hat{z} is not a locally minimizing sequence, as defined in Definition 2.3. Consequently, we use the following.

DEFINITION 2.6. We say that $\hat{z} \in X^s$ is a *local minimizer* for P^s if $\{\hat{x}_i\}_{i=0}^\infty$ is a locally minimizing sequence for P , where $\{\hat{x}_i\}_{i=0}^\infty$ is any sequence in the equivalence class \hat{z} .

Quite clearly, because of Assumption 2.1, local minimizers for P^s are well defined. Furthermore, if \hat{z} is a solution to P^s and $\{\hat{x}_i\}_{i=0}^\infty$ is any sequence in the equivalence class defined by \hat{z} , then $\{\hat{x}_i\}_{i=0}^\infty$ is a minimizing sequence for P .

We are now ready to proceed to the next task: the development of necessary and sufficient conditions for characterizing local minimizers of P^s .

3. Unconstrained minimization. We begin by assuming that $X = \mathcal{X}$ in P , so that $X^s = \mathcal{X}^s$ in P^s . We shall consider both differentiable and nondifferentiable cost functions $f(\cdot)$.

We shall characterize optimality of minimizing sequences by means of first and second order optimality functions which tend to zero along minimizing sequences. Since we will be dealing with bounded sequences which are not necessarily Cauchy, we will have to require that various properties hold uniformly on bounded sets.

We shall denote the first Fréchet derivative of $f(\cdot)$ by $f_x(\cdot)$ and the second Fréchet derivative by $f_{xx}(\cdot)$. We note that f_x maps \mathcal{X} into \mathcal{X}' , the dual of \mathcal{X} and that f_{xx} maps \mathcal{X} into \mathcal{X}'' the dual of \mathcal{X}' .

PROPOSITION 3.1. Suppose that $f(\cdot)$ is uniformly, continuously Fréchet differentiable on bounded sets in \mathcal{X} . Let $\Theta_{uc}^{1s}: \mathcal{X}^s \rightarrow \mathbb{R}$ be defined on bounded $z \in \mathcal{X}^s$ by

$$(3.1) \quad \Theta_{uc}^{1s}(z) \triangleq \liminf_{i \rightarrow \infty} \{(f_x(x_i), h) \mid \|h\| \leq 1\}$$

where $\{x_i\}_{i=0}^\infty$ is any sequence in the equivalence class z . If \hat{z} is a local minimizer for P^s , then $\Theta_{uc}^{1s}(\hat{z}) = 0$.

Proof. First we note that, because of the assumption on $f_x(\cdot)$, the function $\Theta_{uc}^1: \mathcal{X} \rightarrow \mathbb{R}$ defined by

$$(3.2) \quad \Theta_{uc}^1(\bar{x}) \triangleq \inf \{(f_x(\bar{x}), h) \mid \|h\| \leq 1\}$$

is uniformly continuous on bounded sets and hence $\Theta_{uc}^{1s}(\cdot)$ is well defined. Also, $\Theta_{uc}^{1s}(z) \leq 0$ for any $z \in \mathcal{X}^s$. Hence suppose that $\hat{z} (= \{\hat{x}_i\}_{i=0}^\infty)$ is a local minimizer for P^s , with associated radius $\hat{\rho} > 0$, and that $\Theta_{uc}^{1s}(\hat{z}) < 0$. Then there exists a $\delta > 0$ and an infinite subsequence $\{\hat{x}_i\}_{i \in K}$ such that $\Theta_{uc}^1(\hat{x}_i) \leq -\delta$ for all $i \in K$. For $i \in K$, let $h_i \in \mathcal{X}$, be such that $\|h_i\| \leq 1$ and $(f_x(\hat{x}_i), h_i) \leq -\delta/2$. Then, because $f_x(\cdot)$ is uniformly continuous on bounded sets, there exists a $\bar{\lambda} \in (0, \hat{\rho}]$ such that, with $s_i \in (0, 1)$, by the mean value theorem,

$$f(\hat{x}_i + \bar{\lambda} h_i) - f(\hat{x}_i) = \bar{\lambda} (f_x(\hat{x}_i + s_i \bar{\lambda} h_i), h_i) \leq -\bar{\lambda} \delta / 4 \quad \text{for all } i \in K.$$

Consider now the sequence $\{x_i^*\}_{i=0}^\infty$ defined by $x_i^* = \hat{x}_i$ for all $i \notin K$ and $x_i^* = \hat{x}_i + \bar{\lambda} h_i$ for all $i \in K$. Clearly, (2.5) fails for $x_i = x_i^*$, $i \in \mathbb{N}_+$ and hence we get a contradiction. \square

PROPOSITION 3.2. Suppose that $f(\cdot)$ is twice uniformly, continuously Fréchet differentiable on bounded sets in \mathcal{X} . Let $\Theta_{uc}^{2s}: \mathcal{X}^s \rightarrow \mathbb{R}$ be defined on bounded $z \in \mathcal{X}^s$ by

$$(3.3) \quad \Theta_{uc}^{2s}(z) \triangleq \liminf_{i \rightarrow \infty} \{(f_{xx}(x_i)(h), h) \mid \|h\| \leq 1\}$$

where $\{x_i\}_{i=0}^\infty$ is any sequence in the class z . If $\hat{z} \in \mathcal{X}^s$ is a local minimizer for P^s , then $\Theta_{uc}^{2s}(\hat{z}) \geq 0$.

Proof. Let $\Theta_{uc}^2: \mathcal{X} \rightarrow \mathbb{R}$ be defined by

$$(3.4) \quad \Theta_{uc}^2(\bar{x}) \triangleq \inf \{ (f_{xx}(\bar{x})(h), h) \mid \|h\| \leq 1 \}.$$

Since $\Theta_{uc}^2(\cdot)$ is uniformly continuous on bounded sets, $\Theta_{uc}^{2s}(\cdot)$ is well defined. Now suppose that $\hat{z} = \{\hat{x}_i\}_{i=0}^\infty$ is a local minimizer for P^s , with associated radius $\hat{\rho}$, and that $\Theta_{uc}^{2s}(\hat{z}) = -\delta < 0$. Then there exists an infinite subsequence $\{\hat{x}_i\}_{i \in K}$ and a corresponding sequence $\{h_i\}_{i \in K}$, such that $\|h_i\| \leq 1$ and $(f_x(\hat{x}_i), h_i) \leq 0$, satisfying $(f_{xx}(\hat{x}_i)(h_i), h_i) \leq -\delta/2$ for all $i \in K$. Hence, making use of the uniform continuity of $f_{xx}(\cdot)$ and the Taylor formula with remainder, we find that there exist a $\bar{\lambda} \in (0, \hat{\rho}]$ such that

$$(3.5) \quad \begin{aligned} f(\hat{x}_i + \bar{\lambda} h_i) - f(\hat{x}_i) &= \bar{\lambda} (f_x(\hat{x}_i), h_i) + \bar{\lambda}^2 \int_0^1 (1-s) (f_{xx}(\hat{x}_i + s\bar{\lambda} h_i)(h_i), h_i) ds \\ &\leq \bar{\lambda}^2 \delta / 8 \quad \forall i \in K. \end{aligned}$$

Setting $x_i^* = \hat{x}_i$ for $i \notin K$ and $x_i^* = x_i + \bar{\lambda} h_i$ for $i \in K$, we get a contradiction of the fact that \hat{z} is a local minimizer. This completes the proof. \square

It is also quite easy to prove the following result.

PROPOSITION 3.3. *Suppose that $f(\cdot)$ is twice uniformly, continuously Fréchet differentiable on bounded sets in \mathcal{X} . Suppose that $\hat{z} \in \mathcal{X}^s$ is such that (i) $\|\hat{z}\| < \infty$, (ii) $\lim_{i \rightarrow \infty} f(\hat{x}_i) > -\infty$, for $\{x_i\}_{i=0}^\infty$ in the equivalence class \hat{z} , (iii) $\Theta_{uc}^{1s}(\hat{z}) = 0$ and (iv) $\Theta_{uc}^{2s}(\hat{z}) > 0$. Then \hat{z} is a local minimizer for P^s .*

Thus, for the case where $f(\cdot)$ is differentiable, we see that the standard optimality conditions for P lead directly to corresponding optimality conditions for P^s . As we shall shortly see, this is not true for the nondifferentiable case: when $f(\cdot)$ is assumed to be only uniformly Lipschitz continuous on bounded sets in \mathcal{X} . We recall [6] that, the *generalized gradient* $\partial f(\cdot)$ of $f(\cdot)$, is defined as the subset of \mathcal{X}' satisfying for every x, h in \mathcal{X} the relation

$$(3.6) \quad f^0(x; h) \triangleq \lim_{\substack{\lambda \downarrow 0 \\ y \rightarrow 0}} \frac{f(x + y + \lambda h) - f(x + y)}{\lambda} = \sup \{ (\xi, h) \mid \xi \in \partial f(x) \}$$

where $f^0(x; h)$ is called the *generalized directional derivative* of $f(\cdot)$ at x , in the direction h . The sets $\partial f(x)$ are bounded on bounded sets in \mathcal{X} and upper semi-continuous. The standard first order optimality condition for P , see [6], [7], is that if \hat{x} is a local minimizer for P , then $0 \in \partial f(\hat{x})$. A first attempt to convert this statement into an optimality function produces the following candidate $\Theta: \mathcal{X} \rightarrow \mathbb{R}$ defined by

$$(3.7) \quad \Theta(x) \triangleq \sup_{\xi} \{ \inf \{ (\xi, h) \mid \xi \in \partial f(x) \} \mid \|h\| \leq 1 \},$$

which is recognized as being a generalization of the function $\min \{ \|h\| \mid h \in \partial f(x) \}$ in \mathbb{R}^n . Since $\partial f(\cdot)$ is only u.s.c., $\Theta(\cdot)$ is not continuous and it is very easy to construct functions $f(\cdot)$ and sequences $\{x_i\}$ such that $x_i \rightarrow \hat{x}$, a local minimizer, so that $0 \in \partial f(\hat{x})$ and hence $\Theta(\hat{x}) = 0$, but $\Theta(x_i) = -1$ for all i . Clearly, $\Theta(\cdot)$ cannot be used to characterize minimizing sequences for P . However, referring to [2], [8], we find that there are other functions that can. For example, following [2], [8], for any $\varepsilon \geq 0$, and $x \in \mathcal{X}$, let

$$(3.8) \quad \partial_\varepsilon f(x) \triangleq \bigcup_{x' \in B(x, \varepsilon)} \partial f(x')$$

where $B(x, \varepsilon) \triangleq \{x' \in \mathcal{X} \mid \|x' - x\| \leq \varepsilon\}$ and

$$(3.9) \quad \Theta_\varepsilon(x) \triangleq \sup \left\{ \inf_{\xi} \{(\xi, h) \mid \xi \in \partial_\varepsilon f(x)\} \mid \|h\| \leq 1 \right\}.$$

It can be shown [2], [8] that $\partial_\varepsilon f(\cdot)$ is u.s.c. and closed and bounded on bounded sets. Next, with $\beta \in (0, 1)$, let

$$(3.10) \quad \mathcal{E} \triangleq \{0, 1, \beta, \beta^2, \dots\}$$

and let $\varepsilon: \mathcal{X} \rightarrow \mathbb{R}$ be defined by

$$(3.11) \quad \varepsilon(x) = \max \{ \varepsilon \in \mathcal{E} \mid \Theta_\varepsilon(x) \geq \varepsilon \}.$$

The following result can be found in [2], [8].

PROPOSITION 3.4. $0 \in \partial f(x) \Leftrightarrow \theta(x) = 0 \Leftrightarrow \varepsilon(x) = 0$.

However, while $\{\theta(x_i)\}_{i=0}^\infty$ generally will not converge to zero for $\{x_i\}_{i=0}^\infty$, a locally minimizing sequence for P , $\{\varepsilon(x_i)\}_{i=0}^\infty$ must converge to zero along such a sequence, as we shall shortly see. Consequently, we define $\Theta_{ucnd}^s: \mathcal{X}^s \rightarrow \mathbb{R}$ by

$$(3.12) \quad \Theta_{ucnd}^s(z) \triangleq \overline{\lim}_{i \rightarrow \infty} \varepsilon(x_i),$$

with $\{x_i\}_{i=0}^\infty$ any sequence in the equivalence class z .

PROPOSITION 3.5. If $\hat{z} \in \mathcal{X}$ is a bounded local minimizer for P^s , then $\Theta_{ucnd}^s(\hat{z}) = 0$.

Proposition 3.5 is a special case of a result to follow (Proposition 3.8), and hence its proof will not be given.

We now proceed formally.

DEFINITION 3.1. Let $G(\cdot)$ be a map from \mathcal{X} into $2^{\mathcal{X}'}$ (i.e. the class of subsets of \mathcal{X}'). We shall say that $G(\cdot)$ is *uniformly u.s.c. on bounded sets, with respect to $\partial f(\cdot)$* , if for any bounded set $B \subset \mathcal{X}$ and any $\delta > 0$, there exists a $\rho > 0$ such that for all $x, y \in B$ satisfying $\|x - y\| < \rho$ and any $\eta \in \partial f(y)$, there exists a $\xi \in G(x)$ such that $\|\xi - \eta\| < \delta$.

It is easily seen that if $G(\cdot)$ is *any* map which is uniformly u.s.c. on bounded sets w.r.t. $\partial f(\cdot)$, then (i) $\partial f(x) \subset G(x)$, and (ii) for any $\delta > 0$, there exists an $\varepsilon > 0$ such that $\partial_\varepsilon f(x) \subset N_\delta(G(x)) \triangleq \{x' \mid \inf_{y \in G(x)} \|x' - y\| \leq \delta\}$. In fact, we have the following result.

PROPOSITION 3.6. For any $\varepsilon > 0$, the map $\partial_\varepsilon f(\cdot)$ is uniformly u.s.c. on bounded sets, w.r.t. $\partial f(\cdot)$.

Proof. Let $x \in \mathcal{X}$ be arbitrary and let $\delta > 0$. Then, setting $\rho = \varepsilon > 0$, we get for any $y \in B(x, \rho)$ that $\partial f(y) \subset \partial_\varepsilon f(x)$ by definition. Hence the proposition holds. \square

PROPOSITION 3.7. Let $G: \mathcal{X} \rightarrow 2^{\mathcal{X}'}$ be uniformly u.s.c. on bounded sets with respect to $\partial f(\cdot)$. If \hat{z} is a local minimizer for P^s , then $\Theta_G^s(\hat{z}) = 0$, where

$$\Theta_G^s(\hat{z}) \triangleq \overline{\lim}_{i \rightarrow \infty} \Theta_G(\hat{x}_i),$$

and for $x \in \mathcal{X}$

$$(3.13) \quad \Theta_G(x) \triangleq \sup_h \inf_{\xi} \{(\xi, h) \mid \xi \in G(x)\} \mid \|h\| \leq 1\}.$$

Proof. Clearly, $\Theta_G^s(\cdot)$ is well defined. Now suppose that \hat{z} is a local minimizer for P^s , with associated radius $\hat{\rho} > 0$, and that $\Theta_G^s(\hat{z}) = \hat{\delta} > 0$. (Clearly $\Theta_G^s(z) \geq 0$ for all bounded $z \in \mathcal{X}^s$.) Then there exists an infinite subset $K \subset \mathbb{N}_+$ such that $\Theta_G(\hat{x}_i) \geq \hat{\delta}/2$ for all $i \in K$, with $\{\hat{x}_i\}_{i=0}^\infty$ any sequence in the class \hat{z} . Let $\delta = \hat{\delta}/4 > 0$ and a corresponding $\rho > 0$ satisfy the requirements of Definition 3.1. For $i \in K$, let h_i be such that $(\xi, h_i) \geq \hat{\delta}/2$

for all $\xi \in G(\hat{x}_i)$. Then, by the mean value theorem [9],

$$(3.14) \quad f(\hat{x}_i - \lambda h_i) - f(\hat{x}_i) = -\lambda \langle \xi_{i\lambda}, h_i \rangle$$

with $\xi_{i\lambda} \in \partial f(\hat{x}_i - s_i \lambda h_i)$ for some $s_i \in (0, 1)$. Now let $\bar{\lambda} = \min(\hat{\rho}, \rho)$, then (i) $\|(\hat{x}_i - \bar{\lambda} h_i) - \hat{x}_i\| \leq \hat{\rho}$ for all $i \in K$, (ii) with $\eta_i \in G(\hat{x}_i)$ such that $\|\eta_i - \xi_{i\bar{\lambda}}\| \leq \hat{\delta}/4$, we get from (3.14)

$$(3.15) \quad \begin{aligned} f(\hat{x}_i - \bar{\lambda} h_i) - f(\hat{x}_i) &= -\bar{\lambda} \langle \xi_{i\bar{\lambda}}, h_i \rangle = -\bar{\lambda} \langle \eta_i + (\xi_{i\bar{\lambda}} - \eta_i), h_i \rangle \\ &\leq \bar{\lambda} \langle -(\eta_i, h_i) + \|\xi_{i\bar{\lambda}} - \eta_i\|, h_i \rangle \\ &\leq \bar{\lambda} \langle -(\eta_i, h_i) + \hat{\delta}/4, h_i \rangle \\ &\leq -\bar{\lambda} \hat{\delta}/4. \end{aligned}$$

Clearly, the sequence $x_i^* = \hat{x}_i$ for $i \notin K$, $x_i^* = \hat{x}_i - \bar{\lambda} h_i$, violates (2.5) for the local minimizer \hat{z} , and hence we have a contradiction. \square

Referring to Proposition 3.7, we note that since $\partial f(x) \subset G(x)$ is always true, $0 \in \partial f(x) \Rightarrow 0 \in G(x)$, hence \hat{x} optimal for $P \Rightarrow 0 \in G(\hat{x})$. However, $0 \in G(\hat{x})$ is obviously a weaker optimality condition than $0 \in \partial f(\hat{x})$. The condition $0 \in G(\hat{x})$ can be strengthened when $G(\cdot)$, can be parametrized, as follows.

PROPOSITION 3.8 *For $\varepsilon \geq 0$, let $G_\varepsilon: \mathcal{X} \rightarrow 2^{\mathcal{X}}$ be a family of maps that are uniformly u.s.c. on bounded sets w.r.t $\partial f(\cdot)$, such that for all $x \in \mathcal{X}$,*

- (i) $G_0(x) = \partial f(x)$,
- (ii) $0 \leq \varepsilon < \varepsilon' \Rightarrow G_\varepsilon(x) \subset G_{\varepsilon'}(x)$,
- (iii) $G_\varepsilon(x)$ is convex and weak* compact.

Next, let $\Theta: \mathcal{X} \rightarrow \mathbb{R}$ be defined by

$$(3.16) \quad \Theta(x) \triangleq \max \{ \varepsilon \in \mathcal{E} \mid \Theta_{G_\varepsilon}(x) \geq \varepsilon \}$$

(with \mathcal{E} as in (3.10)) and $\Theta^s: \mathcal{X}^s \rightarrow \mathbb{R}$ (with $z = \{x_i\}_{i=0}^\infty$) by

$$(3.17) \quad \Theta^s(z) \triangleq \lim_{i \rightarrow \infty} \overline{\Theta(x_i)}.$$

If \hat{z} is a local minimizer for P^s , then $\Theta^s(\hat{z}) = 0$.

Proof. Suppose that \hat{z} is a local minimizer for P^s , but $\Theta^s(\hat{z}) = \hat{\varepsilon} > 0$, with $\hat{\varepsilon} \in \mathcal{E}$. Then there must exist an infinite $K \subset \mathbb{N}_+$ such that $\Theta(x_i) = \hat{\varepsilon}$ for all $i \in K$, i.e. $\Theta_{G_{\hat{\varepsilon}}}(x_i) \geq \hat{\varepsilon}$ for all $i \in K$. But by Proposition 3.7, this contradicts the local minimality of $\{\hat{x}_i\}_{i \in K}$ and hence the proof is complete. \square

The maps $\partial_\varepsilon f(\cdot)$ are not the only known examples of maps that are uniformly u.s.c. with respect to $\partial f(\cdot)$ and satisfy the assumptions of Proposition 3.8. In [10], [11] we find very different maps in this class that are used for optimization problems with eigenvalue constraints.

4. Constrained minimization. In this section we shall propose a set of optimality functions for a variety of constrained problems of the form P^s . Since the proofs associated with these optimality functions are either quite straightforward, or follow directly from existing results, they will be omitted.

We begin with equality constrained problems (cf. [12, Chap. 8]).

PROPOSITION 4.1. *Suppose that $f(\cdot)$ is uniformly continuously Fréchet differentiable on bounded sets in \mathcal{X} and that*

$$(4.1) \quad X = \{x \mid h^j(x) = 0, j \in \underline{l}\}$$

where $\underline{l} \triangleq \{1, 2, \dots, l\}$ and $h^j: \mathcal{X} \rightarrow \mathbb{R}$ are uniformly continuously differentiable on bounded sets. Furthermore, suppose that the functionals $h_x^j(\bar{x})$, $j \in \underline{l}$, $\bar{x} \in \mathcal{X}$ are linearly

independent for all $\bar{x} \in \mathcal{X}$. If \hat{z} is a local minimizer for P^s , then, with $\hat{z} = \{\hat{x}_i\}_{i=0}^\infty$,

$$(4.2) \quad \hat{z} \in X^s \triangleq \{z \in \mathcal{X}^s \mid \lim_{i \rightarrow \infty} h^j(x_i) = 0, j \in \underline{l}\},$$

$$(4.3) \quad \Theta_{ce}^{1s}(\hat{z}) \triangleq \liminf_{i \rightarrow \infty} \inf_v \{(f_x(\hat{x}_i), v) + \frac{1}{2} \|v\|^2 \mid (h_x^j(\hat{x}_i), v) = 0, j \in \underline{l}\} \\ = 0$$

where $\{\hat{x}_i\}_{i=0}^\infty$ is any sequence in the equivalence class \hat{z} .

When $\mathcal{X} = \mathbb{R}^n$, the inf in (4.3) can be replaced by min and in that case the Lagrange conditions give for the “inner” problem in (4.3)

$$(4.4) \quad \nabla f(\hat{x}_i) + v_i + \frac{\partial h(\hat{x}_i)^T}{\partial x} \psi_i = 0$$

where $h(\cdot)$ is the constraints vector and v_i solves the inner problem.

When x_i solves P , $v_i = 0$ and (4.4) reduces to the standard Lagrange condition. Thus, (4.4) is in one-to-one correspondence with the usual first order conditions for P . We can also obtain a second order condition as follows:

PROPOSITION 4.2. *Suppose that the assumptions of Proposition 4.1 hold and that in addition $f(\cdot)$ and $h^j(\cdot)$, $j \in \underline{l}$, are all twice uniformly continuously differentiable on bounded sets. For any $\bar{x} \in \mathcal{X}$, let $\mu: \mathcal{X} \rightarrow \mathbb{R}^l$ be defined by*

$$(4.5) \quad \mu(\bar{x}) \triangleq \arg \min \left\{ \left\| f_x(\bar{x}) + \sum_{j \in \underline{l}} \mu^j h_x^j(\bar{x}) \right\| \right\}.$$

Let $L: \mathcal{X} \times \mathbb{R}^l \rightarrow \mathbb{R}$ be defined by

$$(4.6) \quad L(\bar{x}, \bar{\mu}) \triangleq f(\bar{x}) + \langle \bar{\mu}, h(\bar{x}) \rangle.$$

If \hat{z} solves P^s , then (i) $\hat{z} \in X^s$, (ii) $\Theta_{ce}^{1s}(\hat{z}) = 0$ and

$$(4.7) \quad \Theta_{ce}^{2s}(\hat{z}) \triangleq \liminf_{i \rightarrow \infty} \inf_v \{(L_{xx}(\hat{x}_i, \mu(x_i)(v)), v) \mid (h_x^j(\hat{x}_i), v) = 0, j \in \underline{l}, \|v\| = 1\} \geq 0$$

where $\{\hat{x}_i\}_{i=0}^\infty$ is any sequence in the equivalence class \hat{z} .

It is also quite easy to establish a second order sufficient condition for P^s under the assumptions of Proposition 4.2, viz., if $\hat{z} \in X^s$, and $\Theta_{ce}^{1s}(\hat{z}) = 0$ and $\Theta_{ce}^{2s}(\hat{z}) > 0$, then \hat{z} is a local minimizer for P^s .

When inequalities are present, it is possible to deduce a broad family of optimality functions for P^s from the literature on algorithms (see e.g. [13, 14]). We shall state the simplest (see [13], [14], [15]).

PROPOSITION 4.3. *Suppose that $f(\cdot)$ is uniformly continuously differentiable on bounded sets in \mathcal{X} and that*

$$(4.8) \quad X \triangleq \{x \mid h^j(x) = 0, j \in \underline{l}; g^k(x) \leq 0, k \in \underline{m}\},$$

where the $h^j: \mathcal{X} \rightarrow \mathbb{R}$ and $g^k: \mathcal{X} \rightarrow \mathbb{R}$ are all uniformly continuously differentiable on bounded sets. Furthermore suppose that the functionals $h_x^j(\bar{x})$, $j \in \underline{l}$, $\bar{x} \in \mathcal{X}$ are linearly independent for all $\bar{x} \in \mathcal{X}$.

Let $\psi: \mathcal{X} \rightarrow \mathbb{R}$ be defined by

$$(4.9) \quad \psi(x)_+ \triangleq \max_{k \in \underline{m}} \{0, g^k(x)\}.$$

If \hat{z} is a local minimizer for P^s , then

$$(4.10) \quad \hat{z} \in X^s \triangleq \{z \mid \lim_{i \rightarrow \infty} h^j(x_i) = 0, j \in \underline{l}, \lim_{i \rightarrow \infty} g^k(x_i) \leq 0, k \in \underline{m}\}$$

and for $\gamma \geq 1$, (with $\{x_i\}_{i=0}^\infty = z$)

$$(4.11) \quad \Theta_{cei}^{1s}(\hat{z}) \triangleq \lim_{i \rightarrow \infty} \inf \{(\frac{1}{2}\|v\|^2 + \max\{(f_x(\hat{x}_i), v) - \gamma\psi(\hat{x}_i)_+, g^k(\hat{x}_i) + (g_x^k(\hat{x}_i), v) - \psi(\hat{x}_i), k \in \underline{m}\})(h^j(x_i), v) = 0, j \in \underline{l}\} = 0.$$

Next we turn to a class of nondifferentiable problems (see [8]).

PROPOSITION 4.4. Suppose that $f(\cdot)$ is uniformly locally Lipschitz continuous on bounded sets and that

$$(4.12) \quad X = \{x \mid g^k(x) \leq 0, k \in \underline{m}\}$$

where the $g^k: \mathcal{X} \rightarrow \mathbb{R}$ are all uniformly locally Lipschitz continuous on bounded sets. Furthermore, suppose that $\partial f(\cdot)$ and all the $\partial g^k(\cdot)$ are weak * compact and uniformly u.s.c. on bounded sets. For any $\varepsilon \geq 0$, and $\bar{x} \in \mathcal{X}$ let

$$(4.13) \quad I_\varepsilon(\bar{x}) \triangleq \{k \in \underline{m} \mid g^k(\bar{x}) \geq \psi(\bar{x})_+ - \varepsilon\},$$

$$(4.14) \quad v_\varepsilon^g(\bar{x}) \triangleq \arg \min \{\|v\| \mid v \in \text{co}_{k \in I_\varepsilon(\bar{x})} \partial_\varepsilon g^k(\bar{x})\},$$

$$(4.15) \quad v_\varepsilon^f(\bar{x}) \triangleq \arg \min \{\|v\| \mid v \in \text{co} \{\partial_\varepsilon f(\bar{x}) \cup \partial_\varepsilon g^k(\bar{x})\}, k \in I_\varepsilon(\bar{x})\},$$

$$(4.16) \quad \gamma_\varepsilon(\bar{x}) \triangleq \max \{e^{-\psi(\bar{x})_+} \|v_\varepsilon^f(\bar{x})\|, (1 - e^{\psi(\bar{x})_+}) \|v_\varepsilon^g(\bar{x})\|\}$$

and

$$(4.17) \quad \Theta_{cind}^1(\bar{x}) = \max \{\varepsilon \in \mathcal{E} \mid \gamma_\varepsilon(\bar{x}) \geq \varepsilon\}.$$

If \hat{z} is a local minimizer for P^s , then, with $z = \{x_i\}_{i=0}^\infty$,

$$(4.18) \quad \hat{z} \in X^s \triangleq \{z \in \mathcal{X}^s \mid \lim_{i \rightarrow \infty} g^k(x_i) \leq 0, k \in \underline{m}\}$$

and

$$(4.19) \quad \Theta_{cind}^{1s}(\hat{z}) \triangleq \lim_{i \rightarrow \infty} \Theta_{cind}^1(\hat{x}_i) = 0$$

where $\{\hat{x}_i\}_{i=0}^\infty$ is any sequence in the equivalence class \hat{z} .

To conclude this section we state an optimality function for a simple optimal control problem, based on the maximum principle and first used in [6].

Consider the dynamical system

$$(4.20) \quad \dot{x}(t) = h(x(t), u(t)), \quad t \in [0, 1]$$

with $x(0) = x_0$ and $h: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ twice uniformly continuously differentiable on bounded sets. We shall denote by $x^u(\cdot)$ the solution of (4.20) corresponding to the control $u(\cdot)$. Let Ω be a compact subset of \mathbb{R}^m and let $\phi: \mathbb{R}^n \rightarrow \mathbb{R}$ be twice uniformly continuously differentiable on bounded sets. We now define

$$(4.21a) \quad \mathcal{X} \triangleq L_\infty^m[0, 1],$$

$$(4.21b) \quad f(u) \triangleq \phi(x^u(1)),$$

$$(4.21c) \quad X \triangleq \{u \in L_\infty^m[0, 1] \mid u(t) \in \Omega \forall t \in [0, 1]\}.$$

Finally, let $\lambda^u(\cdot)$ denote the solution of the adjoint equation

$$(4.22a) \quad \dot{\lambda}(t) = -\frac{\partial h}{\partial x}(x^u(t), u(t))^T \lambda(t),$$

$$(4.22b) \quad \lambda(1) = \nabla \phi(x^u(1)).$$

PROPOSITION 4.5. *Suppose \hat{z} is a local minimizer for P^s , with f , \mathcal{X} and X defined as in (4.21a–b). Then*

$$(4.23) \quad \Theta_{oc}^s(\hat{z}) \triangleq \overline{\lim}_{i \rightarrow \infty} \Theta_{oc}(\hat{u}_i) = 0,$$

where

$$\Theta_{oc}(u) \triangleq \int_0^1 \min \{ \langle (h(x^u(t), v) - h(x^u(t), u(t))), \lambda^u(t) \rangle \mid v \in \Omega \} dt.$$

The proof of Proposition 4.5 follows from the facts that $f(\cdot)$ is uniformly locally Lipschitz continuous on bounded sets (see [16]), the necessary optimality condition for u being a local minimum (see [6], [16], [17]) and the discussion in this paper.

5. Conclusions. We have shown that it is quite straightforward to construct optimality conditions for minimizing sequences by reinterpreting or modifying existing results. We hope that this work will lead to a better understanding of the behavior of optimization algorithms.

REFERENCES

- [1] J. W. BANDLER, P. C. LIU AND H. TROMP, *Nonlinear programming approach to optimal design centering, tolerancing and tuning*, IEEE Trans. Circuits Syst., CAS-23 (1976), pp. 155–165.
- [2] E. POLAK AND A. SANGIOVANNI-VINCENTELLI, *Theoretical and computational aspects of the optimal design centering, tolerancing and tuning problem*, IEEE Trans. Circuit Syst., CAS-26 (1979), pp. 795–813.
- [3] C. JOURON, *Sur un problème d'optimization où la contrainte porte sur la fréquence fondamentale*, RAIRO Analyse Numérique, 12 (1978), pp. 349–374.
- [4] E. HAUG AND J. P. ROUSSELET, *Design sensitivity analysis in structural mechanics II, eigenvalue variations*, Proc. NATO Advanced Study Institute, Univ., Iowa City, May 1980.
- [5] Y. Y. WARDI, *Optimization algorithms for problems with eigenvalue inequality constraints*, Ph.D. thesis, Dept Electrical Engineering and Computer Sciences, Univ. California, Berkeley, 1981.
- [6] F. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 242–262.
- [7] —, *Nonsmooth Analysis and Optimization*, Wiley, Interscience, New York, in press.
- [8] E. POLAK, D. Q. MAYNE AND Y. Y. WARDI, *On the extension of constrained optimization algorithms from differentiable to nondifferentiable problems*, this Journal, 21 (1983), pp. 179–203.
- [9] G. LEBOURG, *Valeur moyenne pour un gradient généralisé*, C.R. Acad. Sci., Paris, 281 (1975), pp. 795–797.
- [10] J. CULLUM, W. E. DONATH AND P. WOLFE, *The minimum of certain non-differentiable sums of eigenvalues of symmetric matrices*, Math. Programming Studies, No. 3, 1975, pp. 35–55.
- [11] E. POLAK AND Y. Y. WARDI, *A nondifferentiable optimization algorithm for design of control systems subject to singular value inequalities over a frequency range*, Automatica, in press.
- [12] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
- [13] E. POLAK, *Computational Methods in Optimization—A Unified Approach*, Academic Press, New York, 1971.
- [14] E. POLAK, R. TRAHAN AND D. Q. MAYNE, *Combined phase I—phase II methods of feasible directions*, Math. Programming. (1979), pp. 32–61.
- [15] O. PIRONEAU AND E. POAK, *On the rate of convergence of certain methods of centers*, Math. Programming (1972), pp. 230–258.
- [16] D. Q. MAYNE AND E. POLAK, *First order, strong variations algorithms for optimal control*, J. Optim. Theory Appl., 16 (1975), pp. 277–301.
- [17] L. J. WILLIAMSON AND E. POLAK, *Relaxed controls and the convergence of optimal control algorithms*, this Journal, 14 (1976), pp. 737–757.

ON OPTIMALITY CONDITIONS IN QUASIDIFFERENTIABLE OPTIMIZATION*

ALEXANDER SHAPIRO†

Abstract. A class of quasidifferentiable functions, whose directional derivatives are representable as a difference of two sublinear functions, is introduced. An optimization problem subject to quasidifferentiable equality and inequality constraints is studied. Optimality conditions for this problem are given in terms of sets inclusion. This refines the results of Demyanov, Rubinov and co-workers.

Key words. nonsmooth optimization, optimality conditions, quasidifferentiable functions

1. Introduction. First-order optimality conditions for extremal problems have been discussed in numerous studies. Some of the most important recent advances in this direction have come as a result of the systematic replacement of smoothness assumptions by convexity. Starting from subdifferential theory in convex analysis this approach has been found suitable to handle many nondifferentiable optimization problems.

Pshenichnyi [11] suggested the concept of quasidifferentiability where the convexity of a considered function is replaced by the convexity of its directional derivative. Introduction by Clarke [5], [6] of the generalized gradient allowed further extension of Lagrange type optimality conditions for locally Lipschitz functions. The generalized gradient method has been shown to be a powerful tool in many problems of nonsmooth and nonconvex optimization. However, in some cases the generalized gradient is too rough to give an adequate presentation of the situation. In particular it does not distinguish between maximum, minimum or saddle points and does not make use of the directional derivatives if they exist. On the other hand the class of quasidifferentiable functions in the sense of Pshenichnyi is insufficient to describe many actual situations. This motivated Demyanov and Rubinov [1] to extend the concept of quasidifferentiability to functions whose directional derivative is representable as a sum of two positively homogeneous functions, one convex and one concave. Quasidifferential calculus and optimality conditions for unconstrained and constrained quasidifferentiable problems has been presented in [1], [2], [3], [4], [9], [10].

The aim of this paper is to refine the optimality conditions presented by Demyanov and Polyakova [2] and Polyakova [9], [10]. In § 2 we introduce the class of quasidifferentiable functions in the sense of Demyanov and Rubinov [1]. Our principal tool will be the theory of support functions and we briefly describe the required definitions and results. In § 3 we prove a version of the reduction theorem of Ioffe [7] which allows us to formulate optimality conditions in § 4.

The following notation will be used. The scalar product of two vectors $x, y \in R^n$ is denoted by $x \cdot y$, $|x| = (x \cdot x)^{1/2}$ and $\|\cdot\|$ denotes a certain specified norm. By $\text{conv}\{\cdot\}$ we denote the convex hull of a subset of R^n and $\nabla f(x)$ denotes the gradient of $f: R^n \rightarrow R$ at x .

2. Quasidifferential calculus. A function $h: R^n \rightarrow R$ is said to be *sublinear* if it is positively homogeneous and subadditive, i.e. $h(tx) = th(x)$ and $h(x+y) \leq h(x) + h(y)$ for all $x, y \in R^n$ and $t \geq 0$. There exists a dual correspondence between sublinear

* Received by the editors April 21, 1981, and in revised form July 26, 1983.

† Department of Mathematics and Applied Mathematics, University of South Africa, PO Box 392, Pretoria 0001, South Africa.

functions and the class Ω of convex, compact subsets of R^n . This duality is carried out by the so-called support functions. In the following proposition we summarize some well-known results from convex analysis about the support functions (see, e.g., Rockafellar [13]).

PROPOSITION 2.1.

(a) *The support function*

$$\sigma(x|A) = \sup \{x \cdot z : z \in A\}$$

of a bounded set $A \subset R^n$ is a sublinear function from R^n to R . Conversely, if $h : R^n \rightarrow R$ is a sublinear function, then it is the support function of a certain set $A \in \Omega$. This set is unique and given by

$$A = \{y : x \cdot y \leq h(x) \text{ for all } x \in R^n\}.$$

(b) *For any $A_1, A_2 \in \Omega$ and $\alpha, \beta \geq 0$:*

$$\sigma(\cdot | \alpha A_1 + \beta A_2) = \alpha \sigma(\cdot | A_1) + \beta \sigma(\cdot | A_2).$$

(c) *For $A_1, A_2 \in \Omega$ the inequality*

$$\sigma(\cdot | A_1) \geq \sigma(\cdot | A_2)$$

holds iff

$$A_1 \supseteq A_2.$$

(d) *If $\{A_i\}$, $i \in I$, is a family of sets such that $\bigcup_{i \in I} A_i$ is bounded, then*

$$\sup \{\sigma(\cdot | A_i) : i \in I\} = \sigma\left(\cdot \mid \text{conv} \left\{ \bigcup_{i \in I} A_i \right\}\right).$$

Let us consider the class of positively homogeneous functions $h : R^n \rightarrow R$ representable as a difference of two sublinear functions. This class with natural algebraic operations and the sup-norm

$$\|h\| \triangleq \sup \{|h(x)| : x \in S^{n-1}\}$$

on the sphere $S^{n-1} = \{x \in R^n : |x| = 1\}$ becomes a normed linear space denoted below by $DSL(R^n)$ or simply DSL . By Proposition 2.1(a) to every $h \in DSL$ corresponds a pair of sets $A_1, A_2 \in \Omega$ such that

$$(2.1) \quad h(\cdot) = \sigma(\cdot | A_1) - \sigma(\cdot | A_2).$$

Consider an equivalence relation \sim defined on $\Omega \times \Omega$ by stating that $(A_1, A_2) \sim (B_1, B_2)$ iff $A_1 + B_2 = A_2 + B_1$, i.e. the pairs (A_1, A_2) and (B_1, B_2) are equivalent iff they represent the same function h in (2.1). The equivalence class containing (A_1, A_2) will be denoted by $[A_1, A_2]$. The linear space $\tilde{\Omega}$ is taken to be the quotient space $\Omega \times \Omega / \sim$ where the algebraic operations in $\tilde{\Omega}$ are imposed by the linear space DSL , i.e.,

$$[A_1, A_2] + [B_1, B_2] = [A_1 + B_1, A_2 + B_2],$$

and if $\alpha \geq 0$ then

$$\alpha[A_1, A_2] \equiv [\alpha A_1, \alpha A_2],$$

while if $\alpha < 0$ then

$$\alpha[A_1, A_2] \equiv [-\alpha A_2, -\alpha A_1].$$

The space $\tilde{\Omega}$ with the norm

$$\|[A_1, A_2]\| \triangleq \rho(A_1, A_2),$$

where $\rho(A_1, A_2)$ is the Hausdorff distance between A_1 and A_2 , becomes a normed space, which has been introduced by Radström [12]. It can be shown that the correspondence

$$(2.2) \quad [A_1, A_2] \leftrightarrow \mathcal{A}(\cdot | A_1) - \mathcal{A}(\cdot | A_2)$$

between the Radström's space $\tilde{\Omega}$ and *DSL* is *isometric* (cf. Demyanov and Rubinov [1]).

Consider a function $f: R^n \rightarrow R$. We say that f is directionally differentiable at x if the directional derivative

$$(2.3) \quad f'_x(y) = \lim_{t \downarrow 0} \frac{f(x + ty) - f(x)}{t}$$

exists for every $y \in R^n$. If convergence in (2.3) is uniform in y , i.e.

$$\lim_{y \rightarrow 0} \frac{f(x + y) - f(x) - f'_x(y)}{|y|} = 0$$

then it will be said that f is *uniformly directionally differentiable* at x .

Now we are prepared to introduce the basic concept of quasidifferentiability (see [1], [2], [3] and [4]).

DEFINITION 2.1. A function $f: R^n \rightarrow R$ is said to be (uniformly) quasidifferentiable at x if it is (uniformly) directionally differentiable at x and $f'_x(\cdot)$ belongs to the space *DSL*, i.e.,

$$f'_x(\cdot) = \mathcal{A}(\cdot | A_1) - \mathcal{A}(\cdot | A_2).$$

The corresponding class $[A_1, A_2] \in \tilde{\Omega}$ is called the quasidifferential of f at x and denoted

$$\mathcal{D}f(x) = [\underline{\partial}f(x), \overline{\partial}f(x)].$$

Note that the pair of sets $(\underline{\partial}f(x), \overline{\partial}f(x))$ is not unique and can be any representative of the corresponding class $\mathcal{D}f(x)$.

Remark 2.1. We use notations slightly different from those proposed in [1] and [4]. There $f'_x(\cdot)$ has been considered representable in the form

$$f'_x(y) = \max \{y \cdot v : v \in \underline{\partial}f(x)\} + \min \{y \cdot w : w \in \overline{\partial}f(x)\}.$$

Clearly two approaches are equivalent except that our $\overline{\partial}f(x)$ is $-\overline{\partial}f(x)$ in their notations.

Remark 2.2. If the directional derivative $f'_x(y)$ is sublinear in y , then $\mathcal{D}f(x) \equiv (\underline{\partial}f(x), \{0\})$. Here the set $\underline{\partial}f(x) \in \Omega$ is uniquely defined and called the subdifferential of f at x . Correspondingly the function f is said to be subdifferentiable at x . The class of subdifferentiable functions has been introduced by Pshenichnyi [11] under the name of quasidifferentiable functions.

The following theorem shows that the class of quasidifferentiable functions is closed under superposition and therefore is much wider than the class of subdifferentiable functions. This result is due to Demyanov and Rubinov [3] and we propose a short proof of it.

THEOREM 2.1. Let $F = (f_1, \dots, f_m): R^n \rightarrow R^m$ and $g: R^m \rightarrow R$. If the functions f_1, \dots, f_m are quasidifferentiable at x and g is uniformly quasidifferentiable at $y = F(x)$, then the superposition $h = g \circ F$ is quasidifferentiable at x . Moreover if the sets $\underline{\partial}g(y)$ and $\overline{\partial}g(y)$ are chosen in such a way that $0 \leq \alpha_i \leq M$, $0 \leq \beta_i \leq M$, $i = 1, \dots, m$, for some

$M \geq 0$ and all $\alpha = (\alpha_1, \dots, \alpha_m) \in \underline{\partial g(y)}$, $\beta = (\beta_1, \dots, \beta_m) \in \overline{\partial g(y)}$, then

$$(2.4) \quad [\underline{\partial h(x)}, \overline{\partial h(x)}] \equiv \left[\text{conv} \left\{ \bigcup_{\alpha \in \underline{\partial g(y)}} \sum_{i=1}^m (\alpha_i \underline{\partial f_i(x)} + (M - \alpha_i) \overline{\partial f_i(x)}) \right\}, \right. \\ \left. \text{conv} \left\{ \bigcup_{\beta \in \overline{\partial g(y)}} \sum_{i=1}^m (\beta_i \underline{\partial f_i(x)} + (M - \beta_i) \overline{\partial f_i(x)}) \right\} \right].$$

Proof. First we note that a representative of the quasidifferential $\mathcal{D}g(y)$ can be always chosen in such a way that every vector from $\underline{\partial g(y)}$ and $\overline{\partial g(y)}$ has nonnegative coordinates. Therefore we can assume without loss of generality that $\underline{\partial g(y)}$ and $\overline{\partial g(y)}$ satisfy the conditions above.

By the chain rule for directional derivatives we have that h is directionally differentiable at x and

$$(2.5) \quad h'_x(\cdot) = \sup \left\{ \sum_{i=1}^m \alpha_i f'_{ix}(\cdot) : \alpha \in \underline{\partial g(y)} \right\} - \sup \left\{ \sum_{i=1}^m \beta_i f'_{ix}(\cdot) : \beta \in \overline{\partial g(y)} \right\}.$$

The sum in the first term of (2.5) can be represented in the following way

$$\sum_{i=1}^m \alpha_i f'_{ix}(\cdot) = \sum_{i=1}^m \alpha_i [\mathcal{J}(\cdot | \underline{\partial f_i(x)}) - \mathcal{J}(\cdot | \overline{\partial f_i(x)})] \\ = \sum_{i=1}^m [\alpha_i \mathcal{J}(\cdot | \underline{\partial f_i(x)}) + (M - \alpha_i) \mathcal{J}(\cdot | \overline{\partial f_i(x)}) - M \mathcal{J}(\cdot | \overline{\partial f_i(x)})].$$

Since α_i and $M - \alpha_i$ are all nonnegative, we obtain by Proposition 2.1(b)

$$\sum_{i=1}^m \alpha_i f'_{ix}(\cdot) = \mathcal{J} \left(\cdot \mid \sum_{i=1}^m [\alpha_i \underline{\partial f_i(x)} + (M - \alpha_i) \overline{\partial f_i(x)}] \right) - \mathcal{J}(\cdot | B)$$

where $B = M \sum_{i=1}^m \overline{\partial f_i(x)}$ is independent of α . Now by Proposition 2.1(d) the supremum in the first term of (2.5) is given by

$$\mathcal{J}(\cdot | A_1) - \mathcal{J}(\cdot | B)$$

where

$$A_1 = \text{conv} \left\{ \bigcup_{\alpha \in \underline{\partial g(y)}} \sum_{i=1}^m [\alpha_i \underline{\partial f_i(x)} + (M - \alpha_i) \overline{\partial f_i(x)}] \right\}.$$

Similarly the second term in (2.5) is

$$\mathcal{J}(\cdot | A_2) - \mathcal{J}(\cdot | B)$$

where

$$A_2 = \text{conv} \left\{ \bigcup_{\beta \in \overline{\partial g(y)}} \sum_{i=1}^m [\beta_i \underline{\partial f_i(x)} + (M - \beta_i) \overline{\partial f_i(x)}] \right\}.$$

Finally we obtain

$$h'_x(\cdot) = \mathcal{J}(\cdot | A_1) - \mathcal{J}(\cdot | A_2)$$

which proves (2.4). \square

If the function g is Fréchet differentiable at y , then expression (2.4) takes the following simple form (see [1, Thm. 3] and [4, p. 197]):

$$\mathcal{D}h(x) = \sum_{i=1}^m \alpha_i \mathcal{D}f_i(x)$$

where $\alpha = (\alpha_1, \dots, \alpha_m)$ is the gradient of g at y .

Another important example is the max-function

$$g(y) = \max \{y_i: i = 1, \dots, m\}.$$

COROLLARY 2.1 (see [3] and [4, Lemma 2.2]). *If functions f_i , $i = 1, \dots, m$, are quasidifferentiable at x , then the max-function*

$$f(x) = \max \{f_i(x): 1 \leq i \leq m\}$$

is quasidifferentiable at x and

$$(2.6) \quad \mathcal{D}f(x) = \left[\text{conv} \left\{ \bigcup_{i \in I(x)} \left(\overline{\partial f_i(x)} + \sum_{\substack{j \neq i \\ j \in I(x)}} \overline{\partial f_j(x)} \right) \right\}, \sum_{j \in I(x)} \overline{\partial f_j(x)} \right]$$

where $I(x) = \{i: f(x) = f_i(x), i = 1, \dots, m\}$.

A similar result holds for the min-function.

3. The reduction theorem. Consider the following optimization problem:

$$(P1) \quad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) \leq 0 \text{ and } F(x) = 0 \end{array}$$

where all functions $f, g: R^n \rightarrow R$ and $F = (f_1, \dots, f_m): R^n \rightarrow R^m$ are supposed to be quasidifferentiable at points considered.

Note that a number of inequality constraints $g_i(x) \leq 0$, $i = 1, \dots, p$, can be always replaced by $g(x) \leq 0$ where $g(x) = \max \{g_i(x): i = 1, \dots, p\}$. Since the max-function g is quasidifferentiable if g_i , $i = 1, \dots, p$, are (see Corollary 2.1), it is not really restrictive to impose only one inequality constraint in Problem (P1).

Under some regularity conditions on the mapping $F(x)$ the constrained problem (P1) is locally reducible to a certain unconstrained problem. We use here a version of the reduction theorem proposed by Ioffe [7]. The second part (sufficiency) of this theorem is simple and does not require any regularity conditions. We formulate it first. In what follows it will be assumed that x^* is a feasible point of Problem (P1) lying on the boundary of the feasible region, i.e. $g(x^*) = 0$ and $F(x^*) = 0$, and we define

$$(3.1) \quad h(x) \triangleq \max \{f(x) - f(x^*), g(x)\}.$$

THEOREM 3.1. *If the function*

$$h(x) + r \|F(x)\|$$

attains a strict local minimum at x^ for some r , then x^* is an isolated local solution of Problem (P1).*

The other part (necessity) is more tricky and requires a sort of regularity conditions. The regularity conditions we use here are slightly different from those proposed by Ioffe and will be formulated in terms of quasidifferentials only.

By $\Delta(y|A)$ we denote the set of contact points of a bounded set A corresponding to a direction y

$$\Delta(y|A) \triangleq \{z \in A: \mathcal{A}(y|A) = y \cdot z\}.$$

Note that $\mathcal{A}(\cdot|A)$ is differentiable at y iff the set $\Delta(y|A)$ is singleton.

DEFINITION 3.1. A point x is said to be regular for F if for every y such that $F'_x(y) = 0$ it follows that

- (a) The sets $\Delta(y|\partial f_i(x)) = \{a_i\}$ and $\Delta(y|\overline{\partial f_i(x)}) = \{b_i\}$, $i = 1, \dots, m$, are singletons.
- (b) The vectors $a_1 - b_1, \dots, a_m - b_m$, are linearly independent.

Note that condition (a), which clearly depends on a choice of representatives $(\partial f_i(x), \overline{\partial f_i(x)})$, is not too restrictive since the contact set $\Delta(y|A)$ of a set $A \in \Omega$ is a singleton for almost every y .

If $\Delta(y|A) = \{a\}$ is a singleton then the corresponding support function $\mathcal{A}(\cdot|A)$ is differentiable at y and its gradient at y is given by a . Therefore Definition 3.1 means that x is regular if $F'_x(\cdot)$ is differentiable and the corresponding Jacobian matrix is of full rank m at every point y such that $F'_x(y) = 0$. If the mapping F is Fréchet differentiable at x , then it becomes a familiar condition for the Jacobian matrix of F to be of full rank m .

It is simple to verify the following assertion (see Ioffe [7]): If x^* is a local solution of Problem (P1), then x^* is a local solution of the following problem

$$(P2) \quad \begin{array}{ll} \text{minimize} & h(x) \\ \text{subject to} & F(x) = 0 \end{array}$$

where h is defined by (3.1).

Now we show that Problem (P2) can be linearized as follows.

THEOREM 2.2. Let h and F be uniformly directionally differentiable at x^* and let x^* be a regular point for F . If x^* is a local solution of Problem (P2), then $y = 0$ is the solution of the following problem

$$(P3) \quad \begin{array}{ll} \text{minimize} & h'_{x^*}(y) \\ \text{subject to} & F'_{x^*}(y) = 0. \end{array}$$

Proof. First we need to show that if $F'_{x^*}(\bar{y}) = 0$, $\bar{y} \neq 0$, then \bar{y} is a feasible direction in the sense that there exists a curve $y(t) : [0, \varepsilon] \rightarrow R^n$ such that $y(0) = 0$, $F(x^* + y(t)) = 0$ and

$$\lim_{t \downarrow 0} \|\bar{y} - t^{-1}y(t)\| = 0.$$

The proof of this is rather standard and is based on Brouwer's fixed point theorem. For example one may proceed precisely as in [11], pp. 114–119. (Note that since x^* is regular, F'_{x^*} is differentiable in a neighborhood of \bar{y} .)

Now we have that if x^* is a local solution of Problem (P2), then $h(x^* + y(t)) \geq h(x^*)$ and consequently $h'_{x^*}(\bar{y}) \geq 0$. We have shown that $h'_{x^*}(y) \geq 0$ whenever $F'_{x^*}(y) = 0$ and the proof is complete. \square

Now we are prepared to formulate the reduction theorem with respect to Problem (P3).

THEOREM 3.3. Let x^* be a regular point for F . If $y = 0$ is the solution of Problem (P3), then for all sufficiently large $r \geq 0$, the function

$$M_r(y) = h'_{x^*}(y) + r\|F'_{x^*}(y)\|$$

attains an unconstrained minimum at $y = 0$.

Proof. Let $y = 0$ be the solution of Problem (P3) and thus $h'_{x^*}(y) \geq 0$ whenever $y \in \Phi$, where $\Phi = \{y \in S^{n-1} : F'_{x^*}(y) = 0\}$. Since h'_{x^*} and F'_{x^*} are positively homogeneous, it is sufficient to show that there exists $r > 0$ such that $M_r(y) \geq 0$ whenever $y \in S^{n-1}$.

Consider a point $y \in \Phi$. Since x^* is a regular point for F , it follows that F'_{x^*} is differentiable in a neighborhood of y and the Jacobian matrix of F'_{x^*} at y is of full rank m . Now we have from the results of Ioffe [7], [8] that there exist $r > 0$ and a neighborhood of y such that M_r is nonnegative in this neighborhood. Since the set Φ is compact, it can be covered by a finite number of such neighborhoods and r can be chosen independently of $y \in \Phi$. It remains to note that outside the union of these neighborhoods the function $\|F'_{x^*}(\cdot)\|$ is greater than a certain positive number and that the continuous function h'_{x^*} is bounded below on S^{n-1} . \square

4. Optimality conditions. In this section we apply quasidifferential calculus and the reduction theorem to give optimality conditions for Problem (P1). Optimality conditions for the unconstrained problem are easily obtainable from the result of Proposition 2.1(c) (see Demyanov and Polyakova [2]).

THEOREM 4.1. *If a quasidifferentiable function $f(x)$ attains local minimum at x^* , then*

$$(4.1) \quad \overline{\partial f(x^*)} \subseteq \underline{\partial f(x^*)}.$$

Conversely, if $\underline{\partial f(x^)}$ is included in the interior of the set $\overline{\partial f(x^*)}$ and f is uniformly quasidifferentiable at x^* , then x^* is a strict local minimizer of f .*

It is worthwhile to note that Theorem 4.1 expresses the fact that $f'_{x^*}(\cdot) \geq 0$ is a necessary and $f'_{x^*}(\cdot) > 0$ is a sufficient condition for a (uniformly) quasidifferentiable function f to attain local minimum at x^* .

Now we formulate the main result.

THEOREM 4.2. *Let f, g and F be uniformly quasidifferentiable at x^* and $F(x^*) = 0$, $g(x^*) = 0$. If x is a local solution of Problem (P1) and x^* is a regular point for F , then, for all sufficiently large $r > 0$,*

$$(4.2) \quad \overline{\partial h(x^*)} + r \overline{\partial \|F(x^*)\|} \subseteq \underline{\partial h(x^*)} + r \underline{\partial \|F(x^*)\|}$$

where the max-function h is defined by (3.1) with

$$\underline{\partial h(x^*)} = \text{conv} \{ \overline{\partial f(x^*)} + \underline{\partial g(x^*)}, \underline{\partial f(x^*)} + \overline{\partial g(x^*)} \}$$

and

$$\overline{\partial h(x^*)} = \overline{\partial f(x^*)} + \overline{\partial g(x^*)}.$$

Conversely, if for some $r > 0$ the left side set is included in the interior of the right side set of (4.2), then x^ is a strict local solution of Problem (P1).*

Proof. First we observe that if x^* is a local solution of Problem (P1) and is a regular point for F , then, by Theorem 3.2, $y = 0$ is the solution of Problem (P3). Moreover, by the reduction theorem 3.3, for all sufficiently large $r > 0$ the function M_r attains an unconstrained minimum at $y = 0$ and consequently

$$\overline{\partial M_r(0)} \subseteq \underline{\partial M_r(0)},$$

which is equivalent to (4.2).

Similarly, the reduction theorem 3.1 together with the second part of Theorem 4.1 implies the sufficiency part of Theorem 4.2. \square

Remark 4.1. It follows from Theorem 2.1 that the function $\|F(\cdot)\|$ is quasidifferentiable at x^* for any norm $\|\cdot\|$. It is convenient to use the l_1 -norm. Then $\mathcal{D}\|F(x^*)\|$ becomes

$$(4.3) \quad \mathcal{D}\|F(x^*)\| = \sum_{i=1}^m \mathcal{D}|f_i(x^*)|$$

where the quasidifferential of the absolute value of a function f is given at x , $f(x) = 0$, by

$$(4.4) \quad \mathcal{D}|f(x)| = [2 \operatorname{conv} \{\underline{\partial f(x)}, \overline{\partial f(x)}\}, \underline{\partial f(x)} + \overline{\partial f(x)}].$$

If Problem (P1) is only inequality constrained, then necessary conditions become

$$(4.5) \quad \overline{\partial f(x^*)} + \underline{\partial g(x^*)} \subseteq \operatorname{conv} \{\underline{\partial f(x^*)} + \underline{\partial g(x^*)}, \underline{\partial f(x^*)} + \overline{\partial g(x^*)}\}.$$

These conditions are related to the necessary conditions (for inequality constrained problems) proposed by Demyanov and Polyakova [2]; (see also [4, Thm. 7.1]). However, unlike in [2], here no regularity assumptions are required and condition (4.5) is seemingly easier to verify (compare with the examples in [4, pp. 224–227]).

If f and g are subdifferentiable at x^* , then (4.5) takes the form

$$0 \in \operatorname{conv} \{\underline{\partial f(x^*)}, \underline{\partial g(x^*)}\},$$

which is equivalent to the familiar condition: There exist nonnegative numbers α, β , not both zero, such that

$$0 \in \alpha \underline{\partial f(x^*)} + \beta \underline{\partial g(x^*)}.$$

If the mapping F is Fréchet differentiable at x^* , then by (4.3) and (4.4) we have

$$\mathcal{D}\|F(x^*)\| = \left[\sum_{i=1}^m \operatorname{conv} \{\nabla f_i(x^*), -\nabla f_i(x^*)\}, \{0\} \right].$$

In this case the optimality conditions of Theorem 4.2 can be formulated in the following form (under the regularity assumption that the gradient vectors $\nabla f_1(x^*), \dots, \nabla f_m(x^*)$, are linearly independent):

For every vector $b \in \underline{\partial h(x^*)}$ there exist numbers (multipliers) $\alpha_1, \dots, \alpha_m$ such that

$$(4.6) \quad b \in \underline{\partial h(x^*)} + \sum_{i=1}^m \alpha_i \nabla f_i(x^*).$$

REFERENCES

- [1] V. F. DEMYANOV AND A. M. RUBINOV, *On quasidifferentiable functionals*, Soviet Math. Dokl., 21 (1980), pp. 14–17.
- [2] V. F. DEMYANOV AND L. N. POLYAKOVA, *Conditions for minimum of a quasidifferentiable function on a quasidifferentiable set*, USSR Comput. Math. Math. Phys., 20 (1980), pp. 34–43.
- [3] V. F. DEMYANOV AND A. M. RUBINOV, *On some approaches to the nonsmooth optimization problem*, Economics and Mathematical Methods, 17 (1981), pp. 1153–1174. (In Russian.)
- [4] V. F. DEMYANOV AND L. V. VASILIEV, *Nondifferentiable Optimization*, Nauka, Moscow, 1981. (In Russian.)
- [5] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.
- [6] ———, *A new approach to Lagrange multipliers*, Math. Oper. Res., 1 (1976), pp. 165–174.
- [7] A. D. IOFFE, *Necessary and sufficient conditions for a local minimum. 1: A reduction theorem and first order conditions*, this Journal, 17 (1979), pp. 245–250.
- [8] ———, *Regular points of Lipschitz mappings*, Trans. Amer. Math. Soc., 251 (1979), pp. 61–69.
- [9] L. N. POLYAKOVA, *Necessary conditions for extremum of a quasidifferentiable function on a quasidifferentiable set*, Vestnik of Leningrad University, 7 (1982), pp. 75–80. (In Russian.)
- [10] ———, *On one problem of nonsmooth optimization*, Cybernetics, 2 (1982), pp. 119–122.
- [11] B. N. PSHENICHNYI, *Necessary Conditions for an Extremum*, Marcel Dekker, New York, 1971.
- [12] H. RADSTRÖM, *An embedding theorem for spaces of convex sets*, Proc. Amer. Math. Soc., 3 (1952), pp. 165–169.
- [13] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.

OPTIMAL INTERPOLATION FOR LINEAR STOCHASTIC SYSTEMS*

MICHELE PAVON†

Abstract. Let y , satisfying $dy = Cx dt + D dw$, $y(0) = 0$, represent the noisy observation of the Gauss–Markov process x generated by $dx = Ax dt + B dw$, $x(0) = x_0$, on the finite interval $[0, T]$. In this paper we solve the following interpolation problem: determine the best least-squares estimate of $x(t)$, $t \in [0, T]$, given the increments of y on the intervals $[0, T_1]$ and $[T_2, T]$, where $0 < T_1 < T_2 < T$, and the statistics of x_0 . Several alternative expressions for the optimal interpolator are derived, some of which are phrased in terms of two Kalman–Bucy estimates. As a by-product, we also solve the interpolation problem for the missing increments of y .

Key words. nonstationary interpolation, Kalman–Bucy model, stochastic realization theory

1. Introduction. Let (Ω, \mathcal{F}, P) be a probability space and $\{w(t); 0 \leq t \leq T\}$ a standard p -dimensional Brownian motion defined on it. Let x_0 be a Gaussian centered n -dimensional random vector independent of w and $\{\mathcal{F}_t; 0 \leq t \leq T\}$ be the nondecreasing family of sub- σ -fields of \mathcal{F} given by $\mathcal{F}_t := \sigma(x_0, w(s); 0 \leq s \leq t)$. Consider the vector processes $\{x(t); 0 \leq t \leq T\}$ and $\{y(t); 0 \leq t \leq T\}$ of dimension n and $m \leq p$, respectively, defined as the solution to the system of stochastic differential equations

$$(1.1a) \quad dx = A(t)x(t) dt + B(t) dw, \quad x(0) = x_0,$$

$$(1.1b) \quad dy = C(t)x(t) dt + D(t) dw, \quad y(0) = 0.$$

Here $A(\cdot)$, $B(\cdot)$, $C(\cdot)$ and $D(\cdot)$ are continuous matrix functions of appropriate dimensions defined on $[0, T]$. It is well known [5] that the processes x and y are zero-mean, jointly Gaussian and nonanticipative with respect to the family $\{\mathcal{F}_t\}$. We think of y as representing the noisy observation of the state—modeled by x —of a dynamical system.

The purpose of this paper is to solve the following *interpolation problem*: find the best least-squares estimate $\hat{x}(t)$ of $x(t)$ $0 \leq t \leq T$, given the increments of y on the intervals $[0, T_1]$ and $[T_2, T]$, where $0 < T_1 < T_2 < T$, and the variance of x_0 . This problem appears to be of considerable importance in many diverse areas of engineering, natural and social sciences, in that it represents the common physical situation in which a black-out has occurred concerning the flow of information about the state of a linear system. To the best of the author's knowledge this problem is tackled and solved here for the first time. As a by-product, we solve the interpolation problem for the missing increments of y .

Our approach hinges on some basic concepts and results from the *stochastic realization theory* (see e.g. [7] and [13] for a bibliography) and on our previous work [1], [9], [10] on the smoothing and interpolation problems. Indeed this paper can be regarded as the natural continuation of the study initiated in [1] and [10] which we take as main references. Some of our results have a compact form, being phrased in terms of Kalman–Bucy estimates, and bear a strong analogy to the smoothing formulas. The derivation, relying on elementary geometry in Hilbert space, is direct and illuminating.

* Received by the editors December 14, 1982, and in revised form June 3, 1983. This research was supported in part by a CNR fellowship at the Department of Statistics, The Florida State University, Tallahassee, Florida and in part by a NATO fellowship at the Mathematisch Instituut, Rijksuniversiteit Groningen, Groningen, the Netherlands.

† LADSEB-CNR, Corso Stati Uniti 4, 35100 Padova, Italy.

The outline of the paper is as follows. In § 2 we introduce the relevant mathematical notation and formulate the problem. In § 3 we construct a stochastic realization framework about (1.1) in order to facilitate the solution of the interpolation problem. This is done much along the same lines as in [1]. In the following section we determine the optimal interpolation estimate for the missing increments of y . We then proceed to derive a representation for \hat{x} . § 5 is devoted to establishing several other formulas for \hat{x} . Some of these representations are particularly attractive in that they allow to get \hat{x} by an economical updating of previously computed smoothing or filtering estimates. Differential equations for \hat{x} are also derived. The paper concludes with some remarks on the convergence properties of \hat{x} and on how it changes as new information becomes available.

2. Basic notation and problem formulation. The mathematical notation is, with a few exceptions, the same as in [1]. This is in order to aid the reader, since we shall quote a number of preliminary results from the latter paper. In particular, if $\{z(t); t \in [0, T]\}$ is a centered vector Gaussian process defined on (Ω, \mathcal{F}, P) , we denote by $H(z(t))$, $H_t^-(dz)$, $H_t^+(dz)$ and $H(dz)$ the *Gaussian spaces* [8] induced by the components of $z(t)$ and by the components of the increments $\{z(r) - z(s); r, s \in I\}$, where I is the interval $[0, t]$, $[t, T]$ and $[0, T]$, respectively. If $H \subset L^2(\Omega, \mathcal{F}, P)$ is a Gaussian space, $E\{\cdot|H\}$ denotes the orthogonal projection onto H . We write $E\{v|H\}$ for the vector with components $E\{v_i|H\}$. If K is a Gaussian space contained in H , we indicate by $H \ominus K$ the orthogonal complement of K in H . When R is a symmetric positive definite matrix we write $R > 0$ and indicate by $R^{1/2}$ its positive square root. Transposition is denoted by prime. All vectors without prime are column vectors.

In view of the Gaussian assumption, the interpolation problem can be formulated as follows. Given $\{y(t) - y(s); t, s \in [0, T_1] \text{ or } t, s \in [T_2, T]\}$ and the statistics of x_0 , determine $\hat{x}(t) := E\{x(t)|H(T_1, T_2)\}$, where $H(T_1, T_2) := H_{T_1}^-(dy) \vee H_{T_2}^+(dy)$, i.e. the smallest Hilbert space containing $H_{T_1}^-(dy)$ and $H_{T_2}^+(dy)$.

3. Preliminaries. Let $R(t) := D(t)D(t)'$. We assume that $R(t) > 0$ for all $t \in [0, T]$. Moreover we assume that the matrix functions $A(\cdot)$, $B(\cdot)$, $C(\cdot)$ and $D(\cdot)$ are analytic on $[0, T]$ and that the representation (1.1) of y is *minimal* in the sense that there is no other Markovian representation of y on $[0, T]$ with a state process x of dimension less than n . The latter assumption can be easily seen to be equivalent to deterministic minimality of the system (1.1) with respect to the input–output map [2, p. 94] *plus* minimality of the state spaces $H(x(t))$, $t \in [0, T]$ in the sense of geometric stochastic realization theory, see e.g. [6]. While we refer the reader to [1, § 1] for a complete justification of the above assumptions, we remark here that they allow us to develop a stochastic realization framework for the interpolation problem without obscuring the key ideas by technicalities and without sacrificing mathematical rigor.

We start by associating to (1.1) the *Kalman–Bucy filter*

$$(3.1a) \quad dx_* = Ax_* dt + B_* dw_*, \quad x_*(0) = 0,$$

$$(3.1b) \quad dy = Cx_* dt + R^{1/2} dw_*, \quad y(0) = 0,$$

where the gain B_* is given by

$$(3.2a) \quad B_* = (Q_* C' + B D') R^{-1/2},$$

the error covariance matrix

$$(3.2b) \quad Q_*(t) = E\{[x(t) - x_*(t)][x(t) - x_*(t)]'\}$$

solving the matrix differential equation

$$(3.2c) \quad \begin{aligned} \dot{Q}_* &= A Q_* + Q_* A' - (Q_* C' + B D') R^{-1} (C Q_* + D B') + B B', \\ Q_*(0) &= P_0 := E\{x_0 x_0'\}. \end{aligned}$$

Here the *innovations process* $\{w_*(t); 0 \leq t \leq T\}$ is a standard m -dimensional Brownian motion satisfying $H_t^-(dw_*) = H_t^-(dy)$ for all $t \in [0, T]$.

We now embed the models (1.1) and (3.1) in the class \mathcal{S} of all Markovian representations of y on $[0, T]$ which are minimal, have analytic system matrices and have (3.1) as Kalman–Bucy filter. Hence, if x is the state process of any such model, we have the filter property

$$(3.3) \quad E\{x(t) | H_t^-(dy)\} = x_*(t), \quad t \in [0, T].$$

Shortly we shall see that there is an important realization in \mathcal{S} (see (3.10) below) whose state variance tends to infinity as $t \rightarrow T$. This explains why we do not require the elements of \mathcal{S} to be defined on the closed interval $[0, T]$. Let $G := P_* C' + B_* R^{1/2}$ with $P_*(t) := E\{x_*(t) x_*(t)'\}$. Then it can be seen [1] that $G = P C' + B R^{1/2}$, $t \in [0, T]$ for any realization in \mathcal{S} , where P is the state variance matrix generated by

$$(3.4) \quad \dot{P} = A P + P A' + B B'$$

with initial condition $P(0) = P_0$. The matrices A , C , G and R are invariant over the class \mathcal{S} (indeed they determine the covariance of y [4]), whereas B , D , P , w , x and the dimension $p \geq m$ of w vary with different representations in \mathcal{S} . As shown in [1], to each model (1.1) in \mathcal{S} there corresponds a backward realization of y defined on $(0, T]$ of the form

$$(3.5a) \quad d\bar{x} = -A' \bar{x} dt + \bar{B} d\bar{w}, \quad \bar{x}(T) = \bar{x}_T,$$

$$(3.5b) \quad dy = G' \bar{x} dt + D d\bar{w},$$

where $\bar{x}_T = \lim_{t \rightarrow T} P(t)^{-1} x(t)$ is orthogonal to $H(d\bar{w})$, $\bar{B} = P^{-1} B$, $\bar{x} = P^{-1} x$ and $d\bar{w} = dw - B' P^{-1} x dt$. The invertibility of the state variance P on $(0, T)$ is ensured by the analyticity assumption [1]. It can be seen that, for each realization in \mathcal{S} , \bar{B} and \bar{x} , which are defined on $(0, T)$, can be extended by continuity to all of $(0, T]$, see [1]. Note that while the Markov property of x is reflected by the orthogonal decomposition

$$(3.6) \quad x(t) = \Phi(t, s) x(s) + \int_s^t \Phi(t, r) B(r) dw, \quad s \leq t,$$

its counterpart for \bar{x} is given by

$$(3.7) \quad \bar{x}(s) = \Phi(t, s)' \bar{x}(t) + \int_t^s \Phi(r, s)' \bar{B}(r) d\bar{w},$$

where Φ is the transition matrix defined by

$$\frac{\partial \Phi}{\partial t}(t, s) = A(t) \Phi(t, s), \quad \Phi(s, s) = I.$$

Among models (3.5) there is their (*backward*) *Kalman–Bucy filter*

$$(3.8a) \quad d\bar{x}_* = -A' \bar{x}_* dt + \bar{B}_* d\bar{w}_*, \quad \bar{x}_*(T) = 0,$$

$$(3.8b) \quad dy = G' \bar{x}_* dt + R^{1/2} d\bar{w}_*$$

driven by the *backward innovations process* \bar{w}_* satisfying $H_t^+(d\bar{w}_*) = H_t^+(dy)$ for all $t \in [0, T]$. The filter property is here expressed by

$$(3.9) \quad E\{\bar{x}(t)|H_t^+(dy)\} = \bar{x}_*(t), \quad t \in (0, T],$$

for any \bar{x} of the type (3.5a). This realization is actually defined on all of $[0, T]$, see [1]. To (3.8) there corresponds a unique forward representation in \mathcal{S} given by

$$(3.10a) \quad dx^* = Ax^* dt + B^* dw^*, \quad x^*(0) = \bar{P}_*(0)^{-1} \bar{x}_*(0),$$

$$(3.10b) \quad dy = Cx^* dt + R^{1/2} dw^*, \quad y(0) = 0,$$

where $\bar{P}_*(t) = E\{\bar{x}_*(t)\bar{x}_*(t)'\}$. Thus, for $t \in [0, T]$, we have

$$(3.11a) \quad E\{x^*(t)x^{*'}(t)\} := P^*(t) = \bar{P}_*(t)^{-1},$$

$$(3.11b) \quad \bar{x}_*(t) = P^*(t)^{-1} x^*(t).$$

We remark that (3.1) and (3.10) are the minimum and maximum variance realizations in \mathcal{S} respectively and that, under the present assumptions, $Q(t) := P^*(t) - P_*(t)$ is positive definite for all $t \in [0, T]$, see [1]. This is due to the fact that the time interval is finite and it contrasts the stationary infinite-interval case where further assumptions on the process y are needed in order to ensure such property, see e.g. [4]. We now like to derive a representation for $x_*(t)$ in terms of the past increments of y and a representation for $\bar{x}_*(t)$ in terms of the future increments of y . To do so we first invert the systems (3.1) and (3.8) and get

$$(3.12) \quad dx_* = \Gamma_* x_* dt + B_* R^{-1/2} dy, \quad x_*(0) = 0,$$

$$(3.13) \quad d\bar{x}_* = -\bar{\Gamma}_* \bar{x}_* dt + \bar{B}_* R^{-1/2} dy, \quad \bar{x}_*(T) = 0,$$

where the feedback matrices Γ_* and $\bar{\Gamma}_*$ are given by $\Gamma_* = A - B_* R^{-1/2} C$ and $\bar{\Gamma}_* = A' + \bar{B}_* R^{-1/2} G'$. Integrating (3.12) and (3.13), we then obtain

$$(3.14) \quad x_*(t) = \Psi(t, s) x_*(s) + \int_s^t \Psi(t, r) B_*(r) R(r)^{-1/2} dy, \quad s \leq t,$$

$$(3.15) \quad \bar{x}_*(s) = \bar{\Psi}(t, s) \bar{x}_*(t) + \int_t^s \bar{\Psi}(r, s) \bar{B}_*(r) R(r)^{-1/2} dy, \quad s \leq t.$$

Here Ψ and $\bar{\Psi}$ are transition matrices satisfying

$$(3.16) \quad \frac{\partial \Psi}{\partial t}(t, s) = \Gamma_*(t) \Psi(t, s), \quad \Psi(s, s) = I,$$

$$(3.17) \quad \frac{\partial \bar{\Psi}}{\partial t}(t, s) = \bar{\Psi}(t, s) \bar{\Gamma}_*(t), \quad \bar{\Psi}(s, s) = I.$$

Let us also recall [1] that $\bar{z}(t) := Q(t)^{-1}(x^*(t) - x_*(t))$ is a Markov process generated by

$$(3.18) \quad d\bar{z} = -\Gamma_*' \bar{z} dt - C' R^{-1/2} dw_*, \quad \bar{z}(T) = 0,$$

and therefore Q^{-1} satisfies

$$(3.19) \quad (\dot{Q}^{-1}) = -\Gamma_*' Q^{-1} - Q^{-1} \Gamma_* - C' R^{-1} C, \quad Q^{-1}(T) = 0.$$

Finally we like to stress the fact that from the knowledge of P_0, A, B, C and D we can compute such quantities as $P_*(t), P^*(t), B_*(t), \bar{B}_*(t), \Gamma_*(t), \bar{\Gamma}_*(t), Q_*(t)$ and $Q^*(t) := P^*(t) - P(t)$, see [1, § 3] for details. This should be understood when we call

solutions to the interpolation problem various expressions in § 4 and 5 which we involve the above mentioned quantities.

4. A solution to the interpolation problem. In this § we solve the interpolation problem for y , which is of independent interest, computing $\hat{y}(t) - \hat{y}(s) := E\{y(t) - y(s) | H(T_1, T_2)\}$ for $T_1 \leq s \leq t \leq T_2$. Next we show that \hat{y} yields a representation for $\hat{x}(t)$, $t \in [0, T]$. Other, more compact, expressions for $\hat{x}(t)$ will be derived in § 5 without resorting to \hat{y} . The proofs of the three lemmas and of the theorem below are trivial modifications of the proofs of the equally numbered results in [10] and are therefore deleted. Let us just remark here that the key step in the derivation consists in showing that the two vectors $\{x_*(T_1), \bar{x}_*(T_2)\}$ represent a “sufficient statistic” for the estimation problem (Lemma 4.2).

LEMMA 4.1. *The space $H(T_1, T_2)$ admits the following orthogonal decomposition:*

$$(4.1) \quad H(T_1, T_2) = N^- \oplus [H(x_*(T_1)) \vee H(\bar{x}_*(T_2))] \oplus N^+,$$

where $N^- := H_{T_1}^-(dy) \ominus H(x_*(T_1))$ and $N^+ := H_{T_2}^+(dy) \ominus H(\bar{x}_*(T_2))$.

LEMMA 4.2. $H(\hat{y}(t) - \hat{y}(s)) \subset [H(x_*(T_1)) \vee H(\bar{x}_*(T_2))]$.

LEMMA 4.3. *The components of the vector $x^*(T_2) - \Phi(T_2, T_1)x_*(T_1)$ are orthogonal to $H(x_*(T_1))$. Moreover its variance*

$$(4.2) \quad \Pi = P^*(T_2) - \Phi(T_2, T_1)P_*(T_1)\Phi(T_2, T_1)'$$

is positive definite.

THEOREM 4.4. *The interpolation estimate \hat{y} is differentiable on $(0, T)$. Its derivative admits the following orthogonal decomposition:*

$$(4.3) \quad \frac{d\hat{y}}{dt} = H(t)x_*(T_1) + K(t)\Pi^{-1}[x^*(T_2) - \Phi(T_2, T_1)x_*(T_1)],$$

where $H(t) := C(t)\Phi(t, T_1)$, $K(t) = G(t)'\Phi(T_2, t)' - C(t)\Phi(t, T_1)P_*(T_1)\Phi(T_2, T_1)'$ and Π is defined by (4.2). Let $\tilde{y}(t) = y(t) - \hat{y}(t)$ denote the estimation error. Then

$$(4.4) \quad \begin{aligned} & E\{[\tilde{y}(t) - \tilde{y}(s)][\tilde{y}(t) - \tilde{y}(s)]'\} \\ &= \int_s^t \int_s^t \int_{T_1}^{\tau \wedge \tau'} [H(\tau - \sigma)B_*(\sigma)B_*(\sigma)'H(\tau' - \sigma)'] d\sigma d\tau d\tau' \\ &+ \int_s^t \int_s^\tau [H(\tau - \sigma)B_*(\sigma)R(\sigma)^{1/2} + R(\sigma)^{1/2}B_*(\sigma)'H(\tau - \sigma)'] d\sigma d\tau \\ &+ \int_s^t R(\sigma) d\sigma - M(t-s)\Pi^{-1}M(t-s)' \end{aligned}$$

where $\tau \wedge \tau' := \min(\tau, \tau')$ and $M(t-s) = \int_s^t K(\tau) d\tau$.

Observe that the last term in (4.3) represents a modification of the prediction estimate due to the extra information $H_{T_2}^+(dy)$. It is easy to derive from (4.3) the following symmetric expression:

$$(4.5) \quad \frac{d\hat{y}}{dt} = \bar{K}(t)\bar{\Pi}^{-1}P_*(T_1)^{-1}x_*(T_1) + K(t)\Pi^{-1}P^*(T_2)\bar{x}_*(T_2),$$

where $\bar{K}(t) := C(t)\Phi(t, T_1) - G(t)'\Phi(T_2, t)'P^*(T_2)\Phi(T_2, T_1)$ and $\bar{\Pi} := P_*(T_1)^{-1} - \Phi(T_2, T_1)'P^*(T_2)^{-1}\Phi(T_2, T_1)$. Then (3.14) and (3.15) yield the following result.

THEOREM 4.5. *The derivative of the interpolation estimate \hat{y} is given by*

$$(4.6) \quad \frac{d\hat{y}}{dt} = \int_0^{T_1} \bar{K}(t) \bar{\Pi}^{-1} P_*(T_1)^{-1} \Psi(T_1, r) B_*(r) R(r)^{-1/2} dy \\ + \int_T^{T_2} K(t) \Pi^{-1} P^*(T_2) \bar{\Psi}(r, T_2) \bar{B}_*(r) R(r)^{-1/2} dy.$$

This solves the interpolation problem for y . Other representations for \hat{y} can also be easily derived from (4.3) and (4.5) along the lines of [10].

We now turn to the problem of estimating $x(t)$, $t \in [0, T]$. Let $\hat{x}_s(t) := E\{x(t) | H(dy)\}$ be the smoothing estimate. Then

$$(4.7) \quad \hat{x}_s(t) = Q^*(t) Q(t)^{-1} x_*(t) + Q_*(t) Q(t)^{-1} P^*(t) \bar{x}_*(t),$$

see [1, Thm. 4.1]. Using (3.14) and (3.15) again, we get

$$(4.8a) \quad \hat{x}_s(t) = \int_0^T T(t, s) dy(s)$$

where

$$(4.8b) \quad T(t, s) = Q^*(t) Q(t)^{-1} \Psi(t, s) B_*(s) R(s)^{-1/2} \quad \text{for } s \leq t,$$

$$(4.8c) \quad T(t, s) = -Q_*(t) Q(t)^{-1} P^*(t) \bar{\Psi}(s, t) \bar{B}_*(s) R(s)^{-1/2} \quad \text{for } s > t.$$

Notice that the kernel $T(t, \cdot)$ is in general discontinuous at $s = t$.

THEOREM 4.6. *The interpolation estimate \hat{x} is given by*

$$(4.9) \quad \hat{x}(t) = \int_0^{T_1} T(t, s) dy(s) + \int_{T_1}^{T_2} T(t, s) d\hat{y}(s) + \int_{T_2}^T T(t, s) dy(s)$$

where T is as in (4.8) and $d\hat{y}$ as in (4.6).

Proof. Since $\hat{x}(t)$ is just the orthogonal projection of $x(t)$ onto $H(T_1, T_2)$ and the right-hand side of (4.9) clearly belongs to such space, it suffices to show that $x(t) - \hat{x}(t)$, with \hat{x} defined by (4.9), is indeed orthogonal to $H(T_1, T_2)$. Note that, in view of (4.8a) we can write $x(t) - \hat{x}(t) = x(t) - \hat{x}_s(t) + \int_{T_1}^{T_2} T(t, s) dy(s) - \int_{T_1}^{T_2} T(t, s) d\hat{y}(s)$. It only remains to observe that $x(t) - \hat{x}_s(t)$ is orthogonal to $H(dy)$ which contains $H(T_1, T_2)$ and the difference between the two integrals is orthogonal to $H(T_1, T_2)$. \square

Now inserting (4.6) into (4.9), we can readily obtain an expression for \hat{x} in terms of the available increments of y . Such a representation, however, is not compact and it takes long and tedious calculations to turn it into a more manageable one. On the other hand we would like to have formulas for \hat{x} in terms of x_* and \bar{x}_* . This because in some situations it might be possible to compute \hat{x} by simple updating of previous estimates phrased in terms of the two filters. We also like, whenever possible, to get a differential equation for \hat{x} . All of this motivates our study in the next section.

5. Main results. We first consider the case when $t \in [T_1, T_2]$.

THEOREM 5.1. *The interpolation estimate $\hat{x}(t)$ is given by*

$$(5.1) \quad \hat{x}(t) = \Phi(t, T_1) x_*(T_1) + U(t) \Phi(T_2, t) \Pi^{-1} [x^*(T_2) - \Phi(T_2, T_1) x_*(T_1)]$$

for all $t \in [T_1, T_2]$, where

$$(5.2) \quad U(T) := P(t) - \Phi(t, T_1) P_*(T_1) \Phi(t, T_1)'$$

and Π is defined by (4.2). Let $\Sigma(t) := E\{[x(t) - \hat{x}(t)][x(t) - \hat{x}(t)]'\}$ be the error variance function. Then

$$(5.3) \quad \Sigma(t) = U(t) - U(t)\Phi(T_2, t)\Pi^{-1}\Phi(T_2, t)U(t).$$

Proof. The proof is very similar to that of Theorem 4.4 and is outlined here for the sake of continuity in exposition. The components of $\hat{x}(t) - \Phi(t, T_1)x_*(T_1)$ are orthogonal to $H_{T_1}^-(dy) \supset N^-$ since $E\{\hat{x}(t)|H_{T_1}^-(dy)\} = E\{x(t)|H_{T_1}^-(dy)\} = \Phi(t, T_1)x_*(T_1)$. Hence $H(\hat{x}(t)) \perp N^-$. Similarly it is seen that $H(\hat{x}(t)) \perp N^+$. It then follows from Lemma 4.1 that $H(\hat{x}(t)) \subset [H(x_*(T_1)) \vee H(\bar{x}_*(T_2))]$. Next observe that, in view of (3.11b), the components of $x_*(T_1)$ and $x^*(T_2) - \Phi(T_2, T_1)x_*(T_1)$ span the latter space. Expression (5.1) is now a consequence of Lemma 4.3 and of a standard projection formula. Finally, the orthogonality between the estimation error $x(t) - \hat{x}(t)$ and the estimate $\hat{x}(t)$ quickly yields (5.3). \square

Formula (5.1) can be written in a symmetric form. Define

$$(5.4) \quad V(t) := \Phi(t, T_2)P^*(T_2)\Phi(t, T_2)' - P(t)$$

and $Z(t) := \Phi(T_2, t)\Pi^{-1}\Phi(T_2, t)$. Then (5.1) becomes

$$(5.5) \quad \hat{x}(t) = V(t)Z(t)\Phi(t, T_1)x_*(T_1) + U(t)Z(t)\Phi(t, T_2)x^*(T_2).$$

Here and in the sequel it should be understood that the vectors $x_*(T_1)$ and $x^*(T_2)$ can be computed from the data via (3.12), (3.13) and (3.11b). Also U can be easily seen to satisfy (3.4) on $[T_1, T_2]$ with initial condition $U(T_1) = Q_*(T_1)$, whereas V is given by

$$(5.6) \quad \dot{V} = AV + VA' - BB', \quad V(T_2) = Q^*(T_2).$$

It is not difficult to derive other representations for the optimal estimate from Theorem 5.1.

COROLLARY 5.2. *Let $t \in (T_1, T_2)$. Then $U(t)$ and $V(t)$, as given by (5.2) and (5.4) respectively, are positive definite and the interpolation estimate \hat{x} is given by*

$$(5.7) \quad \hat{x}(t) = \Sigma(t)[U(t)^{-1}\Phi(t, T_1)x_*(T_1) + V(t)^{-1}\Phi(t, T_2)x^*(T_2)]$$

with the error variance satisfying

$$(5.8) \quad \Sigma(t)^{-1} = U(t)^{-1} + V(t)^{-1}.$$

Before proving this result let us remark here that it brings to light a striking analogy between the interpolation problem for x when $t \in [T_1, T_2]$ and the smoothing problem. Indeed equations (5.7) and (5.8) are quite analogous to the Mayne-Frazer smoothing results, see e.g. (4.14) and (4.15) in [1], with $\Phi(t, T_1)x_*(T_1)$ and $\Phi(t, T_2)x^*(T_2)$ playing the roles of $x_*(t)$ and $x^*(t)$, respectively. Also notice that U and V act here as Q_* and Q^* . Relating quantities as above also exposes the similarity between (5.1), (5.3) and (5.4)—here (5.1) can be rewritten as $\hat{x}(t) = \Phi(t, T_1)x_*(T_1) + U(t)Z(t)[\Phi(t, T_2)x^*(T_2) - \Phi(t, T_1)x_*(T_1)]$ —and equations (4.13), (4.8) and (4.7) in [1] where Z corresponds to Q^{-1} . All of this is hardly surprising since the Hilbert space geometry of the two problems is quite similar.

Proof of Corollary 5.2. Integrating the downstar version of (3.4) between T_1 and t yields $U(t) = Q_*(t) + \int_{T_1}^t \Phi(t, s)B_*(s)B_*(s)'\Phi(t, s)' ds$. The controllability Gramian is positive definite because of total controllability [1, Lemma 2.2]. Thus $U(t) > 0$ for all $t \in (T_1, T_2)$. Similarly, integration of (3.4) on $[t, T_2]$ gives $V(t) = \Phi(t, T_2)[Q^*(T_2) + \int_t^{T_2} \Phi(T_2, s)B(s)B(s)'\Phi(T_2, s)' ds]\Phi(t, T_2)'$ from which the positive

definiteness of $V(t)$ follows. Given the analogy with the smoothing problem, it is clear that the rest can be proven by following the same lines as in [1, Corollary 4.1] or, alternatively, by employing a standard formula [3] for optimal weighting of two estimates with orthogonal errors. In fact, in view of (3.3), (3.11b) and (3.9), we have $E\{[x(t) - \Phi(t, T_1)x_*(T_1)][x(t) - \Phi(t, T_2)x_*(T_2)]'\} = P(t) - \Phi(t, T_1)P_*(T_1)\Phi(t, T_1)' + \Phi(t, T_1)P_*(T_1)\Phi(T_2, T_1)'\Phi(t, T_2)' - P(t) = 0$. \square

COROLLARY 5.3. *The interpolation estimate $\hat{x}(t)$ satisfies on $[T_1, T_2]$ the differential equation*

$$(5.9) \quad \frac{d\hat{x}}{dt} = A\hat{x} + BB'U^{-1}[\hat{x} - \Phi(t, T_1)x_*(T_1)]$$

with boundary condition $\hat{x}(T_2) = Q^*(T_2)\Pi^{-1}\Phi(T_2, T_1)x_*(T_1) + U(T_2)\Pi^{-1}x^*(T_2)$.

Proof. Multiplying $U(t)$ by $\Phi(T_2, t)'$ in (5.1) and then differentiating, we get

$$\begin{aligned} \frac{d\hat{x}}{dt} = & A(t)\Phi(t, T_1)x_*(T_1) + \left[\dot{P}(t)\Phi(T_2, t)' + P(t)\frac{\partial\Phi'}{\partial t}(T_2, t) \right. \\ & \left. - A(t)\Phi(t, T_1)P_*(T_1)\Phi(T_2, T_1)' \right] \\ & \cdot \Pi^{-1}[x^*(T_2) - \Phi(T_2, T_1)x_*(T_1)]. \end{aligned}$$

Now, using (3.4) and the adjoint equation for Φ , the above equation becomes

$$(5.10) \quad \frac{d\hat{x}}{dt} = A(t)\hat{x}(t) + B(t)B(t)'\Phi(T_2, t)'\Pi^{-1}[x^*(T_2) - \Phi(T_2, T_1)x_*(T_1)].$$

Since U is positive definite on (T_1, T_2) , we can solve (5.1) for $\Phi(T_2, t)'\Pi^{-1}[x^*(T_2) - \Phi(T_2, T_1)x_*(T_1)]$. Thus (5.10) reduces to (5.9). The boundary condition follows directly from (5.5). \square

Again we notice a similarity between (5.9) and a classical smoothing formula due to Rauch, Tung and Striebel [11]. It is worthwhile remembering, however, that a rigorous derivation of the Mayne-Frazer and Rauch, Tung and Striebel formulas requires further assumptions on the model (1.1), see [1, § 4] for details.

We now turn to the case when $t \notin (T_1, T_2)$. We shall need a preliminary result—Theorem 5.5 below—which is of interest in its own. Let us introduce $\Gamma^* := A - B^*R^{-1/2}C$.

LEMMA 5.4. *On the interval $[0, T]$ we have*

$$(5.11) \quad \bar{\Gamma}_* = (P^*)^{-1}\dot{P}^* - (P^*)^{-1}\Gamma^*P^*,$$

$$(5.12) \quad \bar{\Gamma}_* = (P^*)^{-1}\dot{P}^* - (P^*)^{-1}\dot{Q}Q^{-1}P^* + (P^*)^{-1}Q\Gamma_*'Q^{-1}P^*.$$

Proof. Since $\bar{B}_* = (P^*)^{-1}B^*$ and $G = P^*C' + B^*R^{1/2}$, we have that $\bar{\Gamma}_* = A' + (P^*)^{-1}B^*B^{*'} + (P^*)^{-1}B^*R^{-1/2}CP^*$. Now use (3.4) for the maximum variance realization to get (5.11). Next observe that because of the invariance of G over \mathcal{S} we have $\Gamma^* = \Gamma_* + QH'R^{-1}H$. It can be easily seen, for example from (3.19), that Q satisfies $\dot{Q} = \Gamma_*Q + Q\Gamma_*' + QH'R^{-1}HQ$. We conclude that

$$(5.13) \quad \Gamma^* = \dot{Q}Q^{-1} - Q\Gamma_*'Q^{-1}.$$

Inserting (5.13) into (5.11) yields (5.12). \square

Let Ψ^* be the transition matrix satisfying

$$(5.14) \quad \frac{\partial\Psi^*}{\partial t}(t, s) = \Gamma^*(t)\Psi^*(t, s), \quad \Psi^*(s, s) = I.$$

THEOREM 5.5. *Let s and t belong to $[0, T)$. Then*

$$(5.15) \quad \bar{\Psi}(s, t) = P^*(t)^{-1} \Psi^*(t, s) P^*(s),$$

$$(5.16) \quad \Psi^*(t, s) = Q(t) \Psi(s, t)' Q(s)^{-1},$$

$$(5.17) \quad \bar{\Psi}(s, t) = P^*(t)^{-1} Q(t) \Psi(s, t)' Q(s)^{-1} P^*(s).$$

Proof. Taking the partial derivative of the right-hand side of (5.15) with respect to s , we get $P^*(t)^{-1} \Psi^*(t, s)[- \Gamma^*(s) P^*(s) + \dot{P}^*(s)]$ which, in view of (5.11), is equal to $P^*(t)^{-1} \Psi^*(t, s) P^*(s) \bar{\Gamma}_*(s)$. By uniqueness of the solution to (3.16) we conclude that (5.15) must hold. Differentiating the right-hand side of (5.16) with respect to t , we get $[\dot{Q}(t) - Q(t) \Gamma_*(t)'] \Psi(s, t)' Q(s)^{-1}$ which is $\Gamma^*(t) Q(t) \Psi(s, t)' Q(s)^{-1}$ because of (5.13). Uniqueness in (5.14) now gives (5.16). Inserting (5.16) into (5.15), we get (5.17). \square

This result, together with Lemma 5.4, can be regarded as a generalization to the nonstationary setting of [14, Lemma 8], which is key to the classification of all solutions to the algebraic Riccati equation, and of a related result of stochastic realization theory [12, p. 53].

THEOREM 5.6. *Let $t \in [0, T_1]$. The interpolation estimate $\hat{x}(t)$ is given by*

$$(5.18) \quad \begin{aligned} \hat{x}(t) = & x_*(t) + Q_*(t) \bar{z}(t; T_1) \\ & + Q_*(t) \Psi(T_1, t)' \Phi(T_2, T_1)' \Pi^{-1} [x^*(T_2) - \Phi(T_2, T_1) x_*(T_1)] \end{aligned}$$

where $\bar{z}(\cdot; T_1)$ is defined as the solution to equation (3.18) on $[0, T_1]$ with initial condition $\bar{z}(T_1; T_1) = 0$. The error variance matrix is given by

$$(5.19) \quad \begin{aligned} \Sigma(t) = & Q_*(t) - Q_*(t) [Q(t)^{-1} - \Psi(T_1, t)'] \\ & \cdot (Q(T_1)^{-1} - \Phi(T_2, T_1)' \Pi^{-1} \Phi(T_2, T_1)) \Psi(T_1, t) Q_*(t). \end{aligned}$$

Proof. First observe that $\hat{x}(t) = E\{\hat{x}_s(t) | H(T_1, T_2)\} = Q^*(t) Q(t)^{-1} x_*(t) + Q_*(t) Q(t)^{-1} P^*(t) E\{\bar{x}_*(t) | H(T_1, T_2)\}$ because of (4.7) and (3.3). Next notice that, in view of (3.15), the components of $\bar{x}_*(t) - \bar{\Psi}(T_1, t) \bar{x}_*(T_1)$ belong to $H(T_1, T_2)$. Thus

$$(5.20) \quad \begin{aligned} \hat{x}(t) = & Q^*(t) Q(t)^{-1} x_*(t) + Q_*(t) Q(t)^{-1} [x^*(t) - P^*(t) \bar{\Psi}(T_1, t) P^*(T_1)^{-1} x^*(T_1)] \\ & + Q_*(t) Q(t)^{-1} P^*(t) \bar{\Psi}(T_1, t) E\{\bar{x}_*(T_1) | H(T_1, T_2)\}, \end{aligned}$$

where we have used (3.11b) twice. It should be clear that one can argue along the same lines as in Theorem 5.1 that $x_*(T_1)$ and $\bar{x}_*(T_2)$ represent a “sufficient statistic” for the projection $E\{\bar{x}_*(T_1) | H(T_1, T_2)\}$. By Lemma 4.3 the latter then becomes

$$(5.21) \quad \begin{aligned} E\{\bar{x}(T_1) | H(T_1, T_2)\} \\ = P^*(T_1)^{-1} [x_*(T_1) + Q(T_1) \Phi(T_2, T_1)' \Pi^{-1} (x^*(T_2) - \Phi(T_2, T_1) x_*(T_1))]. \end{aligned}$$

Inserting (5.21) into (5.20) and taking (5.17) into account, we get

$$\begin{aligned} \hat{x}(t) = & x_*(t) + Q_*(t) [\bar{z}(t) - \Psi(T_1, t)' \bar{z}(T_1)] \\ & + Q_*(t) \Psi(T_1, t)' \Phi(T_2, T_1)' \Pi^{-1} [x^*(T_2) - \Phi(T_2, T_1) x_*(T_1)]. \end{aligned}$$

Expression (5.18) now follows from the fact that $\bar{z}(t) - \Psi(T_1, t)' \bar{z}(T_1)$ satisfies (3.18) on $[0, T_1]$ and is zero at $t = T_1$. The expression (5.19) for Σ finally follows from the usual orthogonality condition. \square

COROLLARY 5.7. *Let $t \in [0, T_1]$ and $\hat{x}_s(t; T_1) := E\{x(t) | H_{T_1}^-(dy)\}$. Then*

$$(5.22) \quad \hat{x}(t) = \hat{x}_s(t; T_1) + Q_*(t) \Psi(T_1, t)' \Phi(T_2, T_1) \Pi^{-1} [x^*(T_2) - \Phi(T_2, T_1) x_*(T_1)].$$

Proof. The result is an immediate consequence of (5.18) and of [1, Thm 4.2]. \square

The usefulness of the decomposition (5.18) is now apparent. Not only are the three terms on the right-hand side pairwise orthogonal—corresponding to the decomposition $H(T_1, T_2) = H_t^-(dy) \oplus H_{[t, T_1]}(dw_*) \oplus [H(T_1, T_2) \cap (H(dy) \ominus H_{T_1}^-(dy))]$ with the obvious significance of the symbols—but we also have the orthogonal decomposition (5.22). Hence, in the case that the smoothing estimate $\hat{x}_s(t; T_1)$ has already been computed and the new information $H_{T_2}^+(dy)$ becomes available, the estimate can be optimally updated according to (5.22).

Let us introduce $\bar{Q} := P_*^{-1} - (P^*)^{-1}$, $\bar{Q}_* := P^{-1} - (P^*)^{-1}$, $\bar{x}^* := P_*^{-1}x_*$ and $\bar{z}_b(\cdot; T_2)$ as the solution on $[T_2, T]$ to the equation

$$(5.23) \quad d\bar{z}_b = \bar{\Gamma}'_* \bar{z}_b dt - GR^{-1/2} d\bar{w}_*, \quad \bar{z}_b(T_2; T_2) = 0.$$

It can be shown, as done for $\bar{z}(t; T_1)$ in Theorem 5.6, that the variance \bar{Q}_b of \bar{z}_b is given by

$$(5.24) \quad \bar{Q}_b(t) = \bar{Q}(t)^{-1} - \bar{\Psi}(t, T_2)' \bar{Q}^{-1} \bar{\Psi}(t, T_2).$$

THEOREM 5.8. Let $t \in [T_2, T]$. The interpolation estimate $\hat{x}(t)$ is given by

$$(5.25) \quad \begin{aligned} \hat{x}(t) = P(t)[\bar{x}_*(t) + \bar{Q}_*(t)\bar{z}_b(t; T_2) + \bar{Q}_*(t)\bar{\Psi}(t, T_2)'\Phi(T_2, T_1) \\ \times \bar{\Pi}^{-1}[\bar{x}^*(T_1) - \Phi(T_2, T_1)'\bar{x}_*(T_2)]] \end{aligned}$$

where \bar{z}_b satisfies (5.23). The error variance matrix is given by

$$(5.26) \quad \begin{aligned} \Sigma(t) = P(t)[\bar{Q}_*(t) - \bar{Q}_*(t)[\bar{Q}(t)^{-1} - \bar{\Psi}(t, T_2)'(\bar{Q}(t)^{-1} \\ - \Phi(T_2, T_1)\bar{\Pi}^{-1}\Phi(T_2, T_1)')\bar{\Psi}(t, T_2)]\bar{Q}_*(t)]P(t). \end{aligned}$$

Proof. Following exactly the same argument as in [1, §4], but employing backward quantities, one can readily establish the result $\hat{x}_s(t) = P(t)[\bar{Q}^*(t)\bar{Q}(t)^{-1}\bar{x}_*(t) + \bar{Q}_*(t)\bar{Q}(t)^{-1}\bar{x}^*(t)]$ where $\bar{Q}^* := P_*^{-1} - P^{-1}$. The proof is now completely analogous to that of Theorem 5.6. \square

Although the structure of (5.25) is the same as (5.18), this representation does not appear to be as useful as the previous one except in the case when the “old” information $H_{T_2}^-(dy)$ becomes available after the smoothing estimate corresponding to $H_{T_2}^+(dy)$, represented here by $P(t)[\bar{x}_*(t) + \bar{Q}_*(t)\bar{z}_b(t; T_2)]$, has been computed. The following expression, phrased in terms of forward quantities, can alternatively be employed.

PROPOSITION 5.9. Let $t \in [T_2, T]$. The interpolation estimate $\hat{x}(t)$ is given by

$$(5.27) \quad \begin{aligned} \hat{x}(t) = Q^*(t)Q(t)^{-1}\Psi(t, T_2)Q(T_2)\Pi^{-1}\Phi(T_2, T_1)x_*(T_1) \\ + \int_{T_2}^t Q^*(t)Q(t)^{-1}\Psi(t, s)B_*(s)R(s)^{-1/2} dy + Q_*(t)Q(t)^{-1}x^*(t) \\ + Q^*(t)Q(t)^{-1}\Psi(t, T_2)(I - Q(T_2)\Pi^{-1})x^*(T_2). \end{aligned}$$

Proof. By (4.7) we have $\hat{x}(t) = Q_*(t)Q(t)^{-1}x^*(t) + E\{Q^*(t)Q(t)^{-1}x_*(t)|H(T_1, T_2)\}$. Observing that the components of $x_*(t) - \Psi(t, T_2)x_*(T_2)$ belong to $H(T_1, T_2)$ and that $x_*(T_1)$ and $\bar{x}_*(T_2)$ represent a sufficient statistic for the projection $E\{x_*(T_2)|H(T_1, T_2)\}$, we get

$$\begin{aligned} \hat{x}(t) = Q_*(t)Q(t)^{-1}x^*(t) + Q^*(t)Q(t)^{-1}[x_*(t) - \Psi(t, T_2)x_*(T_2)] \\ + Q^*(t)Q(t)^{-1}\Psi(t, T_2)[\Phi(T_2, T_1)x_*(T_1) \\ + (I - Q(T_2)\Pi^{-1})(x^*(T_2) - \Phi(T_2, T_1)x_*(T_1))]. \end{aligned}$$

Taking (3.14) into account, we get (5.27). \square

The following result is a corollary to Theorems 5.6 and 5.8.

PROPOSITION 5.10. *The interpolation estimate $\hat{x}(t)$ satisfies on $[0, T_1]$ the stochastic differential equation*

$$(5.28) \quad d\hat{x} = A\hat{x} dt + B(I - D'R^{-1}D)B'\xi dt + BD'R^{-1/2}(dy - C\hat{x} dt),$$

with initial condition $\hat{x}(T_1) = x_*(T_1) + Q_*(T_1)\Phi(T_2, T_1)'\Pi^{-1}(x^*(T_2) - \Phi(T_2, T_1)x_*(T_1))$ and on $[T_2, T]$ the stochastic differential equation

$$(5.29) \quad d\hat{x} = (A' + BB'P^{-1})\hat{x} dt + \bar{B}(I - D'R^{-1}D)\bar{B}'\bar{\xi} dt + \bar{B}D'R^{-1/2}(dy - G'P^{-1}\hat{x} dt)$$

with initial condition $\hat{x}(T_2) = x^*(T_2) - Q^*(T_2)\Pi^{-1}[x^*(T_2) - \Phi(T_2, T_1)x_*(T_1)]$. The vector processes ξ and $\bar{\xi}$ are given by

$$\xi(t) := \bar{z}(t; T_1) + \Psi(T_1, t)'\Phi(T_2, T_1)'\Pi^{-1}[x^*(T_2) - \Phi(T_2, T_1)x_*(T_1)],$$

$$\bar{\xi}(t) := \bar{z}_b(t; T_2) + \bar{\Psi}(t, T_2)'\Phi(T_2, T_1)\bar{\Pi}^{-1}[\bar{x}^*(T_1) - \Phi(T_2, T_1)'\bar{x}_*(T_2)].$$

Proof. Differentiating (5.22), we get

$$\begin{aligned} d\hat{x}(t) &= d\hat{x}_s(t; T_1) + [\dot{Q}_*(t) - \Gamma_*(t)'Q_*(t)]\Psi(T_1, t)' \\ &\quad \times \Phi(T_2, T_1)'\Pi^{-1}[x^*(T_2) - \Phi(T_2, T_1)x_*(T_1)]. \end{aligned}$$

Rewriting (3.2c), we see that $\dot{Q}_* - \Gamma_*'Q_* = \Gamma_*'Q_* + Q_*C'R^{-1}CQ_* + B(I - DR^{-1}D')B'$. By (3.2a) we also have $\Gamma_* + Q_*C'R^{-1}C = A - BD'R^{-1/2}$. Taking (5.22) and [1, Corollary 4.2] into account, we conclude that (5.28) must hold. In a completely dual way one can argue that $\hat{x} := P^{-1}\hat{x}$ satisfies $d\hat{x} = -A'\hat{x} dt + \bar{B}(I - D'R^{-1}D)\bar{B}'\bar{\xi}(t) + \bar{B}D'R^{-1/2}(dy - G'\hat{x} dt)$ from which (5.29) quickly follows using (3.4). \square

Once more the analogy with the smoothing problem [1, Corollary 4.2] is apparent and it carries over to some special cases. First let $BD' = 0$, which corresponds to the standard assumption in the filtering literature of uncorrected state and observation noises. Then \hat{x} is differentiable and (5.28) reads

$$(5.30) \quad \frac{d\hat{x}}{dt} = A\hat{x} + BB'\xi.$$

Under the extra assumption that $Q_*(t) > 0$, (5.30) reduces to

$$(5.31) \quad \frac{d\hat{x}}{dt} = A\hat{x} + BB'Q_*^{-1}(\hat{x} - x_*).$$

Secondly, suppose that D is a square matrix. Then $D'R^{-1}D = I$ and (5.28) reduces to

$$(5.32) \quad d\hat{x} = A\hat{x} dt + BD'R^{-1/2}(dy - C\hat{x} dt).$$

Equation (5.30) should be compared to (4.30) in [1], whereas (5.31) and (5.32) coincide with (4.30) and (4.31) in the same paper except for the initial conditions which differ by the amount $Q_*(T_1)\Phi(T_2, T_1)'\Pi^{-1}[x^*(T_2) - \Phi(T_2, T_1)x_*(T_1)]$. Similarly, (5.29) reduces to simpler equations in the cases considered above.

6. Closing comments. Theorems 5.1, 5.6 and 5.8, together with Corollary 5.2, Corollary 5.3, Proposition 5.9 and Proposition 5.10, furnish several procedures to calculate $\hat{x}(t)$, $t \in [0, T]$. It is apparent from these results that, as expected, \hat{x} is continuous on $[0, T]$ in $L^2_\pi(\Omega, \mathcal{F}, P)$ and that it converges to the smoothing estimate as $|T_1 - T_2|$ tends to zero. It also converges to the prediction estimate for $t \in (T_1, T)$ and to $\hat{x}_s(t; T_1)$ for $t \in [0, T_1]$ as T_2 tends to T .

Finally we like to remark that the expressions obtained for \hat{x} are easy to update as T increases. Indeed, as new information is acquired, it suffices to change x^* and related quantities in the formulas. This should be done considering how the backward filter changes. For instance, if $\bar{x}_*(t; T+h)$ denotes such a filter for the interval $[0, T+h]$, we simply have $\bar{x}_*(t; T+h) = \bar{x}_*(t) + \int_T^{T+h} \Phi(s, t)' \bar{B}_*(s) d\bar{w}_*$. In particular, if $t \in [0, T_2]$, only $x^*(T_2)$ and $P^*(T_2)$ need to be changed in (5.18) and (5.1).

REFERENCES

- [1] F. BADAWI, A. LINDQUIST AND M. PAVON, *A stochastic realization approach to the smoothing problem*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 878–888.
- [2] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
- [3] R. DEUTSCH, *Estimation Theory*, Prentice-Hall, Englewood, Cliffs, NJ, 1965.
- [4] P. FAURRE, M. CLERGET AND F. GERMAIN, *Opérateurs rationnels positifs*, Dunod, Paris, 1979.
- [5] I. I. GIKHMAN AND A. V. SKOROKHOD, *Introduction to the Theory of Random Processes*, Saunders, Philadelphia, 1969.
- [6] A. LINDQUIST, G. PICCI AND G. RUCKEBUSCH, *On splitting subspaces and Markovian representations*, Math. Systems Theory, 12 (1979), pp. 271–279.
- [7] A. LINDQUIST, M. PAVON AND G. PICCI, *Recent trends in stochastic realization theory*, in Harmonic Analysis and Prediction Theory—the Masani Volume, V. Mandrekar and H. Salehi, eds., North-Holland, Amsterdam, 1983, pp. 201–224.
- [8] J. NEVEU, *Discrete-Parameter Martingales*, North-Holland, Amsterdam, 1975.
- [9] M. PAVON, *A new algorithm for optimal interpolation of discrete-time stationary processes*, in Analysis and Optimization of Systems, Lecture Notes in Control and Information Sciences 44, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, New York, 1982, pp. 701–718.
- [10] —, *New results on the interpolation problem for continuous-time stationary-increments processes*, this Journal, 22 (1984), pp. 133–142.
- [11] H. E. RAUCH, F. TUNG AND C. T. STRIEBEL, *Maximum likelihood estimates of linear dynamic systems*, AIAA, J., 3 (1965), pp. 1445–1450.
- [12] G. RUCKEBUSCH, *Représentations Markoviennes de processus Gaussiens stationnaires*, Thèse de 3ème cycle, Université de Paris VI, May 1975.
- [13] —, *Théorie géométrique de la représentation Markovienne*, Ann. Inst. H. Poincaré, 3 (1980), pp. 225–297.
- [14] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 621–634.

STATE DEADBEAT RESPONSE AND OBSERVABILITY IN MULTI-MODAL SYSTEMS*

LUTHER T. CONNER, JR.† AND DAVID P. STANFORD†

Abstract. This paper deals with two aspects of multi-modal systems. First we show that any completely controllable multi-modal system, with state dimension n not exceeding 3, is capable, through feedback, of state deadbeat response. We conjecture that the result holds for all n , as is the case for the classical single-mode system.

Certain properties of multi-modal systems indicate that they differ significantly from the single-mode systems. For example, the controllable set is not in general a subspace, and furthermore, the number of steps necessary to reach all states in the controllable set is not bounded by the state dimension. In this paper, we obtain bounds for this number in the case of a completely controllable system with $n \leq 3$, and use them to establish state deadbeat response.

The second portion of this paper refines the controllability canonical form for a multi-modal system. This is accomplished through the introduction of a notion of observability, dual to controllability for these systems.

Key words. multi-modal system, controllability, state deadbeat response, observability, controllability canonical form

Introduction. This paper discusses linear discrete-time systems of the form

$$x_{k+1} = C_p x_k + D_p u_k,$$

where the pair (C_p, D_p) is selected from a finite set $\{(C_i, D_i)\}_{i=1}^N$ of pairs, with C_i a real $n \times n$ matrix and D_i a real $n \times m$ matrix. We will refer to such a system as a *multi-modal system*. In contrast to the usual time-varying discrete-time system $x_{k+1} = C_k x_k + D_k u_k$, the pair (C_p, D_p) employed at time k is not, in general, determined by k or dependent upon the control u_k selected. In fact, the basic problem studied in this paper, and in the authors' previous papers on this system, can be stated as follows. Given a system $L = \{(C_i, D_i)\}_{i=1}^N$, what effects can be produced by various choices of control laws for the system? We assume feedback controls of the form $u_k = F_p x_k$, so that the choice of the pair (C_p, D_p) at time k also determines the feedback matrix F_p . Thus, a control law would consist of a scheme for selecting an index p for each time k .

Multi-modal systems arise naturally in the study of multi-rate sampled-data systems (see [3]), and may have applications in variable structure systems and in switched capacitor circuits.

The stabilizability of multi-modal systems through feedback has been investigated in [1]. Pre-contractiveness and contractiveness of the closed-loop system are introduced and the selection of feedbacks is discussed. In [3] and [4], the concept of controllability is extended to these systems, and the set of points reachable from zero (the "controllable set") is investigated. It is shown that the controllable set is a subspace under certain hypotheses, but not always. When this is the case, an extended version of the controllability canonical form is obtained, and it is applied to the study of state deadbeat response and more general forms of stabilizability.

This paper deals with two aspects of multi-modal systems. First, we place in the standard literature results which previously appeared only in technical reports, but which are referred to extensively in [3]. We show that any completely controllable

* Received by the editors April 1, 1981, and in revised form March 15, 1983. This research was supported by NASA-Langley Research Center under grants NAS1-14972 and NAS1-16042.

† Department of Mathematics and Computer Science, College of William and Mary, Williamsburg, Virginia 23185.

system, with $n \leq 3$, allows a choice of feedback matrices resulting in a state deadbeat response. Some parts of our work are valid for arbitrary n , and we conjecture that for all n the state deadbeat response can be obtained under the hypothesis of complete controllability. Although this fact follows easily from the pole-placement theorem for a single-node system, we have been unable to prove it for general n in a multi-modal system. Our work, which includes the determination of a bound on the number of iterations necessary to reach an arbitrary state, is very cumbersome. The fact that the bound is in general greater than n indicates the need for an entirely different approach to the problem from that used in the single-mode case.

The second portion of the paper refines the controllability canonical form for a multi-modal system (see [3]). This is accomplished through the introduction of a notion of observability, dual to controllability, for these systems.

2. Controllability in multi-modal systems. We list here for convenience the definitions, notation and results from [3] which will be used in this paper. Throughout the paper, n , m and N denote positive integers, with $m < n$.

A *multi-modal system* is an indexed set of pairs $L = \{(C_i, D_i)\}_{i=1}^N$, in which C_i and D_i are real $n \times n$ and $n \times m$ matrices respectively. L represents a discrete-time control system of the form $x_{k+1} = C_p x_k + D_p u_k$, as described in the introduction. We will denote by

$$\Sigma^0(n, m, N)$$

the set of all such systems L . Those systems in which all C_i 's are nonsingular and all D_i 's are of full column rank are of particular importance in our work, and this subset of $\Sigma^0(n, m, N)$ will be denoted by

$$\Sigma(n, m, N).$$

Multi-modal systems arising from multi-rate sampled-data systems belong to $\Sigma(n, m, N)$.

For each positive integer k we let $\bar{k} = \{1, 2, \dots, k\}$. A control law for a system L in $\Sigma^0(n, m, N)$ consists of a selection of $m \times n$ feedback matrices $\{F_i\}_{i=1}^N$ and a sequence of indices from \bar{N} . In this paper we will be concerned with finite sequences of indices, and so we let Γ_k denote the collection of all k -termed sequences with terms in \bar{N} .

For any $L \in \Sigma^0(n, m, N)$ and any index-sequence $\gamma \in \bigcup_k \Gamma_k$, we denote by $S(L, \gamma)$ the set of states reachable from 0 using γ ; that is,

$S(L, \gamma)$ contains x provided there is a sequence $\{u_i\}_{i=1}^N$ from R^m such that $x_{k+1} = x$, where $x_1 = 0$ and $x_{i+1} = C_{\gamma(i)} x_i + D_{\gamma(i)} u_i$, $i \in \bar{k}$.

We let $S(L)$ denote the set of all states reachable from 0; that is

$$S(L) = \bigcup_k \bigcup_{\gamma \in \Gamma_k} S(L, \gamma).$$

For $L = \{(C_i, D_i)\}_{i=1}^N \in \Sigma^0(n, m, N)$ and $\gamma \in \Gamma_k$, $S(L, \gamma)$ can be described as the column space of a controllability matrix as follows. Given $i, j \in \bar{k}$ with $i \leq j$, we define

$$C(\gamma, j, i) = C_{\gamma(j)} C_{\gamma(j-1)} \cdots C_{\gamma(i)}.$$

$C(\gamma, i)$ will denote $C(\gamma, k, i)$, $C(\gamma; i)$ will denote $C(\gamma, i, 1)$, and $C(\gamma)$ will denote $C(\gamma, k, 1)$. We define the $n \times km$ controllability matrix

$$P(L, \gamma) = [D_{\gamma(k)}, C(\gamma, k)D_{\gamma(k-1)}, C(\gamma, k-1)D_{\gamma(k-2)}, \dots, C(\gamma, 2)D_{\gamma(1)}].$$

Then $S(L, \gamma) = CS(P(L, \gamma))$, where CS denotes column space, so that $S(L) = \bigcup_k \bigcup_{\gamma \in \Gamma_k} CS(P(L, \gamma))$.

The set of reachable states $S(L)$ is not, in general, a subspace of R^n (see the example in [3]). If, however, each C_i is nonsingular (as is the case when $L \in \Sigma(n, m, N)$), then $S(L)$ is a subspace of R^n . Furthermore, whenever $S(L)$ is a subspace, there is a finite index-sequence γ such that $S(L) = S(L, \gamma)$.

A system $L \in \Sigma^0(n, m, N)$ is *completely controllable* provided $S(L) = R^n$. It follows that L is completely controllable if and only if there is a finite index-sequence γ such that $\text{rank}(P(L, \gamma)) = n$.

3. Bounds on the minimal length of a controlling index-sequence. In this section we derive bounds for the number of steps required to control a completely controllable multi-modal system in which $n \leq 3$, and we provide some information for general n . The problem of determining whether or not a multi-modal system is completely controllable is made finite by these bounds. Furthermore, these bounds will be used to establish our results on state deadbeat response.

We will employ the following notation. If $L \in \Sigma^0(n, m, N)$, then $b(L)$ is the smallest number of steps which is sufficient to reach any reachable state. If no such number exists, then $b(L) = +\infty$. More precisely,

$$b(L) = \min \{k | x \in S(L) \text{ implies there is } i \in \bar{k} \text{ and } \gamma \in \Gamma_i \text{ such that } x \in S(L, \gamma)\}.$$

It is not known whether $L \in \Sigma^0(n, m, N)$ implies $b(L)$ finite. However, it is shown in [3], that $b(L)$ is finite if and only if $S(L)$ is the union of finitely many subspaces of R^n .

Restricting our attention to completely controllable systems in $\Sigma(n, m, N)$, we will denote by $B(n, m, N)$ the maximum value of $b(L)$ overall completely controllable $L \in \Sigma(n, m, N)$. Thus for any completely controllable system in $\Sigma(n, m, N)$, any state can be reached in $B(n, m, N)$ or fewer steps. Thus for the classical system ($N = 1$), we have $B(n, m, 1) \leq n$.

Many of the arguments in this section and the next require knowledge of the rank of the matrix $[D_1, D_2, \dots, D_N]$ formed from $L = \{(C_i, D_i)\}_{i=1}^N$. We define, therefore, for $m \leq p \leq n$,

$$\Sigma_p(n, m, N) = \{L \in \Sigma(n, m, N) | \text{rank}([D_1, D_2, \dots, D_N]) = p\},$$

$$B_p(n, m, N) = \max \{b(L) | L \in \Sigma_p(n, m, N) \text{ is completely controllable}\}.$$

THEOREM 1. *If $L \in \Sigma_n(n, m, N)$, then L is completely controllable and $b(L) \leq n - m + 1$. Thus $B_n(n, m, N) \leq n - m + 1$.*

Proof. A controlling index-sequence will be constructed inductively. Let $\beta_1 = 1$, so that

$$m = \text{rank}(D_{\beta_1}) < n.$$

Suppose $\beta_1, \beta_2, \dots, \beta_r$ have been chosen so that

$$m + r - 1 \leq \text{rank}([D_{\beta_1}, C_{\beta_1}D_{\beta_2}, \dots, C_{\beta_1}C_{\beta_2} \dots C_{\beta_{r-1}}D_{\beta_r}]) < n.$$

Since $\text{rank}(C_{\beta_1}C_{\beta_2} \dots C_{\beta_r}[D_1, D_2, \dots, D_N]) = n$, there is a β_{r+1} such that

$$m + r \leq \text{rank}([D_{\beta_1}, C_{\beta_1}D_{\beta_2}, \dots, C_{\beta_1}C_{\beta_2} \dots C_{\beta_r}D_{\beta_{r+1}}]) \leq n.$$

Clearly, for some $k \leq n - m + 1$,

$$\text{rank}([D_{\beta_1}, C_{\beta_1}D_{\beta_2}, \dots, C_{\beta_1}C_{\beta_2} \dots C_{\beta_{k-1}}D_{\beta_k}]) = n,$$

and so $\gamma = (\beta_k, \beta_{k-1}, \dots, \beta_1) \in \Gamma_k$ is a controlling index-sequence. \square

THEOREM 2. *If $L \in \Sigma(n, n-1, N)$ is completely controllable, then $b(L) = 2$, and thus $B(n, n-1, N) = 2$.*

Proof. Let $L = \{(C_i, D_i)\}_{i=1}^N \in \Sigma(n, n-1, N)$ be completely controllable. Certainly $b(L) \neq 1$. Assume $b(L) > 2$. Then for all $\beta = (i, j) \in \Gamma_2$,

$$\text{rank}(P(L, \beta)) = \text{rank}([D_j, C_j D_i]) = n-1,$$

which implies $CS(C_j D_i) = CS(D_j)$. It follows that, for any $\gamma \in \Gamma_k$ with $k > 2$, $CS(C(\gamma, 2) D_{\gamma(1)}) = CS(D_{\gamma(k)})$, and so $\text{rank}(P(L, \gamma)) = n-1$, which contradicts the complete controllability of L . Hence $b(L) = 2$. \square

Since we are concerned here primarily with $n \leq 3$, we observe that, in particular, $B(2, 1, N) = 2$ and $B(3, 2, N) = 2$.

LEMMA 1. *If $L \in \Sigma(3, 1, N)$ is completely controllable, $b(L) = k$, and $\gamma \in \Gamma_k$ with $\text{rank}(P(L, \gamma)) = 3$, then all interior columns of $P(L, \gamma)$ are multiples of some fixed nonzero vector.*

Proof. Deleting the last column of $P(L, \gamma)$ results in $P(L, \sigma)$, where $\sigma \in \Gamma_{k-1}$. Alternatively, deleting the first column of $P(L, \gamma)$ results in $C_{\gamma(k)} P(L, \tau)$, where $\tau \in \Gamma_{k-1}$. Since each of $P(L, \sigma)$ and $C_{\gamma(k)} P(L, \tau)$ has rank less than 3, the lemma follows. \square

THEOREM 3. $B(3, 1, N) = 4$ for $N \geq 2$.

Proof. Theorem 1 implies $B_3(3, 1, N) = 3$.

We next prove that $B_2(3, 1, N) = 4$ for $N \geq 2$.

Let $L = \{(C_i, d_i)\}_{i=1}^N \in \Sigma_2(3, 1, N)$ be completely controllable with $N \geq 2$. Assume $b(L) = k \geq 5$. Then for each $\sigma \in \Gamma_t$ with $t < k$, $\text{rank}(P(L, \sigma)) < 3$. Let $\gamma \in \Gamma_k$ with

$$\text{rank}(P(L, \gamma)) = \text{rank}([d_{\gamma(k)}, C_{\gamma(k)} d_{\gamma(k-1)}, \dots, C_{\gamma(k)} \cdots C_{\gamma(2)} d_{\gamma(1)}]) = 3.$$

It follows, using Lemma 1, that

- (1) $C_{\gamma(i)} d_{\gamma(i-1)} \neq \lambda d_{\gamma(i)}$ for any $\lambda \in R$, for $i = k$ and $i = 2$;
- (2) $C_{\gamma(i)} d_{\gamma(i-1)} = \alpha_i d_{\gamma(i)}$ for some nonzero $\alpha_i \in R$, $2 < i < k$;
- (3) $C_{\gamma(k)} C_{\gamma(k-1)} \cdots C_{\gamma(2)} d_{\gamma(1)} \notin T = \text{span}\{C_{\gamma(k)} d_{\gamma(k-1)}, d_{\gamma(k)}\}$.

We will obtain a contradiction to (3), thus proving that $b(L) \leq 4$. Let $S = \text{span}\{d_1, d_2, \dots, d_N\}$. We will prove that:

- a) S is invariant under $C_{\gamma(i)}$, $2 < i < k$;
- b) $C_{\gamma(2)} d_{\gamma(1)} \in S$;
- c) $C_{\gamma(k)} C_{\gamma(k-1)}$ maps S into T .

Once these have been established, it follows that $C_{\gamma(k)} C_{\gamma(k-1)} \cdots C_{\gamma(2)} d_{\gamma(1)} \in T$, contradicting (3).

To prove a), let $2 < i < k$ and select $p \in \bar{N}$ so that $\text{rank}([d_{\gamma(i-1)}, d_p]) = 2$. Then using (2), $\text{rank}([d_{\gamma(i)}, C_{\gamma(i)} d_p]) = 2$.

For $j \in \bar{N}$, since $b(L) \geq 5$, $\text{rank}([d_{\gamma(i)}, C_{\gamma(i)} d_p, C_{\gamma(i)} C_p d_j]) = 2$, which implies $C_{\gamma(i)} C_p d_j = \mu_1 C_{\gamma(i)} d_p + \mu_2 d_{\gamma(i)}$ for some $\mu_1, \mu_2 \in R$. Using (2), $C_p d_j = \mu_1 d_p + \mu_2 \alpha_i^{-1} d_{\gamma(i-1)} \in S$. Hence, S is invariant under C_p and so S is also invariant under C_p^{-1} .

Again for $j \in \bar{N}$

$$\text{rank}([d_{\gamma(i)}, C_{\gamma(i)} d_p, C_{\gamma(i)} C_p d_{\gamma(i)}, C_{\gamma(i)} C_p C_{\gamma(i)} d_j]) = 2,$$

and so $C_{\gamma(i)} C_p C_{\gamma(i)} d_j = \mu_3 C_{\gamma(i)} d_p + \mu_4 d_{\gamma(i)}$. Using (2), $C_p C_{\gamma(i)} d_j = \mu_3 d_p + \mu_4 \alpha_i^{-1} d_{\gamma(i-1)} \in S$. Since S is invariant under C_p^{-1} , $C_{\gamma(i)} d_j \in S$, and so a) is proved.

We next prove b). First suppose there is an i , $2 < i < k$, with $\text{rank}([d_{\gamma(i)}, d_{\gamma(2)}]) = 2$. Then using (2), we see that $d_{\gamma(3)}$ and $C_{\gamma(3)} d_{\gamma(i)}$ are linearly independent. Since $b(L) \geq 5$,

$$\text{rank}([d_{\gamma(3)}, C_{\gamma(3)} d_{\gamma(i)}, C_{\gamma(3)} C_{\gamma(i)} d_{\gamma(2)}, C_{\gamma(3)} C_{\gamma(i)} C_{\gamma(2)} d_{\gamma(1)}]) = 2,$$

and so $C_{\gamma(3)}C_{\gamma(i)}C_{\gamma(2)}d_{\gamma(1)} = \mu_5 C_{\gamma(3)}d_{\gamma(i)} + \mu_6 d_{\gamma(3)}$. Thus $C_{\gamma(i)}C_{\gamma(2)}d_{\gamma(1)} = \mu_5 d_{\gamma(i)} + \mu_6 \alpha_3^{-1} d_{\gamma(2)} \in S$. By a), S is invariant under $C_{\gamma(i)}^{-1}$, and so $C_{\gamma(2)}d_{\gamma(1)} \in S$.

Next, suppose $\text{rank}([d_{\gamma(i)}, d_{\gamma(2)}]) = 1$ for $2 < i < k$. Then each of $d_{\gamma(2)}$ and $d_{\gamma(3)}$ is a multiple of $d_{\gamma(k-1)}$, and, using (1), $\text{rank}([d_{\gamma(k)}, C_{\gamma(k)}d_{\gamma(i)}]) = 2$ for $i = 2, 3$. Select q so that $\text{rank}([d_q, d_{\gamma(2)}]) = 2$. Then $\text{rank}([d_{\gamma(k)}, C_{\gamma(k)}d_{\gamma(3)}, C_{\gamma(k)}C_{\gamma(3)}d_q]) = 2$, and so $C_{\gamma(k)}C_{\gamma(3)}d_q = \mu_7 C_{\gamma(k)}d_{\gamma(3)} + \mu_8 d_{\gamma(k)}$ with $\mu_8 \neq 0$. Hence, $d_q = \mu_7 \alpha_3^{-1} d_{\gamma(2)} + \mu_8 C_{\gamma(3)}^{-1} C_{\gamma(k)}^{-1} d_{\gamma(k)}$, and so $C_{\gamma(3)}^{-1} C_{\gamma(k)}^{-1} d_{\gamma(k)} \in S$. Since S is invariant under $C_{\gamma(3)}$, $C_{\gamma(k)}^{-1} d_{\gamma(k)} \in S$. Now $\text{rank}([d_{\gamma(k)}, C_{\gamma(k)}d_{\gamma(2)}, C_{\gamma(k)}C_{\gamma(2)}d_{\gamma(1)}]) = 2$, and so $C_{\gamma(k)}C_{\gamma(2)}d_{\gamma(1)} = \mu_9 C_{\gamma(k)}d_{\gamma(2)} + \mu_{10} d_{\gamma(k)}$. Thus $C_{\gamma(2)}d_{\gamma(1)} = \mu_9 d_{\gamma(2)} + \mu_{10} C_{\gamma(k)}^{-1} d_{\gamma(k)} \in S$, and b) is established.

For $j \in \bar{N}$, $\text{rank}([d_{\gamma(k)}, C_{\gamma(k)}d_{\gamma(k-1)}, C_{\gamma(k)}C_{\gamma(k-1)}d_j]) = 2$, and so $C_{\gamma(k)}C_{\gamma(k-1)}d_j \in T$. This proves c), so $b(L) \leq 4$ and thus $B_2(3, 1, N) \leq 4$.

On the other hand, let $L = \{(C_i, d_i)\}_{i=1}^N \in \Sigma_2(3, 1, N)$ be defined by

$$C_1 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad d_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

$$C_i = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad d_i = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad i = 2, 3, \dots, N.$$

Computation shows that $\text{rank}(P(L, \sigma)) < 3$ for $\sigma \in \Gamma_3$ and $\text{rank}(P(L, \gamma)) = 3$ for $\gamma = (1, 2, 1, 1)$. Thus L is completely controllable and $b(L) = 4$. Hence $B_2(3, 1, N) = 4$.

It can be shown in a similar manner (see [2]) that $B_1(3, 1, N) = 4$ for $N \geq 2$, and the theorem follows. \square

The following discussion shows that $B(n, 1, 2) \geq 2n - 3$ for $4 \leq n \leq 8$. It is trivial to extend these examples to $N > 2$, and we conjecture that $B(n, 1, N) \geq 2n - 3$ for all $n, N > 1$.

Suppose $n \in \mathbb{Z}$, $n \geq 6$, and $\{e_1, e_2, \dots, e_n\}$ is the standard basis for R^n . Let C_1 and C_2 be the $n \times n$ matrices defined by

$$C_1 e_1 = e_1, \quad C_1 e_2 = e_{n-1}, \quad C_1 e_3 = e_n,$$

$$C_1 e_j = e_{n-j+2} \quad \text{for } 4 \leq j \leq n-2,$$

$$C_1 e_{n-1} = e_2, \quad C_1 e_n = e_3, \quad \text{and}$$

$$C_2 e_j = e_{n-j+1} \quad \text{for } 1 \leq j \leq n.$$

Then $L = \{(C_1, e_1), (C_2, e_2)\} \in \Sigma(n, 1, 2)$, and computation shows that $b(L) = 2n - 3$ for $6 \leq n \leq 8$. In fact, $\gamma = (1, 2, 1, 2, \dots, 1, 2, 1, 1, 2) \in \Gamma_{2n-3}$ implies $\text{rank}(P(L, \gamma)) = n$ for all $n \geq 6$. Similar examples produce $b(L) = 2n - 3$ for $4 \leq n \leq 5$.

4. State deadbeat response. In this section we will show that every completely controllable multi-modal system with $n \leq 3$ is capable through feedback of a state deadbeat response; that is, each state vector can be brought to zero in finitely many steps. Results of this type are first obtained for arbitrary n under certain restrictive hypotheses.

When feedback controls are to be applied to a system $L = \{(C_i, D_i)\}_{i=1}^N$, a fixed set $\{F_1, F_2, \dots, F_N\}$ of $n \times m$ feedback matrices is selected at the outset. When state x_p is reached, the choice of the next index i determines not only the pair (C_i, D_i) to be applied, but also the control $U_p = F_i x_p$. Thus $x_{p+1} = C_i x_p + D_i U_p = (C_i + D_i F_i) x_p$.

DEFINITION. $L = \{(C_i, D_i)\}_{i=1}^N \in \Sigma(n, m, N)$ has a *state deadbeat response* provided there are feedback matrices F_i , a positive integer k and a $\gamma \in \Gamma_k$ such that

$$H(\gamma) = H_{\gamma(k)} H_{\gamma(k-1)} \cdots H_{\gamma(1)} = 0,$$

where $H_i = C_i + D_i F_i$, $i \in \bar{N}$.

If L has a state deadbeat response and γ satisfies $H(\gamma) = 0$ for some choice of F_i 's, we write $L \in \text{SDR}(\gamma)$. It will be shown that if $n \leq 3$ (or other restrictive hypotheses are assumed), complete controllability of L implies the existence of an index-sequence γ satisfying both $\text{rank}(P(L, \gamma)) = n$ and $L \in \text{SDR}(\gamma)$.

We will need to transform systems in $\Sigma(n, m, N)$ in the following way.

DEFINITION. Let $L = \{(C_i, D_i)\}_{i=1}^N \in \Sigma(n, m, N)$. Let G be an $n \times n$ nonsingular matrix, and let J_i be an $m \times m$ nonsingular matrix for $i \in \bar{N}$. Then

$$L_{G,J} = \{(GC_i G^{-1}, GD_i J_i)\}_{i=1}^N \quad \text{and}$$

$$L_G = \{(GC_i G^{-1}, GD_i)\}_{i=1}^N = L_{G,J} \quad \text{with each } J_i = I.$$

THEOREM 4. Let $L = \{(C_i, D_i)\}_{i=1}^N \in \Sigma(n, m, N)$, k be a positive integer, and $\gamma \in \Gamma_k$. Then

$$\text{rank}(P(L, \gamma)) = \text{rank}(P(L_{G,J}, \gamma)) \quad \text{and}$$

$$\text{rank}([D_1, D_2, \dots, D_N]) = \text{rank}([GD_1 J_1, GD_2 J_2, \dots, GD_N J_N]).$$

Hence, $L \in \Sigma_p(n, m, N)$ is completely controllable if and only if $L_{G,J}$ is completely controllable, and in this case $b(L) = b(L_{G,J})$. Furthermore, $L \in \text{SDR}(\gamma)$ if and only if $L_{G,J} \in \text{SDR}(\gamma)$.

Theorem 4 is a natural extension of [5, Prop. 1.1] and will not be proved.

LEMMA 2. If $L_1 \in \Sigma_m(n, m, N)$ is completely controllable and $b(L) = 2$, then L contains a completely controllable pair (C_i, D_i) with $\text{rank}([D_i, C_i D_i]) = n$.

Proof. Since $CS(D_i) = CS(D_j)$ for all $i, j \in \bar{N}$, the lemma follows. \square

THEOREM 5. If $L \in \Sigma(n, n-1, N)$ is completely controllable, then there is a $\gamma \in \Gamma_2$ such that $\text{rank}(P(L, \gamma)) = n$ and $L \in \text{SDR}(\gamma)$.

Proof. First suppose that L contains a completely controllable pair (C_i, D_i) . Let G be $n \times n$ and nonsingular such that

$$GD_i = [e_1, e_2, \dots, e_{n-1}] = E,$$

where $\{e_1, e_2, \dots, e_n\}$ is the standard basis for R^n . Let $GC_i G^{-1} = K = [\alpha_{ij}]$, so that (K, E) is completely controllable.

Then $\text{rank}([E, KE]) = n$, and so there is an $l \in \overline{n-1}$ such that $\alpha_{nl} \neq 0$. Define the $(n-1) \times n$ matrix F by

$$f_{ij} = \begin{cases} -\alpha_{ij}, & i \neq l, \\ -\alpha_{ij} - \frac{\alpha_{nj}\alpha_{nn}}{\alpha_{nl}}, & i = l. \end{cases}$$

Computation then shows that $(K + EF)^2 = 0$. Thus $L_G \in \text{SDR}(\gamma)$, where $\gamma = (t, t)$, and so by Theorem 4, $L \in \text{SDR}(\gamma)$. Also since $\text{rank}(P(L_G, \gamma)) = n$, we have $\text{rank}(P(L, \gamma)) = n$.

On the other hand, suppose L does not contain a completely controllable pair. Using Theorem 2 and Lemma 2, we see that $L \in \Sigma_n(n, n-1, N)$, and so there is a $t \in \bar{N}$ such that $\text{rank}([D_t, D_t]) = n$. Select G $n \times n$ and nonsingular, and select J_t $(n-1) \times (n-1)$ and nonsingular, so that $GD_t = [e_1, e_2, \dots, e_{n-1}] = E_1$ and the first

column of $GD_n J_i = E_i$ is e_n . Let $J_i = I$ for $i \in \bar{n}$ and $i \neq t$. We now have $L_{G,J} = \{(GC_i G^{-1}, GD_i J_i)\}_{i=1}^n = \{(K_i, E_i)\}_{i=1}^n$. Since $L_{G,J}$ contains no completely controllable pair, $\text{rank}([E_1, K_1 E_1]) < n$ which implies $\text{ent}_{nj}(K_1) = 0$ for $j \in \overline{n-1}$. Let F_1 be the $(n-1) \times n$ matrix satisfying $\text{Row}_i(F_1) = -\text{Row}_i(K_1)$ for $i \in \overline{n-1}$, to obtain $H_1 = K_1 + E_1 F_1$ with $\text{ent}_{ij}(H_1) = 0$ for $(i, j) \neq (n, n)$. Let F_t be the $(n-1) \times n$ matrix satisfying $\text{Row}_1(F_t) = -\text{Row}_n(K_t)$, and $\text{Row}_i(F_t) = 0$ for $i = 2, 3, \dots, n-1$. Then $H_t = K_t + E_t F_t$ satisfies $\text{Row}_n(H_t) = 0$. Hence $H_1 H_t = 0$, so $L_{G,J} \in \text{SDR}(\gamma)$ where $\gamma = (t, 1)$. Since $\text{rank}([E_1, K_1 E_1]) = n-1$, each of the first $n-1$ columns of K_1 is in $\text{CS}(E_1)$. Thus $\text{Col}_n(K_1) \in \text{CS}(E_1)$ and so $\text{rank}([E_1, K_1 E_1]) = n$. Hence $\text{rank}(P(L_{G,J}, \gamma)) = n$ and by Theorem 4 the desired result is obtained. \square

Since we are primarily interested in the case $n \leq 3$, we observe for L completely controllable in $\Sigma(2, 1, N)$ or in $\Sigma(3, 2, N)$, there is a $\gamma \in \Gamma_2$ such that $P(L, \gamma)$ has full rank and $L \in \text{SDR}(\gamma)$.

DEFINITION. Let $L \in \Sigma(n, m, N)$ be completely controllable. L is *reducible* if some pair (C_i, D_i) may be discarded to produce a completely controllable system. Otherwise, L is *irreducible*.

Clearly, if L' is a reduction of L and $\gamma' \in \Gamma_k$ for some k , then

- (i) $\text{rank}(P(L', \gamma')) = n$ if and only if $\text{rank}(P(L, \gamma)) = n$;
- (ii) $L' \in \text{SDR}(\gamma')$ if and only if $L \in \text{SDR}(\gamma)$, where γ is appropriately formed from γ' .

THEOREM 6. If $L \in \Sigma_n(n, 1, n)$ is completely controllable and irreducible, then there is $\gamma \in \Gamma_n$ such that $\text{rank}(P(L, \gamma)) = n$ and $L \in \text{SDR}(\gamma)$.

Proof. Let $L = \{(C_i, d_i)\}_{i=1}^n \in \Sigma_n(n, 1, n)$ be completely controllable and irreducible. Let $Gd_i = e_i$ for each $i \in \bar{n}$. Then $L_G = \{(K_i, e_i)\}_{i=1}^n$ is completely controllable and irreducible.

We claim that for each $i \in \bar{n}$ all off-diagonal entries in K_i are 0 except possibly those in the i th row. To establish this, fix $i \in \bar{n}$ and let $j \in \bar{n}$ with $j \neq i$. Assume $\text{ent}_{ji}(K_i) \neq 0$. Then $\text{rank}([e_i, K_i e_i]) = 2$. Since any product of K_i 's is nonsingular, we may choose $t_1, t_2, \dots, t_{n-2} \in \bar{n}$ so that

$$\text{rank}([e_i, K_i e_i, K_i K_i e_{t_1}, \dots, K_i K_i K_i \dots K_i e_{t_{n-2}}]) = n.$$

This contradicts the irreducibility of L_G , and thus the claim is established.

For each $i \in \bar{n}$, select $F_i = -\text{Row}_i(K_i)$ to obtain $H_i = K_i + e_i F_i$ with H_i diagonal and $\text{ent}_{ii}(H_i) = 0$. We see that if σ is any permutation of \bar{n} then $H_{\sigma(n)} H_{\sigma(n-1)} \dots H_{\sigma(1)} = 0$. Thus, since L is irreducible, there is a $\gamma \in \Gamma_n$ which is a permutation of \bar{n} , and for which $\text{rank}(P(L_G, \gamma)) = n$. Then $L_G \in \text{SDR}(\gamma)$ and this establishes the theorem. \square

For the remainder of this article whenever K_1 and K_2 are 3×3 matrices, we will use the notation

$$(4) \quad K_1 = \begin{bmatrix} \alpha_1 & \alpha_4 & \alpha_7 \\ \alpha_2 & \alpha_5 & \alpha_8 \\ \alpha_3 & \alpha_6 & \alpha_9 \end{bmatrix} \quad \text{and} \quad K_2 = \begin{bmatrix} \beta_1 & \beta_4 & \beta_7 \\ \beta_2 & \beta_5 & \beta_8 \\ \beta_3 & \beta_6 & \beta_9 \end{bmatrix}.$$

The following computational lemma will be employed several times. Its proof consists of eight computations, each of which is straightforward. We need the following notation:

$$\text{for } x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in R^3, \quad \Psi_1(x) = [0 -x_3 \ x_2] \quad \text{and} \quad \Psi_2(x) = [x_3 \ 0 -x_1].$$

LEMMA 3. If each of i, j and k is 1 or 2, and A and B are 3×3 , then

$$\det([e_i, Ae_j, AB e_k]) = \det(M),$$

where $\text{Row}_p(M) = \text{Row}_p(A)$ for $p \neq i$ and $\text{Row}_i(M) = \Psi_j(\text{Col}_k(B))$.

LEMMA 4. If $L \in \Sigma_1(3, 1, 2)$ is completely controllable and irreducible, then there is a $\gamma \in \Gamma_k$, with $k = b(L)$, such that $\text{rank}(P(L, \gamma)) = 3$ and $L \in \text{SDR}(\gamma)$.

Proof. Using Theorem 4, we may assume $L = \{(C_1, d), (C_2, d)\}$. Furthermore, in view of Theorem 3, $b(L)$ is 3 or 4.

First we consider the case $b(L) = 3$. We may select G 3×3 and nonsingular with $Gd = e_1$. Let $K_i = GC_iG^{-1}$ for $i = 1, 2$, so that

$$L_G = \{(K_1, e_1), (K_2, e_1)\}.$$

Using Lemma 3 and the fact that L contains no completely controllable pair we obtain (in the notation of (4)),

$$(5) \quad \det \begin{bmatrix} 0 & -\alpha_3 & \alpha_2 \\ \alpha_2 & \alpha_5 & \alpha_8 \\ \alpha_3 & \alpha_6 & \alpha_9 \end{bmatrix} = \det \begin{bmatrix} 0 & -\beta_3 & \beta_2 \\ \beta_2 & \beta_5 & \beta_8 \\ \beta_3 & \beta_6 & \beta_9 \end{bmatrix} = 0.$$

Either $\text{rank}([e_1, K_1 e_1, K_1 K_2 e_1]) = 3$ or else $\text{rank}([e_1, K_2 e_1, K_2 K_1 e_1]) = 3$. By renumbering pairs, if necessary, we obtain $\text{rank}([e_1, K_2 e_1, K_2 K_1 e_1]) = 3$, which implies

$$(6) \quad \det \begin{bmatrix} 0 & -\alpha_3 & \alpha_2 \\ \beta_2 & \beta_5 & \beta_8 \\ \beta_3 & \beta_6 & \beta_9 \end{bmatrix} \neq 0.$$

From (6) we see that $\text{rank}([e_1, K_1 e_1]) = \text{rank}([e_1, K_2 e_1]) = 2$. Thus, we assume G satisfies the additional condition $GC_1 d = e_2$, so that $K_1 e_1 = e_2$ which implies $\alpha_2 = 1$ and $\alpha_1 = \alpha_3 = 0$. Condition (5) now implies $\alpha_6 = 0$. Let

$$Q = [d, C_1 d, C_2 d].$$

We see that $\text{rank}(Q) = \text{rank}(GQ) = \text{rank}([e_1, e_2, K_2 e_1]) \geq 2$. Suppose $\text{rank}(Q) = 2$. Then $K_2 e_1 = \lambda e_1 + \mu e_2$, for some $\lambda, \mu \in R$, with $\mu \neq 0$. Thus $\beta_3 = 0$ and $\beta_2 = \mu \neq 0$. Using (5), we obtain $\beta_6 = 0$, which contradicts (6). Thus $\text{rank}(Q) = 3$. We may thus assume that G satisfies the further condition $GC_2 d = e_3$, so we have selected $G = Q^{-1}$. It follows that $K_2 e_1 = e_3$ which implies $\beta_1 = \beta_2 = 0$ and $\beta_3 = 1$. From (6) $\beta_5 \neq 0$, and from (5) $\beta_8 = 0$. If we select

$$F_1 = [0 \quad -\alpha_4 \quad -\alpha_7 - \alpha_9 \beta_9] \quad \text{and} \quad F_2 = [-\alpha_8 \quad -\beta_4 - \alpha_5 \beta_5 - \alpha_8 \beta_6 \quad -\beta_7 - \alpha_8 \beta_9],$$

we find that $H_2 H_1 H_2 = 0$, where $H_i = K_i + e_1 F_i$. Thus $\text{rank}(P(L_G, \gamma)) = 3$ and $L_G \in \text{SDR}(\gamma)$, where $\gamma = (2, 1, 2)$. This concludes the proof for $b(L) = 3$.

Now suppose $b(L) = 4$. Let $\gamma \in \Gamma_4$ with $\text{rank}(P(L, \gamma)) = 3$. Clearly we may assume $\gamma(1) = 1$, and, renumbering pairs if necessary so that $\gamma(2) = 1$, we obtain $\gamma \in \{(1, 1, 1, 2), (1, 1, 2, 2), (1, 1, 2, 1)\}$. Using Lemma 1, we see that $\gamma \notin \{(1, 1, 1, 2), (1, 1, 2, 2)\}$, and so $\gamma = (1, 1, 2, 1)$.

In a manner similar to the case $b(L) = 3$, we find that $Q = [d, C_2 C_1 d, C_1 d]$ has rank 3, and we select $G = Q^{-1}$ to produce $L_G = \{(K_1, e_1), (K_2, e_1)\}$. Employing the notation of (4) we choose

$$F_1 = [-\alpha_9 \quad -\alpha_4 - \alpha_5^2 \beta_5 - \alpha_5 \alpha_6 - \alpha_6 \alpha_9 \quad -\alpha_7 - \alpha_9^2] \quad \text{and} \quad F_2 = [-\beta_1 \quad -\beta_4 - \alpha_9 \beta_6 \quad 0],$$

and we find that $H_1 H_2 H_1 H_1 = 0$, where $H_i = K_i + e_1 F_i$. Thus L_G , and hence L , is in $\text{SDR}(\gamma)$. \square

LEMMA 5. If $L \in \Sigma_2(3, 1, 2)$ is completely controllable and irreducible, then there is a $\gamma \in \Gamma_k$, with $k = b(L)$, such that $\text{rank}(P(L, \gamma)) = 3$ and $L \in \text{SDR}(\gamma)$.

Proof. Let $L = \{(C_1, d_1), (C_2, d_2)\}$. Again we have $b(L)$ is 3 or 4.

Suppose $b(L) = 3$. For some nonconstant $\gamma \in \Gamma_3$, $\text{rank}(P(L, \gamma)) = 3$. By renumbering if necessary, we assume $\gamma(1) = 1$, and hence

$$\gamma \in \{(1, 2, 1), (1, 2, 2), (1, 1, 2)\}.$$

Let G be 3×3 and nonsingular such that $Gd_1 = e_1$ and $Gd_2 = e_2$, to produce $L_G = \{(K_i, e_i)\}_{i=1}^2$. Using Lemma 3 and the irreducibility of L , we obtain

$$(7) \quad \det \begin{bmatrix} 0 & -\alpha_3 & \alpha_2 \\ \alpha_2 & \alpha_5 & \alpha_8 \\ \alpha_3 & \alpha_6 & \alpha_9 \end{bmatrix} = \det \begin{bmatrix} \beta_1 & \beta_4 & \beta_7 \\ \beta_6 & 0 & -\beta_4 \\ \beta_3 & \beta_6 & \beta_9 \end{bmatrix} = 0.$$

We consider separately the three possible sequences γ .

Assume $\gamma = (1, 2, 1)$, so that $\text{rank}([e_1, K_1 e_2, K_1 K_2 e_1]) = 3$. By Lemma 3,

$$(8) \quad \det \begin{bmatrix} \beta_3 & 0 & -\beta_1 \\ \alpha_2 & \alpha_5 & \alpha_8 \\ \alpha_3 & \alpha_6 & \alpha_9 \end{bmatrix} \neq 0.$$

If all of $K_1 e_1$, $K_1 e_2$ and $K_2 e_1$ belong to $\text{span}\{e_1, e_2\}$, (8) is contradicted. Hence not all of $C_1 d_1$, $C_1 d_2$ and $C_2 d_1$ belong to $\text{span}\{d_1, d_2\}$.

First suppose $C_1 d_1 \notin \text{span}\{d_1, d_2\}$. Then we may assume G satisfies the additional condition $GC_1 d_1 = e_3$, so that $K_1 e_1 = e_3$. Thus $\alpha_1 = \alpha_2 = 0$ and $\alpha_3 = 1$, so that (7) implies $\alpha_8 = 0$. Hence (8) implies $\beta_1 + \alpha_9 \beta_3 \neq 0$. If we select

$$F_1 = \begin{bmatrix} -\frac{\beta_7 + \alpha_9 \beta_9}{\beta_1 + \alpha_9 \beta_3} & -\alpha_4 - \frac{\alpha_5 \beta_4 + \alpha_6 \beta_7 + \alpha_5 \alpha_9 \beta_6 + \alpha_6 \alpha_9 \beta_9}{\beta_1 + \alpha_9 \beta_3} & -\alpha_7 - \frac{\alpha_9 \beta_7 + \alpha_9^2 \beta_9}{\beta_1 + \alpha_9 \beta_3} \end{bmatrix}$$

and

$$F_2 = [-\beta_2 \quad -\beta_5 \quad -\beta_8],$$

we find that $H_1 H_2 H_1 = 0$, where $H_i = K_i + e_i F_i$.

Next suppose $C_1 d_1 \in \text{span}\{d_1, d_2\}$ and $C_1 d_2 \notin \text{span}\{d_1, d_2\}$. In this case, $\alpha_3 = 0$ and we may assume that G satisfies the additional condition $GC_1 d_2 = e_3$, so that $K_1 e_2 = e_3$, which implies $\alpha_4 = \alpha_5 = 0$ and $\alpha_6 = 1$. Hence (7) implies $\alpha_2 = 0$ and (8) implies $\beta_3 \neq 0$. Selecting

$$F_1 = \begin{bmatrix} -\alpha_1 & -\frac{\beta_9}{\beta_3} & -\alpha_7 - \frac{\alpha_8 \beta_6 + \alpha_9 \beta_9}{\beta_3} \end{bmatrix} \quad \text{and} \quad F_2 = [-\beta_2 \quad -\beta_5 \quad -\beta_8],$$

we find that $H_1 H_2 H_1 = 0$, where $H_i = K_i + e_i F_i$.

Finally, suppose $C_1 d_1, C_1 d_2 \in \text{span}\{d_1, d_2\}$. Then $C_2 d_1 \notin \text{span}\{d_1, d_2\}$. In this case $\alpha_3 = \alpha_6 = 0$ and we may assume that G satisfies the additional condition $GC_2 d_1 = e_3$, so that $K_2 e_1 = e_3$, which implies $\beta_1 = \beta_2 = 0$ and $\beta_3 = 1$. Hence (8) implies $\alpha_5 \neq 0$. Selecting

$$F_1 = [-\alpha_1 - \alpha_2 \beta_6 \quad -\alpha_4 - \alpha_5 \beta_6 \quad -\alpha_7 - \alpha_8 \beta_6 - \alpha_9 \beta_9]$$

and

$$F_2 = \begin{bmatrix} 0 & -\beta_5 - \frac{\alpha_2 \beta_4}{\alpha_5} & -\beta_8 - \frac{\alpha_2 \beta_7}{\alpha_5} \end{bmatrix},$$

we find that $H_1 H_2 H_1 = 0$, where $H_i = K_i + e_i F_i$, and we have shown $L \in \text{SDR}(\gamma)$ for $\gamma = (1, 2, 1)$.

Now assume $\gamma = (1, 2, 2)$, so that $\text{rank}([e_2, K_2 e_2, K_2 K_2 e_1]) = 3$. By Lemma 3,

$$(9) \quad \det \begin{bmatrix} \beta_1 & \beta_4 & \beta_7 \\ \beta_3 & 0 & -\beta_1 \\ \beta_3 & \beta_6 & \beta_9 \end{bmatrix} \neq 0.$$

We first show that $C_2 d_2 \notin \text{span}\{d_1, d_2\}$. If $K_2 e_2 \in \text{span}\{e_1, e_2\}$ and $K_2 e_1 \notin \text{span}\{e_1, e_2\}$, then $\beta_6 = 0$, and we may require that G satisfy the condition $GC_2 d_1 = e_3$. Then $K_2 e_1 = e_3$, which implies $\beta_1 = 0$ and $\beta_3 = 1$. Thus (7) implies $\beta_4 = 0$. But $\beta_4 = \beta_6 = 0$ contradicts (9), and so $K_2 e_2 \in \text{span}\{e_1, e_2\}$ implies $K_2 e_1 \in \text{span}\{e_1, e_2\}$. But this, in turn, implies $\beta_3 = \beta_6 = 0$ contradicting (9). Thus $K_2 e_2 \notin \text{span}\{e_1, e_2\}$ and so $C_2 d_2 \notin \text{span}\{d_1, d_2\}$.

We may now assume $GC_2 d_2 = e_3$, so that $K_2 e_2 = e_3$ which implies $\beta_4 = \beta_5 = 0$ and $\beta_6 = 1$. Now (7) implies $\beta_7 = 0$, and so the selection

$$F_1 = [-\alpha_1 \quad -\alpha_4 \quad -\alpha_7] \quad \text{and} \quad F_2 = [-\beta_2 \quad -\beta_9 \quad -\beta_8 - \beta_5^2]$$

leads to $H_2 H_2 H_1 = 0$, where $H_i = K_i + e_i F_i$, and $L \in \text{SDR}(\gamma)$ for $\gamma = (1, 2, 2)$.

Finally assume $\gamma = (1, 1, 2)$, so that $\text{rank}([e_2, K_2 e_1, K_2 K_1 e_1]) = 3$. By Lemma 3

$$(10) \quad \det \begin{bmatrix} \beta_1 & \beta_4 & \beta_7 \\ 0 & -\alpha_3 & \alpha_2 \\ \beta_3 & \beta_6 & \beta_9 \end{bmatrix} \neq 0.$$

We first show that not both of $C_1 d_1$ and $C_2 d_2$ are in $\text{span}\{d_1, d_2\}$. For suppose $K_1 e_1, K_2 e_2 \in \text{span}\{e_1, e_2\}$, so that $\alpha_3 = \beta_6 = 0$. If $\beta_3 = 0$, then (10) is contradicted, so $\beta_3 \neq 0$. Thus $C_2 d_1 \notin \text{span}\{d_1, d_2\}$ so we may require $GC_2 d_1 = e_3$. Then $K_2 e_1 = e_3$, so $\beta_1 = 0$ and $\beta_3 = 1$. Now (7) implies $\beta_4 = 0$, and (10) is contradicted. Hence at least one of $C_1 d_1$ and $C_2 d_2$ is not in $\text{span}\{d_1, d_2\}$.

We will now show that $\text{rank}(P(L, \sigma)) = 3$ for $\sigma = (1, 2, 2)$ (in one case a renumbering of pairs being necessary), and so by the previous argument, the lemma is proved.

If $C_2 d_2 \notin \text{span}\{d_1, d_2\}$ we may require $GC_2 d_2 = e_3$, so that $K_2 e_2 = e_3$ and we have $\beta_4 = 0$ and $\beta_6 = 1$. Thus (7) implies $\beta_7 = 0$, and so (10) implies $\beta_1 \neq 0$. Hence

$$\det \begin{bmatrix} \beta_1 & \beta_4 & \beta_7 \\ \beta_3 & 0 & -\beta_1 \\ \beta_3 & \beta_6 & \beta_9 \end{bmatrix} = \beta_1^2 \neq 0,$$

which by Lemma 3 implies $\text{rank}(P(L, \sigma)) = 3$ with $\sigma = (1, 2, 2)$.

If $C_2 d_2 \in \text{span}\{d_1, d_2\}$, then $C_1 d_1 \notin \text{span}\{d_1, d_2\}$. In this case we may require $GC_1 d_1 = e_3$, so that $K_1 e_1 = e_3$, and we have $\alpha_2 = 0$ and $\alpha_3 = 1$. Then (7) implies $\alpha_8 = 0$, and so by the nonsingularity of K_1 , $\alpha_5 \neq 0$. Hence

$$\det \begin{bmatrix} 0 & -\alpha_6 & \alpha_5 \\ \alpha_2 & \alpha_5 & \alpha_8 \\ \alpha_3 & \alpha_6 & \alpha_9 \end{bmatrix} = -\alpha_5^2 \neq 0,$$

and by Lemma 3, $\text{rank}(P(L, \tau)) = 3$ for $\tau = (2, 1, 1)$. Thus a renumbering of the pairs gives $\text{rank}(P(L, \sigma)) = 3$ for $\sigma = (1, 2, 2)$. This concludes the proof for $b(L) = 3$.

The case $b(L) = 4$ is handled in a similar fashion. It is shown in [2] that the only possible $\gamma \in \Gamma_4$ with $\gamma(1) = 1$ and $\text{rank}(P(L, \gamma)) = 3$ are $(1, 2, 1, 1)$ and $(1, 2, 1, 2)$. For

either of these cases, we may transform by $Gd_1 = e_1$, $Gd_2 = e_2$, $GC_2d_1 = e_3$ and select

$$F_1 = [-\alpha_1 \quad -\alpha_4 \quad -\alpha_7 - \alpha_9\beta_9] \quad \text{and} \quad F_2 = [0 \quad -\beta_5 \quad -\beta_8]$$

to obtain the desired result. \square

THEOREM 7. *If $L \in \Sigma(3, 1, N)$ is completely controllable, then there is a $\gamma \in \Gamma_k$, with $k = b(L)$, such that $\text{rank}(P(L, \gamma)) = 3$ and $L \in \text{SDR}(\gamma)$.*

Proof. Let $L = \{(C_i, d_i)\}_{i=1}^N \in \Sigma(3, 1, N)$ be completely controllable.

First suppose $L \in \Sigma_1(3, 1, N)$ with $b(L) = 3$. We may assume $L = \{(C_i, d)\}_{i=1}^N$ using Theorem 4. Since $b(L) = 3$, L is reducible when $N \geq 3$. Thus either L contains a completely controllable pair, or else Lemma 4 applies to some reduction of L . In either case, the theorem holds.

Now suppose $L \in \Sigma_1(3, 1, N)$ with $b(L) = 4$. Clearly L is reducible for $N \geq 4$, and we now show that if $N = 3$, L is reducible. By assuming L irreducible, and renumbering pairs if necessary, we obtain

$$(11) \quad \text{rank}([d, C_3d, C_3C_2d, C_3C_2C_1d]) = 3.$$

If $\text{rank}([C_1d, C_3d]) = 1$, then $C_1d = \lambda C_3d$ with $\lambda \neq 0$, and so using (11), $\text{rank}([d, C_3d, C_3C_2d, C_3C_2C_3d]) = 3$, contradicting the irreducibility of L . Thus $\text{rank}([C_1d, C_3d]) = 2$, so that $\text{rank}([d, C_1d, C_3d]) \geq 2$. If this rank is 3, it follows that $C_1C_3d \notin \text{span}\{C_1d, C_1C_1d\}$. By (11), $\text{rank}([d, C_1d]) = 2$, so $\text{rank}([d, C_1d, C_1C_1d]) = 2$, and $\text{span}\{C_1d, C_1C_1d\} = \text{span}\{d, C_1d\}$. Thus, $\text{rank}([d, C_1, C_1C_3d]) = 3$, which contradicts $b(L) = 4$. Hence $\text{rank}([d, C_1d, C_3d]) = 2$, so that $C_3d = \lambda_1 C_1d + \lambda_2 d$. By Lemma 1, $\text{rank}([d, C_3d]) = 2$, so $\lambda_1 \neq 0$. Therefore, using (11), $\text{rank}([d, C_3d, C_3C_2d, C_3C_2C_3d]) = 3$, contradicting the irreducibility of L . Hence L is reducible for $N \geq 3$. Either L contains a completely controllable pair, or else Lemma 4 applies to some reduction of L . In either case, the theorem holds.

Next suppose $L \in \Sigma_2(3, 1, N)$ with $b(L) = 3$. Clearly L is reducible when $N \geq 4$, and we now show that L is reducible when $N = 3$. By assuming L is irreducible and renumbering pairs if necessary, we obtain

$$(12) \quad \text{rank}([d_1, C_1d_2, C_1C_2d_3]) = 3.$$

If $\text{rank}([d_1, d_2]) = 2$, then $C_1C_2d_3 \in \text{span}\{C_1C_2d_1, C_1C_2d_2\}$. However, $C_1C_2d_j \in \text{span}\{d_1, C_1d_2\}$ for $j = 1, 2$, since L is irreducible. Hence $C_1C_2d_3 \in \text{span}\{d_1, C_1d_2\}$ which contradicts (12). Thus $\text{rank}([d_1, d_2]) = 1$. Suppose $\text{rank}([d_1, C_1d_2, d_3]) = 3$, so that $\text{rank}([d_1, C_1d_1, d_3]) = 3$. Then $C_1C_1d_3 \notin \text{span}\{C_1C_1d_1, C_1C_1C_1d_1\} = \text{span}\{d_1, C_1d_1\}$, and so $\text{rank}([d_1, C_1d_1, C_1C_1d_3]) = 3$, contradicting the irreducibility of L . Thus $\text{rank}([d_1, C_1d_2, d_3]) = 2$, and so $C_1C_2d_3 \in \text{span}\{C_1C_2d_1, C_1C_2C_1d_2\} = \text{span}\{d_1, C_1d_2\}$, which contradicts (12). Hence L is reducible for $N \geq 3$. Either L contains a completely controllable pair, or else one of Lemmas 4 and 5 may be applied to some reduction of L . In either case, the theorem holds.

Now suppose $L \in \Sigma_2(3, 1, N)$ with $b(L) = 4$. Clearly L is reducible for $N \geq 5$, and we now show that L is reducible for $3 \leq N \leq 4$. By assuming L irreducible, we obtain $\gamma \in \Gamma_4$ satisfying $\{\gamma(1), \gamma(2), \gamma(3), \gamma(4)\} = \bar{N}$ and $\text{rank}(P(L, \gamma)) = 3$. Since $\text{rank}([d_{\gamma(4)}, C_{\gamma(4)}d_{\gamma(3)}, C_{\gamma(4)}C_{\gamma(3)}d_i]) < 3$ for $i \in \bar{N}$ we obtain, by applying Lemma 1 to $P(L, \gamma)$, $C_{\gamma(2)}d_{\gamma(1)} \notin \text{span}\{d_1, d_2, \dots, d_N\}$. Choose $j \in \bar{N}$ so that $\text{rank}([d_{\gamma(2)}, d_j]) = 2$, which implies $\text{rank}([C_{\gamma(2)}d_{\gamma(1)}, d_{\gamma(2)}, d_j]) = 3$; it follows that

$$C_{\gamma(2)}C_{\gamma(1)}d_j \notin \text{span}\{C_{\gamma(2)}C_{\gamma(1)}d_{\gamma(2)}, C_{\gamma(2)}C_{\gamma(1)}C_{\gamma(2)}d_{\gamma(1)}\} = \text{span}\{d_{\gamma(2)}, C_{\gamma(2)}d_{\gamma(1)}\}.$$

The last equality follows from the irreducibility of L . Thus $\text{rank}([d_{\gamma(2)}, C_{\gamma(2)}d_{\gamma(1)}, C_{\gamma(2)}C_{\gamma(1)}d_j]) = 3$ which contradicts $b(L) = 4$. Hence L is reducible for $N \geq 3$. As in the previous case, the theorem holds.

Finally suppose $L \in \Sigma_3(3, 1, N)$. By Theorem 1, $b(L) \leq 3$, and so there is an irreducible reduction L_1 of L with $L_1 \in \Sigma_j(3, 1, k)$ with $j, k \in \{1, 2, 3\}$. If $j < 3$, previous cases show $k < 3$ and the theorem follows. If $j = 3$, then $k = 3$ and Theorem 6 applies. \square

In the preceding proofs, feedback matrices producing state deadbeat response for the transformed system $L_{G,J}$ were constructed. Since it may be useful to construct feedback matrices for the original system L , we note that the feedback F_i for the system $L_{G,J}$ corresponds to the feedback $J_i F_i G$ for the original system L .

In most cases delineated in the proofs of our theorems, $L \in \text{SDR}(\gamma)$ followed from the assumption that $\text{rank}(P(L, \gamma)) = n$. However, this did not occur for $L \in \Sigma_2(3, 1, 2)$ with $\gamma = (1, 1, 2)$, and for this γ , the following example is an instance in which $\text{rank}(P(L, \gamma)) = 3$ and $L \notin \text{SDR}(\gamma)$.

$$C_1 = \begin{bmatrix} 0 & 1 & 0 \\ 2 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad d_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad d_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

5. Observability and controllability canonical form. In this section, we introduce observability for multi-modal systems and use it to refine the controllability form presented in [3]. For this purpose, we need to expand the multi-modal system to one of the form

$$L: \begin{cases} x_{k+1} = C_i x_k + D_i u_k, \\ y_k = E_i x_k \end{cases}$$

with x_k, C_i, D_i, u_k as before, E_i a real $p \times n$ matrix, and $y_k \in R^p$. Hence we define

$$\Sigma^0(n, m, p, N) = \{L = \{(C_i, D_i, E_i)\}_{i=1}^N \mid C_i \text{ real } n \times n, D_i \text{ real } n \times m, E_i \text{ real } p \times n\}.$$

Throughout this paper U_k denotes the set of all k -termed sequences with terms in R^m , and Γ denotes $\bigcup \{\Gamma_k \mid k \in \mathbb{Z}^+\}$. Suppose $L \in \Sigma^0(n, m, p, N)$, $u \in U_k$ and $\alpha \in \Gamma_k$. For $x \in R^n$, the *trajectory of x under u and α* , denoted $T(L, x, u, \alpha)$, is the sequence $\{x_i\}_{i=1}^{k+1}$, where $x_1 = x$, and $x_{i+1} = C_{\alpha(i)}x_i + D_{\alpha(i)}u_i$ for $i \in \bar{k}$. For $i \in \overline{k+1}$, the i th term of $T(L, x, u, \alpha)$ is denoted by $T_i(L, x, u, \alpha)$. The *output trajectory* $Y(L, x, u, \alpha)$ is defined by $Y_i(L, x, u, \alpha) = E_{\alpha(i)}T_i(L, x, u, \alpha)$, $i \in \bar{k}$.

DEFINITION. For $L \in \Sigma^0(n, m, p, N)$ and $\gamma \in \Gamma_k$, L is γ -observable provided for each $u \in U_k$ the map

$$x \rightarrow Y(L, x, u, \gamma)$$

is one-to-one on R^n . That is L is γ -observable provided each $x \in R^n$ can be discerned as the initial state from knowledge of the input $u \in U_k$ and the corresponding output $Y(L, x, u, \gamma)$.

We observe that for any $x \in R^n$, $u \in U_k$ and $\gamma \in \Gamma_k$, $Y(L, x, u, \gamma)$ is given by

$$\begin{aligned} Y_1(L, x, u, \gamma) &= E_{\gamma(1)}x, \\ (*) \quad Y_j(L, x, u, \gamma) &= E_{\gamma(j)}C(\gamma; j-1)x + E_{\gamma(j)}D_{\gamma(j-1)}u_{j-1} \\ &\quad + \sum_{t=1}^{j-2} E_{\gamma(j)}C(\gamma; j-1, t+1)D_{\gamma(t)}u_t \quad \text{for } j = 2, 3, \dots, k. \end{aligned}$$

Thus, for any subspace V of R^n , the map $x \rightarrow Y(L, x, u, \gamma)$ can be viewed as an affine

transformation from R^n into R^{pk} , and hence is one-to-one on V , if and only if the map $x \rightarrow Y(L, x, \Theta, \gamma)$ is one-to-one on V , where Θ denotes the sequence of zero inputs.

For $L \in \Sigma^0(n, m, p, N)$ and $\gamma \in \Gamma_k$, we define the observability matrix

$$W(L, \gamma) = [E_{\gamma(1)}^T, C^T(\gamma; 1)E_{\gamma(2)}^T, C^T(\gamma; 2)E_{\gamma(3)}^T, \dots, C^T(\gamma; k-1)E_{\gamma(k)}^T],$$

where $C^T(\gamma; i)$ is the transpose of $C(\gamma; i)$.

The following two theorems parallel the standard theory for single-mode systems and we state them without proof.

THEOREM 8. *If $L \in \Sigma^0(n, m, p, N)$ and $\gamma \in \Gamma_k$, then the map $x \rightarrow Y(L, x, \Theta, \gamma)$ is one-to-one on $CS(W(L, \gamma))$.*

THEOREM 9. *Suppose $L \in \Sigma^0(n, m, p, N)$ and $\gamma \in \Gamma_k$. L is γ -observable if and only if $\text{rank}(W(L, \gamma)) = n$.*

For $L \in \Sigma^0(n, m, p, N)$ we define

$$Z(L) = \bigcup \{CS(W(L, \gamma)) \mid \gamma \in \Gamma\}.$$

We will see that, in general, $Z(L)$ is not a subspace of R^n . In case it is, it follows from Theorem 8 and the forthcoming discussion that for each $x \in R^n$, the orthogonal projection of x on $Z(L)$ can be determined from the knowledge of $Y(L, x, \Theta, \gamma)$ for some γ , and that γ may be chosen independently of x .

DEFINITION. For $L \in \Sigma^0(n, m, p, N)$, L is *completely observable* provided $Z(L) = R^n$.

We wish to demonstrate the duality of controllability and observability for our systems.

DEFINITION. For $L = \{(C_i, D_i, E_i)\}_{i=1}^N \in \Sigma^0(n, m, p, N)$, the *dual* of L is the system $L^* = \{(C_i^T, E_i^T, D_i^T)\}_{i=1}^N$, which belongs to $\Sigma^0(n, p, m, N)$.

THEOREM 10. *If $L \in \Sigma^0(n, m, p, N)$, then $Z(L) = S(L^*)$ and $S(L) = Z(L^*)$. In particular a system is completely observable if and only if its dual is completely controllable.*

Proof. The theorem follows, since for each $k \in Z^+$ and each $\gamma \in \Gamma_k$, $W(L, \gamma) = P(L^*, \tilde{\gamma})$, where $\tilde{\gamma} \in \Gamma_k$ with $\tilde{\gamma}(i) = \gamma(k - i + 1)$ for $i \in \bar{k}$. \square

Using this duality theorem and the results concerning the controllability set in [3], we now see that, in general, $Z(L)$ is not a subspace of R^n . The following theorems are the duals of [3, Thm. 1] and part of [3, Thm. 2].

THEOREM 11. *Suppose $L = \{(C_i, D_i, E_i)\}_{i=1}^N \in \Sigma^0(n, m, p, N)$. If each C_i is nonsingular, then $Z(L)$ is a subspace of R^n .*

THEOREM 12. *If $L \in \Sigma^0(n, m, p, N)$ and $Z(L)$ is a subspace of R^n , then there exists $\gamma \in \Gamma$ such that $Z(L) = CS(W(L, \gamma))$. Thus L is completely observable if and only if L is γ -observable for some $\gamma \in \Gamma$.*

DEFINITION. For $L \in \Sigma^0(n, m, p, N)$ and $\gamma \in \Gamma$, the γ -unobservable space of L is defined by

$$T(L, \gamma) = (CS(W(L, \gamma)))^\perp = NS(W^T(L, \gamma)).$$

The *unobservable space* of L is defined by

$$T(L) = \bigcap \{T(L, \gamma) \mid \gamma \in \Gamma\}.$$

We note that for each $L \in \Sigma^0(n, m, p, N)$, $T(L) = (Z(L))^\perp$. In the case when $Z(L)$ is a subspace of R^n , $R^n = Z(L) \oplus T(L)$, and we see that L is completely observable if and only if $T(L) = \{0\}$.

We now apply duality to obtain information about the controllability canonical form. We must first extend the transformation of Theorem 4 to $L \in \Sigma^0(n, m, p, N)$. If

$L = \{(C_i, D_i, E_i)\}_{i=1}^N$, then $L_G = \{(GC_iG^{-1}, GD_i, E_iG^{-1})\}_{i=1}^N$. Using Theorem 4 and duality, we obtain:

THEOREM 13. Suppose $L \in \Sigma^0(n, m, p, N)$. Then, for each $\gamma \in \Gamma$, $Z(L_G, \gamma) = (G^{-1})^T Z(L, \gamma)$ so that $Z(L_G) = (G^{-1})^T Z(L)$. Thus L_G is γ -observable if and only if L is γ -observable.

THEOREM 14. If $L \in \Sigma^0(n, m, p, N)$, then, for each $\gamma \in \Gamma$, $T(L_G, \gamma) = GT(L, \gamma)$, and so $T(L_G) = GT(L)$.

Proof. For each $\gamma \in \Gamma$, $W(L_G, \gamma) = (G^{-1})^T W(L, \gamma)$, and so $W^T(L_G, \gamma) = W^T(L, \gamma)G^{-1}$. The theorem follows. \square

We need the following lemma to prove Theorems 15 and 16, dealing with the controllability canonical form.

LEMMA 6. If $L \in \Sigma^0(n, m, p, N)$, then $T(L)$ is C_i -invariant and $Z(L)$ is C_i^T -invariant for each $i \in \bar{N}$.

Proof. Suppose $x \in T(L)$ and $i \in \bar{N}$. For $\gamma \in \Gamma_k$, let (i, γ) denote the sequence $(i, \gamma(1), \gamma(2), \dots, \gamma(k))$. For each $\gamma \in \Gamma$, $x \in T(L, (i, \gamma)) = NS(W^T(L, (i, \gamma)))$ and thus $C_i x \in T(L, \gamma)$. Hence $C_i x \in T(L)$. $Z(L) = S(L^*)$ which is C_i^T -invariant for each $i \in \bar{N}$, as shown in [3]. \square

The following theorem and its proof are obvious extensions of the discussion of the case $N = 1$ found in [6, Chapt. 11].

THEOREM 15. Suppose $L = \{(C_i, D_i, E_i)\}_{i=1}^N \in \Sigma^0(n, m, p, N)$ with $S(L)$ a subspace of R^n . Let $M_1(L) = S(L) \cap T(L)$. Let $M_2(L)$, $M_3(L)$ and $M_4(L)$ be subspaces of R^n such that $S(L) = M_1(L) \oplus M_2(L)$, $T(L) = M_1(L) \oplus M_3(L)$, and $R^n = M_1(L) \oplus M_2(L) \oplus M_3(L) \oplus M_4(L)$. If $r_j = \dim(M_j(L)) > 0$, $j = 1, 2, 3, 4$, then there is an $n \times n$ nonsingular matrix G such that

$$L_G = \left\{ \left(\begin{bmatrix} C_{i11} & C_{i12} & C_{i13} & C_{i14} \\ 0 & C_{i22} & 0 & C_{i24} \\ 0 & 0 & C_{i33} & C_{i34} \\ 0 & 0 & 0 & C_{i44} \end{bmatrix}, \begin{bmatrix} D_{i1} \\ D_{i2} \\ 0 \\ 0 \end{bmatrix}, [0 \quad E_{i2} \quad 0 \quad E_{i4}] \right) \right\}_{i=1}^N,$$

where C_{ij} is $r_j \times r_j$ for $j = 1, 2, 3, 4$, D_{ij} is $r_j \times m$ for $j = 1, 2$, and E_{ij} is $p \times r_j$ for $j = 2, 4$. The system

$$(L_G)_1 = \left\{ \left(\begin{bmatrix} C_{i11} & C_{i12} \\ 0 & C_{i22} \end{bmatrix}, \begin{bmatrix} D_{i1} \\ D_{i2} \end{bmatrix}, [0 \quad E_{i2}] \right) \right\}_{i=1}^N$$

is completely controllable.

Proof. Let $B = \{\beta_{11}, \beta_{12}, \dots, \beta_{1r_1}, \beta_{21}, \dots, \beta_{2r_2}, \beta_{31}, \dots, \beta_{3r_3}, \beta_{41}, \dots, \beta_{4r_4}\}$ be a basis for R^n with $\{\beta_{j1}, \beta_{j2}, \dots, \beta_{jr_j}\}$ a basis for $M_j(L)$ for $j = 1, 2, 3, 4$. Let G be the matrix of transition from the standard basis of R^n to B . Since $M_1(L)$, $S(L)$ and $T(L)$ are all C_i -invariant and $CS(D_i) \subset S(L)$ for each $i \in \bar{N}$,

$$GS(L) = \left\{ \begin{bmatrix} x \\ 0 \end{bmatrix} \middle| x \in R^{r_1+r_2} \right\} \quad \text{and} \quad GT(L) = \left\{ \begin{bmatrix} x_1 \\ 0 \\ x_3 \\ 0 \end{bmatrix} \middle| x_1 \in R^{r_1}, x_3 \in R^{r_3} \right\},$$

and the theorem follows. \square

THEOREM 16. Suppose $L = \{(C_i, D_i, E_i)\}_{i=1}^N \in \Sigma^0(n, m, p, N)$ with $S(L)$ a subspace of R^n and with each C_i symmetric. Let $V(L) = \cap \{NS(P^T(L, \gamma)) \mid \gamma \in \Gamma\}$. If $\dim(V(L)) =$

$r < n$, then there is a nonsingular G such that

$$L_G = \left\{ \left(\begin{bmatrix} C_{i1} & 0 \\ 0 & C_{i2} \end{bmatrix}, \begin{bmatrix} D_{i1} \\ 0 \end{bmatrix}, [E_{i1}, E_{i2}] \right) \right\}_{i=1}^N$$

with each $C_{i1}(n-r) \times (n-r)$ and each $D_{i1}(n-r) \times m$. The system $(L_G)_1 = \{(C_{i1}, D_{i1}, E_{i1})\}_{i=1}^N$ is completely controllable.

Proof. In [3] we showed that $S(L)$ is C_i -invariant for each $i \in \bar{N}$. Now $V(L) = \cap \{NS(P^T(L, \gamma)) | \gamma \in \Gamma\} = \cap \{NS(W^T(L^*, \gamma)) | \gamma \in \Gamma\} = T(L^*)$ and thus, by Lemma 6, $V(L)$ is C_i^T -invariant, and hence C_i -invariant for each $i \in \bar{N}$. Also, since $Z(L^*) = S(L)$ is a subspace of R^n , $R^n = Z(L^*) \oplus T(L^*) = S(L) \oplus V(L)$. Let $B = \{\beta_1, \beta_2, \dots, \beta_n\}$ be a basis for R^n with $\{\beta_1, \beta_2, \dots, \beta_{n-r}\}$ a basis for $S(L)$ and $\{\beta_{n-r+1}, \dots, \beta_n\}$ a basis for $V(L)$. If G is the matrix of transition from the standard basis of R^n to B , then L_G has the required form. \square

REFERENCES

- [1] D. P. STANFORD, *Stability for a multi-rate sampled-data system*, this Journal, 17 (1979), pp. 390–399.
- [2] D. P. STANFORD AND L. T. CONNER, JR., *Controllability in multi-rate sampled-data systems*, Final Report: Task Order NAS1-14972-9, September 1978.
- [3] ———, *Controllability and stabilizability in multi-pair systems*, this Journal, 18 (1980), pp. 488–497.
- [4] ———, *Addendum: controllability and stabilizability in multi-pair systems*, this Journal, 19 (1981), pp. 708–709.
- [5] W. M. WONHAM, *Linear Multivariable Control*, Springer-Verlag, New York, 1974.
- [6] LOTFI A. ZADEH AND CHARLES A. DESOER, *Linear System Theory*, McGraw-Hill, New York, 1963.

THEORIE GENERALE DU CONTROLE IMPULSIONNEL MARKOVIEEN*

J. P. LEPELTIER† AND B. MARCHAL‡

Abstract. This paper is concerned with the optimal control of a Markov process which is submitted to a sequence of changes of state and law at times $T_1, T_2, \dots, T_n, \dots$. The cost is made with an impulse cost which is positive, and a continuous cost. The modelization is deeply inspired by the theory of renewal for Markov processes by P. A. Meyer. The main result is the Markovian form of the value function of this problem. Then under smooth additional conditions, we prove the existence of an optimal control.

Key words. stochastic control, impulse-control, Markov process, right process, regeneration of Markov processes

Introduction. Historiquement, le problème du contrôle impulsionnel a été introduit par A. Bensoussan et J. L. Lions [1] de la manière suivante: on considère un stock soumis à une demande aléatoire, il s'agit d'optimiser les instants de réapprovisionnement et les quantités commandées à ces instants, la structure du coût étant la suivante: coût fixe à chaque commande et coût continu de stockage. Leur technique de résolution d'un tel problème fondamentalement du type analyse numérique a l'avantage essentiel de pouvoir obtenir une approche effective par des algorithmes de la stratégie optimale, et présente donc un intérêt évident dans la pratique. Par contre d'un point de vue théorique, cette méthode impose des hypothèses fortes de régularité, d'où les limites de celle-ci en vue de la recherche d'un théorème d'existence très général.

Les premières méthodes générales probabilistes qui permettent notamment de se dégager du cadre des diffusions ont été introduites par M. Robin [16], où les processus aléatoires envisagés sont essentiellement des processus de Markov fellériens.

Dans ce travail où les processus traités sont des processus de Markov droits [8] très généraux, nous considérons un système dont la loi est régie par un opérateur, qui peut à tout moment changer l'état du système et sa loi, ce que nous appellerons son régime, ce dernier choix ayant lieu parmi un certain nombre de régimes permis.

Considérant que fondamentalement le problème de contrôle impulsionnel consiste en une succession de changements de loi à des instants T_1, \dots, T_m, \dots , nous avons choisi de le modéliser suivant un modèle entièrement canonique dont la construction est basée sur la théorie de la renaissance des processus de Markov de P. A. Meyer [14] en considérant qu'à chaque instant d'impulsion nous tuons le système, puis nous le faisons renaître suivant une nouvelle loi choisie parmi les régimes permis.

Dans le second chapitre nous établissons les propriétés de ce modèle nécessaires pour poursuivre l'étude. Nous définissons essentiellement une propriété de stabilité par arrêt des stratégies, fondamentale pour établir dans la troisième partie un principe d'optimalité qui est la version probabiliste du principe de Bellman. Ce principe nous conduit par une méthode adaptée de C. Striebel [17], [6], à un critère d'optimalité analogue probabiliste des I. Q. V. de A. Bensoussan, J. L. Lions.

Le quatrième chapitre est la partie la plus importante et la plus originale de notre travail. Nous y établissons la dépendance markovienne des coûts minimaux conditionnels sous des hypothèses très générales sur les coûts. Notre technique consiste à percevoir le contrôle impulsionnel comme une suite d'arrêts optimaux et ainsi d'itérer

* Received by the editors January 30, 1981, and in revised form December 20, 1982.

† Université du Maine, Département de Mathématiques, Route de Laval, 72017 Le Mans Cedex, France.

‡ Université Paris Val de Marne, U.E.R. Sciences Economiques et Gestion, 58 avenue Didier, 94210 La Varenne St. Hilaire, France.

un problème d'arrêt optimal, technique d'itération similaire à celle employée par P. A. Meyer dans les jeux de hasard [13]. Nous utilisons des résultats très fins de l'arrêt optimal [6] ce qui nous permet d'obtenir le caractère très général de nos résultats. Cette technique jointe à notre modélisation avec à la fois changement de régime et d'état nous permet d'établir dans un autre article [7] le caractère markovien de la fonction de valeur du contrôle continu markovien, problème qui était jusqu'ici resté ouvert.

Enfin dans le dernier chapitre, nous indiquons sous quelles hypothèses supplémentaires nous savons construire une stratégie optimale.

En conclusion, cette méthode nous permet de résoudre un problème qui englobe les principaux exemples de contrôle impulsif qui existent dans la littérature (pour plus de précisions voir [9] ou [12]).

Nous remercions le rapporteur de cet article pour ses remarques pertinentes en vue de rendre ce travail plus pédagogique.

1. Modélisation.

A. Nous allons utiliser l'idée de la renaissance des processus de Markov d'après P. A. Meyer [14] en considérant qu'avant la première impulsion la loi du système est celle d'un processus de Markov tué au moment de cette impulsion, puis en le faisant renaître suivant une nouvelle loi d'un processus de Markov tué prenant en compte l'impulsion dans la renaissance, et ainsi de suite

Décrivons tout d'abord l'espace canonique associé aux lois tuées. Nous supposons que le système dont l'évolution est continue à droite et limitée à gauche (càdlàg) entre deux impulsions, est à valeurs dans un espace E , sous-ensemble universellement mesurable d'un métrique compact; nous noterons \mathbb{E} la tribu borélienne associée. Soit \bar{E} l'espace obtenu par adjonction à E d'un point isolé Δ et $\bar{\mathbb{E}}$ sa tribu associée; nous appellerons $\bar{\Omega}$ l'ensemble des applications $\bar{\omega}$ de R_+ dans \bar{E} qui possèdent les propriétés suivantes:

- ou l'ensemble $\{t: \bar{\omega}(t) = \Delta\}$ est la demi-droite ouverte $]a, +\infty[$, et $\bar{\omega}$ est càdlàg sauf au point a ;
- ou $\bar{\omega}(t) = \Delta$ pour tout t , ce point est noté $[\Delta]$.

Le nombre a s'appelle la durée de vie et est noté $\zeta(\bar{\omega})$; nous prolongeons ζ au point $[\Delta]$ en posant $\zeta([\Delta]) = 0$. Nous noterons \bar{X}_t les applications coordonnées, $\bar{\mathbf{F}}_t^0$ la filtration engendrée par les applications coordonnées, $\bar{\mathbf{F}}_{t+}$ la filtration précédente rendue continue à droite, et $\bar{\mathbf{F}}_t^*$ la complétée par rapport à tous les ensembles négligeables pour toute loi P de $\bar{\mathbf{F}}_\infty^0 = \bar{\mathbf{F}}$.

Dans ce problème, une impulsion signifie non seulement changement d'état mais aussi changement de régime; par suite nous nous donnons un ensemble U sous-ensemble universellement mesurable d'un métrique compact, \mathbb{U} la tribu borélienne associée. L'ensemble U désignera l'ensemble des régimes permis. Comme précédemment nous adjoignons à U un point isolé noté ∂ ; à cet indice nous ferons correspondre la loi qui maintient le système en Δ .

Nous noterons \bar{U} l'espace $U \cup \{\partial\}$ et $\bar{\mathbb{U}}$ sa tribu associée.

L'espace canonique sera l'ensemble produit $W = (\bar{U} \times \bar{\Omega})^N$ muni de la tribu \mathbf{G}^* complétée universelle de la tribu $\mathbf{G} = \bigotimes_N (\bar{\mathbb{U}} \otimes \bar{\mathbf{F}}^*)$. Posons pour tout élément $w = (u_n, \bar{\omega}_n)$ de W :

$$\tau_{-1}(w) = 0, \quad \tau_0(w) = \zeta(\bar{\omega}_0), \quad \dots, \quad \tau_n(w) = \zeta(\bar{\omega}_0) + \dots + \zeta(\bar{\omega}_n), \dots$$

Ces instants représentent les instants d'impulsion canoniques du système. L'évolution du système impulsé est alors décrite naturellement sur l'espace canonique (W, \mathbf{G}^*) par

le processus limité à gauche et à droite (làdlàg)

$$(Y_t)_{t \geq 0} \text{ où: } Y_0(w) = \bar{X}_0(\bar{\omega}_0),$$

$$Y_t(w) = \begin{cases} \bar{X}_{t-\tau_n(w)}(\bar{\omega}_{n+1}) & \text{si } \tau_n(w) < t \leq \tau_{n+1}(w) < \infty, \\ \Delta & \text{si } t > \tau_n(w) \text{ pour tout } n. \end{cases}$$

De même l'évolution du régime au cours du temps est décrite par le processus càdlàg $(U_t)_{t \geq 0}$ où:

$$U_0(w) = u_0,$$

$$U_t(w) = \begin{cases} u_n & \text{si } \tau_n(w) < t \leq \tau_{n+1}(w) < \infty, \\ \partial & \text{si } t > \tau_n(w) \text{ pour tout } n. \end{cases}$$

Enfin si $\tau = \lim \nearrow \tau_n$, τ désigne le temps de mort du système.

Sur l'espace (W, \mathbf{G}^*) , nous pouvons construire plusieurs filtrations qui joueront un rôle différent et qu'il est important de distinguer. Soit \mathbf{G}^0 la sous-tribu de \mathbf{G}^* engendrée par les applications $U_t, Y_t, t \geq 0$:

a) Pour tout $t \geq 0$, \mathbf{G}_t est la sous-tribu de \mathbf{G} constituée des ensembles A de \mathbf{G}^0 vérifiant:

Pour tout $n \geq 0$ l'ensemble $A \cap (\tau_{n-1} < t \leq \tau_n)$ ne dépend que des $(n+1)$ premières coordonnées, à $(u_n, \bar{\omega}_n)$ fixés il appartient à $\otimes_n (\bar{\mathbf{U}} \otimes \bar{\mathbf{F}})$ et à $(u_k, \bar{\omega}_k)_{0 \leq k \leq n-1}$ fixés à $\bar{\mathbf{U}} \otimes \bar{\mathbf{F}}_{(t-\tau_{n-1})+}$, enfin $A \cap (\cap_k (\tau_k > t))$ est ou $\cap_k (\tau_k > t)$ ou vide; nous noterons pour tout $t \geq 0$ \mathbf{G}_t^a la complétée à l'aide des ensembles négligeables pour toute loi P de \mathbf{G} de la sous-tribu \mathbf{G}_t de \mathbf{G} .

On peut facilement établir en généralisant le test de Galmarino [5] à notre espace W une caractérisation des \mathbf{G}_t -temps d'arrêt et, puisque les τ_n sont des \mathbf{G}_t -temps d'arrêt, obtenir que tout élément de \mathbf{G}_{τ_n} ne dépend que des $(n+1)$ premières coordonnées, résultat utilisé constamment dans la suite de ce chapitre.

Enfin \mathbf{G}_t représente à tout instant t la connaissance de l'observateur (observation complète).

b) Pour tout $t \geq 0$, \mathbf{G}_t^u est la tribu $\mathbf{G}_t^a \vee \sigma(U_t)$.

c) \mathbf{G}_{t+} (resp. \mathbf{G}_t^*) représente la filtration \mathbf{G}_t (resp. \mathbf{G}_t^a) rendue continue à droite.

B. Nous allons maintenant construire les lois sur (W, \mathbf{G}^*) associées aux suites d'impulsion. Tout d'abord, afin de décrire l'évolution sans impulsion, nous introduisons l'ensemble $\Omega = D(R_+, E)$ des applications càdlàg de R_+ dans E , les applications coordonnées étant désignées par $(X_t)_{t \geq 0}$. Nous notons $(\mathbf{F}_t^0)_{t \geq 0}$ la famille croissante des tribus engendrées par les applications coordonnées, $(\mathbf{F}_t)_{t \geq 0}$ la filtration précédente rendue continue à droite, et $(\mathbf{F}_t^*)_{t \geq 0}$ la complétée par rapport à tous les ensembles négligeables pour toute loi P de \mathbf{F}_∞^0 . L'évolution du système sans impulsion est alors décrite sous le régime $u - u$ variant dans U —par la famille $P_{u,x}, x \in E$, telle que:

- i) $P_{u,x}$ soit une probabilité de transition de $(U \times E, \mathbf{U} \otimes \mathbf{E})$ vers (Ω, \mathbf{F}^*) .
- ii) Pour tout u appartenant à U , $X^u = (\Omega, X_t, \mathbf{F}_t^*, \mathbf{F}^*, (P_{u,x})_{x \in E})$ est un processus de Markov droit à durée de vie infinie.

Un contrôle sera alors décrit par deux suites:

- un système d'arrêt qui permettra de construire la loi des durées de vie canonique successives.
- un système de renaissance qui sera la suite des noyaux de renaissance.

Commençons par construire les systèmes d'arrêt à partir de la notion de noyau d'arrêt. Plus précisément étant donnée $\tilde{\mathbf{G}}^*$ sous-tribu de \mathbf{G}^* , nous appelons $\tilde{\mathbf{G}}^*$ noyau d'arrêt une application mesurable S de $(W \times \Omega, \tilde{\mathbf{G}}^* \otimes \mathbf{F}^*)$ dans (R_+, \mathbf{B}_{R_+}) telle que pour tout w fixé, $S(w, \cdot)$ soit un \mathbf{F}_{t+} temps d'arrêt.

DEFINITION 1. Nous appelons système d'arrêt toute famille $(S_n)_{n \geq 0}$ vérifiant:

- i) Pour tout $n \geq 0$, S_n est un $\mathbf{G}_{\tau_{n-1}}^u$ -noyau d'arrêt.
- ii) Pour tout $n \geq 0$, $S_n(w, \cdot)$ est identiquement nul si $u_n = \partial$.
- iii) Pour tout $n \geq 1$, $S_n(w, \cdot)$ est nul s'il existe $k \leq n - q$ pour lequel $\zeta(\bar{\omega}_k)$ est infini.
- iv) Pour tout $n \geq 1$, S_n est strictement positif excepté dans les cas considérés en ii) et iii).

Nous noterons \mathcal{S} l'ensemble de ces systèmes.

Remarque. L'introduction des tribus $\mathbf{G}_{\tau_n}^u$ provient du fait que pour décider la loi de la nouvelle durée de vie, l'opérateur connaît non seulement l'histoire du système jusqu'à l'impulsion, mais aussi le nouveau régime choisi.

D'autre part il est clair que si l'on connaît l'évolution du système jusqu'à l'instant τ_n , changement de régime à cet instant inclus, la politique d'arrêt n'est pas nécessairement markovienne (\mathbf{F}_t temps d'arrêt quelconque).

Passons maintenant à la construction des systèmes de renaissance, qui se fait à partir des lois de renaissance permises. Nous nous restreindrons, en vue d'avoir un modèle entre deux impulsions totalement markovien, aux familles de renaissance markoviennes, c'est-à-dire intuitivement à celles où la renaissance ne dépend que de l'état du système à l'instant où on l'a "tué". Plus précisément notons $\bar{U} \times \bar{E}$ l'ensemble $(U \times E) \cup \{\partial, \Delta\}$.

DEFINITION 2. Nous appelons maison de renaissance markovienne toute partie M de $\bar{U} \times \bar{E} \times P$ vérifiant:

Pour tout (u, x) , $\varepsilon_{(u,x)}$ appartient à $M_{u,x}$ et $M_{\partial,\Delta}$ est l'ensemble réduit à $\varepsilon_{(\partial,\Delta)}$ où $M_{u,x}$ (resp. $M_{\partial,\Delta}$) désigne la section en (u, x) (resp. (∂, Δ)) de M , et $\varepsilon_{(u,x)}$ (resp. $\varepsilon_{(\partial,\Delta)}$) désigne la mesure de Dirac en (u, x) (resp. (∂, Δ)).

Nous aurions pu adopter un système de renaissance différent à chaque instant d'impulsion; pour simplifier l'écriture nous adopterons le même système à ces instants. D'où la:

DEFINITION 3. Nous appelons système admissible de renaissance associé à la maison de renaissance markovienne R , tout élément $N = (N_n)$ vérifiant pour tout $n \geq 1$:

- i) N_n est une probabilité de transition de $(W, \mathbf{G}_{\tau_{n-1}}^a)$ vers $(\bar{U} \times \bar{E}, \bar{\mathbf{U}} \times \bar{\mathbf{E}})$.
- ii) Pour tout $w \in W$, $N_n(w, \cdot)$ appartient à $R_{u_{n-1}, Y_{\tau_{n-1}}(w)}$ si $\tau_n(w)$ est fini, et est égale à $\varepsilon_{(\partial,\Delta)}$ sinon.

Nous noterons \mathcal{N} l'ensemble de ces systèmes.

Etant donné un couple $\delta = (S, N)$ appartenant à $\mathcal{S} \times \mathcal{N}$, nous allons maintenant nous intéresser à la construction de la probabilité $P_{u,x}^\delta$ sur (W, \mathbf{G}^*) qui sera sous δ la loi du processus $(U_n, Y_t)_{t \geq 0}$ avec état initial (u, x) .

Donnons nous une stratégie $\delta = (S_0, N_1, S_1, \dots)$ appartenant à $\mathcal{S} \times \mathcal{N}$ et (u, x) appartenant à $U \times E$, et considérons la famille H_n^δ , $n \geq 0$, telle que:

- i) H_0^δ soit la probabilité sur $(\bar{U} \times \bar{\Omega}, \bar{\mathbf{U}} \otimes \bar{\mathbf{F}}^*)$ égale à $\varepsilon_u \otimes P_{u,x}^{S_0(u, \cdot)}$ (où $P_{u,x}^T$ désigne la loi du processus X_n tué en T avec départ x et régime u).
- ii) Pour tout $n \geq 1$, H_n^δ soit la probabilité de transition de $(W, \mathbf{G}_{\tau_{n-1}}^a)$ vers $(\bar{U} \times \bar{\Omega}, \bar{\mathbf{U}} \otimes \bar{\mathbf{F}}^*)$ égale à:

$$(1) \quad H_n^\delta(w, \Gamma) = \int_{\bar{U} \times \bar{E}} N_n(w, d(v, y)) \int_{\bar{\Omega}} P_{u,y}^{S_n(w, \cdot)}(d\bar{\omega}) \mathbb{1}_\Gamma(v, \bar{\omega}).$$

Grâce au théorème de Ionescu-Tulcea [15] nous pouvons "recoller" ces lois afin de décrire la loi de $(U_n, Y_t)_{t \geq 0}$ soumis au contrôle δ à savoir:

THÉORÈME 4. Pour toute stratégie δ de $\underline{S} \times \underline{N}$ et tout couple (u, x) il existe une unique probabilité $P_{u,x}^\delta$ sur (W, \mathbf{G}^*) dont la valeur sur tout pavé mesurable $\prod_n \Gamma_n$ est:

$$(2) \quad P_{u,x}^\delta \left(\prod_n \Gamma_n \right) = \int_{\Gamma_0(u, \cdot)} H_0^\delta(d(u_0, \bar{\omega}_0)) \int_{\Gamma_0(u_1, \cdot)} \dots \int_{\Gamma_N(u_N, \cdot)} H_N^\delta(u_0, \bar{\omega}_0, \dots, u_{N-1}, \bar{\omega}_{N-1}, d(u_N, \bar{\omega}_N))$$

où N est tel que pour tout $p > N$, $\Gamma_p = \bar{U} \times \bar{\Omega}$.

Comme il est naturel de supposer que dans tout intervalle de temps fini il n'y ait qu'un nombre fini d'actions, nous sommes amenés à restreindre la classe $\underline{S} \times \underline{N}$ de la façon suivante:

DEFINITION 5. Nous appelons stratégie admissible tout élément δ de $\underline{S} \times \underline{N}$ satisfaisant à:

$$(3) \quad P_{u^0, x^0}^\delta(\tau_n < \tau < \infty \forall n) = 0.$$

Nous noterons \underline{D} l'ensemble des stratégies admissibles.

Remarque. Les résultats connus d'existence aboutissent toujours à une stratégie optimale à renaissance "déterministe". Nous avons toutefois choisi d'admettre la possibilité de renaissances "randomisées", pour pouvoir obtenir dans toute sa généralité notre résultat essentiel, à savoir la forme markovienne de la fonction de valeur de notre problème de contrôle.

Soient f et c deux fonctions boréliennes positives définies respectivement sur $(\overline{U \times E}, \overline{U \otimes E})$ et sur $((U \times E)^2, (U \otimes E)^2)$, à chaque stratégie admissible δ nous associons un coût moyen égal à:

$$(4) \quad K^\delta = E_{u^0, x^0}^\delta \left(\int_0^\tau e^{-\alpha t} f(U_t, Y_t) dt + \sum_n e^{-\alpha \tau_n} c(U_{\tau_n}, Y_{\tau_n}, U_{\tau_n}^+, Y_{\tau_n}^+) \right) = E_{u^0, x^0}^\delta(k),$$

où f et c sont telles que:

$$(5) \quad f(\partial, \Delta) = c(\partial, \Delta, u, x) = 0 \quad \forall (u, x),$$

$$(6) \quad c(u, x, u, x) = 0 \quad \forall (u, x)$$

où (5) exprime que le système arrivé au cimetière n'évolue plus, et (6) que s'il n'y a pas d'impulsion il ne peut y avoir coût.

Dans toute la suite, pour simplifier l'écriture, nous omettrons les indices u^0, x^0 dans $P_{u^0, x^0}^\delta = P^\delta$.

2. Propriétés du modèle.

A. Comme dans de nombreux problèmes de contrôle stochastique pour accéder à un principe d'optimalité du type Bellman il est nécessaire de pouvoir arrêter l'action d'une stratégie à un instant quelconque et de laisser ensuite le système évoluer librement. Par suite nous allons construire ce que nous appellerons "les contrôles arrêtés" et ensuite vérifier que cette opération ne fait pas sortir de la classe des contrôles admissibles: c'est la notion de stabilité par arrêt.

Etant donnés une stratégie admissible δ et un \mathbf{G}_t temps d'arrêt T , nous allons annuler l'influence de δ sur le système à partir de l'instant T . Pour le choix de cette nouvelle stratégie nous sommes guidés par l'intuition: sur toutes les trajectoires à durée de vie infinie qui se comportent jusqu'en S_0 comme les trajectoires tuées en S_0

nous choisissons d'impulser normalement en τ_0 sur $(T > \tau_0)$, d'annuler l'impulsion et ne plus impulser sur le complémentaire. Aussi est-il nécessaire d'introduire les opérateurs de meurtre.

DEFINITION 6. Soit S un \mathbf{F}_{t+} -temps d'arrêt, nous appelons opérateur de meurtre en S l'application m_S de Ω dans $\bar{\Omega}$ définie pour tout ω de $\bar{\Omega}$ par:

$$m_S(\omega) = \begin{cases} \omega_t & \text{si } t \leq S(\omega), \\ \Delta & \text{si } t > S(\omega). \end{cases}$$

Il est alors facile de remarquer que m_S est mesurable de $(\Omega, \mathbf{F}_{S+})$ dans $(\bar{\Omega}, \bar{\mathbf{F}})$, et nous pouvons définir les contrôles arrêtés:

DEFINITION 7. Soient T un \mathbf{G}_t -temps d'arrêt et $\delta = ((S_n)_{n \geq 0}, (N_n)_{n \geq 1})$ une stratégie admissible, nous appelons contrôle arrêté en T (resp. à droite en T) la famille $\delta^T = ((S_n^T)_{n \geq 0}, (N_n^T)_{n \geq 1})$ (resp. $\delta^{T+} = ((S_n^{T+})_{n \geq 0}, (N_n^{T+})_{n \geq 1})$ où pour tout $n \geq 0$:

$$(7) \quad S_n^T = \begin{cases} S_n & \text{sur } m_{S_n}^{-1}(\tau_n < T), \\ +\infty & \text{sur } m_{S_n}^{-1}(\tau_{n-1} < T \leq \tau_n), \\ 0 & \text{ailleurs} \end{cases}$$

$$\left(\text{resp. } S_n^{T+} = \begin{cases} S_n & \text{sur } m_{S_n}^{-1}(\tau_n \leq T), \\ +\infty & \text{sur } m_{S_n}^{-1}(\tau_{n-1} \leq T < \tau_n) \cap (S_n \neq 0) \\ 0 & \text{ailleurs} \end{cases} \right)$$

et pour tout $n \geq 1$:

$$(8) \quad N_n^T = \begin{cases} N_n & \text{sur } (\tau_{n-1} < T), \\ \varepsilon_{\{u_{n-1}, \bar{X}_t(\bar{\omega}_{n-1})\}} & \text{sur } (T = \tau_{n-1}), \\ \varepsilon_{\{\partial, \Delta\}} & \text{sur } (T < \tau_{n-1}) \end{cases}$$

$$\left(\text{resp. } N_n^{T+} = \begin{cases} N_n & \text{sur } (\tau_{n-1} \leq T) \\ \varepsilon_{\{\partial, \Delta\}} & \text{sur } (T < \tau_{n-1}) \end{cases} \right).$$

Nous établissons que les opérations d'arrêt ne font pas sortir de la classe des stratégies admissibles. Ce résultat repose essentiellement sur le lemme technique suivant, dont la preuve (facile) peut être trouvée dans [9] ou [12]:

LEMME 8. Soit δ une stratégie et T un \mathbf{G}_t -temps d'arrêt, pour tout Γ appartenant à \mathbf{G}^* nous avons:

$$(9) \quad P^{\delta^T} \left(\Gamma \cap \left(\bigcap_k (\tau_k < T) \right) \right) = P^\delta \left(\Gamma \cap \bigcap_k (\tau_k < T) \right)$$

et

$$(10) \quad P^{\delta^{T+}} (\Gamma \cap (\tau \leq T)) = P^\delta (\Gamma \cap (\tau \leq T)).$$

D'où la:

PROPOSITION 9. Pour tout \mathbf{G}_t -temps d'arrêt T , si la stratégie δ est admissible, alors les stratégies δ^T et δ^{T+} sont admissibles.

Preuve. En appliquant (9) à l'ensemble $\Gamma = \bigcap_k (\tau_k < \tau)$ nous obtenons:

$$(11) \quad P^{\delta^T} \left(\bigcap_n (T > \tau_n) \cap \bigcap_k (\tau_k < \tau) \right) = P^\delta \left(\bigcap_n (T > \tau_n) \cap \bigcap_k (\tau_k < \tau) \right).$$

D'autre part par construction du contrôle arrêté, il est facile de voir que pour tout $n \geq 0$:

$$(12) \quad P^{\delta^T} \left((T \leq \tau_n) \cap_k (\tau_k \leq \tau) \right) = 0, \quad \text{soit en sommant sur } n:$$

$$P^{\delta^T} \left(\bigcup_n (T \leq \tau_n) \cap_k (\tau_k < \tau) \right) = 0.$$

En combinant (11) et (12) nous obtenons le résultat cherché pour δ^T . Le résultat pour δ^{T+} s'obtient de la même manière à partir de (10).

B. La preuve des résultats suivants s'obtient facilement à partir de la construction de P^δ et des contrôles arrêtés. Nous obtenons en particulier une propriété de filtration décroissante (proposition 11) nécessaire en vue du critère d'optimalité [17].

PROPOSITION 10. *Le système est compatible avec les opérations d'arrêt, c'est-à-dire que pour toute stratégie admissible δ et tout \mathbf{G}_t -temps d'arrêt T nous avons:*

$$(13) \quad P^{\delta^T}(\Gamma) = P^\delta(\Gamma) \quad \forall \Gamma \in \mathbf{G}_T^a$$

et

$$(14) \quad P\delta^{T+}(\Gamma) = P^\delta(\Gamma) \quad \forall \Gamma \in \mathbf{G}_T^*.$$

PROPOSITION 11. *Etant donnés T un \mathbf{G}_t -temps d'arrêt et deux stratégies admissibles δ_1, δ_2 telles que δ_1^T soit égale à δ_2^T (resp. δ_1^{T+} à δ_2^{T+}), pour toute fonction positive \mathbf{G}^* -mesurable ϕ il existe une stratégie admissible δ telle que:*

$$(15) \quad \delta^T = \delta_1^T = \delta_2^T \quad (\text{resp. } \delta^{T+} = \delta_1^{T+} = \delta_2^{T+}),$$

$$E^\delta(\phi / \mathbf{G}_T^a) = E^{\delta_1}(\phi / \mathbf{G}_T^a) \wedge E^{\delta_2}(\phi / \mathbf{G}_T^a) \quad P^{\delta^T} p.s.$$

$$(16) \quad (\text{resp. } E^\delta(\phi / \mathbf{G}_T^*) = E^{\delta_1}(\phi / \mathbf{G}_T^*) \wedge E^{\delta_2}(\phi / \mathbf{G}_T^*) \quad P^{\delta^{T+}} p.s.).$$

3. Critères d'optimalité. Le problème consiste à prouver l'existence d'une stratégie admissible $\hat{\delta}$ qui minimise le coût moyen K^δ . La stratégie $\hat{\delta}$ sera dite optimale dans la classe \underline{D} des stratégies admissibles.

Grâce à la forme particulière du coût k , nous pouvons introduire les coûts minimaux suivants:

DEFINITION 12. Pour toute stratégie admissible δ et pour tout \mathbf{G}_t -temps d'arrêt T nous appelons coût minimal conditionnel après T (resp. à droite après T) la variable aléatoire \mathbf{G}_T^a (resp. \mathbf{G}_T^*)-mesurable définie par:

$$W_T^\delta = P^\delta - \text{ess inf}_{\mu^T = \delta^T} E^\mu \left[\int_{T \wedge \tau}^\tau e^{-\alpha t} f(U_t, Y_t) dt \right. \\ \left. + \sum_n \mathbb{1}_{(\tau_n \geq T)} e^{-\alpha \tau_n} c(U_{\tau_n}, Y_{\tau_n}, U_{\tau_n}^+, Y_{\tau_n}^+) / \mathbf{G}_T^a \right] \\ \left(\text{resp. } W_T^{\delta+} = P^\delta - \text{ess inf}_{\mu^{T+} = \delta^{T+}} E^\mu \left[\int_{T \wedge \tau}^\tau e^{-\alpha t} f(U_t, Y_t) dt \right. \right. \\ \left. \left. + \sum_n \mathbb{1}_{(\tau_n > T)} e^{-\alpha \tau_n} c(U_{\tau_n}, Y_{\tau_n}, U_{\tau_n}^+, Y_{\tau_n}^+) / \mathbf{G}_T^* \right] \right).$$

La propriété de filtration décroissante de la proposition 11 permettant d'inverser ess inf. et espérance conditionnelle, nous amène facilement au principe de la programmation dynamique [17].

THÉORÈME 13. *Pour toute stratégie δ et tout couple (S, T) de \mathbf{G}_t -temps d'arrêt, $S \leq T$, nous avons P^δ p.s.:*

$$(17) \quad W_S^\delta \leq E^\delta \left(\int_{S \wedge \tau}^{T \wedge \tau} e^{-\alpha t} f(U_t, Y_t) dt + \sum_n \mathbb{1}_{(S \leq \tau_n < T)} e^{-\alpha \tau_n} c(U_{\tau_n}, Y_{\tau_n}, U_{\tau_n}^+, Y_{\tau_n}^+) / \mathbf{G}_S^a \right) + E^\delta(W_S^\delta / \mathbf{G}_S^a).$$

De plus δ est optimale si et seulement si l'égalité a lieu P^δ p.s. pour tout couple (S, T) .

Afin d'obtenir un critère plus "performant" nous allons examiner ce qui se passe entre deux instants de changement de régime et d'état et en un instant de ce type; d'où notre second critère d'optimalité dont la preuve est identique à celle du théorème III.2.6 de [11].

THÉORÈME 14. *Pour tout contrôle admissible δ nous avons P^δ p.s. les inégalités:*

$$(18) \quad W_0^\delta \leq E^\delta \left(\int_0^{\tau_0} e^{-\alpha s} f(U_s, Y_s) ds / \mathbf{G}_0^a \right) + E^\delta(W_{\tau_0}^\delta / \mathbf{G}_0^a),$$

pour tout $n \geq 0$:

$$(19) \quad W_{\tau_n}^\delta \leq e^{-\alpha \tau_n} \int_{U \times E} c(U_{\tau_n}, Y_{\tau_n}, u, x) N_{n+1}^\delta(\cdot, d(u, x)) + E^\delta(W_{\tau_n}^{\delta+} / \mathbf{G}_{\tau_n}^a),$$

$$(17') \quad W_{\tau_n}^{\delta+} \leq E^\delta \left(\int_{\tau_n}^{\tau_{n+1}} e^{-\alpha s} f(U_s, Y_s) ds / \mathbf{G}_{\tau_n}^* \right) + E^\delta(W_{\tau_{n+1}}^\delta / \mathbf{G}_{\tau_n}^*).$$

De plus δ est optimale si et seulement si l'égalité a lieu simultanément dans (18), (19), (17') pour tout $n \geq 0$.

4. Etude des coûts conditionnels.

A. Le critère d'optimalité—théorème 14—tel qu'il est présenté est insuffisant pour aider à la construction d'une stratégie, car les variables qui interviennent— $W_{\tau_n}^\delta$ et $W_{\tau_n}^{\delta+}$ —dépendent de la stratégie admissible δ d'une manière inconnue. Il nous faut donc étudier la dépendance en δ de ces coûts conditionnels. Or remarquons le caractère markovien et homogène entre deux instants de mort du système, ainsi que la forme markovienne de chaque renaissance. Nous pouvons donc espérer obtenir que les coûts conditionnels ne dépendent que de l'état de système à l'instant du conditionnement. Plus précisément si nous posons:

$$\rho(u, x) = \inf_{\delta \in \mathcal{D}} E_{u,x}^\delta(k) \quad \text{et} \quad \rho^+(u, x) = \inf_{\{\delta: S_0^\delta \neq 0\}} E_{u,x}^\delta(k),$$

nous établirons P^δ p.s.:

$$(20) \quad W_{\tau_n}^\delta = e^{-\alpha \tau_n} \rho(U_{\tau_n}, Y_{\tau_n})$$

et

$$(21) \quad W_{\tau_n}^{\delta+} = e^{-\alpha \tau_n} \rho^+(U_{\tau_n}^+, Y_{\tau_n}^+).$$

Ces réflexions nous amènent à l'étude des propriétés des applications ρ et ρ^+ , en particulier leur mesurabilité. Pour cela il est naturel d'introduire par exemple les applications ρ_n^+ approximant ρ^+ :

$$(22) \quad \rho_n^+(u, x) = \inf_{S_0^\delta \neq 0, S_n^\delta = \{+\infty\}} E_{u,x}^\delta(k)$$

où $S_n^\delta = \{+\infty\}$ signifie que $\rho_n(x)$ est la valeur sur les contrôles qui sautent *au plus* n fois.

Par analogie avec les méthodes introduites par M. Robin dans sa thèse [16], il est raisonnable d'espérer que les applications ρ_n^+ soient liées par la relation de récurrence:

$$(23) \quad \rho_n^+(u, x) = \inf_{s \geq 0} E_{u,x} \left(\int_0^s e^{-\alpha s} f(u, X_s) ds + e^{-\alpha s} m \rho_{n-1}^+(u, X_s) \right)$$

où pour toute fonction borélienne ϕ :

$$(24) \quad m\phi(u, x) = \inf_{v \in M(u,x)} \int_{U \times E} \nu(d(v, y)) (c(u, x, v, y) + \phi(v, y)),$$

d'où l'étude en premier lieu des propriétés de l'arrêt optimal, de l'opération m et du procédé itératif de ces opérations.

Commentaire. 1) La méthode qui consiste à percevoir le contrôle impulsif comme une suite de problèmes d'arrêt optimaux apparaît clairement pour la première fois dans la thèse de M. Robin [16].

2) Notre modèle étant complètement canonique, il n'est pas surprenant de voir apparaître dans les égalités (22) et (23) l'indépendance P^δ p.s. en δ des coûts conditionnels $W_{\tau_n}^\delta$ et $W_{\tau_{n+1}}^\delta$.

B. Pour tout u de U le processus de Markov $X^u = (\Omega, X_t, \mathbf{F}_t^*, \mathbf{F}^*, P_{u,x}, x \in E)$ est un processus droit très général. Nous ne pouvons pas directement utiliser les résultats de l'arrêt optimal sur les processus droits en raison de l'introduction du paramètre u , car nous désirons une mesurabilité en u . Nous allons utiliser un artifice de calcul qui consiste à regarder la famille de semi-groupe $(P_t^u)_{t \geq 0}$ indexée par u comme un semi-groupe sur $(U \times E, \mathbf{U} \times \mathbf{E}^*)$, que nous notons $(\tilde{P}_t)_{t \geq 0}$, tel que pour toute fonction ϕ universellement mesurable définie sur $U \times E$

$$(25) \quad \tilde{P}_t \phi(u, x) = P_t^u \phi_u(x).$$

Soient $\tilde{\Omega}$ l'ensemble des applications càdlàg de \mathbb{R}_+ dans $U \times E$, $(\tilde{U}_t, \tilde{X}_t)_{t \geq 0}$ les applications coordonnées et $(\mathbf{H}_t^*)_{t \geq 0}$ la filtration engendrée par celles-ci, rendue continue à droite et complétée par rapport à tous les ensembles négligeables pour toute loi P de $\mathbf{H}_\infty = \sigma(\tilde{U}_t, \tilde{X}_t, t \geq 0)$, alors:

LEMME 15. *Pour toute loi μ sur $(U \times E, \mathbf{U} \times \mathbf{E})$ il existe une unique probabilité sur $(\tilde{\Omega}, \mathbf{H}^*)$, notée \tilde{P}_μ , rendant constante en t la première application coordonnée telle que le processus $(\tilde{U}_t, \tilde{X}_t, \mathbf{H}_t^*, \mathbf{H}^*, \tilde{P}_\mu)$ soit un processus de Markov droit de semi-groupe de transition $(\tilde{P}_t)_{t \geq 0}$ et de loi initiale μ . De plus, si ϕ est une fonction β -excessive, l'application $\phi(\tilde{U}_t, \tilde{X}_t)$ est \tilde{P}_μ p.s. continue à droite en t .*

Preuve. Considérons la probabilité \tilde{P}_μ définie sur $(\tilde{\Omega}, \mathbf{H}^*)$ par

$$\begin{aligned} \tilde{P}_\mu(\tilde{U}_{t_1} \in B_1, \tilde{X}_{t_1} \in \Gamma_1, \dots, \tilde{U}_{t_n} \in B_n, \tilde{X}_{t_n} \in \Gamma_n) \\ = \int_{U \times E} \mu(d(u, x)) \mathbb{1}_{B_1 \cap \dots \cap B_n}(u) P_{u,x}(X_{t_1} \in \Gamma_1, \dots, X_{t_n} \in \Gamma_n) \end{aligned}$$

pour tout n , tout B_i de \mathbf{U} et tout Γ_i de \mathbf{E} , $i = 1, \dots, n$; alors il est facile de voir que la probabilité \tilde{P}_μ satisfait aux conditions du lemme grâce aux hypothèses droites des processus X^u , u parcourant U .

Commentaire. 1) L'intérêt de l'emploi de l'artifice précédent réside dans le fait que le processus $(\tilde{U}_t, \tilde{X}_t, \mathbf{H}_t^*, \mathbf{H}^*, \tilde{P}_\mu)$ est un processus de Markov droit et donc possède toutes les propriétés de ces processus, en particulier nous pourrions appliquer les résultats connus sur l'arrêt optimal, et que d'autre part la probabilité \tilde{P}_μ ne charge que les points \tilde{u} tels que $\tilde{U}_t(\tilde{u}) = u$ pour tout t , ce qui permettra de nous ramener immédiatement aux seuls espaces introduits dans notre modèle.

2) Pour toute précision sur les processus droits, ainsi que sur la compactification de Ray–Knight, sur les processus de Ray et sur le lien entre ces deux types de processus markoviens, il est conseillé de consulter le remarquable livre de R. K. Gettoor [8].

Avant d'énoncer les résultats sur l'arrêt optimal, il nous faut faire deux remarques dont nous devons absolument tenir compte: le problème du contrôle impulsif se présente comme une itération de problèmes d'arrêt optimal, et donc nous devons avoir une mesurabilité de l'enveloppe de Snell suffisante, car par exemple l'universelle mesurabilité nous empêcherait toute réitération; d'autre part notre problème se présente comme un inf sur une certaine famille, alors que l'arrêt optimal est un sup sur une famille de temps d'arrêt, ce qui nous oblige à énoncer les résultats pour des fonctions de signe quelconque. La mesurabilité adaptée à notre problème est la suivante:

DEFINITION 16. Nous dirons qu'une fonction v minorée définie sur $U \times E$ est Ray-analytique si elle est la trace sur $U \times E$ d'une fonction \tilde{v} analytique sur le compactifié de Ray–Knight $\tilde{U} \times \tilde{E}$.

Dans la suite nous élargirons, si nécessaire, la définition habituelle des fonctions analytiques [4], [5] aux fonctions de signe quelconque minorées en disant que v est analytique si l'ensemble $(v > a)$ est analytique pour tout a réel. Rappelons que toute fonction borélienne est analytique et toute fonction analytique universellement mesurable.

THÉORÈME 17. Pour toute fonction v positive Ray-analytique la fonction

$$(26) \quad qv(u, x) = \sup_{T \in \tilde{T}} \tilde{E}_{u,x}(e^{-\alpha T} v(u, \tilde{X}_T))$$

est Ray-analytique et α -fortement surmédiane, où \tilde{T} est l'ensemble des \mathbf{H}_t^* -temps d'arrêt. Sa régularisée α -excessive \bar{q} est Ray-analytique et presque borélienne.

De plus pour toute loi μ sur $(U \times E, \mathbf{U} \times \mathbf{E})$ et tout \mathbf{H}_t^* -temps d'arrêt T :

$$(27) \quad \bar{q}v(\tilde{U}_0, \tilde{X}_T) = \tilde{P}_\mu - \text{ess sup}_{\substack{S > T \\ S \in \tilde{T}}} \tilde{E}_\mu(e^{-\alpha S} v(\tilde{U}_0, \tilde{X}_S) / \mathbf{H}_T^*) \quad \tilde{P}_\mu p.s.$$

Transposons ces résultats sur Ω en tenant compte du commentaire 1 précédent. Nous pouvons restreindre le sup dans (26) et (27) à la classe des \mathbf{U} -noyaux d'arrêt, que nous noterons \underline{S}_{-1} , c'est à dire aux applications T de $(U \times \Omega, \mathbf{U} \times \mathbf{F}^*)$ dans $(\bar{\mathbb{R}}_+, B_{\bar{\mathbb{R}}_+})$ telle que pour tout u fixé $T(u, \cdot)$ soit un \mathbf{F}_t^* -temps d'arrêt, d'où:

COROLLAIRE 18. Pour tout (u, x) de $U \times E$

$$(28) \quad qv(u, x) = \sup_{T \in \underline{S}_{-1}} E_{u,x}(e^{-\alpha T(u, \cdot)} v(u, X_{T(u, \cdot)})),$$

et pour tout noyau d'arrêt de \underline{S}_{-1} , toute loi μ sur $(U \times E, \mathbf{U} \times \mathbf{E})$ nous avons $P_\mu p.s.$:

$$(29) \quad \bar{q}v(U_0, X_{T(U_0, \cdot)}) = P_\mu - \text{ess sup}_{\substack{S > T \\ S \in \underline{S}_{-1}}} E_\mu(e^{-\alpha S(U_0, \cdot)} v(U_0, X_{S(U_0, \cdot)}) / (\mathbf{U} \times \mathbf{F}^*)_T).$$

Commentaire. 1) Dans l'expression des égalités (26) et (27) nous avons pris en considération le fait que \tilde{P}_μ rend constante en t l'application coordonnée \tilde{U}_t .

2) L'ensemble des résultats du théorème sont établis dans le cours de N. El Karoui [6]. Ceux du corollaire sont aussi établis et utilisés dans [7].

3) Les propriétés de mesurabilité et les égalités (26) à (29) sont encore satisfaites pour toute fonction v Ray-analytique de signe quelconque.

4) La variable aléatoire U_0 qui apparaît dans (29) est l'application identité de U dans U .

Pour terminer l'étude préliminaire, il nous faut établir des propriétés de mesurabilité similaires pour l'opération m . Nous avons besoin d'une hypothèse supplémentaire sur le système de renaissance, qui nous permettra d'appliquer les propriétés des projections des fonctions analytiques.

HYPOTHÈSE 2. Posons $M^* = \{(u, x, \mu): \mu(\{u, x\}) < 1\}$, alors M^* est un borélien de $U \times E \times \mathbb{P}$; de plus il existe un compact \mathbb{K} dans l'ensemble des probabilités sur $(\bar{U} \times \bar{E}, \bar{U} \times \bar{E})$ muni de la convergence étroite, tel que pour tout (u, x) appartenant à $U \times E$:

$$M_{\{u, x\}}^* \subset \mathbb{K}.$$

Posons pour toute fonction ϕ universellement mesurable minorée:

$$\bar{m}\phi(u, x) = \sup_{\nu \in M_{\{u, x\}}} \int_{U \times E} \nu(d(v, y))(\phi(v, y) - c(u, x, v, y))$$

alors:

PROPOSITION 19. L'application $\bar{m}\phi$ est Ray-analytique dès que ϕ l'est. De plus pour toute probabilité μ sur $\bar{U} \times \bar{E}$ et tout $\varepsilon > 0$ il existe un noyau permis $r(u, x)$ borélien tel que:

$$(30) \quad \bar{m}\phi(u, x) \leq \int_{U \times E} r(u, x; d(v, y))(\phi(v, y) - c(u, x, v, y)) + \varepsilon.$$

Preuve. ϕ étant Ray-analytique, elle peut être prolongée sur $(\bar{U} \times \bar{E}) \cup \{\partial, \Delta\}$ en une application $\tilde{\phi}$ analytique, de même $c(u, x, v, y)$ peut-être prolongée en une application $\tilde{c}(u, x, v, y)$ borélienne sur $(\bar{U} \times \bar{E} \times U \times E)$, enfin M^* borélien de $U \times E \times \mathbb{P}$ peut lui aussi être prolongé en un borélien \tilde{M} de $(\bar{U} \times \bar{E}) \times \mathbb{P}$ inclus dans $(\bar{U} \times \bar{E}) \times \mathbb{K}$ grâce à l'hypothèse 2, où $\bar{U} \times \bar{E}$ est le compactifié de Ray-Knight de $U \times E$. Alors l'application:

$$m\tilde{\phi}(u, x) = \left(\sup_{\nu \in \mathbb{K}} \int_{\tilde{M}} \nu(d(v, y))(\tilde{\phi}(v, y) - \tilde{c}(u, x, v, y)) \right) \vee \tilde{\phi}(u, x)$$

admet comme restriction sur $U \times E$ l'application $\bar{m}\phi$. L'application qui à ν, u, x fait correspondre $\int_{U \times E} \nu(d(v, y))(\phi(v, y) - c(u, x, v, y))$ est analytique ([13] lemme 1). $\bar{m}\phi$ est donc la projection sur $U \times E$ d'une fonction analytique définie sur le compact $\bar{U} \times \bar{E} \times \mathbb{K}$ (hypothèse 2), par suite elle est analytique ([4]); nous en déduisons immédiatement que $\bar{m}\phi$ est Ray-analytique.

La démonstration de l'existence d'un noyau permis est identique à celle de P. A. Meyer ([13] lemme 2).

Commentaire. 1) L'opération m est l'inf d'une certaine famille alors que les propriétés utilisées sont relatives à des sup—par exemple la projection d'une fonction analytique est analytique—d'où l'introduction de l'opérateur \bar{m} .

2) Pour appliquer la propriété relative à la projection des fonctions analytiques nous avons besoin que les lois ν permises appartiennent à un compact, ce qui a nécessité l'introduction de l'hypothèse 2.

C. Revenons à l'étude du contrôle impulsif. Nous pouvons établir un résultat important à partir des propriétés énoncées dans la partie B, résultat qui traduit le lien entre l'arrêt optimal et le contrôle impulsif.

THÉORÈME 20. Pour toute application ϕ telle que— ϕ soit Ray-analytique, l'application— $J\phi$ est elle-même Ray-analytique, où

$$(31) \quad J\phi(u, x) = \inf_{\substack{T > 0 \\ T \in \mathcal{S}_{-1}}} E_{u, x} \left(\int_0^{T(u, \cdot)} e^{-\alpha s} f(u, X_s) ds + e^{-\alpha T(u, \cdot)} m\phi(u, X_{T(u, \cdot)}) \right).$$

De plus pour tout couple (u, x) de $U \times E$ et toute stratégie admissible δ nous avons P^δ p.s.:

$$(32) \quad \begin{aligned} e^{-\alpha\tau_n} J\phi(U_{\tau_n}^+, Y_{\tau_n}^+) &= P_{u,x}^\delta - \operatorname{ess\,inf}_{\mu_{\tau_n}^+ = \delta_{\tau_n}^+} E_{u,x}^\mu \left(\int_{\tau_n}^{\tau_{n+1}} e^{-\alpha s} f(U_s, Y_s) \, ds \right. \\ &\quad \left. + e^{-\alpha\tau_{n+1}} (c(U_{\tau_{n+1}}, Y_{\tau_{n+1}}, U_{\tau_{n+1}}^+, Y_{\tau_{n+1}}^+) \right. \\ &\quad \left. + \phi(U_{\tau_{n+1}}^+, Y_{\tau_{n+1}}^+)) / \mathbf{G}_{\tau_n}^* \right). \end{aligned}$$

Preuve. Nous pouvons écrire:

$$J\phi(u, x) = - \sup_{\substack{T > 0 \\ T \in \mathcal{S}_{-1}}} E_{u,x} (e^{-\alpha T(u, \cdot)} (U^\alpha f(u, X_{T(u, \cdot)}) - m\phi(u, X_{T(u, \cdot)}))) - U^\alpha f(u, x)).$$

La fonction $U^\alpha f(u, x)$ est borélienne, donc est la trace d'une fonction borélienne sur le compactifié de Ray-Knight $\widetilde{U \times E}$ ([8] proposition 11-3) et par suite elle est Ray-analytique. Nous en déduisons que l'application $J\phi$ est Ray-analytique grâce au théorème 17, à son corollaire et à la proposition 19.

Considérons une stratégie admissible $\mu = ((S_n)_{n \geq 0}, (N_n)_{n \geq 1})$ vérifiant $\mu_{\tau_n}^+ = \delta_{\tau_n}^+$. Par définition de $J\phi$ nous avons sur $(\tau_n < \tau)$:

$$\begin{aligned} J\phi(U_{\tau_n}^+, Y_{\tau_n}^+) &\leq E_{U_{\tau_n}^+, Y_{\tau_n}^+} \left(\int_0^{S_{n+1}(\cdot)} e^{-\alpha s} f(U_{\tau_n}^+, X_s) \, ds \right. \\ &\quad \left. + e^{-\alpha S_{n+1}(\cdot)} \int_{U \times E} N_{n+2}(\cdot, d(u, x)) \right. \\ &\quad \left. \times (c(U_{\tau_n}^+, X_{S_{n+1}(\cdot)}, u, x) + \phi(u, x)) \right) \end{aligned}$$

soit:

$$\begin{aligned} J\phi(U_{\tau_n}^+, Y_{\tau_n}^+) &\leq E_{U_{\tau_n}^+, Y_{\tau_n}^+} \left(\left(\int_0^\zeta e^{-\alpha s} f(U_{\tau_n}^+, \bar{X}_s) \, ds \right. \right. \\ &\quad \left. \left. + e^{-\alpha \zeta} \int_{U \times E} N_{n+2}(\cdot, d(u, x)) (c(U_{\tau_n}^+, \bar{X}_\zeta, u, x) + \phi(u, x)) \right) \right); \end{aligned}$$

appliquant la construction de P^μ , nous obtenons: $P_{u,x}^\mu$ p.s., soit encore $P_{u,x}^\delta$ p.s. (proposition 10):

$$\begin{aligned} e^{-\alpha\tau_n} J\phi(U_{\tau_n}^+, Y_{\tau_n}^+) &\leq E_{u,x}^\mu \left(\int_{\tau_n}^{\tau_{n+1}} e^{-\alpha s} f(U_s, Y_s) \, ds \right. \\ &\quad \left. + e^{-\alpha\tau_{n+1}} (c(U_{\tau_{n+1}}, Y_{\tau_{n+1}}, U_{\tau_{n+1}}^+, Y_{\tau_{n+1}}^+) \right. \\ &\quad \left. + \phi(U_{\tau_{n+1}}^+, Y_{\tau_{n+1}}^+)) / \mathbf{G}_{\tau_n}^* \right). \end{aligned}$$

Remarquant que les deux membres de l'inégalité précédente sont nuls sur $(\tau_n = \tau)$ (puisque nous prolongeons toute fonction définie sur $U \times E$ à $\widetilde{U \times E}$ en posant

$\phi(\partial, \Delta) = 0$), nous obtenons P^δ p.s.:

$$(33) \quad e^{-\alpha\tau_n} J\phi(U_{\tau_n}^+, Y_{\tau_n}^+) \leq P_{u,x}^\delta - \text{ess inf}_{\mu^{\tau_n^+} = \delta^{\tau_n^+}} E_{u,x}^\mu \left(\int_{\tau_n}^{\tau_{n+1}} e^{-\alpha s} f(U_s, Y_s) ds \right. \\ \left. + e^{-\alpha\tau_{n+1}} (c(U_{\tau_{n+1}}, Y_{\tau_{n+1}}, U_{\tau_{n+1}}^+, Y_{\tau_{n+1}}^+) \right. \\ \left. + \phi(U_{\tau_{n+1}}^+, Y_{\tau_{n+1}}^+)) / \mathbf{G}_{\tau_n}^* \right).$$

Démontrons l'inégalité inverse. Un résultat sur l'arrêt optimal ([6]) nous indique que pour toute loi ν sur $\overline{U \times E}$ nous avons:

$$\langle \nu, J\phi \rangle = \inf_{\substack{T > 0 \\ T \in \mathbb{S}_{-1}}} \int_{\overline{U \times E}} \nu(d(u, x)) E_{u,x} \left(\int_0^{T(u, \cdot)} e^{-\alpha s} f(u, X_s) + e^{-\alpha T(u, \cdot)} m\phi(u, X_{T(u, \cdot)}) \right).$$

Remarquons que cette égalité s'établit facilement en inversant l'inf et la loi ν , propriété qui s'établit à l'aide d'une propriété de filtration décroissante qui est immédiate grâce à la nature des U-noyau d'arrêt. Nous en déduisons grâce à la définition du $P_{u,x}^\delta - \text{ess inf}$ l'égalité:

$$E_{u,x}^\delta (J\phi(U_{\tau_n}^+, Y_{\tau_n}^+)) \\ = \inf_{\substack{T > 0 \\ T \in \mathbb{S}_{-1}}} E_{u,x}^\delta \left(E_{U_{\tau_n}^+, Y_{\tau_n}^+} \left(\int_0^{T(U_{\tau_n}^+, \cdot)} e^{-\alpha s} f(U_{\tau_n}^+, X_s) + e^{-\alpha T(U_{\tau_n}^+, \cdot)} m\phi(U_{\tau_n}^+, X_{T(U_{\tau_n}^+, \cdot)}) \right) \right) \\ \cong E_{u,x}^\delta \left(P_{u,x}^\delta - \text{ess inf}_{\substack{T > 0 \\ T \in \mathbb{S}_{-1}}} E_{U_{\tau_n}^+, Y_{\tau_n}^+}^{T(U_{\tau_n}^+, \cdot)} \left(\int_0^\zeta e^{-\alpha s} f(U_{\tau_n}^+, \bar{X}_s) + e^{-\alpha \zeta} m\phi(U_{\tau_n}^+, \bar{X}_\zeta) \right) \right)$$

soit par définition de $J\phi$ $P_{u,x}^\delta$ p.s.:

$$(34) \quad J\phi(U_{\tau_n}^+, Y_{\tau_n}^+) = P_{u,x}^\delta - \text{ess inf}_{\substack{T > 0 \\ T \in \mathbb{S}_{-1}}} E_{U_{\tau_n}^+, Y_{\tau_n}^+}^{T(U_{\tau_n}^+, \cdot)} \left(\int_0^\zeta e^{-\alpha s} f(U_{\tau_n}^+, \bar{X}_s) + e^{-\alpha \zeta} m\phi(U_{\tau_n}^+, \bar{X}_\zeta) \right).$$

D'autre part, grâce à l'existence de noyau permis (proposition 19) pour tout $\varepsilon > 0$, il existe un noyau borélien $r(u, x)$ tel que:

i) $r(\partial, \Delta) = \varepsilon_{\partial, \Delta}$.

ii) Pour tout (u, x) de $\overline{U \times E}$

$$m\phi(u, x) \geq \int_{\overline{U \times E}} r(u, x; d(v, y)) (c(u, x, v, y) + \phi(v, y)) - \varepsilon.$$

Par suite à partir de (34) nous obtenons $P_{u,x}^\delta$ p.s.:

$$J\phi(U_{\tau_n}^+, Y_{\tau_n}^+) + \varepsilon \\ \geq P_{u,x}^\delta - \text{ess inf}_{\substack{T > 0 \\ T \in \mathbb{S}_{-1}}} E_{U_{\tau_n}^+, Y_{\tau_n}^+}^{T(U_{\tau_n}^+, \cdot)} \left(\int_0^\zeta e^{-\alpha s} f(U_{\tau_n}^+, \bar{X}_s) \right. \\ \left. + e^{-\alpha \zeta} \int_{\overline{U \times E}} r(U_{\tau_n}^+, \bar{X}_\zeta, d(v, y)) \right. \\ \left. \times (c(U_{\tau_n}^+, \bar{X}_\zeta, v, y) + \phi(v, y)) \right).$$

Nous en déduisons pour tout ε , $P_{u,x}^\delta p.o.$:

$$\begin{aligned} J\phi(U_{\tau_n^+}, Y_{\tau_n^+}) + \varepsilon \\ \geq P_{u,x}^\delta - \operatorname{ess\,inf}_{\mu_{\tau_n^+} = \delta_{\tau_n^+}} E_{U_{\tau_n^+}, Y_{\tau_n^+}}^{S_{\tau_n^+}^{\mu_{\tau_n^+}}} \left(\int_0^\zeta e^{-\alpha s f}(U_{\tau_n^+}, \bar{X}_s) \right. \\ \left. + e^{-\alpha \zeta} \int_{U \times E} N_{n+2}(\cdot, d(v, y)) \right. \\ \left. \times (c(U_{\tau_n^+}, \bar{X}_\zeta, v, y) + \phi(v, y)) \right) \end{aligned}$$

et par un calcul déjà effectué précédemment:

$$\begin{aligned} J\phi(U_{\tau_n^+}, Y_{\tau_n^+}) e^{-\alpha \tau_n} \\ \geq P_{u,x}^\delta - \operatorname{ess\,inf}_{\mu_{\tau_n^+} = \delta_{\tau_n^+}} E_{u,x}^\mu \left(\int_{\tau_n}^{\tau_{n+1}} e^{-\alpha s f}(U_s, Y_s) ds \right. \\ \left. + e^{-\alpha \tau_{n+1}} (c(U_{\tau_{n+1}}, Y_{\tau_{n+1}}, U_{\tau_{n+1}^+}, Y_{\tau_{n+1}^+}) \right. \\ \left. + \phi(U_{\tau_{n+1}^+}, Y_{\tau_{n+1}^+})) / \mathbf{G}_{\tau_n}^* \right) \end{aligned}$$

c'est à dire l'inégalité cherchée, d'où l'égalité (32).

Comme nous l'avons déjà remarqué $U^\alpha f(u, x)$ est borélienne et donc— $U^\alpha f(u, x)$ est Ray-analytique, par suite grâce au théorème précédent nous pouvons définir par récurrence pour tout $k \geq 1$:

$$(35) \quad J^{k+1} U^\alpha f(u, x) = \inf_{\substack{T > 0 \\ T \in S_{-1}}} E_{u,x} \left(\int_0^{T(u, \cdot)} e^{-\alpha s f}(u, X_s) ds + e^{-\alpha T(u, \cdot)} m J^k U^\alpha f(u, X_{T(u, \cdot)}) \right);$$

alors:

COROLLAIRE 21. *L'application $J^{n+1} U^\alpha f(u, x)$ est presque borélienne. Soit δ une stratégie admissible, pour tout $n \geq -1$ nous avons $P_{u,x}^\delta p.s.$:*

$$(36) \quad e^{-\alpha \tau_n} J^k U^\alpha f(U_{\tau_n^+}, Y_{\tau_n^+}) = P_{u,x}^\delta - \operatorname{ess\,inf}_{\substack{\mu_{\tau_n^+} = \delta_{\tau_n^+} \\ S_{n+k+1}^\mu = \ell_{+\infty}^0}} E_{u,x}^\mu (k_{\tau_n^+} / \mathbf{G}_{\tau_n}^*) \quad \forall k \geq 1,$$

en particulier

$$(37) \quad J^k U^\alpha f(u, x) = \rho_k^+(u, x) = \inf_{\substack{S_0^\mu > 0 \\ S_k^\mu = \ell_{+\infty}^0}} E_{u,x}^\mu (k).$$

Commentaire. La restriction " $S_{n+k+1}^\mu = 0$ ou $+\infty$ " signifie que la stratégie admissible μ n'a au plus que k impulsions effectives après n . Grâce à la définition (35) nous avons établi la relation de récurrence annoncée en (24).

Preuve. Si nous prenons pour ϕ , $U^\alpha f$ dans l'égalité (32), nous voyons immédiatement apparaître l'égalité (36) pour $k = 1$. En appliquant de nouveau plusieurs fois l'égalité (32) nous obtiendrions facilement en procédant d'une manière similaire l'égalité (36) pour tout $k \geq 2$. L'égalité (37) se déduit de l'égalité (36) avec $n = -1$ en choisissant pour δ une stratégie admissible vérifiant ($S_0 > 0$) et en remarquant que " $\mu^{0+} = \delta^{0+}$ " est équivalent à " $S_0^\mu > 0$ ", puis en prenant l'espérance par rapport à $E_{u,x}^\delta$, en inversant cette espérance et le $P_{u,x}^\sigma$ —ess inf et en utilisant la définition (23).

L'égalité (37) nous montre que la suite d'applications positives $J^k U^\alpha f$ est décroissante lorsque k tend vers l'infini, l'inf portant sur des ensembles croissants. Elle

admet donc une limite ponctuelle, que nous notons $J^\infty U^\alpha f$. Nous allons établir en particulier que $J^\infty U^\alpha f$ est ρ^+ .

PROPOSITION 22. *L'application $J^\infty U^\alpha f$ est presque borélienne et est égale à $\rho^+(u, x)$. De plus pour toute stratégie admissible δ et tout $n \geq -1$ P^δ p.s.:*

$$e^{-\alpha\tau_n} \rho^+(U_{\tau_n}^+, Y_{\tau_n}^+) = W_{\tau_n}^{\delta,+},$$

ainsi l'égalité (21) annoncée est établie.

Preuve. Il suffit d'établir pour tout (u, x) de $U \times E$ l'égalité $P_{u,x}^\delta$ p.s.:

$$(38) \quad e^{-\alpha\tau_n} J^\infty U^\alpha f(U_{\tau_n}^+, Y_{\tau_n}^+) = P_{u,x}^\delta - \operatorname{ess\,inf}_{\mu^{\tau_n^+} = \delta^{\tau_n^+}} E_{u,x}^\mu(k_{\tau_n}^+ / \mathbf{G}_{\tau_n}^*);$$

en effet à partir de cette égalité pour un raisonnement identique à celui de l'égalité (37) nous obtenons l'égalité:

$$J^\infty U^\alpha f(u, x) = \inf_{\{\mu: S_0 > 0\}} E^\mu(k) = \rho^+(u, x) \quad (\text{égalité (20)}),$$

alors (21) n'est autre que (38) prise au point (u^0, x^0) .

Nous déduisons immédiatement de l'égalité (36) $P_{u,x}^\delta$ p.s.:

$$e^{-\alpha\tau_n} J^k U^\alpha f(U_{\tau_n}^+, Y_{\tau_n}^+) \geq P_{u,x}^\delta - \operatorname{ess\,inf}_{\mu^{\tau_n^+} = \delta^{\tau_n^+}} E_{u,x}^\mu(k_{\tau_n}^+ / \mathbf{G}_{\tau_n}^*)$$

pour tout k ; d'où en passant à la limite:

$$(39) \quad e^{-\alpha\tau_n} J^\infty U^\alpha f(U_{\tau_n}^+, Y_{\tau_n}^+) \geq P_{u,x}^\delta - \operatorname{ess\,inf}_{\mu^{\tau_n^+} = \delta^{\tau_n^+}} E_{u,x}^\mu(k_{\tau_n}^+ / \mathbf{G}_{\tau_n}^*).$$

Inversement soit une stratégie admissible μ vérifiant $\mu^{\tau_n^+} = \delta^{\tau_n^+}$, nous allons comparer les coûts associés à la stratégie μ et à la stratégie arrêtée à droite de τ_p : $\mu^{\tau_p^+}$, pour tout $p > n$. Nous pouvons écrire $P_{u,x}^\delta$ p.s.:

$$\begin{aligned} E_{u,x}^{\mu^{\tau_p^+}}(k_{\tau_n}^+ / \mathbf{G}_{\tau_n}^*) &= E_{u,x}^\mu \left(\int_{\tau_n}^{\tau_p} e^{-\alpha s} f(U_s, Y_s) ds + \sum_{0 < n \leq p} c(U_{\tau_k}, Y_{\tau_k}, U_{\tau_k}^+, Y_{\tau_k}^+) \right. \\ &\quad \left. + \mathbb{1}_{(\tau_p < \tau)} E_{U_{\tau_p}^+, Y_{\tau_p}^+} \left(\int_{\tau_p}^{+\infty} e^{-\alpha s} f(U_{\tau_p}^+, X_s) ds \right) / \mathbf{G}_{\tau_n}^* \right), \end{aligned}$$

soit en ajoutant le coût associé à μ après τ_p :

$$\begin{aligned} E_{u,x}^{\mu^{\tau_p^+}}(k_{\tau_n}^+ / \mathbf{G}_{\tau_n}^*) &\leq E_{u,x}^\mu(k_{\tau_n}^+ / \mathbf{G}_{\tau_n}^*) \\ &\quad + E_{u,x}^\mu \left(\mathbb{1}_{(\tau_p < \tau)} E_{U_{\tau_p}^+, Y_{\tau_p}^+} \left(\int_{\tau_p}^{+\infty} e^{-\alpha s} f(U_{\tau_p}^+, X_s) ds \right) / \mathbf{G}_{\tau_n}^* \right) \end{aligned}$$

d'où en majorant $f P_{u,x}^\delta$ p.s.:

$$E_{u,x}^{\mu^{\tau_p^+}}(k_{\tau_n}^+ / \mathbf{G}_{\tau_n}^*) \leq E_{u,x}^\mu(k_{\tau_n}^+ / \mathbf{G}_{\tau_n}^*) + C E_{u,x}^\mu(\mathbb{1}_{(\tau_p < \tau)} / \mathbf{G}_{\tau_n}^*).$$

Grâce à l'admissibilité de μ , étant donné $\varepsilon > 0$, nous pouvons trouver p_ε tel que pour tout $q \geq p_\varepsilon$:

$$E_{u,x}^{\mu^{\tau_q^+}}(k_{\tau_n}^+) \leq E_{u,x}^\mu(k_{\tau_n}^+) + \varepsilon.$$

Remarquant que $\mu^{\tau_q^+}$ vérifie la condition $S_{q+1} = +\infty$, nous en déduisons l'inégalité:

$$E_{u,x}^\delta(e^{-\alpha\tau_n} J^\infty U^\alpha f(U_{\tau_n}^+, Y_{\tau_n}^+)) \leq E_{u,x}^\mu(k_{\tau_n}^+) + \varepsilon$$

pour toute stratégie μ vérifiant $\mu^{\tau_n^+} = \delta^{\tau_n^+}$ et tout $\varepsilon > 0$, par suite:

$$E_{u,x}^\delta (e^{-\alpha\tau_n} J^\infty U^\alpha f(U_{\tau_n^+}, Y_{\tau_n^+})) \leq \inf_{\mu^{\tau_n^+} = \delta^{\tau_n^+}} E_{u,x}^\delta (E_{u,x}^\mu (k_{\tau_n}^+ / \mathbf{G}_{\tau_n}^*)),$$

d'où par inversion de l'ess inf et de l'espérance:

$$(40) \quad E_{u,x}^\delta (e^{-\alpha\tau_n} J^\infty U^\alpha f(U_{\tau_n^+}, Y_{\tau_n^+})) \leq E_{u,x}^\delta (P_{u,x}^\delta - \text{ess inf}_{\mu^{\tau_n^+} = \delta^{\tau_n^+}} E_{u,x}^\mu (k_{\tau_n}^+ / \mathbf{G}_{\tau_n}^*)).$$

A partir des inégalités (39) et (40) nous obtenons immédiatement le résultat.

Nous allons maintenant établir un résultat analogue pour les coûts conditionnels après τ_n , en particulier l'égalité (20).

PROPOSITION 23. *Le coût moyen $\rho(u, x)$ est égal à $M\rho^+(u, x)$, et de plus pour toute stratégie admissible δ et tout $n \geq -1$ P^δ p.s.:*

$$W_{\tau_n}^\delta = e^{-\alpha\tau_n} m\rho^+(U_{\tau_n}, Y_{\tau_n}).$$

Preuve. Comme dans la proposition 22, il suffit d'établir pour tout (u, x) de $U \times E$ l'égalité $P_{u,x}^\delta$ p.s.:

$$(41) \quad e^{-\alpha\tau_n} m\rho^+(U_{\tau_n}, Y_{\tau_n}) = P_{u,x}^\delta - \text{ess inf}_{\mu^{\tau_n} = \delta^{\tau_n}} E_{u,x}^\mu (k_{\tau_n} / \mathbf{G}_{\tau_n}^a).$$

Donnons-nous une stratégie μ telle que $\mu^{\tau_n} = \delta^{\tau_n}$, alors P^δ p.s.:

$$\begin{aligned} e^{-\alpha\tau_n} m\rho^+(U_{\tau_n}, Y_{\tau_n}) &\leq e^{-\alpha\tau_n} \int_{U \times E} N_{n+1}(\cdot, d(u, x))(c(U_{\tau_n}, Y_{\tau_n}, u, x) + \rho^+(u, x)) \\ &\leq e^{-\alpha\tau_n} E_{u,x}^\mu (c(U_{\tau_n}, Y_{\tau_n}, U_{\tau_n^+}, Y_{\tau_n^+}) + \rho^+(U_{\tau_n^+}, Y_{\tau_n^+}) / \mathbf{G}_{\tau_n}^a). \end{aligned}$$

Grâce à l'égalité (21) nous obtenons $P_{u,x}^\delta$ p.s.:

$$e^{-\alpha\tau_n} m\rho^+(U_{\tau_n}, Y_{\tau_n}) \leq E_{u,x}^\mu (e^{-\alpha\tau_n} c(U_{\tau_n}, Y_{\tau_n}, U_{\tau_n^+}, Y_{\tau_n^+}) + E_{u,x}^\mu (k_{\tau_n}^+ / \mathbf{G}_{\tau_n}^*) / \mathbf{G}_{\tau_n}^a)$$

soit encore:

$$e^{-\alpha\tau_n} m\rho^+(U_{\tau_n}, Y_{\tau_n}) \leq E_{u,x}^\mu (k_{\tau_n} / \mathbf{G}_{\tau_n}^a).$$

Nous en déduisons donc l'inégalité dans un sens.

Inversement pour tout $\varepsilon > 0$ il existe un noyau permis borélien (proposition 19) tel que $P_{u,x}^\delta$ p.s.:

$$\begin{aligned} \varepsilon + e^{-\alpha\tau_n} m\rho^+(U_{\tau_n}, Y_{\tau_n}) &\geq e^{-\alpha\tau_n} \int_{U \times E} r(U_{\tau_n}, Y_{\tau_n}; d(u, x)) \\ &\quad \cdot (c(U_{\tau_n}, Y_{\tau_n}, u, x) + \rho^+(u, x)) \\ &\geq P_{u,x}^\delta - \text{ess inf}_{\mu^{\tau_n} = \delta^{\tau_n}} E_{u,x}^\mu (c(U_{\tau_n}, Y_{\tau_n}, U_{\tau_n^+}, Y_{\tau_n^+}) \\ &\quad + \rho^+(U_{\tau_n^+}, Y_{\tau_n^+}) / \mathbf{G}_{\tau_n}^a); \end{aligned}$$

par suite d'après la forme de ρ^+ :

$$\begin{aligned} e^{-\alpha\tau_n} m\rho^+(U_{\tau_n}, Y_{\tau_n}) &\geq P_{u,x}^\delta - \text{ess inf}_{\mu^{\tau_n} = \delta^{\tau_n}} E_{u,x}^\mu (c(U_{\tau_n}, Y_{\tau_n}, U_{\tau_n^+}, Y_{\tau_n^+}) \\ &\quad + P_{u,x}^\mu - \text{ess inf}_{\nu^{\tau_n^+} = \mu^{\tau_n^+}} E_{u,x}^\nu (k_{\tau_n}^+ / \mathbf{G}_{\tau_n}^*) / \mathbf{G}_{\tau_n}^a). \end{aligned}$$

En inversant le $P_{u,x}^\mu$ - ess inf et l'espérance conditionnelle et en remarquant que:

$$P_{u,x}^\delta - \text{ess inf}_{\mu^{\tau_n} = \delta^{\tau_n}} P_{u,x}^\mu - \text{ess inf}_{\nu^{\tau_n} = \mu^{\tau_n}} \text{ est égal à } P_{u,x}^\delta - \text{ess inf}_{\mu^{\tau_n} = \delta^{\tau_n}}$$

nous obtenons l'inégalité inverse et donc l'égalité (41).

En inversant l'espérance conditionnelle et l'inf, et grâce aux égalités (38) et (41) ou par une démonstration du même type que celle qui nous a conduit à (41) (en particulier en employant l'existence de noyaux permis) nous obtenons:

PROPOSITION 24. *L'application $\rho^+(u, x)$ satisfait aux systèmes*

$$(42) \quad \rho^+(u, x) = \inf_{\substack{T > 0 \\ T \in S_{-1}}} E_{u,x} \left(\int_0^{T(u,\cdot)} e^{-\alpha s} f(u, X_s) ds + e^{-\alpha T(u,\cdot)} m\rho^+(u, X_{T(u,\cdot)}) \right)$$

et

$$(43) \quad \rho^+(u, x) = \inf_{\substack{T > 0 \\ T \in S_{-1}}} E_{u,x} \left(\int_0^{T(u,\cdot)} e^{-\alpha s} f(u, X_s) ds + e^{-\alpha T(u,\cdot)} m^*\rho^+(u, X_{T(u,\cdot)}) \right);$$

où

$$m^*\rho^+(u, x) = \inf_{\nu \in M_{\{u,x\}}^*} \int_{U \times E} \nu(d(v, y)) (c(x, u, v, y) + \rho^+(v, y)).$$

Commentaire. 1) Dans la définition m^* l'inf porte sur tous les noyaux de renaissance permis, excepté "la renaissance identique", c'est à dire que la mesure de Dirac $\varepsilon_{\{u,x\}}$ est exclue. Nous verrons au paragraphe suivant l'utilité de l'introduction de m^* .

2) Les égalités (42) et (43) sont assez naturelles vue la signification de ρ^+ , $m\rho^+$ et $m^*\rho^+$, car dans la partie droite de ces égalités l'opération effectuée revient à ajouter à chaque stratégie admissible une impulsion supplémentaire à un temps strictement positif, qui est effective dans (43) et qui ne l'est pas forcément dans (42).

Nous pouvons exprimer le critère d'optimalité (théorème 13) à l'aide de l'application ρ^+ qui est indépendante de toute stratégie admissible, but essentiel de toute l'étude effectuée dans ce paragraphe.

THÉORÈME 25. *Pour toute stratégie admissible δ nous avons $P^\delta p.s.$ les inégalités:*

$$(44) \quad m\rho^+(u^0, x^0) \leq E_{u^0, x^0} \left(\int_0^{S_0^\delta(u,\cdot)} e^{-\alpha s} f(u, X_s) ds + e^{-\alpha S_0^\delta(u,\cdot)} m\rho^+(u, X_{S_0^\delta(u,\cdot)}) \right);$$

pour tout $n \geq 0$:

$$(45) \quad m\rho^+(U_{\tau_n}, Y_{\tau_n}) \leq \int_{U \times E} N_{n+1}^\delta(\cdot, d(u, x)) (c(U_{\tau_n}, Y_{\tau_n}, u, x) + \rho^+(u, x)),$$

et pour tout $n \geq 0$:

$$(46) \quad \rho^+(U_{\tau_n}^+, Y_{\tau_n}^+) \leq E_{U_{\tau_n}^+, Y_{\tau_n}^+} \left(\int_0^{S_{n+1}^\delta} e^{-\alpha s} f(U_{\tau_n}^+, X_s) ds + e^{-\alpha S_{n+1}^\delta} m\rho^+(U_{\tau_n}^+, X_{S_{n+1}^\delta}) \right).$$

De plus la stratégie $\hat{\delta}$ est optimale si et seulement si l'égalité a lieu simultanément dans (44), (45) et (46).

Preuve. Ce théorème s'obtient immédiatement à partir du théorème 13 en exploitant les égalités (21) et (22), puis en interprétant la loi P^δ et en remarquant de plus pour l'inégalité (44) que l'optimalité est équivalente à l'optimalité conditionnelle après 0.

Commentaire. Cette méthode originale, qui nous a permis d'expliciter la forme des fonctions de coût, peut être utilisée dans le contrôle dit continu, notamment dans le contrôle instantané des diffusions (voir par exemple [2], [10]), pour établir dans le cadre markovien une forme analogue pour la fonction de valeur, problème ouvert jusqu'ici (voir M. H. A. Davis et C. B. Wan [3]) et qui est ainsi résolu par les auteurs et N. El Karoui [7]. Le problème est ici simplifié car nous ne nous intéressons à la forme de la fonction de valeur qu'aux instants canoniques d'impulsion.

5. Résolution. Si nous regardons l'inégalité (45) du critère d'optimalité, une stratégie optimale doit nécessairement vérifier pour tout $n \geq 0$ sur l'ensemble $(\tau_n < \tau)$, P^δ p.s.:

$$\rho(U_{\tau_n}, Y_{\tau_n}) = m\rho^+(U_{\tau_n}, Y_{\tau_n}) = \int_{U \times E} N_{n+1}^\delta(\cdot, d(u, x))(c(U_{\tau_n}, Y_{\tau_n}, u, x) + \rho^+(u, x))$$

où le noyau N_{n+1}^δ est pris dans l'ensemble $M_{\{U_{\tau_n}, Y_{\tau_n}\}}^*$ sur $(\tau_n < \tau)$. Par suite il est naturel d'introduire:

$$m^*\rho^+(u, x) = \inf_{v \in M_{\{u, x\}}^*} \int_{U \times E} V(d(v, y))(c(u, x, v, y) + \rho^+(v, y))$$

et pour construire la suite de noyaux d'arrêt admissible en vue d'une stratégie optimale, nous devons nécessairement utiliser le temps:

$$T^* = \inf(t \geq 0: e^{-\alpha t} \rho(u, X_t) = e^{-\alpha t} m^*\rho^+(y, X_t)).$$

Pour son existence nous sommes conduits à introduire l'hypothèse suivante d'existence de temps d'arrêt optimaux.

HYPOTHÈSE 3. *Le processus $m^*\rho^+(u, X_t)$ est $\mathbf{U} \otimes \mathbf{0}$ mesurable ($\mathbf{0}$ tribu optionnelle) et $m^*\rho^+$ est s.c.s. sur les noyaux d'arrêt, c'est-à-dire que pour tout (u, x) de $U \times E$ et toute suite monotone T_n de noyaux d'arrêt de \underline{S}_{-1} de limite T*

$$\overline{\lim} E_{u,x}(m^*\rho^+(u, X_{T_n(u,\cdot)})) \leq E_{u,x}(m^*\rho^+(u, X_{T(u,\cdot)})).$$

De même il nous faut construire un noyau de renaissance qui satisfasse à l'égalité:

$$(47) \quad \rho(u, x) = m\rho^+(u, x) = \int_{U \times E} \nu(d(v, y))(c(u, x, v, y) + \rho^+(v, y)),$$

d'où:

HYPOTHÈSE 4. *Pour tout (u, x) de I où I désigne l'ensemble des couples (u, x) vérifiant $\rho(u, x) = m^*\rho^+(u, x)$, l'application qui à ν de $M_{\{u, x\}}^*$ fait correspondre*

$$\int_{U \times E} \nu(d(v, y))(c(u, x, v, y) + \rho^+(v, y))$$

atteint son minimum sur $M_{\{u, x\}}^$.*

Remarque. Ces hypothèses sont en particulier vérifiées dans le modèle Fellerien de M. Robin [16]. L'hypothèse 3 telle qu'elle est posée est purement technique. Il nous paraît raisonnable par analogie avec la méthode de pénalisation de conjecturer que $\rho^+(u, x)$ est continue en (u, x) dans un modèle fellérien. Alors si pour tout (u, x) , $M_{\{u, x\}}^*$ est "déterministe", et l'ensemble des impulsions permises envoie dans un compact de $U \times E$, les hypothèses faites sur c entraînent que $m^*\rho^+$ est s.c.s. et donc que (H_3) est satisfaite.

Grâce à une démonstration identique à celle de la proposition 19 qui conduit à l'existence de noyaux permis, nous avons:

LEMME 26. *Sous les hypothèses 1 à 4, pour toute loi ν sur $(U \times E, \mathbf{U} \otimes \mathbf{E})$ il existe un noyau borélien $r^*(u, x)$ à valeurs dans $M_{\{u, x\}}^*$ pour tout (u, x) de $U \times E$, tel que pour tout (u, x) de I :*

$$(48) \quad m^* \rho^+(u, x) = \int_{\overline{U \times E}} r^*(u, x, d(v, y))(c(u, x, v, y) + \rho^+(v, y)).$$

Posons pour tout $n \geq 0$:

$$S_n^*(w, \omega) = \begin{cases} T^*(u_n, \omega) & \text{sur } (u_n \neq \partial) \cap (T^*(u_n, \omega) > 0), \\ +\infty & \text{sur } (u_n \neq \partial) \cap (T^*(u_n, \omega) = 0), \\ 0 & \text{sur } (u_n = \partial) \end{cases}$$

et

$$N_{n+1}^*(w, \cdot) = \begin{cases} r^*(u_n, Y_{\tau_n}; \cdot) & \text{sur } (u_n \neq \partial) \cap (0 < T^*(u_n, \omega) < +\infty), \\ \varepsilon_{\{\partial, \Delta\}} & \text{sinon.} \end{cases}$$

Nous allons établir que $\delta^* = ((S_n^*)_{n \geq 0}, (N_n^*)_{n \geq 1})$ est une stratégie admissible optimale. Pour l'admissibilité nous devons encore faire trois hypothèses supplémentaires sur le coût c et la maison de renaissance.

HYPOTHÈSE 5. *Pour tous noyaux boréliens r_1, r_2 de M^* le noyau borélien*

$$r(u_1, x_1, d(u_3, x_3)) = \int_{\overline{U \times E}} r_1(u_1, x_1, d(u_2, x_2)) r_2(u_2, x_2, d(u_3, x_3))$$

appartient à $M_{\{u_1, x_1\}}$ pour tout (u_1, x_1) de $U \times E$ (où nous prolongeons toute fonctions ϕ définie sur $U \times E$ à $\overline{U \times E}$ en posant $\phi(\partial, \Delta) = 0$).

HYPOTHÈSE 6. *Pour tout (u, x) de $U \times E$ et tout ν de $M_{\{u, x\}}^*$ nous avons quel que soit (v, y) de $\overline{U \times E}$*

$$\int_{\overline{U \times E}} \nu(d(w, z))(c(u, x, w, z) + c(w, z, v, y)) > c(u, x, v, y).$$

HYPOTHÈSE 7. *Il existe une constante C strictement positive telle que pour tout (u, x) de $U \times E$ et tout ν de $M_{\{u, x\}}^*$*

$$\int_{\overline{U \times E}} \nu(d(v, y))c(u, x, v, y) \geq C.$$

Nous avons le:

THÉORÈME 27. *Sous les hypothèses 1 à 7 la famille $\delta^* = ((S_n^*)_{n \geq 0}, (N_n^*)_{n \geq 1})$ est une stratégie admissible optimale, qui par construction est à la fois indépendante de la loi initiale et "markovienne".*

Preuve. Montrons tout d'abord que $(S_n^*)_{n \geq 0}$ appartient à \mathcal{S}_2 . Soit (u, x) élément de I ; par construction (48) nous avons pour tout couple (u, x) de I .

$$\begin{aligned} \rho(u, x) &= m^* \rho^+(u, x) = \int_{\overline{U \times E}} r^*(u, x; d(v, y))(c(u, x, v, y) + \rho^+(v, y)) \\ &\geq \int_{\overline{U \times E}} r^*(u, x, d(v, y)) \\ &\quad \times (c(u, x, v, y) + \mathbb{1}_{I^c}(v, y)\rho^+(v, y) + \mathbb{1}_I(v, y)\rho(v, y)). \end{aligned}$$

Soit en appliquant de nouveau (48):

$$\begin{aligned} \rho(u, x) \cong & \int_{U \times E} r^*(u, x, d(v, y))(c(u, x, v, y) + \mathbb{1}_I^c(v, y)\rho^+(v, y)) \\ & + \mathbb{1}_I(v, y) \int_{U \times E} r^*(v, y, d(w, z))(c(v, y, w, z) + \rho^+(w, z)). \end{aligned}$$

Si $r^*(u, x, I)$ est strictement positif nous avons en vertu de l'hypothèse 6:

$$\begin{aligned} \rho(u, x) > & \int_{U \times E} r^*(u, x, d(v, y)) \int_{U \times E} (\mathbb{1}_I^c(v, y)\varepsilon_{\{v, y\}}(d(w, z))) \\ & + \mathbb{1}_I(v, y)r^*(v, y, d(w, z))(c(u, x, w, z) + \rho^+(w, z)). \end{aligned}$$

Appliquant l'hypothèse 5 nous en déduisons:

$$\rho(u, x) > m\rho(u, x) = \rho(u, x).$$

Il y a donc une contradiction et $r^*(u, x, I)$ est nul. Ce résultat entraîne immédiatement que $(S_n^*)_{n \geq 0}$ appartient à \mathcal{S} .

La construction de δ^* entraîne facilement que les égalités (44), (45), (46) sont satisfaites. Il reste donc à établir que la stratégie δ^* est admissible.

Nous obtenons facilement de la construction même de δ^* que pour tout $n \geq 1$:

$$\begin{aligned} \rho(u^0, x^0) = E^\delta \Bigg(& \sum_{0 \leq k \leq n} e^{-\alpha \tau_k} c(U_{\tau_k}, Y_{\tau_k}, U_{\tau_k}^+, Y_{\tau_k}^+) \\ & + \int_0^{\tau_n} e^{-\alpha s} f(U_s, Y_s) ds + e^{-\alpha \tau_n} \rho^+(U_{\tau_n}^+, Y_{\tau_n}^+) \Bigg). \end{aligned}$$

Par suite:

$$\begin{aligned} E^\delta \Bigg(& \sum_{0 \leq k \leq n} c(U_{\tau_k}, Y_{\tau_k}, U_{\tau_k}^+, Y_{\tau_k}^+) e^{-\alpha \tau_k} \Bigg) \\ & \leq \rho(u^0, x^0) \leq E_{u^0, x^0} \Bigg(\int_0^{+\infty} e^{-\alpha s} f(u^0, X_s) ds \Bigg) \leq h \end{aligned}$$

puisque l'application f est majorée.

Soit en utilisant la construction de P^δ et l'hypothèse 7:

$$(n+1)CE^\delta(e^{-\alpha \tau_n} \mathbb{1}_{(\tau_n < \tau)}) \leq E^\delta \Bigg(\sum_{0 \leq k \leq n} c(U_{\tau_k}, Y_{\tau_k}, U_{\tau_k}^+, Y_{\tau_k}^+) \Bigg) e^{-\alpha \tau_k} \leq h,$$

c'est-à-dire:

$$E^\delta(e^{-\alpha \tau_n} \mathbb{1}_{(\tau_n < \tau)}) \leq \frac{h}{C(n+1)},$$

ce qui entraîne l'admissibilité.

6. Conclusion. Ainsi le point de vue adopté en considérant un problème de type impulsif comme une suite de changements de régime et d'état d'un système nous paraît être le plus général possible. Par des techniques analogues à celles du paragraphe 4 nous avons résolu le contrôle continu markovien [7], et nous pensons utiliser les mêmes techniques en vue de la résolution d'autres problèmes (contrôle continu avec arrêt, contrôle partiellement observable, ...).

BIBLIOGRAPHIE

- [1] A. BENSOUSSAN AND J. L. LIONS, *Nouvelle méthode en contrôle impulsionnel*, Applied Math. Optim., 1 (1975), pp. 289–312.
- [2] M. H. A. DAVIS AND P. P. VARAIYA, *Dynamic programming condition for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.
- [3] M. H. A. DAVIS AND C. B. WAN, *The principle of optimality for Markov jump processes*, IMA Conference on Analysis and Optimization of Stochastic Systems, Oxford, 1978.
- [4] C. DELLACHÉRIE, *Ensembles analytiques, capacités–mesures de Hausdorff*, Lecture Notes in Mathematics 295, Springer, New York, 1972.
- [5] C. DELLACHÉRIE AND P. A. MEYER, *Probabilités et potentiel*, Hermann, Paris, 1975.
- [6] N. EL KAROUI, *Contrôle stochastique*, Ecole d'été de Probabilités de St Flour, 1979.
- [7] N. EL KAROUI, J. P. LEPELTIER AND B. MARCHAL, *Arrêt optimal dépendant d'un paramètre et contrôle continu markovien*, to appear.
- [8] R. K. GETTOOR, *Markov Processes, Ray Processes and Right Processes*, Lecture Notes in Mathematics 440, Springer-Verlag, New York, 1973.
- [9] J. P. LEPELTIER, *Contrôle stochastique et jeux impulsionnels*, Thèse d'état Le Mans, 1980.
- [10] J. P. LEPELTIER AND B. MARCHAL, *Sur l'existence de politiques optimales dans le contrôle intégral-différentiel*, An I.H.P. Section B, 13 (1977), pp. 45–97.
- [11] ———, *Technique probabiliste dans le contrôle impulsionnel*, Stochastics, 2 (1979), pp. 243–286.
- [12] B. MARCHAL, *Contrôle impulsionnel et contrôle continu markovien*, Thèse d'état Paris VI, 1980.
- [13] P. A. MEYER, *Réduite et jeux de hasard*, Séminaire Proba. VII, Lecture Notes in Mathematics 321, Springer, New York, 1973, pp. 155–171.
- [14] ———, *Renaissance, recollements, mélanges, Ralentissement de processus de Markov*, Ann. Inst., Fourier Grenoble, XXV 3–4 (1975), pp. 465–495.
- [15] J. NEVEU, *Bases mathématiques du calcul de probabilités*, Gauthier Villars, 1968.
- [16] M. ROBIN, *Contrôle impulsionnel des processus de Markov*, Thèse d'état Paris IX (1978).
- [17] C. STRIEBEL, *Martingale conditions for the optimal control of continuous time stochastic systems*, International Workshop on Stochastic Filtering and Control, Los Angeles, 1974.

EXISTENCE AND UNIQUENESS OF MINIMAL REALIZATIONS FOR A CLASS OF C^∞ SYSTEMS*

J. P. GAUTHIER† AND G. BORNARD†

Abstract. The main result is relative to the quotient of a manifold M by a closed discrete equivalence relation when the first homotopy group of M has no infinite order element.

As a direct consequence of the main theorem, the existence and uniqueness of minimal realizations is proved for a class of C^∞ completely controllable weakly observable systems.

This extends:

- Some results of Sussmann available in the analytic or symmetric cases.
- Some results of the authors obtained in the compact case.

Key words. nonlinear, manifold, quotient, homotopy, realization

1. Introduction. In a previous work (Gauthier and Bornard [4]) the following result was proved:

THEOREM 1. *Let M be a C^∞ connected compact manifold of dimension n , and let R be a closed discrete equivalence relation on M . Then the following properties are equivalent:*

- (i) R is regular.
- (ii) There exists a (completely) controllable set of C^∞ vector fields on M which are positive symmetry vector fields for R .
- (iii) M is a covering space of M/R with π_R (the canonical mapping from M to M/R) as covering mapping.

A vector field X is called a symmetry vector field for R if $xRy \rightarrow X_t(x)RX_t(y)$, $\forall t \in \mathbb{R}$, where X_t is the one parameter subgroup generated by X . If t is restricted to be positive in the preceding definition, X becomes a *positive* symmetry vector field.

From now on, condition (ii) of Theorem 1 will be referred to as the property (P) for R . The same condition when the manifold M is analytic and the vector fields are complete analytic will be called the property (P) -analytic for R .

In Gauthier and Bornard [4], some interesting consequences in terms of realization theory were derived, extending to the C^∞ compact case some results obtained by Sussmann [8], [9], [11] for the symmetric or analytic cases.

The purpose of the present paper is to extend Theorem 1 to a class of noncompact cases, and to derive the consequences for realization theory.

Most of the ideas developed here come from the following example:

Example 1. Let $N = \mathbb{R}^2$, and consider the analytic vector fields X_1 and X_2 given by

$$X_1 = \frac{\partial}{\partial x_1} + \alpha x_2 \frac{\partial}{\partial x_2}, \quad X_2 = \frac{\partial}{\partial x_2} \quad \text{with } \alpha = \frac{\log 2}{2\pi}.$$

Consider, on N , the equivalence relation \sim

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix} \quad \text{iff} \quad \begin{cases} x'_1 = x_1 + 2k\pi, \\ x'_2 = x_2, \end{cases} \quad K \in \mathbb{Z},$$

* Received by the editors October 12, 1982, and in revised form March 31, 1983.

† Laboratoire d'automatique de Grenoble, Institut National Polytechnique de Grenoble, École Nationale Supérieure d'Ingénieurs Electriciens de Grenoble, B.P. 46, 38402-Saint-Martin-d'Hères, France.

and let M be the quotient cylinder N/\sim . From now on we shall consider it as the product $M = [0, 2\pi[\times \mathbb{R}$.

Clearly X_1 and X_2 are symmetry vector fields for \sim . They induce on M , through the canonical projection π_\sim , two analytic complete vector fields \tilde{X}_1 and \tilde{X}_2 . Moreover, the family $\Gamma = \{\tilde{X}_1, \pm \tilde{X}_2\}$ is completely controllable on M .

Let us consider on M the equivalence relation R given by

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} R \begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix} \quad \text{iff} \quad \begin{cases} x_1 = x'_1, \\ x_2 - x'_2 = K e^{\alpha x_1}, \quad K \in \mathbb{Z}, \end{cases} \quad 0 \leq x_1 < 2\pi.$$

One can check that we now have the following situation:

- (i) M is analytic.
- (ii) R is a closed discrete equivalence relation on M .
- (iii) Γ is a controllable family of analytic complete vector fields that are positive symmetry vector fields for R . Then R has the property (P)-analytic.

However, M is not compact and Theorem 1 fails to apply, because of the following remarks:

(i) $-\tilde{X}_1$ does not comply with R . Then \tilde{X}_1 is not a symmetry vector field. (Moreover, if it were, R would be regular from Sussmann's theorem [9], [11].)

(ii) R is not regular since the quotient space M/R (see Fig. 1) is not a manifold.

M/R can be obtained as follows: Take a cylinder with the boundaries $C = [0, 2\pi[\times [0, 1]$. The two boundaries are circles C_1 and C_2 . Move C in such a way that C_2 becomes a two-fold covering of C_1 , and patch C_1 and C_2 .

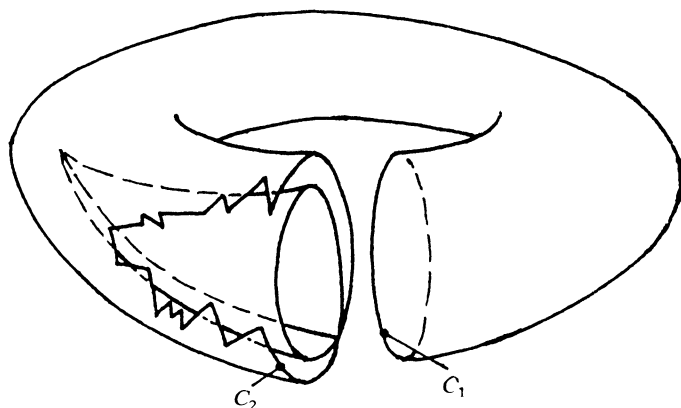


FIG. 1

This is an analytic counterexample of Theorem 1 in the noncompact case.

The main result of this paper is the following theorem:

THEOREM 2 (main result). *Let M be an analytic connected manifold whose first homotopy group $\pi_1(M)$ has no infinite order element.*

Assume that R is a closed discrete equivalence relation on M and has the property (P)-analytic. Then R is regular.

Remark. In Example 1, Theorem 2 fails to apply because $\pi_1(M) = \mathbb{Z}$. This fact points out the importance of the topological assumption made in this theorem.

In § 2 the proof of Theorem 2 will be given. The consequences in terms of C^∞ realization theory are discussed in § 3, and Example 1 is completed in this context.

2. Proof of the main result. The following lemma will be necessary to achieve the proof of Theorem 2.

LEMMA 1. *Let M be an analytic manifold of dimension n and let R be a closed discrete equivalence relation on M meeting the property (P)-analytic.*

Consider on M a controllable family Γ of complete analytic positive symmetry vector fields for R . Γ induces in $M \times M$ the family $\tilde{\Gamma}$ of analytic vector fields of the form $X \oplus X$, $X \in \Gamma$.

Let σ be the orbit in $M \times M$ of $\tilde{\Gamma}$ through some point (x_0, y_0) such that $x_0 R y_0$, and let p_1 be the canonical mapping from $M \times M$ to M : $(x, y) \rightarrow^{p_1} x$.

Then σ is a covering space of M with p_1 (restricted to σ) as covering mapping.

The proof of the lemma will be given later.

Proof of Theorem 2. We shall prove that complete positive symmetry vector fields for R are also symmetry vector fields for R . Then from the fact that Γ is analytic and controllable, and from the results of Sussmann [9], [11], R will be regular.

Consider $x_0, y_0 \in M$ such that $x_0 R y_0$, $X \in \Gamma$, and $x_1 = X_{-t}(x_0)$, $y_1 = X_{-t}(y_0)$ for some $t > 0$. We have to show that $x_1 R y_1$.

Γ being controllable, there exists a triple (l, \S, T_0) such that $\psi_{\S}(x_0, T_0) = x_1$ where

$$\S = \{X^i, i = 1, \dots, l \mid X^i \in \Gamma\},$$

$$T_0 \in (\mathbb{R}^*)^l,$$

$$\psi_{\S}(x, T) = X_{t^l}^l \circ \dots \circ X_{t^1}^1(x).$$

Consider now the triple $\gamma = (l+1, \S', T_1)$ obtained by concatenating (l, \S, T_0) and $(1, X, t)$:

$$\S' = \{X^1, \dots, X^l, X\},$$

$$T_1 = (t_1, \dots, t_l, t) (T_1 \in (\mathbb{R}_*^+)^{l+1}).$$

One has $\psi_{\S'}(x_0, T_1) = x_0$.

Clearly γ defines a loop in M from x_0 to x_0 . Let us denote by γ^i the loop obtained by concatenating i times the loop γ .

Since the first homotopy group $\pi_1(M)$ has no infinite order element, there exists a positive integer k such that the loop γ^k is in the homotopy class of the zero loop.

Let σ be the orbit of $\tilde{\Gamma}$ in $M \times M$ through (x_0, y_0) . Any loop in M through x_0 which is the concatenation of pieces of integral curves of elements of Γ can be lifted in σ in a unique way, through the lifting of Γ , with (x_0, y_0) as initial point. Let us call $\tilde{\gamma}$ this lift of γ in σ . (Note that the elements of Γ and $\tilde{\Gamma}$ are complete.)

The initial point of $\tilde{\gamma}^k$ is (x_0, y_0) . The lift \tilde{I} in σ of the zero loop I at x_0 is (x_0, y_0) , and then it has the same initial point as $\tilde{\gamma}^k$.

From Lemma 1, σ is a covering space for M , and all the lifts were made through the covering mapping.

Then, I and γ^k being in the same homotopy class, and because of Boothby [2, Thm. 9.3, p.288], \tilde{I} and $\tilde{\gamma}^k$ also have the same end point, which is (x_0, y_0) .

One therefore has the following relations:

$$\psi_{\S'}^k(y_0, T_1) = y_0,$$

$$\psi_{\S}(\psi_{\S'}^{k-1}(x_0, T_1), T_0) = x_1,$$

$$\psi_{\S}(\psi_{\S'}^{K-1}(y_0, T_1), T_0) = y_1.$$

The elements of Γ being positive symmetry vector fields for R , then $x_1 R y_1$. The result is obtained.

Proof of Lemma 1. From Sussmann [12], σ is an analytic submanifold of $M \times M$. We shall first show that $\dim \sigma = n$ (dimension of M).

Let L be the Lie algebra of vector fields on M generated by Γ , and \tilde{L} the Lie algebra of vector fields on $M \times M$ generated by $\tilde{\Gamma}$. Since Γ is controllable, L is of rank n on M . This implies that \tilde{L} has rank at least n on σ . Then $\dim \sigma \geq n$.

Let $D^+(x_0, y_0)$ be the accessibility set by $\tilde{\Gamma}$ from $(x_0, y_0) \in \sigma$, $x_0 R y_0$. One has $D^+(x_0, y_0) \subset \sigma$, and since the situation is analytic, $D^+(x_0, y_0)$ has a nonempty interior in σ . Moreover, $D^+(x_0, y_0) \subset G_R$, the graph of R in $M \times M$.

Assume that $\dim \sigma > n$, and consider the set $J(x, y) = p_1^{-1}(x) \cap \text{Int } D^+(x_0, y_0)$ for $(x, y) \in \text{Int } D^+(x_0, y_0)$. The connected component of $J(x, y)$ containing (x, y) is locally a manifold of dimension > 0 for almost every $(x, y) \in \text{Int } D^+(x_0, y_0)$. To show this, one can write the expression of σ in some coordinate neighborhood of (x, y) in $M \times M$, and apply the implicit function theorem (Auslander and Mackenzie [1, pp. 29, 30]).

This property of $J(x, y)$, with the fact that $J(x, y) \subset G_R$, is clearly not compatible with the assumption that R is discrete. Then $\dim \sigma = n$.

Now, consider on σ the equivalence relation \bar{R} defined by: $x \bar{R} y$ iff $p_1(x) = p_1(y)$. \bar{R} is closed.

Let us assume that the following property (S) is true:

(S): There exists a set of complete analytic symmetry vector fields for \bar{R} which is transitive on σ .

With property (S) and \bar{R} being closed, Sussmann [9, Thm. 11] implies that p_1 , restricted to σ , is a fiber mapping. Since $\dim \sigma = n$, σ is a covering space for M with p_1 as covering mapping.

To achieve the proof it is now sufficient to prove (S).

Consider on M the saturated family Γ' of vector fields defined by $\Gamma' = \{d\psi_s \cdot X \mid s \in \Gamma, X \in \Gamma\}$, where

$$\psi_s = X_{t_1}^1 \circ \dots \circ X_{t_i}^1(X^i \in \Gamma), \quad t_i \in R.$$

Clearly, since the elements of Γ are complete and analytic, the same is true for the elements of Γ' . One has also $\Gamma \subset \Gamma'$.

Consider on $M \times M$ the family $\tilde{\Gamma}'$ of complete analytic vector fields of the form $Y \oplus Y$, $Y \in \Gamma'$. Clearly, the orbit of $\tilde{\Gamma}'$ through (x, y) is σ , and from the arguments developed by Sussmann [12], $\tilde{\Gamma}'$ is transitive on σ . In fact, σ is an integral submanifold of the distribution generated by $\tilde{\Gamma}'$ in $M \times M$.

By construction, the elements of $\tilde{\Gamma}'$ are symmetry vector fields for \bar{R} , and the proof of (S) is achieved.

3. Consequences for realization theory. A direct consequence of the main result is the following theorem:

THEOREM 3. Consider the system $\Sigma = \{M, H_\Sigma = \{X_u \mid u \in U\}, h\}$ where:

(i) M is an analytic manifold whose first homotopy group has no infinite order elements,

(ii) H_Σ is a family of complete analytic vector fields on M ,

(iii) h is a C^∞ mapping from M to \mathbb{R}^p .

Assume that Σ is completely controllable, weakly observable (in the sense of Hermann and Krener [5]).

Then there exists a system Σ' unique up to a diffeomorphism, having the same input-output properties as Γ , and which is controllable and observable.

Proof. Consider $R = J$, the indistinguishability relation for Σ . J is closed because of the completeness of the elements of H_Σ , and discrete because Σ is weakly observable.

The conditions of Theorem 2 are satisfied, and J is then regular. The quotient $M' = M/J$ is a manifold, and the canonical mapping p_J is a submersion from M on M' . Moreover, the elements of H_Σ and the function h pass to the quotient, resulting in a quotient system Σ' on M' , which is controllable, observable and has the same input-output properties as Σ .

The uniqueness of Σ' comes from Gauthier and Bornard [4, Thm. 3], for which the compactness property was not assumed.

Example 1 (continued). Consider the system $\Sigma = (M, \Gamma, h)$, where M, Γ are defined as in the first part of the example, h being a C^∞ function defined by:

$$h(x_1, x_2) = f(x_1) \sin \left(\frac{2\pi x_2}{e^{\alpha x_1}} \right)$$

with

$$f(x) = \begin{cases} e^{-1/x} e^{-1/(2\pi-x)} & \text{if } 0 < x < 2\pi, \\ 0 & \text{if } x = 0. \end{cases}$$

Σ is controllable on M and weakly observable (however, the observability rank condition is not satisfied at any x such that $x_1 = 0$).

One can check that the indistinguishability relation J of Σ is exactly the equivalence relation R taken in the first part of the example.

This constitutes a counterexample of Theorem 3, since $\pi_1(M) = Z$ has all its nonzero elements of infinite order.

REFERENCES

- [1] L. AUSLANDER AND R. E. MACKENZIE, *Introduction to Differentiable Manifolds*, Dover, New York, 1963.
- [2] W. M. BOOTHBY, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.
- [3] J. P. GAUTHIER AND G. BORNARD, *Uniqueness of weakly minimal analytic realizations*, IEEE Trans. Aut. Control, 28 (1983), pp. 111–113.
- [4] ———, *Existence and uniqueness of minimal realizations in the C^∞ case*, Systems and Control Letters, 1 (May 1982), pp. 395–398.
- [5] A. J. KRENER AND R. HERMANN, *Nonlinear controllability and observability*, IEEE Trans. Aut. Contr., 22, (1977), pp. 728–740.
- [6] C. LODBRY, *Bases mathématiques de la théorie des asservissements nonlinéaires*, Université de Bordeaux 1975.
- [7] J. P. SERRE, *Lie Groups and Lie Algebras*, Benjamin, New York, 1975.
- [8] H. J. SUSSMANN, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. System Theory, 10 (1977), pp. 263–284.
- [9] ———, *A generalization of the closed subgroup theorem to quotients of an arbitrary manifold*, J. Differential Geometry, 10 (1975), pp. 151–166.
- [10] ———, *Some properties of vector field systems that are not altered by small perturbations*, J. Differential Equations (1976), pp. 292–315.
- [11] ———, *On quotients of manifolds, a generalization of their closed subgroup theorem*, Bull. Amer. Math. Society, 80 (1974), pp. 573–575.
- [12] ———, *Orbits of families of vector fields and integrability of distributions*, Trans. Amer. Math. Soc., 120 (1979).

EQUILIBRIUM POINT THEOREMS FOR TWO-PERSON GAMES*

JÜRGEN KINDLER†

Abstract. In this paper (ε) -equilibrium point theorems for finitely additive and σ -additive mixed extensions of noncooperative two-person games are derived. Especially, some well-known minimax theorems of Fenstad, Teh Tjoe-Tie, Wald, Young, and others are generalized to the nonzero-sum case.

Key words. two-person game, (ε) -equilibrium point, finitely additive mixed strategy, discrete mixed extension, double limit condition

1. Introduction. In the following, let $\Gamma = (X, Y, a_1, a_2)$ denote a (noncooperative two-person) *game*: X and Y are the (nonvoid) strategy sets of player 1 and player 2, and $a_i: X \times Y \rightarrow \mathbb{R}$ are the *bounded* payoff functions: If both players choose strategies $x \in X$ and $y \in Y$, respectively, then player i receives the (possibly negative) amount $a_i(x, y)$.

A game $\Gamma' = (S, T, a'_1, a'_2)$ is called a *subgame* of Γ if S and T are nonvoid subsets of X and Y and if the functions $a'_i: S \times T \rightarrow \mathbb{R}$ are the restrictions of the a_i . Sometimes it is convenient to write (S, T, a_1, a_2) instead of (S, T, a'_1, a'_2) .

For $\varepsilon \geq 0$ a pair $(\hat{x}, \hat{y}) \in X \times Y$ is called an ε -*equilibrium point* if for every $x \in Y$ and $y \in Y$ we have

$$a_1(\hat{x}, \hat{y}) \geq a_1(x, \hat{y}) - \varepsilon$$

and

$$a_2(\hat{x}, \hat{y}) \geq a_2(\hat{x}, y) - \varepsilon.$$

0-equilibrium points are usually called (*Nash*) *equilibrium points*. (ε) -equilibrium points are characterized by the fact that neither player can increase his payoff (up to ε) by a unilateral deviation.

Example 1.1. A strategy $x^* \in X$ ($y^* \in Y$) is called an *equalizer* if the function $a_2(x^*, \cdot)$ ($a_1(\cdot, y^*)$) is constant. Of course, every pair (x^*, y^*) of equalizers is an equilibrium point. Pairs of equalizers are called *simple equilibrium points*.

In the present paper some theorems on the existence of (ε) -equilibrium points are derived. Our starting point is the following fundamental theorem due to Nikaido and Isoda [43]:

THEOREM 1.1. *Let X and Y be two convex and compact subsets of two topological vector spaces. Assume that the functions $a_1(\cdot, y)$, $y \in Y$ and $a_2(x, \cdot)$, $x \in X$ are concave and the functions $a_1(x, \cdot)$, $x \in X$ and $a_2(\cdot, y)$, $y \in Y$ are continuous. If, moreover, the function $a_1 + a_2$ is continuous¹ then the game $\Gamma = (X, Y, a_1, a_2)$ has an equilibrium point.*

Remark 1.2. Nikaido and Isoda [43] used Brouwer's fixed point theorem to prove Theorem 1.1. On the other hand, this fixed point theorem is an easy consequence of Theorem 1.1. Let X be a compact and convex subset of \mathbb{R}^n and $g: X \rightarrow X$ a continuous function. Consider the game $\Gamma = (X, X, a_1, a_2)$ with $a_1(x, y) = \sum_{i=1}^n |y_i - g(y)_i| - \sum_{i=1}^n |x_i - g(y)_i|$ and $a_2(x, y) = -\sum_{i=1}^n |x_i - y_i|$, $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$. By Theorem 1.1, Γ has an equilibrium point (\hat{x}, \hat{y}) . It is easily seen that $\hat{y} = \hat{x} = g(\hat{y})$ must hold.

* Received by the editors October 5, 1982, and in revised form September 15, 1983.

† Department of Mathematics, Technische Hochschule, 6100 Darmstadt, Federal Republic of Germany.

¹ In the following, products of topological spaces will always be endowed with the product topology.

This observation, which is in essence due to Fan [15], illustrates the philosophy that various problems can be transformed into a game such that an (ε) -equilibrium point of the game provides a (ε) -solution of the original problem. Compare § 6 in the book of Parthasarathy and Raghavan [45] and the papers [9], [14], [29], [30], [31], [36], [41], [52] for further examples of this type.

Of course, there are also more practical applications. Suppose that two decision makers with (partially) conflicting interests can take influence into the realisation of a process. If their preference of possible results can be expressed in terms of a real function, then such a situation may be interpreted as a two-person game. Let us sketch some characteristic examples.

a) *Duopoly*. In a competitive economy two firms try to maximize their gain. A typical example due to Wald is described in Burger's book [11, § 5]. Further interrelations between game theory and economic situations can also be found in the books of Aubin [1], Friedman [17], Marshak and Selten [35], Nikaido [42], and Rosenmüller [49].

b) *Bargaining*. Two individuals, whose interests are neither completely opposed nor completely coincident, may either agree on cooperation for mutual benefit or they may choose noncooperative actions if cooperation fails. In [37], [40] Nash proposed a solution concept for such situations where the cooperative "bargaining game" is reduced to a noncooperative "threat game." We refer to the book of Rauhut, Schmitz and Zachow [48] and the survey paper of Jansen and Tijs [23] for further information.

c) *Inspection*. An inspection authority tries to prevent an operator, a firm, or a country from an illegal action. A typical example is the development of nuclear safeguards methods. A game theoretic treatment of such problems was first proposed by Bierlein [7], [8] and developed further by Avenhaus, Frick, Höpfinger and others [2], [3], [4], [21], [22].

d) *Decision making under uncertainty*. The classical example is the general statistical decision problem which, according to Wald [60], may be interpreted as a game between the statistician and "nature," a second fictive player.

Other examples are search problems such as the repair of a complex system where "nature hides" a defect, inventory problems where a quantity of a specific item has to be stocked without knowing future demand [47], [50], or the problem of choosing optimal insurance contracts [5], [10].

If a game does not satisfy such concavity conditions as in Theorem 1.1, then one usually passes to a mixed extension, where the players use probability measures or, more generally, probability contents as "mixed strategies." It is a fundamental result in game theory that the mixed extension of a game with finite strategy sets does always have an equilibrium point. This theorem, which is due to Nash [38], [39], is an immediate consequence of Theorem 1.1.

Because the set of probability *contents* is compact, it will prove useful to derive first an equilibrium point theorem for *finitely additive* mixed extensions. By combining this theorem with an appropriate integral representation theorem or with an approximation theorem, we are able to formulate various (ε) -equilibrium point theorems. (A similar technique was used by the author [28] in the zero-sum case.)

For other proof techniques compare the survey paper [57].

If S is a nonvoid set, then let M_S denote the set of all probability contents on the power set $\mathfrak{P}(S)$, i.e. the set of all additive $\rho: \mathfrak{P}(S) \rightarrow [0, 1]$ with $\rho(S) = 1$. We embed M_S into $[0, 1]^{\mathfrak{P}(S)}$. Then, by Tikhonov's theorem, M_S is compact and for every $f \in B(S) := \{f \in \mathbb{R}^S: \|f\| := \sup_{x \in X} |f(x)| < \infty\}$ the integral $\rho \rightarrow \int_S f(s) \rho(ds)$ is continuous on M_S .

As usual, we embed S into M_S by identifying s with the Dirac measure ε_s . The convex hull of the set of Dirac measures, i.e. the set of all probability measures on $\mathfrak{P}(S)$ with finite support, will be denoted by P_S .

For every game $\Gamma = (X, Y, a_1, a_2)$ we define a game $\tilde{\Gamma} = (M_X, M_Y, A_1, A_2)$ according to

$$A_1(\mu, \nu) = \int_Y \int_X a_1(x, y) \mu(dx) \nu(dy)$$

and

$$A_2(\mu, \nu) = \int_X \int_Y a_2(x, y) \nu(dy) \mu(dx)$$

for $\mu \in M_X$, $\nu \in M_Y$. Then Γ is a subgame of $\tilde{\Gamma}$. The elements of M_X and M_Y are called *finitely additive mixed strategies* and $\tilde{\Gamma}$ is called the (asymmetric) *finitely additive mixed extension* of Γ . The subgame $\Gamma^d = (P_X, P_Y, A_1, A_2)$ of $\tilde{\Gamma}$ is the *discrete mixed extension* of Γ . The elements of P_X and P_Y are called *discrete mixed strategies*.

To every game $\Gamma = (X, Y, a_1, a_2)$ we associate two pseudometrics on Y and X , respectively, according to

$$d_1(\Gamma)(y_1, y_2) = \sup_{x \in X} |a_1(x, y_1) - a_1(x, y_2)|,$$

$$d_2(\Gamma)(x_1, x_2) = \sup_{y \in Y} |a_2(x_1, y) - a_2(x_2, y)|.$$

$d_1(\Gamma)$ and $d_2(\Gamma)$ are known as *intrinsic*, *Helly*, or *natural pseudometrics* ([60], [34], [59]).

2. Zero-sum games. In this section we consider (two-person) *zero-sum games*, i.e. games $\Gamma = (X, Y, a_1, a_2)$ with $a_1 + a_2 = 0$. It is convenient to write (X, Y, a) with $a := a_1 = -a_2$ instead of (X, Y, a_1, a_2) .

A pair $(\hat{x}, \hat{y}) \in X \times Y$ is an equilibrium point in (X, Y, a) if and only if it is a saddle point, i.e.

$$\max_{x \in X} a(x, \hat{y}) = \min_{y \in Y} a(\hat{x}, y),$$

and (X, Y, a) has an ε -equilibrium point for every $\varepsilon > 0$ if and only if it is *strictly determined*, i.e.

$$\inf_{y \in Y} \sup_{x \in X} a(x, y) = \sup_{x \in X} \inf_{y \in Y} a(x, y).$$

ε -equilibrium point theorems for zero-sum games (commonly called “minimax theorems”) have been studied extensively. For details, the reader may consult Parthasarathy and Raghavan [45, Chap. 5] and Yanovskaya’s survey papers [62], [63].

The sets of functions $a(X, \cdot) = \{a(x, \cdot) : x \in X\}$ and $a(\cdot, Y) = \{a(\cdot, y) : y \in Y\}$ are the *risk sets* of player 1 and player 2, respectively.

If X and Y are topological spaces and if $a(\cdot, Y) \subset C(X) := \{f \in B(X) : f \text{ continuous}\}$ and $a(X, \cdot) \subset C(Y)$ holds, then a will be called *bicontinuous*.

For a zero-sum game $\Gamma = (X, Y, a)$ we have

$$A_1(\mu, \nu) = \int_Y \int_X a(x, y) \mu(dx) \nu(dy)$$

and

$$A_2(\mu, \nu) = - \int_X \int_Y a(x, y) \nu(dy) \mu(dx)$$

for the finitely additive mixed extension $\tilde{\Gamma} = (M_X, M_Y, A_1, A_2)$. Of course, we have $A_1 = -A_2 =: A$ on $M_X \times P_Y$ and on $P_X \times M_Y$. However, the finitely additive mixed extension of a zero-sum game need not be zero-sum as the following example shows.

Example 2.1. The game “More Money”. The zero-sum game $\Gamma = (X, Y, a)$ with $X = Y = \mathbb{N}$ and $a(x, y) = 1, 0, -1$ for $x > y, x = y, x < y$ has been introduced by Wald [60]. Related versions of this game were studied by Ville [58] and Bierlein [6] (compare [32]). Simple calculation shows that for $0 \leq \varepsilon < 1$ the set of ε -equilibrium points in Γ^d is empty.

For $\hat{\mu} \in M_X$ with $\hat{\mu}(\{x\}) = 0$ for all $x \in X$ we get

$$A_1(\hat{\mu}, \nu) = 1 \quad \text{for all } \nu \in M_Y,$$

and for $\hat{\nu} \in M_Y$ with $\hat{\nu}(\{y\}) = 0$ for all $y \in Y$ we have

$$A_2(\mu, \hat{\nu}) = 1 \quad \text{for all } \mu \in M_X.$$

Therefore, each such pair $(\hat{\mu}, \hat{\nu})$ is a simple equilibrium point in $\tilde{\Gamma}$. Now, let us show that there are no other equilibrium points in $\tilde{\Gamma}$. Assume that (μ^*, ν^*) is an equilibrium point in $\tilde{\Gamma}$ with $\nu^*(\{n\}) = \delta > 0$ for some $n \in \mathbb{N}$. Then we have $A_1(\mu^*, \nu^*) \geq A_1(\hat{\mu}, \nu^*) = 1$. This implies

$$\begin{aligned} 0 &= 1 - A_1(\mu^*, \nu^*) \\ &= \int_Y [\mu^*(\{y\}) + 2\mu^*(\{1, \dots, y-1\})] \nu^*(dy) \\ &\geq [\mu^*(\{n\}) + 2\mu^*(\{1, \dots, n-1\})] \delta \geq 0. \end{aligned}$$

In particular, we have $\mu^*(\{1, \dots, n\}) = 0$. Now we distinguish two cases.

Case 1. There is an $m \in \mathbb{N}$ with $\mu^*(\{m\}) > 0$. As above, we can show $\nu^*(\{1, \dots, m\}) = 0$. But $\mu^*(\{1, \dots, n\}) = 0$ and $\mu^*(\{m\}) > 0$ imply $m > n$ in contradiction to $\nu^*(\{1, \dots, m\}) = 0$ and $\nu^*(\{n\}) > 0$.

Case 2. $\mu^*(\{m\}) = 0$ for all $m \in \mathbb{N}$. From $A_2(\mu^*, \nu^*) \geq A_2(\mu^*, \hat{\nu}) = 1$ we conclude:

$$\begin{aligned} 0 &= 1 - A_2(\mu^*, \nu^*) \\ &= \int_{\{n, n+1, \dots\}} [\nu^*(\{1, \dots, x\}) + \nu^*(\{1, \dots, x-1\})] \mu^*(dx) \\ &\geq \nu^*(\{n\}) \mu(\{n, n+1, \dots\}) = \delta, \end{aligned}$$

a contradiction.

So we have shown $\nu^*(\{n\}) = 0$ for all $n \in \mathbb{N}$. By symmetry we get $\mu^*(\{n\}) = 0$ for all $n \in \mathbb{N}$.

DEFINITION. Let X and Y be nonvoid sets. Then

a) SDLC ($X \times Y$) denotes the set of all $a \in B(X \times Y)$ which satisfy the following *strong double limit condition* (SDLC):

For all sequences (x_m) in X and (y_n) in Y

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} a(x_m, y_n) = \lim_{n \rightarrow \infty} \lim_{m \rightarrow \infty} a(x_m, y_n)$$

holds whenever the iterated limits exist.

b) $ULC(X \times Y)$ denotes the set of all $a \in B(X \times Y)$ which satisfy the *uniform limit condition* (ULC):

For all sequences (x_m) in X and (y_n) in Y such that $\alpha_n := \lim_{m \rightarrow \infty} a(x_m, y_n)$ exists for every $n \in \mathbb{N}$ we have

$$\lim_{m \rightarrow \infty} \sup_{n \in \mathbb{N}} |a(x_m, y_n) - \alpha_n| = 0.$$

c) $SEP(X \times Y)$ denotes the set of all $a \in B(X \times Y)$ of the form

$$a(x, y) = \sum_{j=1}^k f_j(x)g_j(y), \quad f_j \in B(X), \quad g_j \in B(Y), \quad j \leq k \in \mathbb{N}.$$

Zero-sum games (X, Y, a) with $a \in SEP(X \times Y)$ are called *separable* or *polynomial-like* [12], [13], [18], [25].

Remark 2.1.

a) $SEP(X \times Y) \subset ULC(X \times Y) \subset SDLC(X \times Y)$.

b) $SEP(X \times Y) = B(X \times Y)$ for X (or Y) finite.

c) $a \in SDLC(X \times Y)$ if and only if $a' \in SDLC(S \times T)$ for every subgame (S, T, a') of (X, Y, a) .

d) The same as c) for ULC.

In the following, SDLC and ULC play a fundamental role. Therefore, it is desirable to have some equivalent definitions which are easy to handle in applications. We first recall some topological notions.

DEFINITION. Let U be a topological space and $F \subset C(U)$.

a) F is called *equicontinuous* if for every $\varepsilon > 0$ and every $u_0 \in U$ there exists a neighbourhood W of u_0 such that $|f(u) - f(u_0)| < \varepsilon$ for all $u \in W, f \in F$.

b) F is called *quasi-equicontinuous* if for every net $(u_i: i \in I)$ in U , for every cluster point u_0 of this net, and for every $\varepsilon > 0$ there is a finite $K \subset I$ such that $\sup_{f \in F} \min_{i \in K} |f(u_i) - f(u_0)| < \varepsilon$.

c) If U is endowed with a pseudometric d , then F is called *uniformly equicontinuous* if for every $\varepsilon > 0$ there is a $\delta > 0$ such that $\sup_{f \in F} |f(u_1) - f(u_2)| < \varepsilon$ for all $u_1, u_2 \in U$ with $d(u_1, u_2) < \delta$.

DEFINITION. Let U be a topological space. Then

a) U is called *pseudocompact* if every zero-sum game (U, \mathbb{N}, b) with $b(\cdot, 1) \geq b(\cdot, 2) \geq \dots$ and $b(\cdot, \mathbb{N}) \in C(U)$ is strictly determined.

b) U is called *countably compact* if every zero-sum game (U, \mathbb{N}, b) with $b(\cdot, 1) \geq b(\cdot, 2) \geq \dots$ and upper-semicontinuous $b(\cdot, n), n \in \mathbb{N}$ is strictly determined.

c) U is called *compact* if every zero-sum game (U, V, b) for which $b(\cdot, V)$ is filtering downwards (i.e. for $v_1, v_2 \in V$ there is a $v_0 \in V$ such that $b(\cdot, v_0) \leq \min\{b(\cdot, v_1), b(\cdot, v_2)\}$) and the functions $b(\cdot, v), v \in V$ are upper-semicontinuous, is strictly determined.

The well-known ([19], [28], [33], [54]) fact that our “game theoretic” definitions are equivalent with the usual ones is commonly called “Dini’s theorem.”

PROPOSITION 2.1. For a zero-sum game $\Gamma = (X, Y, a)$ the following conditions are equivalent:

(a) $a \in ULC(X \times Y)$.

(b) X is totally bounded with respect to the pseudometric $d_2(\Gamma)$.

(c) There exists a pseudometric on X such that X is totally bounded and $a(\cdot, Y)$ is uniformly equicontinuous.

(d) A_1 is continuous.

(e) $a(\cdot, Y)$ is a relatively compact subset of $B(X)$ in the norm topology.

(f) Γ is a subgame of a zero-sum game (U, V, b) such that U is a pseudocompact topological space and $b(\cdot, V)$ is equicontinuous.

(g) Γ is a subgame of a zero-sum game (U, V, b) such that U is a pseudocompact and V is a compact topological space and b is continuous.

(h) a is the uniform limit of a sequence $a_n \in \text{SEP}(X \times Y)$.

(*) Further conditions (a)*–(g)* with the roles of the players reversed.

Zero-sum games which satisfy condition (b) have been studied by Wald [60], Fenstad [16], and by Kretkowski and Telgárski [34] who called these games “totally bounded.” In [34, Thm. 2] a further equivalent condition can be found.

Proof. We shall show that Proposition 2.1 is a reformulation of well-known results. If X and Y are endowed with the discrete topology, then a is continuous on $X \times Y$. Furthermore X , say, is completely regular, $C(X) = B(X)$, and the topological dual $C(X)'$ of $C(X)$ is the linear hull of M_X . Hence, the assumptions of [46, Thm. (6.2)] are satisfied. Our conditions (a), (b), (e), and (h) coincide with Pták’s conditions 6° , 2° , 5° and 11° in the above-mentioned theorem and therefore are equivalent. For the proof of (g) \Rightarrow (f) compare [46, p. 575]. Applying Glicksberg’s version of Ascoli’s theorem [19, p. 259], we may conclude from (f) that $b(\cdot, V)$ is relatively compact in $C(U)$ or in $B(U)$, respectively, and by (e) \Rightarrow (a) we have $b \in \text{ULC}(U \times V)$. But this implies $a \in \text{ULC}(X \times Y)$ by Remark 2.1d). Thus we have shown (f) \Rightarrow (a). As (b) \Rightarrow (c) and (h) \Rightarrow (d) \Rightarrow (g) is trivial, everything is proved.

Remark 2.2. Let $\Gamma = (X, Y, a)$ be a zero-sum game with $a \in \text{ULC}(X \times Y)$. For $\mu \in M_X$, $\nu \in M_Y$ we define a precontent $\mu \otimes \nu$ on the semialgebra of all rectangles in $X \times Y$ according to

$$\mu \otimes \nu(A \times B) = \mu(A) \cdot \nu(B), \quad A \in \mathfrak{P}(X), \quad B \in \mathfrak{P}(Y).$$

Then we have

$$\int_{X \times Y} s(x, y) \mu \otimes \nu(d(x, y)) = \int \int_{XY} s(x, y) \nu(dy) \mu(dx) = \int \int_{YX} s(x, y) \mu(dx) \nu(dy)$$

for every $s \in \text{SEP}(X \times Y)$. From (a) \Rightarrow (h) we conclude that the same is true for $s = a$. This is Fenstad’s “Fubini-theorem” [16].

The following characterization of SDLC summarizes, among others, results of Pták [46], Simons [51], Young [64] and the author [29], [30].

PROPOSITION 2.2. *For a zero-sum game $\Gamma = (X, Y, a)$ the following conditions are equivalent.*

- (i) $a \in \text{SDLC}(X \times Y)$.
- (k) P_X is $d_2(\tilde{\Gamma})$ -dense in M_X .
- (l) A_1 is bicontinuous.
- (m) $a(\cdot, Y)$ is a relatively compact subset of $B(X)$ in the $\sigma(B(X), B(X)')$ -topology.
- (n) Γ is a subgame of a zero-sum game (U, V, b) such that U is a countably compact topological space and $b(\cdot, V)$ is quasi-equicontinuous.
- (o) Γ is a subgame of a zero-sum game (U, V, b) such that U is a pseudocompact and V is a countably compact topological space and b is bicontinuous.
- (p) $\tilde{\Gamma}$ is a zero-sum game.
- (q) For every subgame (S, T, a) of Γ the games (P_S, P_T, A_i) , $i \in \{1, 2\}$, are strictly determined.

(*) Further conditions (i)*–(o)* with the roles of the players reversed.

Another equivalent condition was given by Young [64, Thm. 7].

Proof. The equivalence of conditions (i), (m), (p), and (q) was shown in [30, Corollary].

(o) \Rightarrow (i) is true by [29, 3, 12]. (Compare [46], [51], [64] for related results.)

(i) \Rightarrow (k). By [30, Corollary] (i) implies

$$(1) \quad \forall \varepsilon > 0 \quad \forall \mu \in M_X \quad \exists p \in P_X \quad \forall y \in Y: |A(p, y) - A(\mu, y)| < \varepsilon$$

and

$$(2) \quad \forall \varepsilon > 0 \quad \forall \nu \in M_Y \quad \exists q \in P_Y \quad \forall x \in X: |A(x, q) - A(x, \nu)| < \varepsilon.$$

But (2) implies

$$d_2(\tilde{\Gamma})(\mu_1, \mu_2) = \sup_{q \in P_Y} |A(\mu_1, q) - A(\mu_2, q)|, \quad \mu_1, \mu_2 \in M_X.$$

If for $\varepsilon > 0$, $\mu \in M_X$ we choose a $p \in P_X$ according to (1), then we have

$$d_2(\tilde{\Gamma})(\mu, p) = \sup_{q \in P_Y} |A(\mu, q) - A(p, q)| \leq \sup_{q \in P_Y} \sum_{y \in Y} q(\{y\}) |A(\mu, y) - A(p, y)| \leq \varepsilon.$$

(k) \Rightarrow (p). For $\mu \in M_X$ and $p \in P_X$ we have

$$\sup_{\nu \in M_Y} |A_1(\mu, \nu) + A_2(\mu, \nu)| \leq 2d_2(\tilde{\Gamma})(\mu, p).$$

(n) \Rightarrow (i). Let (x_m) and (y_n) be sequences with existing double limits $\alpha = \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} a(x_m, y_n)$ and $\beta = \lim_{n \rightarrow \infty} \lim_{m \rightarrow \infty} a(x_m, y_n)$, and let $\varepsilon > 0$. Without loss of generality we may assume

$$(3) \quad |\alpha - \lim_{n \rightarrow \infty} a(x_m, y_n)| < \varepsilon \quad \text{for all } m \in \mathbb{N}.$$

As U is countably compact, the sequence (x_m) has a cluster point $u_0 \in U$, and we get $\beta = \lim_{n \rightarrow \infty} b(u_0, y_n)$. As $b(\cdot, V)$ is quasi-equicontinuous, there exist natural numbers m_1, \dots, m_r such that

$$\sup_{v \in V} \min_{\rho \leq r} |b(x_{m_\rho}, v) - b(u_0, v)| < \varepsilon.$$

Hence

$$\min_{\rho \leq r} |a(x_{m_\rho}, y_n) - b(u_0, y_n)| < \varepsilon \quad \text{for all } n \in \mathbb{N}.$$

In particular, we have

$$\min_{\rho \leq r} |\lim_{n \rightarrow \infty} a(x_{m_\rho}, y_n) - \beta| < \varepsilon,$$

i.e., there is an $s \in \mathbb{N}$ such that

$$(4) \quad |\lim_{n \rightarrow \infty} a(x_s, y_n) - \beta| < \varepsilon.$$

From (3), (4), and $\varepsilon \rightarrow 0$ we get $\alpha = \beta$.

(o) \Rightarrow (n)*. Choose (U, V, b) according to (o). Let (v_0) be a cluster point of a net $(v_i: i \in I)$ in V . For the zero-sum game (U, I, c) with $c(u, i) = |b(u, v_i) - b(u, v_0)|$ we have

$$(5) \quad \sup_{p \in P_U} \inf_{q \in P_I} C(p, q) = \sup_{p \in P_U} \inf_{i \in I} \sum_{u \in U} p(\{u\}) c(u, i) = 0.$$

By (o) \Rightarrow (i) we get $b \in \text{SDLC}(U \times V)$. This implies $c \in \text{SDLC}(U \times I)$. From (i) \Rightarrow (q) and (5) we conclude

$$\inf_{q \in P_I} \sup_{p \in P_U} C(p, q) = 0;$$

i.e. for each $\varepsilon > 0$ there is a $q \in P_I$ with (finite) support $K \subset I$ such that

$$\varepsilon \geq \sup_{u \in U} \sum_{i \in K} q(\{i\}) c(u, i) \geq \sup_{u \in U} \min_{i \in K} |b(u, v_i) - b(u, v_0)|.$$

Thus we have shown that $b(U, \cdot)$ is quasi-equicontinuous.

(n) $^* \Rightarrow$ (i). Compare the proof of (n) \Rightarrow (i). Now, as (p) \Rightarrow (l) \Rightarrow (o) and (i) \Leftrightarrow (i) * are trivially satisfied, everything is proved.

Remark 2.3. $\text{ULC}(X \times Y)$ and $\text{SDLC}(X \times Y)$ are closed vector sublattices of $B(X \times Y)$, and $\text{SEP}(X \times Y)$ is dense in $\text{ULC}(X \times Y)$.

This follows from (a) \Leftrightarrow (d) \Leftrightarrow (g) \Leftrightarrow (h) and (i) \Leftrightarrow (l) \Leftrightarrow (o).

3. The main results. The following equilibrium point theorem for the (asymmetric) finitely additive mixed extension of a game is fundamental for our further investigations.

THEOREM 3.1. *Let $\Gamma = (X, Y, a_1, a_2)$ be a game such that $a_1 \in \text{SDLC}(X \times Y)$ and $(a_1 + a_2) \in \text{ULC}(X \times Y)$. Then the finitely additive mixed extension $\tilde{\Gamma} = (M_X, M_Y, A_1, A_2)$ has an equilibrium point.*

Proof. From (i) \Rightarrow (p), applied to the zero-sum game (X, Y, a_1) , we get $A_1(\mu, \nu) = \int_X \int_Y a_1(x, y) \nu(dy) \mu(dx)$, $\mu \in M_X, \nu \in M_Y$. Hence, by (a) \Rightarrow (d) * , $A_1 + A_2$ is continuous, and Theorem 1.1 may be applied to $\tilde{\Gamma}$.

Remark 3.1. In the zero-sum case $a_1 + a_2 = 0$ in Theorem 3.1 the set of equilibrium points of $\tilde{\Gamma}$ coincides with the set of “finitely additive solutions” which were defined in [32] for every two-person zero-sum game. Moreover, this special case is a generalization of Fenstad’s saddle point theorem [16, Thm. II]; a related result is due to Karlin [24, Thm. 10]).

LEMMA 3.1. *Let $\Gamma = (X, Y, a_1, a_2)$ be a game with $a_i \in \text{SDLC}(X \times Y)$, $i \in \{1, 2\}$, and let $\delta > 0$. Then:*

a) *For each $\hat{\mu} \in M_X$ there exists a $\hat{p} \in P_X$ such that*

$$\sup_{\nu \in M_Y} |A_i(\hat{\mu}, \nu) - A_i(\hat{p}, \nu)| \leq \delta, \quad i \in \{1, 2\}.$$

b) *For each $\hat{\nu} \in M_Y$ there exists a $\hat{q} \in P_Y$ such that*

$$\sup_{\mu \in M_X} |A_i(\mu, \hat{\nu}) - A_i(\mu, \hat{q})| \leq \delta, \quad i \in \{1, 2\}.$$

Proof. a) Consider the zero-sum game $\Gamma' = (X, Z, a)$ with $Z = Y \times \{1, 2\}$ and $a(x, (y, i)) = a_i(x, y)$. $a_i \in \text{SDLC}(X \times Y)$, $i \in \{1, 2\}$ implies $a \in \text{SDLC}(X \times Z)$. By applying (i) \Rightarrow (k) to Γ' we conclude that for each $\hat{\mu} \in M_X$ there exists a $\hat{p} \in P_X$ such that $d_2(\tilde{\Gamma}')(\hat{\mu}, \hat{p}) \leq \delta$. In particular, we have

$$|A_i(\hat{\mu}, y) - A_i(\hat{p}, y)| \leq \delta \quad \text{for all } y \in Y, i \in \{1, 2\}.$$

Together with (i) \Rightarrow (p) the assertion follows. The proof of b) is similar.

The following theorem generalizes Young’s minimax theorem [64, Thm. 14]. Further generalizations of Young’s minimax theorem can be found in [26], [30], and [31].

THEOREM 3.2. Let $\Gamma = (X, Y, a_1, a_2)$ be a game such that $a_1 \in \text{SDLC}(X \times Y)$ and $(a_1 + a_2) \in \text{ULC}(X \times Y)$. Then, for every $\varepsilon > 0$, the discrete mixed extension $\Gamma^d = (P_X, P_Y, A_1, A_2)$ has an ε -equilibrium point.

Proof. By Theorem 3.1, $\tilde{\Gamma}$ has an equilibrium point $(\hat{\mu}, \hat{\nu})$. From Remarks 2.1a) and 2.3 we conclude $a_2 \in \text{SDLC}(X \times Y)$. Now, for $\delta = \varepsilon/3$, choose \hat{p} and \hat{q} according to Lemma 3.1. Then

$$A_1(\hat{p}, \hat{q}) \geq A_1(\hat{\mu}, \hat{q}) - \delta \geq A_1(\hat{\mu}, \hat{\nu}) - 2\delta \geq A_1(p, \hat{\nu}) - 2\delta \geq A_1(p, \hat{q}) - \varepsilon$$

for all $p \in P_X$, and

$$A_2(\hat{p}, \hat{q}) \geq A_2(\hat{\mu}, \hat{q}) - \delta \geq A_2(\hat{\mu}, \hat{\nu}) - 2\delta \geq A_2(\hat{\mu}, q) - 2\delta \geq A_2(\hat{p}, q) - \varepsilon$$

for all $q \in P_Y$.

4. Examples. We now present some consequences of the foregoing results. It is obvious that, by combining Propositions 2.1 and 2.2 with Theorems 3.1 and 3.2, a great variety of (ε) -equilibrium point theorems may be constructed for games with (quasi-)equicontinuous or with (bi-)continuous payoffs. We list some characteristic examples, and we invite the reader to derive further results by the same method.

Example 4.1. Let the game $\Gamma = (U, V, b_1, b_2)$ satisfy the following conditions:

(α) U is a countably compact topological space.

(β) $b_1(\cdot, V)$ is quasi-equicontinuous.

(γ) $\{b_1(\cdot, v) + b_2(\cdot, v) : v \in V\}$ is equicontinuous.

Then, for every $\varepsilon > 0$, the discrete mixed extension (P_X, P_Y, A_1, A_2) of every subgame (X, Y, a_1, a_2) of Γ has an ε -equilibrium point.

Proof. Let (X, Y, a_1, a_2) be a subgame of (U, V, b_1, b_2) . From (α) and (β) we conclude via (n) \Rightarrow (i) that $a_1 \in \text{SDLC}(X \times Y)$, and from (α) and (γ) we get via (f) \Rightarrow (a) that $(a_1 + a_2) \in \text{ULC}(X \times Y)$. Hence Theorem 3.2 may be applied.

Example 4.2. (Tijs [56]). Let $\Gamma = (X, Y, a_1, a_2)$ be a game with finite X . Then, for every $\varepsilon > 0$, Γ^d has an ε -equilibrium point.

Proof. By Remark 2.1 we have $a_i \in \text{ULC}(X \times Y)$, and we may apply Theorem 3.2.

Example 4.3. Let $\Gamma = (X, Y, a_1, a_2)$ be a game such that, for every $\varepsilon > 0$, X has a finite ε -complete subset. (A subset $S \subset X$ is called ε -complete if for each $x \in X$ there exists an $s \in S$ such that $a_1(s, y) \geq a_1(x, y) - \varepsilon$ for all $y \in Y$.) Then Γ^d has an ε -equilibrium point for every $\varepsilon > 0$.

Similar results were obtained by Teh Tjoe-Tie [53] and the author [29] in the zero-sum case.

Proof. For $\varepsilon > 0$ choose an $\varepsilon/2$ -complete finite subset $S \subset X$. According to Example 4.2, (P_S, P_Y, A_1, A_2) has an $\varepsilon/2$ -equilibrium point (\hat{p}, \hat{q}) . This point is an ε -equilibrium point of Γ^d .

In the above proof we implicitly used an observation of Tijs [56, (2.6)].

The following "Fubini-type" theorem is, in essence, well known ([30], [46], [51], [55]). It is an easy consequence of the implications (o) \Rightarrow (k) and (o) \Rightarrow (p).

Example 4.4. Let $\Gamma = (U, V, b_1, b_2)$ be a game. Assume that U is a pseudocompact and V is a countably compact topological space and that b_1 and b_2 are bicontinuous. Let P_U^B (P_V^B) denote the set of all Baire probability measures on U (V). Then, for all $p \in P_U^B$, $q \in P_V^B$, $i \in \{1, 2\}$ we have

(α) $\int_U b_i(u, \cdot) p(du) \in C(V)$.

(β) $\int_V b_i(\cdot, v) q(dv) \in C(U)$.

(γ) $\int_U \int_V b_i(u, v) q(dv) p(du) = \int_V \int_U b_i(u, v) p(du) q(dv) =: B_i(p, q)$.

The game $\Gamma^B = (P_U^B, P_V^B, B_1, B_2)$ is called the *Baire mixed extension* of Γ . Observe that by the Hahn–Banach theorem every $p \in P_U^B$ can be extended to a $\mu \in M_U$. Hence, Γ^B may be considered as a subgame of $\tilde{\Gamma}$.

Example 4.5. Let the assumptions of Example 4.4 be satisfied. Moreover, let V be compact and let $b_1 + b_2$ be continuous. Then:

- a) The Baire mixed extension Γ^B has an equilibrium point.
- b) For every nonvoid $Y \subset V$ there exists a $\hat{p} \in P_U^B$, and for every $\varepsilon > 0$ there exists a $\hat{q} \in P_Y$ such that (\hat{p}, \hat{q}) is an ε -equilibrium point in the subgame (P_U^B, P_Y, B_1, B_2) of Γ^B .
- c) The discrete mixed extension of every subgame of (U, V, b_1, b_2) has an ε -equilibrium point for every $\varepsilon > 0$.

Special cases of this result are due to Glicksberg [20] and Owen [44].

Proof. Let $\phi \neq Y \subset V$ and $\Gamma' = (U, Y, b_1, b_2)$. From (o) \Rightarrow (i) and (g) \Rightarrow (a) we obtain $b_i \in \text{SDL}C(U \times Y)$, $i \in \{1, 2\}$, and $(b_1 + b_2) \in \text{ULC}(U \times Y)$.

By Theorem 3.1, $\tilde{\Gamma}'$ has an equilibrium point $(\hat{\mu}, \hat{\nu})$. For $\hat{\nu}$ and $\delta = \varepsilon/2$, choose $\hat{q} \in P_Y$ according to Lemma 3.1b). Then $(\hat{\mu}, \hat{q})$ is an ε -equilibrium point of $\tilde{\Gamma}'$. By Glicksberg's integral representation theorem [19] for every $\hat{\mu} \in M_U$, U pseudocompact, there is a $\hat{p} \in P_U^B$ such that $\int_U f d\hat{\mu} = \int_U f d\hat{p}$ for all $f \in C(U)$. Hence, (\hat{p}, \hat{q}) is an ε -equilibrium point of (P_U^B, P_Y, B_1, B_2) . In case $Y = V$, $\hat{\nu}$ does also have a representing measure $\hat{q} \in P_V^B$ and, in combination with Example 4.4, we see that (\hat{p}, \hat{q}) is an equilibrium point of Γ^B . Thus, a) and b) are proved.

For the proof of c) apply Theorem 3.2.

We note a converse to the above result:

Example 4.6. Assume that U is a topological space such that for every zero-sum game $\Gamma = (U, V, b)$ in which V is a compact topological space and b is bicontinuous either assertion a), b) or c) of Example 4.5 hold. Then U must be pseudocompact.

Proof. We have to show that every zero-sum game (U, \mathbb{N}, b) with $b(\cdot, 1) \geq b(\cdot, 2) \geq \dots$ and $b(\cdot, \mathbb{N}) \in C(U)$ is strictly determined. Without loss of generality we may assume $\inf_{n \in \mathbb{N}} b(u, n) = 0$ for every $u \in U$. (Otherwise consider (U, \mathbb{N}, c) with $c(u, n) = \max\{b(u, n) - \delta, 0\}$, $\delta > \sup_{u \in U} \inf_{n \in \mathbb{N}} b(u, n)$.) Now we apply assertion c) to the zero-sum game (U, V, \bar{b}) where V is the one point compactification $\mathbb{N} \cup \{\infty\}$ of \mathbb{N} , \mathbb{N} endowed with the discrete topology, and \bar{b} is the extension of b to $U \times V$ with $\bar{b}(u, \infty) = 0$ for every $u \in U$. According to c), for every $\varepsilon > 0$ there exists an ε -equilibrium point $(p_\varepsilon, q_\varepsilon)$ of $(P_U, P_{\mathbb{N}}, B)$, the discrete mixed extension of (U, \mathbb{N}, b) . For $k \geq \max\{n \in \mathbb{N} : n \in \text{support}(q_\varepsilon)\}$ we have

$$\sup_{u \in U} b(u, k) \leq \sup_{u \in U} B(u, q_\varepsilon) \leq \inf_{n \in \mathbb{N}} B(p_\varepsilon, n) + 2\varepsilon = \lim_{n \rightarrow \infty} \int_U b(u, n) p_\varepsilon(du) + 2\varepsilon = 2\varepsilon.$$

From $\varepsilon \rightarrow 0$ we infer

$$\inf_{n \in \mathbb{N}} \sup_{u \in U} b(u, n) \leq 0 = \sup_{u \in U} \inf_{n \in \mathbb{N}} b(u, n),$$

i.e. (U, \mathbb{N}, b) is strictly determined.

Secondly, for arbitrary $f \in C(U)$ we apply a) or b) to the zero-sum game (U, V, b) , $V = \{f\}$ and $b(u, f) = f(u)$. For $\hat{p} \in P_U^B$ chosen according to a) or b) we get $\int f d\hat{p} = \sup_{u \in U} f(u)$. In particular, every $f \in C(X)$ attains its supremum. But this implies [19] that U must be pseudocompact.

Example 4.7. Let $\Gamma = (X, Y, a_1, a_2)$ be a game such that X and Y are topological spaces. If a_1 and a_2 are continuous with compact support, then Γ^d has an ε -equilibrium point for every $\varepsilon > 0$.

Proof. Apply Example 4.5c) to the game (U, V, b_1, b_2) , where $U = X \cup \{\infty\}$ and $V = Y \cup \{\infty\}$ are the one point compactifications, $b_i = a_i$ on $X \times Y$ and $b_i = 0$ on $U \times V - X \times Y$ for $i \in \{1, 2\}$.

Next, we note a generalization of Theorem 3.1:

Example 4.8. Let the game $\Gamma = (X, Y, a_1, a_2)$ satisfy the assumptions of Theorem 3.1. Let $\Gamma' = (U, V, A_1, A_2)$ be a subgame of $\tilde{\Gamma} = (M_X, M_Y, A_1, A_2)$. If U and V are convex, then Γ' has an ε -equilibrium point for every $\varepsilon > 0$. If furthermore U and V are closed, then Γ' has an equilibrium point.

Proof. As was shown in the proof of Theorem 3.1 A_1 is bicontinuous and $A_1 + A_2$ is continuous. Hence, we may apply Example 4.5c) to the game $\tilde{\Gamma}$ to conclude that $(U, V, A_1, A_2) = (P_U, P_V, A_1, A_2)$ has an ε -equilibrium point for every $\varepsilon > 0$. If, moreover, U and V are closed, then we may apply Theorem 1.1 to Γ' .

Example 4.9. Let $\Gamma = (X, Y, a_1, a_2)$ be a game. Let \mathfrak{A} and \mathfrak{B} be σ -algebras on X and Y , respectively, such that a_1 is $\mathfrak{A} \otimes \mathfrak{B}$ -measurable. Assume that

$$a_1(x, y) + a_2(x, y) = \sum_{i=1}^k f_i(x)g_i(y), \quad x \in X, \quad y \in Y$$

where the f_i 's (g_i 's) are bounded \mathfrak{A} -measurable (\mathfrak{B} -measurable) real functions. Furthermore, let $\mu|_{\mathfrak{A}}$ and $\nu|_{\mathfrak{B}}$ be finite measures. For $c > 0$ we denote by P_X^c the set of all probability measures on \mathfrak{A} which have a μ -density f such that $f(x) \leq c$ for all $x \in X$. For $d > 0$ define P_Y^d similarly. Write

$$A_i(p, q) = \int_{X \times Y} a_i dp \otimes q, \quad p \in P_X^c, \quad q \in P_Y^d, \quad i \in \{1, 2\}.$$

Then, for all nonvoid convex subsets $U \subset P_X^c$ and $V \subset P_Y^d$, the game (U, V, A_1, A_2) has an ε -equilibrium point for every $\varepsilon > 0$. If P_X^c and P_Y^d are nonvoid, then (P_X^c, P_Y^d, A_1, A_2) has an equilibrium point.

This example generalizes results of Wald [61] and the author [27, Satz 9]. In applications, X and Y often are time intervals and the value of the density at time t may be interpreted as intensity of the activities of the corresponding player at time t .

Proof. Let $L_1(X, \mathfrak{A}, \mu)$ and the topological dual $L_\infty(X, \mathfrak{A}, \mu)$ be defined as usual. By identifying each $p \in U$ with the equivalence class of its densities we embed U into the ball

$$\Phi_c = \{\varphi \in L_\infty(X, \mathfrak{A}, \mu) : \|\varphi\|_\infty \leq c\},$$

which is $\sigma(L_\infty, L_1)$ -compact by Alaoglu's theorem. Let Ψ_d be defined similarly. Put

$$B_i(\varphi, \psi) = \int_X \int_Y a_i(x, y) \varphi(x) \psi(y) \nu(dy) \mu(dx), \quad \varphi \in \Phi_c, \quad \psi \in \Psi_d, \quad i \in \{1, 2\}.$$

Then B_1 is bicontinuous and $B_1 + B_2$ is continuous. Now, as (U, V, A_1, A_2) is a subgame of $(\Phi_c, \Psi_d, B_1, B_2)$, Example 4.5c) implies that $(U, V, A_1, A_2) = (P_U, P_V, A_1, A_2)$ has an ε -equilibrium point for every $\varepsilon > 0$. As P_X^c and P_Y^d are closed (convex) subsets of Φ_c and Ψ_d , respectively, we may apply Theorem 1.1 to the game (P_X^c, P_Y^d, A_1, A_2) .

REFERENCES

- [1] J. P. AUBIN, *Mathematical Methods of Game and Economic Theory*, Studies in Mathematics and Its Applications, 7, North-Holland, Amsterdam, 1982.
- [2] R. AVENHAUS AND E. HÖPFINGER, *Optimal sampling for safeguarding nuclear material*, Operations Research Verfahren, 11 (1972), pp. 1-12.

- [3] R. AVENHAUS AND H. FRICK, *Analyse von Fehlalarmen in Überwachungssystemen mit Hilfe von Zweipersonen-Nichtnullsummenspielen*, Operations Research Verfahren, 26 (1977), pp. 629–639.
- [4] ———, *Game theoretical treatment of material accountability problems I*, Int. J. Game Theory, 5 (1976), pp. 117–135.
- [5] D. BIERLEIN, *Spieltheoretische Modelle für Entscheidungssituationen des Versicherers*, Bl. Dt. Gesellsch. f. Versicherungsmath., 3 (1958), pp. 461–476.
- [6] ———, *Spiele mit mehr als einem Spielwert*, Arch. Math., 19 (1968), pp. 330–336.
- [7] ———, *Direkte Überwachungssysteme*, Operations Research Verfahren, 6 (1969), pp. 57–69.
- [8] ———, *Auf Bilanzen und Inventuren basierende Safeguards-Systeme*, Operations Research Verfahren, 8 (1970), pp. 36–43.
- [9] D. BLACKWELL, *Minimax and irreducible matrices*, J. Math. Anal. Appl., 3 (1961), pp. 37–39.
- [10] K. BORCH, *Equilibrium in a reinsurance market*, Econometrica, 30 (1962), pp. 424–444.
- [11] E. BURGER, *Einführung in die Theorie der Spiele*, De Gruyter, Berlin, 1966.
- [12] M. DRESHER, *Solution of polynomial-like games*, Proc. Int. Congr. Math., I, 1950, AMS, Providence, RI, 1952, pp. 334–335.
- [13] M. DRESHER, S. KARLIN AND L. S. SHAPLEY, *Polynomial games*, Ann. Math. Studies, 24 (1950), pp. 161–180.
- [14] K. FAN, *Existence theorems and extreme solutions for inequalities concerning convex functions or linear transformations*, Math. Z., 68 (1957), pp. 205–216.
- [15] ———, *Sur un théorème minimax*, C.R. Acad. Sci. Paris, 259 (1964), pp. 3925–3928.
- [16] J. E. FENSTAD, *Good strategies in general games*, Math. Z., 101 (1967), pp. 322–330.
- [17] J. W. FRIEDMAN, *Oligopoly and the Theory of Games*, North-Holland, Amsterdam, 1979.
- [18] D. GALE AND O. GROSS, *A note on polynomial and separable games*, Pacific J. Math., 8 (1958), pp. 735–744.
- [19] I. GLICKSBERG, *The representation of functionals by integrals*, Duke Math. J., 19 (1952), pp. 253–261.
- [20] ———, *A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points*, J. Amer. Math. Soc., 3 (1952), pp. 170–174.
- [21] E. HÖPFINGER, *Zuverlässige Inspektionsstrategien*, Z. Wahrsch. Verw. Gebiete, 31 (1974), pp. 35–46.
- [22] ———, *Reliable Inspection Strategies*, Math. Systems in Economics 17, A. Hain, Meisenheim am Glan, West Germany, 1975.
- [23] M. J. M. JANSEN AND S. H. TIJS, *Arbitration games. A survey*, Operations Research Proceedings, Springer, Berlin, 1980.
- [24] S. KARLIN, *Operator treatment of minimax principle*, Ann. Math. Studies, 24 (1950), pp. 133–154.
- [25] ———, *Mathematical Methods and Theory in Games, Programming, and Economics*, vol. II: *The Theory of Infinite Games*, Addison-Wesley, Reading, MA, 1959.
- [26] J. KINDLER, *Über ein Minimaxtheorem von Young*, Math. Operationsforsch. Statist., 7 (1976), pp. 477–480.
- [27] ———, *Über Spiele auf konvexen Mengen*, Methods Oper. Res., 26 (1977), pp. 695–704.
- [28] ———, *Minimaxtheoreme und das Integraldarstellungsproblem*, Manuscripta Math., 29 (1979), pp. 277–294.
- [29] ———, *Minimaxtheoreme für die diskrete gemischte Erweiterung von Spielen und ein Approximationssatz*, Math. Operationsforsch. Statist., Ser. Optimization, 11 (1980), pp. 473–485.
- [30] ———, *Some consequences of a double limit condition*, Game Theory and Mathematical Economics, O. Moeschlin and D. Pallaschke, eds., North-Holland, Amsterdam, 1981, pp. 73–82.
- [31] ———, *A minimax version of Pták's combinatorial lemma*, J. Math. Anal. Appl., 94 (1983), pp. 454–459.
- [32] ———, *A general solution concept for two-person zero-sum games*, J. Optim. Theory Appl., 40 (1983), pp. 105–119.
- [33] H. KÖNIG, *Über das von Neumannsche Minimax-Theorem*, Arch. Math., 19 (1968), pp. 482–487.
- [34] R. KRETZKOWSKI AND R. TELGÁRSKY, *On totally bounded games*, Math. Slovaca, to appear.
- [35] T. MARSHAK AND R. SELTEN, *General equilibrium with price-making firms*, Lecture Notes in Economics and Mathematical Systems 91, Springer, Berlin, 1974.
- [36] G. MINTY, *On the extension of Lipschitz, Lipschitz-Hölder continuous, and monotone functions*, Bull. Amer. Math. Soc., 76 (1970), pp. 334–339.
- [37] J. F. NASH, *The bargaining problem*, Econometrica, 18 (1950), pp. 155–162.
- [38] ———, *Equilibrium points in n-person games*, Proc. Nat. Acad. Sci. USA, 36 (1950), pp. 48–49.
- [39] ———, *Non-cooperative games*, Ann. Math., 54 (1951), pp. 286–295.
- [40] ———, *Two-person cooperative games*, Econometrica, 21 (1953), pp. 128–140.
- [41] H. NIKAIDO, *On a minimax theorem and its applications to functional analysis*, J. Math. Soc. Japan, 5 (1953), pp. 86–94.
- [42] ———, *Convex Structures and Economic Theory*, Academic Press, New York, 1968.

- [43] H. NIKAIDO AND K. ISODA, *Note on non-cooperative convex games*, Pacific J. Math., 5 (1955), pp. 807–815.
- [44] G. OWEN, *Existence of equilibrium pairs in continuous games*, Internat. J. Game Theory, 5 (1976), pp. 97–105.
- [45] T. PARTHASARATHY AND T. E. S. RAGHAVAN, *Some Topics in Two-Person Games*, American Elsevier, New York, 1971.
- [46] V. PTÁK, *An extension theorem for separately continuous functions and its application to functional analysis*, Czech. Math. J., 14(89) (1964), pp. 562–581.
- [47] B. RAUHUT, *Minimaxstrategien in einperiodischen Lagermodellen*, Operations Research Verfahren, 8 (1970), pp. 227–250.
- [48] B. RAUHUT, N. SCHMITZ AND E.-W. ZACHOW, *Spieltheorie*, Teubner, Stuttgart, 1979.
- [49] J. ROSENMÜLLER, *The Theory of Games and Markets*, North-Holland, Amsterdam, 1981.
- [50] H. SCARF, *A min-max solution of an inventory problem*, in Studies in the Mathematical Theory of Inventory and Production, K. J. Arrow, S. Karlin, and H. Scarf, eds., Stanford Univ. Press, Stanford, CA, 1958, pp. 201–209.
- [51] S. SIMONS, *The iterated limit condition, a Fubini theorem, and weak compactness*, Math. Annalen, 176 (1968), pp. 87–95.
- [52] ———, *Maximinimax, minimax, and antiminimax theorems and a result of R. C. James*, Pacific J. Math. 40 (1972), pp. 709–718.
- [53] TEH TJOE-TIE, *Minimax theorems on conditionally compact sets*, Ann. Statist., 34 (1963), pp. 1536–1540.
- [54] F. TERKELSEN, *Some minimax theorems*, Math. Scand., 31 (1972), pp. 405–413.
- [55] W. THOMSEN, *On a Fubini-type theorem and its application to game theory*, Math. Operationsforsch. Statist., Ser. Statistics, 9 (1978), pp. 419–423.
- [56] S. H. TIJS, *ϵ -equilibrium point theorems for two-person games*, Methods Oper. Res., 26 (1977), pp. 755–766.
- [57] ———, *Nash equilibria for noncooperative n -person games in normal form*, SIAM Rev., 23 (1981), pp. 225–237.
- [58] J. VILLE, *Note sur la théorie générale des jeux où intervient l'habileté des joueurs*, Traité du calcul des probabilités et ses applications, E. Borel et collaborateurs, Paris, 2 (1938), pp. 105–113.
- [59] N. N. VOROB'EV, *Game Theory*, Lectures for Economists and Systems Scientists, Springer, Berlin, 1977.
- [60] A. WALD, *Statistical Decision Functions*, John Wiley, New York, 1950.
- [61] ———, *Note on zero sum two person games*, Ann. Math., 52 (1950), pp. 739–742.
- [62] E. B. YANOWSKAYA, *Infinite zero-sum two-person games*, J. Soviet Math., 2 (1974), pp. 520–541.
- [63] ———, *Antagonistic games*, Problemy Kibernetiki, 34 (1978), pp. 221–246. (In Russian.)
- [64] N. YOUNG, *Admixtures of two-person games*, Proc. London Math. Soc., Ser. III, 25 (1972), pp. 736–750.

THE LINEAR REGULATOR PROBLEM FOR PARABOLIC SYSTEMS*

H. T. BANKS† AND K. KUNISCH‡

Abstract. We present an approximation framework for computation (in finite dimensional spaces) of Riccati operators that can be guaranteed to converge to the Riccati operator in feedback controls for abstract evolution systems in a Hilbert space. It is shown how these results may be used in the linear optimal regulator problem for a large class of parabolic systems.

Key words. parabolic systems, feedback controls, approximation schemes

1. Introduction. In this note we consider feedback controls for parabolic partial differential equations and the related Riccati operator theory when an infinite horizon integral quadratic cost functional is optimized. A general convergence framework for approximation ideas which can be used in computational techniques is developed in the context of the regulator problem theory pursued by Gibson in several recent investigations [8], [9], [10]. To illustrate our ideas we shall consider a specific model problem: The infinite horizon regulator problem for the parabolic control system

$$(1.1) \quad \frac{\partial y}{\partial t} = \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(a_{ij}(x) \frac{\partial y}{\partial x_j} \right) + \sum_{i=1}^n b_i(x) \frac{\partial y}{\partial x_i} + cy + Bu(t),$$

for $t > 0$, $x \in \Omega \subset R^n$, with Dirichlet boundary conditions $y|_{\partial\Omega} = 0$ and known initial data $y|_{t=0} = \phi$. Consideration of this model problem is motivated by our desire to develop efficient computational schemes for optimal control problems in connection with the insect dispersal investigations detailed in [1], [2]. These problems (involving parabolic partial differential equations) will entail distributed controls (e.g., spraying of pesticides over a region with frequency and intensity of spraying constituting important control variables). We expect the theoretical results presented in this paper to form a sound foundation for the development in the near future of computational procedures for feedback controls in such problems.

In § 2, we state carefully a convergence theory for approximate Riccati operators that is essentially a modification and refinement of the theory presented by Gibson in [8]. (In an appendix, we indicate details as to how our framework follows from the results of Gibson.) We then in § 3 state precisely our control problem for the system (1.1) and show that under reasonable assumptions (which imply a certain "preservation of exponential stability under approximation" condition) the abstract framework of § 2 can be used to guarantee convergence of approximate solutions in the event the basic approximation scheme enjoys rather fundamental convergence properties. These

* Received by the editors July 1, 1983, and in revised form January 2, 1984. This research was supported in part by NSF grant MCS-8205335, by AFOSR contract 81-0198, and ARO contract ARO-DAAG-29-79-C-0161. Parts of the research were carried out while both authors were visitors at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, VA, which is operated under NASA contracts NAS1-15810 and NAS1-16394.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912 and Department of Mathematics, Southern Methodist University, Dallas, Texas 75275.

‡ Institut für Mathematik, Technische Universität Graz, A-8010 Graz, Austria and Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, 02912. This author gratefully acknowledges support from the Max Kade Foundation. His research was supported in part by the Fonds zur Förderung der Wissenschaftlichen Forschung, Austria, no. P4534, and by the Steiermärkischen Wissenschafts- und Forschungsförderungsfonds.

are sufficiently relaxed to allow a generous number of practical schemes (modal, splines of several types and orders) to fall within our treatment.

A concluding section contains remarks on the potential usefulness of the results in this presentation.

2. Approximation of an abstract linear regulator problem. In this section we summarize approximation results for an abstract linear optimal regulator problem that we shall subsequently employ in our treatment of parabolic systems. The results given here involve an important practical modification of the abstract theory developed by Gibson in [8]. Specifically our presentation is formulated so as to facilitate approximation of the regulator problem by a sequence of finite dimensional state space problems, each defined on a subspace of the state space of the original problem. Gibson's presentation [8] requires the approximating problems each be defined on the entire original state space and as we shall explain below, this can lead to some tedious technical considerations. Our modified framework of this section really follows directly from that of Gibson, but we shall defer to an appendix a detailed explanation of this aspect of our considerations.

We suppose throughout that H and U are Hilbert spaces, that $A: \text{dom } A \subset H \rightarrow H$ is the infinitesimal generator of a strongly continuous or C_0 semigroup $T(t)$ on H and that $B \in \mathcal{L}(U, H)$. We consider a control system in H given by

$$(2.1) \quad \begin{aligned} \dot{y}(t) &= Ay(t) + Bu(t), \quad t > 0, \\ y(0) &= y_0, \end{aligned}$$

and an associated performance measure

$$(2.2) \quad J(y_0, u) = \int_0^\infty \{ \langle Dy(t), y(t) \rangle + \langle Qu(t), u(t) \rangle \} dt,$$

where $D \in \mathcal{L}(H)$, $Q \in \mathcal{L}(U)$ are selfadjoint and satisfy $D \geq 0$, $Q > 0$. Our fundamental abstract linear optimal regulator problem can then be stated as

(\mathcal{R}) Minimize $J(y_0, u)$ over $u \in L^2(0, \infty; U)$ subject to $y = y(\cdot; u)$ satisfying (2.1).

We shall say that a function $u \in L^2(0, \infty; U)$ is an *admissible control* for the initial state $y_0 \in H$ if $J(y_0, u)$ is finite. As usual, a certain algebraic Riccati equation will play a fundamental role in our analysis and an operator $\Pi \in \mathcal{L}(H)$ is called a *solution* of the *algebraic Riccati equation* (A.R.E) if Π maps $\text{dom } A$ into $\text{dom } A^*$ and satisfies on H the equation

$$(2.3) \quad A^*\Pi + \Pi A - \Pi BQ^{-1}B^*\Pi + D = 0.$$

Here A^* is the Hilbert space adjoint of A and we recall [4, p. 51] that it is the generator of the C_0 semigroup $T(t)^*$ which is adjoint to $T(t)$. We note that if Π satisfies (2.3) on $\text{dom } A$ then (2.3) can be taken as an equation on H since $\Pi BQ^{-1}B^*\Pi - D$ is bounded so that $A^*\Pi + \Pi A$ has a bounded extension to all of H .

The following result is taken from [8, Thms. 4.11, 4.6].

THEOREM 2.1. *Let A, B, Q, D be as given above. Then there exists a nonnegative selfadjoint solution Π of the algebraic Riccati equation (2.3) if and only if for each $y_0 \in H$, there exists an admissible control. If this latter holds, then the unique optimal control and corresponding trajectory for (\mathcal{R}) are given by*

$$(2.4) \quad \bar{u}(t) = -Q^{-1}B^*\Pi_\infty \bar{y}(t),$$

$$(2.5) \quad \bar{y}(t) = S(t)y_0,$$

where Π_∞ is the minimal nonnegative selfadjoint solution of the A.R.E. (2.3) and $S(t)$ is the C_0 semigroup generated by $A - BQ^{-1}B^*\Pi_\infty$. If $|y(t; u)| \rightarrow 0$ as $t \rightarrow \infty$ for any admissible control (this is guaranteed for example by the condition $D > 0$), then Π_∞ is the unique nonnegative selfadjoint solution of the A.R.E. If $D > 0$, then we also have that $S(t)$ is uniformly exponentially stable.

In this theorem the term minimal for a selfadjoint operator is in reference to the usual ordering of selfadjoint nonnegative operators on a Hilbert space. We note that the minimal solution Π_∞ of (2.3) can be obtained as the limit of a sequence of Riccati operators for associated finite interval regulator problems (see [8, Thms. 4.10, 4.11]) in a manner analogous to the usual procedure for finite dimensional state space regulator problems [13].

We next formulate a sequence of approximate regulator problems and present a convergence result for the corresponding Riccati operators. Let H^N , $N = 1, 2, \dots$, be a sequence of finite dimensional linear subspaces of H and $P^N: H \rightarrow H^N$ be the canonical orthogonal projections. Assume that $T^N(t)$ is a sequence of C_0 semigroups on H^N with infinitesimal generators $A^N \in \mathcal{L}(H^N)$. Given operators $B^N \in \mathcal{L}(U, H^N)$ and $D^N \in \mathcal{L}(H^N)$ we consider the family of regulator problems:

$$(\mathcal{R}^N) \text{ Minimize } J^N(y^N(0), u) \text{ over } u \in L^2(0, \infty; U),$$

where

$$(2.6) \quad \begin{aligned} \dot{y}^N(t) &= A^N y^N(t) + B^N u(t), \quad t > 0, \\ y^N(0) &= y_0^N \equiv P^N y_0, \end{aligned}$$

and

$$(2.7) \quad J^N(y^N(0), u) = \int_0^\infty \{ \langle D^N y^N(t), y^N(t) \rangle + \langle Qu(t), u(t) \rangle \} dt.$$

We note that since $B^N: U \rightarrow H^N$, the trajectories of (2.6) evolve in H^N and consequently (\mathcal{R}^N) is a linear regulator problem in the finite dimensional state space H^N so that finite dimensional control theory is applicable here. We shall need several assumptions in a convergence statement regarding solutions of (\mathcal{R}^N) and (\mathcal{R}) .

- (H1). For each $y_0^N \in H^N$ there exists an admissible control $u^N \in L^2(0, \infty; U)$ for (\mathcal{R}^N) and any admissible control for (2.6), (2.7) drives the state of (2.6) to zero asymptotically.
- (H2). (i) For each $z \in H$, we have $T^N(t)P^N z \rightarrow T(t)z$ with the convergence uniform in t on bounded subsets of $[0, \infty)$.
(ii) For each $z \in H$, we have $T^N(t)^*P^N z \rightarrow T(t)^*z$ with the convergence uniform in t on bounded subsets of $[0, \infty)$.
(iii) For each $v \in U$, $B^N v \rightarrow Bv$ and for each $z \in H$, $B^{N*}z \rightarrow B^*z$.
(iv) For each $z \in H$, $D^N P^N z \rightarrow Dz$.

We remark that (H2)(i) implies in particular (take $t=0$) that $P^N z \rightarrow z$ for each $z \in H$ and in this sense we have the subspaces H^N approximate H .

If assumption (H1) holds, then the optimal control \bar{u}^N for (\mathcal{R}^N) is given in feedback form by

$$(2.8) \quad \bar{u}^N(t) = -Q^{-1}B^{N*}\Pi^N \bar{y}^N(t)$$

where $\Pi^N \in \mathcal{L}(H^N)$ is the unique nonnegative selfadjoint solution of the algebraic Riccati equation on H^N

$$(2.9) \quad A^{N*}\Pi^N + \Pi^N A^N - \Pi^N B^N Q^{-1} B^{N*} \Pi^N + D^N = 0,$$

and \bar{y}^N is the corresponding solution of (2.6) with $u = \bar{u}^N$. Moreover

$$J^N(y_0^N, \bar{u}^N) = \langle \Pi^N y_0^N, y_0^N \rangle.$$

We also have the following fundamental convergence results.

THEOREM 2.2. *Suppose (H1), (H2) hold, $Q > 0$, $D \geq 0$ and $D^N \geq 0$ and let Π^N denote the unique nonnegative selfadjoint Riccati operators on H^N for the problems (\mathcal{R}^N) . Further assume that a unique nonnegative selfadjoint Riccati operator on H for the problem (\mathcal{R}) exists. Let $S(t)$ and $S^N(t)$ be the semigroups generated by $A - BQ^{-1}B^*\Pi$ and $A^N - B^NQ^{-1}B^{N*}\Pi^N$ on H and H^N , respectively and suppose $|S(t)z| \rightarrow 0$, $t \rightarrow \infty$, for all $z \in H$. If there are positive constants M_1, M_2 and ω independent of N and t such that*

$$(2.10) \quad |S^N(t)|_{H^N} \leq M_1 e^{-\omega t} \quad \text{for } t \geq 0, \quad N = 1, 2, \dots,$$

and

$$(2.11) \quad |\Pi^N|_{H^N} \leq M_2,$$

then

$$(2.12) \quad \Pi^N P^N z \rightarrow \Pi z \quad \text{for every } z \in H,$$

$$(2.13) \quad S^N(t) P^N z \rightarrow S(t) z \quad \text{for every } z \in H,$$

where the convergence is uniform in t on bounded subsets of $[0, \infty)$, and

$$(2.14) \quad |S(t)| \leq M_1 e^{-\omega t} \quad \text{for } t \geq 0.$$

We present a proof of Theorem 2.2 in the Appendix. Meanwhile we remark that under the hypotheses of this theorem, $\Pi^N P^N$ is an extension of $\Pi^N \in \mathcal{L}(H^N)$ to all of H . If D^N, A^N are replaced by $D^N P^N, A^N P^N$, respectively and (2.9) is considered as an equation on H , then $\Pi^N P^N$ is its unique minimal nonnegative selfadjoint solution.

Theorem 2.2 is essentially contained in [8]. The main difference between the theorem here and the result in [8] is, as stated earlier, that here each of the finite dimensional state problems (\mathcal{R}^N) is defined in the subspace H^N only, whereas in [8], Gibson requires that the approximate regulator problems be defined on the entire space H . This causes some unnecessary technical difficulties: First note that if $D > 0$ and $D^N = P^N D$ (as an operator in H^N), then $D^N > 0$ on H^N . But $D^N = 0$ on $H^{N\perp}$. This difficulty can be circumvented by considering instead $\hat{D}^N = P^N D + I - P^N$ as an operator on H —see [9, p. 698].

To explain a second disadvantage to the formulation of (\mathcal{R}^N) on all of H , let us assume that $|T^N(t)|_{H^N} \leq M e^{-\alpha t}$ for positive constants M and α . This allows one to infer existence of Riccati operators Π^N on H^N ; however, if the semigroups $T^N(t)$ are extended to H by taking $T^N(t)P^N + I - P^N$, then these extensions are not uniformly exponentially stable and the existence of feedback solutions to (\mathcal{R}^N) on H is not guaranteed. However there is a more subtle difficulty regarding verification of the analogue of (2.10) on H (e.g., see [8, Thm. 4.3, condition (5.17)] if the approximate problems are defined on H . Even if one has the condition $|T^N(t)|_{H^N} \leq M e^{-\alpha t}$ with $T^N(t)$ extended to H as mentioned above, the feedback operators $S^N(t)$ on H satisfy $S^N(t)z = z$ for $z \in H^{N\perp}$ and hence it is not possible to satisfy directly the stability requirement (2.10) on H . In [9] this difficulty is handled by taking $T^N(t)z = e^{-t}z$ for $z \in H^{N\perp}$. But then $T(t)$ and $T^N(t)$ are essentially unrelated on $H^{N\perp}$.

Remark 2.1. We point out that the requirement $H^N \subset H$ is not essential to the development of an approximation framework and theory such as that culminating in Theorem 2.2. Of course in this case the hypotheses (H2) must be modified. More

precisely, suppose that $(H, |\cdot|)$, $(H^N, |\cdot|_N)$ are Hilbert spaces (in general $H^N \not\subset H$) with $T(t)$, $T^N(t)$ strongly continuous semigroups on H , H^N , respectively, and assume that the following hypotheses hold:

- (H2'). (0) There exist bounded linear operators $P^N: H \rightarrow H^N$ satisfying $|P^N z|_N \rightarrow |z|$ as $N \rightarrow \infty$ for all $z \in H$.
- (i) There exist constants M, ω such that $|T^N(t)|_N \leq M e^{\omega t}$ for all N and for each $z \in H$, $|T^N(t)P^N z - P^N T(t)z|_N \rightarrow 0$ as $N \rightarrow \infty$, uniformly in t on bounded subsets of $[0, \infty)$.
- (ii) For each $z \in H$, $|T^N(t)^* P^N z - P^N T(t)^* z|_N \rightarrow 0$ as $N \rightarrow \infty$, uniformly in t on bounded subsets of $[0, \infty)$.
- (iii) For each $v \in U$, the operators $B \in \mathcal{L}(U, H)$, $B^N \in \mathcal{L}(U, H^N)$ satisfy $|B^N v - P^N B v|_N \rightarrow 0$. For each $z \in H$ we have $|B^{N*} P^N z - B^* z|_U \rightarrow 0$.
- (iv) There exist operators $D^N \in \mathcal{L}(H^N)$ with $|D^N|_N$, $N = 1, 2, \dots$, bounded and for each $z \in H$, $|D^N P^N z - P^N D z|_N \rightarrow 0$.
- (v) There exist operators $G^N \in \mathcal{L}(H^N)$ with $|G^N|_N$ bounded and for each $z \in H$, $|G^N P^N z - P^N G z|_N \rightarrow 0$.
- (vi) For each N , the operators D^N , G^N are nonnegative selfadjoint.

Under these assumptions, Theorem 2.2 is valid where the convergences in (2.12), (2.13) are replaced by $|\Pi^N P^N z - P^N \Pi z|_N \rightarrow 0$ for every $z \in H$ and $|S^N(t)P^N z - P^N S(t)z|_N \rightarrow 0$ for $z \in H$, uniformly in t on bounded subintervals of $[0, \infty)$. This version of the approximation results follows from modifications of arguments in the Appendix once one has proven an analogue to Theorem A.2 which permits $H^N \not\subset H$. Such a version is useful in developing certain types of approximation schemes—e.g., finite differences, spectral methods.

In the next two sections we shall see that the version of approximation results given in our Theorem 2.2 lends itself to easy verification for certain classes of approximation schemes for parabolic systems.

3. Convergence of approximate Riccati operators for parabolic systems. We use the framework summarized in the previous section to treat the optimization of integral quadratic cost functionals for parabolic systems of the form given in (1.1). We shall follow the notation introduced in § 2 and for our state space H we choose $H^0(\Omega)$ with Ω a bounded domain in R^n possessing a piecewise C^1 boundary $\partial\Omega$. Unless otherwise indicated, all of the function spaces below are to be understood as spaces of functions with domain Ω and range R^1 .

For B , D , and Q as given in defining problem (\mathcal{R}) of § 2, with $D > 0$ and $Q > 0$, we consider the regulator problem (\mathcal{R}) with the system (2.1) for the state $y(t) = y(t, \cdot)$ in $H = H^0(\Omega)$ taken as the parabolic system

$$(3.1) \quad \begin{aligned} y_t &= \sum_{i,j=1}^n D^i (a_{ij} D^j y) + \sum_{i=1}^n b_i D^i y + cy + Bu, & t > 0, \\ y(0, \cdot) &= \phi, & y(t, \cdot)|_{\partial\Omega} = 0, \end{aligned}$$

where $u \in L^2(0, \infty; U)$ and $D^i = \partial/\partial x_i$ denotes differentiation with respect to the i th spatial variable x_i .

We make the following standing assumptions on the coefficients in (3.1):

There exist positive constants γ and μ such that

$$\gamma \sum_i \xi_i^2 \leq \sum_{i,j} a_{ij} \xi_i \xi_j \leq \mu \sum_i \xi_i^2$$

for every $\xi \in R^n$; $a_{ij} = a_{ji}$, and $a_{ij} \in L^\infty(\Omega)$, $b_i \in L^\infty(\Omega)$, $c \in L^\infty(\Omega)$ for every $i, j = 1, \dots, n$.

Throughout our discussions the concept of a solution of (3.1) will be that of a weak solution (i.e., in the sense of distributional derivatives).

We introduce the sesquilinear form $\sigma: H_0^1(\Omega) \times H_0^1(\Omega) \rightarrow \mathbb{C}$ defined by

$$\sigma(z, v) = \int_{\Omega} \left\{ \sum_{i,j=1}^n a_{ij} D^j z D^i \bar{v} - \left(\sum_{i=1}^n b_i D^i z + \tilde{c}z \right) \bar{v} \right\} dx,$$

where

$$\tilde{c}(x) = c(x) - k,$$

with the constant $k = k(\Omega, \gamma) > 0$ determined so that the inequality

$$(3.2) \quad \operatorname{Re} \sigma(z, z) \geq c_1 |z|_1^2, \quad z \in H_0^1(\Omega)$$

holds for some positive constant c_1 independent of z (see [14, p. 144]). Here and throughout $|\cdot|_1$ and $|\cdot|$ denote the $H^1(\Omega)$ and $H^0(\Omega)$ norms respectively. Furthermore, to allow use of the theory of sectorial operators and sesquilinear forms in discussing the spectra of various operators, we assume in defining σ that the functions in $H_0^1(\Omega)$ are complex valued. For the sesquilinear form σ it can be shown (see [14, p. 143]) that there is also a constant $c_2 = c_2(\Omega, \mu)$ so that

$$(3.3) \quad |\sigma(z, v)| \leq c_2 |z|_1 |v|_1$$

for all $z, v \in H_0^1(\Omega)$. Furthermore, it follows from the bounds (3.2), (3.3) and well-known results on sesquilinear forms (see, e.g., [12, p. 101]) that there exist operators A_k, A_k^* in $H^0(\Omega)$ such that

$$(3.4) \quad \sigma(z, v) = \langle -A_k z, v \rangle \quad \text{for } z \in \operatorname{dom} A_k, v \in H_0^1(\Omega),$$

and

$$(3.5) \quad \overline{\sigma(z, v)} = \langle -A_k^* v, z \rangle \quad \text{for } v \in \operatorname{dom} A_k^*, z \in H_0^1(\Omega).$$

In addition $\operatorname{dom} A_k$ and $\operatorname{dom} A_k^*$ are dense in $H_0^1(\Omega)$, and we have $A_k \operatorname{dom} A_k = H^0(\Omega)$, $A_k^* \operatorname{dom} A_k^* = H^0(\Omega)$.

From (3.2), (3.4) and (3.5), we find that

$$(3.6) \quad \begin{aligned} \operatorname{Re} \langle A_k z, z \rangle &\leq -c_1 |z|_1^2, & z \in \operatorname{dom} A_k, \\ \operatorname{Re} \langle A_k^* z, z \rangle &\leq -c_1 |z|_1^2, & z \in \operatorname{dom} A_k^*. \end{aligned}$$

In view of (3.6) and the range statements above for A_k and A_k^* , we may invoke standard results from linear semigroup theory [15, p. 16, Thm. 4.5] to assert that A_k and A_k^* are the infinitesimal generators of linear C_0 -semigroups $T_k(t)$ and $T_k^*(t)$, respectively. (As we have noted above, in fact $T_k^*(t) = T_k(t)^*$.)

We note that the solution semigroup $T(t)$ for (3.1) is given by

$$(3.7) \quad T(t) = e^{kt} T_k(t)$$

with the infinitesimal generator A of $T(t)$ given by $A = A_k + kI$ and $\operatorname{dom} A = \operatorname{dom} A_k$. Similarly, we have

$$(3.8) \quad T(t)^* = e^{kt} T_k^*(t)$$

with the infinitesimal generator A^* of $T(t)^*$ given by $A^* = A_k^* + kI$ and $\operatorname{dom} A^* = \operatorname{dom} A_k^*$. We also have $|T(t)| \leq e^{(k-c_1)t}$ for $t \geq 0$.

Turning next to approximations for (\mathcal{R}) , we suppose we have a sequence of finite dimensional (real) subspaces $H^N \subset H_0^1(\Omega)$, $N = 1, 2, \dots$, which satisfy the

approximation condition:

(C1). For each $z \in H_0^1(\Omega)$, there exists an element \tilde{z}^N in H^N such that $|z - \tilde{z}^N|_1 \leq \varepsilon(N)$, where $\varepsilon(N) \rightarrow 0$ as $N \rightarrow \infty$.

We remark that this condition is fulfilled in the event H^N is chosen as in many classes of finite element approximation schemes [5, Chap. III, 3.2]. In particular, (C1) holds for the case where Ω is a rectangle in R^2 and the H^N are the usual linear spans of tensor products of standard one dimensional piecewise linear splines [16] with mesh size approaching zero as $N \rightarrow \infty$.

Proceeding in standard fashion, we observe that the restriction of σ to $H^N \times H^N$ defines, in a unique manner, bounded linear operators A_k^N, A_k^{*N} on H^N such that

$$(3.9) \quad \alpha(z, v) = \langle -A_k^N z, v \rangle, \quad z, v \in H^N,$$

and

$$(3.10) \quad \sigma(z, v) = \langle -A_k^{*N} v, z \rangle, \quad z, v \in H^N.$$

Here $A_k^{*N} = A_k^{N*}$. We let $A^N = A_k^N + kI$, $A^{N*} = A_k^{N*} + kI$ with domains H^N and note that A^N, A^{N*} generate C_0 -semigroups $T^N(t)$, $T^N(t)^*$ on H^N , with $T^N(t)^*$ the adjoint of $T^N(t)$. For the finite dimensional approximating problems (\mathcal{R}^N) we choose

$$(3.11) \quad B^N = P^N B, \quad D^N = P^N D$$

where $P^N: H^0(\Omega) \rightarrow H^N$ is, as in § 2, the canonical orthogonal projection. We have thus specified all of the needed entities for the problem (\mathcal{R}^N) of § 2. As we noted previously the trajectories of this problem evolve in H^N and hence it is a finite dimensional regulator problem for which computational techniques are readily available (assuming of course that one has made an appropriate decision in defining the H^N).

We turn to a verification of (H2) of § 2 for the approximations at hand. Since it is a trivial matter to see that (C1) implies that

$$(3.12) \quad P^N z \rightarrow z \quad \text{as } N \rightarrow \infty, \quad \text{for } z \in H^0(\Omega),$$

the conditions (H2) (iii), (H2) (iv) follow at once from (3.11). We next argue that

$$(3.13) \quad T_k^N(t) P^N z \rightarrow T_k(t) z,$$

$$(3.14) \quad T_k^N(t)^* P^N z \rightarrow T_k(t)^* z,$$

for $z \in H^0(\Omega)$, with the convergence uniform in t on bounded subsets of $[0, \infty)$. This taken with (3.7), (3.8) will imply conditions (H2)-(i), (H2)-(ii). First we note that from (3.2), (3.9), (3.10) and the fact that $H^N \subset H_0^1(\Omega)$, we have

$$(3.15) \quad \operatorname{Re} \langle A_k^N z, z \rangle \leq -c_1 |z|_1^2,$$

$$(3.16) \quad \operatorname{Re} \langle A_k^{N*} z, z \rangle \leq -c_1 |z|_1^2$$

for all $z \in H^N$. Moreover, we shall demonstrate that the following convergence statements hold:

$$(3.17) \quad (I - A_k^N)^{-1} P^N z \rightarrow (I - A_k)^{-1} z, \quad z \in H^0(\Omega),$$

$$(3.18) \quad (I - A_k^{N*})^{-1} P^N z \rightarrow (I - A_k^*)^{-1} z, \quad z \in H^0(\Omega).$$

We then may use the Trotter-Kato theorem (see, e.g., the version given in [18]) to conclude that (3.15)–(3.18) imply at once the statements (3.13), (3.14).

Thus we turn to establish (3.17) and (3.18). We shall employ a result given in [7, p. 756, Lemma 3.3], which we state without proof here.

LEMMA 3.1. *There exist a positive constant δ_1 and a constant θ_1 in $(0, \pi/2)$ such that*

$$(3.19) \quad |\lambda||z|^2 + |z|_1^2 \leq \delta_1 |\lambda| |z|^2 - \sigma(z, z)$$

for all $z \in H_0^1(\Omega)$ and $\lambda \in \{\xi \in \mathbb{C} : \theta_1 \leq |\arg \xi| \leq \pi\}$.

We use this to show that (3.17) holds. For $z \in H^0(\Omega)$, define $w = (I - A_k)^{-1}z$ and $w^N = (I - A_k^N)^{-1}P^N z$. Then we have for all $z^N \in H^N$

$$\begin{aligned} \langle w, z^N \rangle + \sigma(w, z^N) &= \langle z, z^N \rangle, \\ \langle w^N, z^N \rangle + \sigma(w^N, z^N) &= \langle z, z^N \rangle. \end{aligned}$$

Consequently, defining $e^N = w - w^N$, we find

$$\langle e^N, z^N \rangle + \sigma(e^N, z^N) = 0$$

for all $z^N \in H^N$. Taking $\lambda = -1$ and $z = e^N$ in (3.19)—note that $e^N \in H_0^1(\Omega)$ —we obtain using this last equation

$$|e^N|^2 + |e^N|_1^2 \leq \delta_1 |e^N|^2 - \sigma(e^N, e^N) = \delta_1 |e^N|^2 - \langle e^N, e^N + z^N \rangle - \sigma(e^N, e^N + z^N)$$

for all $z^N \in H^N$. Let $z^N = w^N - \tilde{w}^N$, where \tilde{w}^N is an approximation for w chosen according to (C1). (Here we again use the fact that $w \in \text{dom } A_k \subset H_0^1(\Omega)$.) We thus obtain the estimate

$$|e^N|^2 + |e^N|_1^2 \leq \delta_1 |\langle e^N, w - \tilde{w}^N \rangle + \sigma(e^N, w - \tilde{w}^N)| \leq c_2 \varepsilon(N) \delta_1 \{|e^N| + |e^N|_1\},$$

where we have, without loss of generality, assumed that $c_2 \geq 1$. This last estimate implies $e^N \rightarrow 0$ in $H^1(\Omega)$ and, in particular, (3.17) holds.

Turning to (3.18), we recall that $\overline{\sigma}(z, v) = \langle -A_k^* v, z \rangle$ and define $\tau: H_0^1(\Omega) \times H_0^1(\Omega) \rightarrow \mathbb{C}$ by $\tau(z, v) = \overline{\sigma}(z, v)$. Then τ satisfies the same inequalities (3.2), (3.3) as σ . We may therefore verify (3.18) by referring to the analysis for (3.17). We summarize our discussions to this point.

LEMMA 3.2. *Let (C1) hold. Then (H2) obtains with $B^N, D^N, T^N(t)$, and $T^N(t)^*$ defined as in and just above (3.11).*

To use Theorems 2.1, 2.2 of § 2, we shall make the following stabilizability hypothesis.

(C2). The pair (A, B) is exponentially stabilizable, i.e., there exists a bounded linear operator $K: H^0(\Omega) \rightarrow U$ such that the semigroup $T_s(t)$ generated by $A + BK$ satisfies $|T_s(t)| \leq M_1 e^{-\omega_1 t}$ for some positive constants M_1 and ω_1 .

For a discussion of (C2), we refer the reader to [17] and the references given there.

To make use of the theory of § 2, we need to verify that a certain *preservation of exponential stabilizability under approximation* condition holds for our problem. This condition can be stated as:

(POES). Suppose that condition (C2) holds. Then there exists an integer N_0 such that for all $N \geq N_0$ the pairs $(A^N, P^N B)$ are uniformly exponentially stabilizable by the operator K of (C2), i.e., there exist positive constants (independent of N) M_s and ω_s such that the semigroups $T_s^N(t)$ generated by $A^N + P^N B K$ satisfy $|T_s^N(t)| \leq M_s e^{-\omega_s t}$ for all $N \geq N_0$ and $t \geq 0$.

Before returning to the theoretical results of § 2, we argue that the class of approximations for our system (3.1) does indeed satisfy the preservation of stabilizability condition (POES).

LEMMA 3.3. *Let (C1), (C2) hold. Then the approximations defined through (3.9), (3.10), (3.11) yield systems that satisfy (POES).*

Proof. We define a sesquilinear form $\sigma_B: H_0^1(\Omega) \times H_0^1(\Omega) \rightarrow \mathbb{C}$ by $\sigma_B(z, v) = \sigma(z, v) + \langle -BKz, v \rangle + \langle \tilde{k}z, v \rangle$, where \tilde{k} is chosen so that

$$(3.20) \quad \operatorname{Re} \sigma_B(z, z) \geq c_3 |z|_1^2, \quad z \in H_0^1(\Omega)$$

and

$$(3.21) \quad |\sigma_B(z, v)| \leq c_4 |z|_1 |v|_1, \quad z, v \in H_0^1(\Omega),$$

for some positive constants c_3, c_4 (recall (3.2), (3.3)). Then arguments similar to those in [7, p. 756, Lemmas 3.2, 3.3] can be used to establish that the numerical range of σ_B is contained in a right sector $S_{\theta, \gamma} = \{\lambda \in \mathbb{C}: |\arg(\lambda - \gamma)| \leq \theta\}$ where $0 < \theta < \pi/2$, γ real.

We next consider the restriction of σ_B to $H^N \times H^N$ and, in a manner already discussed, this gives rise to bounded linear operators $\hat{A}_B^N, \hat{A}_B^{N*}$ on H^N such that $\hat{A}_B^N = A^N + P^N BK - \hat{k}$, where $\hat{k} = k + \tilde{k}$ (see the definition of k in σ , \tilde{k} in σ_B). Indeed $\sigma_B(z, v) = \langle -\hat{A}_B^N z, v \rangle$, $\overline{\sigma_B(z, v)} = \langle -\hat{A}_B^{N*} v, z \rangle$ for $z, v \in H^N$, with \hat{A}_B^{N*} the adjoint of \hat{A}_B^N . Furthermore, the numerical range of \hat{A}_B^N (and \hat{A}_B^{N*}) is contained in the left sector $S_{\theta, \gamma}^- = \{\lambda: -\lambda \in S_{\theta, \gamma}\}$, uniformly in N . Thus the numerical range and hence the spectrum (see [11, p. 280]) of $A_B^N = A^N + P^N BK = \hat{A}_B^N + \hat{k}$ are contained in a left sector $\hat{S} = S_{\theta, \gamma}^- + \hat{k}$, uniformly in N . It follows that the set of all eigenvalues λ of A_B^N with $\operatorname{Re} \lambda > -\delta$ is bounded, uniformly in N , for any fixed δ .

Using arguments similar to those behind (3.17), (3.18)—see Lemma 3.1 and the proof of Lemma 3.2—it is easily shown that for some $\zeta \in \mathbb{R}^1, \zeta > 0$ sufficiently large, we have ζ in all the resolvent sets $\rho(A_B^N), \rho(A + BK), N = 1, 2, \dots$, and

$$(3.22) \quad (\zeta - A_B^{N*})^{-1} P^N z \rightarrow (\zeta - (A + BK)^*)^{-1} z \quad \text{for } z \in H^0(\Omega).$$

Next we claim that for some positive integer N_0 , $\bigcup_{N=N_0}^\infty \sigma(A_B^N)$ is contained in a left half plane $\operatorname{Re}(z) \leq -\varepsilon, \varepsilon > 0$, and hence in the interior of a left sector \hat{S} with vertex $-\varepsilon/2$. If not then there exists a sequence N_j with $N_j \rightarrow \infty$ and λ^{N_j} an eigenvalue of $A_B^{N_j}$ satisfying $\operatorname{Re} \lambda^{N_j} > -1/j$. From our findings on the spectrum of $A_B^N, N = 1, 2, \dots$, we know there exists a limit point $\hat{\lambda}$ of $\{\lambda^{N_j}\}$ with $\operatorname{Re} \hat{\lambda} \geq 0, \hat{\lambda} \in \hat{S}$. We shall argue that $\hat{\lambda}$ is an eigenvalue of $A + BK$, which is a contradiction since (C2) implies $\operatorname{Re} \lambda \leq -\omega_1$ for λ in the spectrum of $A + BK$ (see [4, p. 32]).

For convenience, we relabel and drop the subsequential notation, assuming henceforth that $\lambda^N \rightarrow \hat{\lambda}$, with λ^N an eigenvalue of A_B^N . Let ϕ^N be an eigenvector with $|\phi^N| = 1$ and $A_B^N \phi^N = \lambda^N \phi^N$ for all N sufficiently large. Then we have

$$\lambda^N \phi^N - \zeta \phi^N = A_B^N \phi^N - \zeta \phi^N$$

and hence

$$(A_B^N - \zeta)^{-1} (\lambda^N \phi^N - \zeta \phi^N) = \phi^N.$$

Let χ in $\operatorname{dom} A^*$ be arbitrary and put $\psi = (A + BK - \zeta)^* \chi$. Then

$$(3.23) \quad \langle \lambda^N \phi^N - \zeta \phi^N, ((A_B^N - \zeta)^{-1})^* P^N \psi \rangle = \langle \phi^N, P^N \psi \rangle.$$

Using (3.2), one can readily show that the set $\{\phi^N\}$ is a bounded set in $H_0^1(\Omega)$. Consequently, there exists a subsequence, again denoted by $\{\phi^N\}$, converging strongly in $H^0(\Omega)$ to some nontrivial $w \in H^0(\Omega)$. Thus from (3.22), (3.23) and (3.12)—a result of (C1)—we have

$$\langle (\hat{\lambda} - \zeta) w, ((A + BK - \zeta)^{-1})^* \psi \rangle = \langle w, \psi \rangle$$

or

$$\langle (\hat{\lambda} - \zeta) w, \chi \rangle = \langle w, (A + BK - \zeta)^* \chi \rangle$$

for all χ in $\text{dom } A^*$. Thus $w \in \text{dom } A$ and

$$\langle (\hat{\lambda} - \zeta) w, \chi \rangle = \langle (A + BK - \zeta) w, \chi \rangle$$

or

$$\langle (A + BK - \hat{\lambda}) w, \chi \rangle = 0$$

for all χ in $\text{dom } A^*$. It follows that w is an eigenvector corresponding to the eigenvalue $\hat{\lambda}$ for $A + BK$ and this yields the desired contradiction.

Having established existence of the sector \tilde{S} , the uniform exponential bound in (POES) now follows from the representation

$$T_s^N(t) = \int_{\Gamma} e^{\lambda t} (\lambda + A_B^N)^{-1} d\lambda$$

where Γ is the positively oriented boundary of \tilde{S} . To see this, note that an estimate like (3.19) for σ_B yields $|(\lambda + A_B^N)^{-1}| \leq \hat{\delta}_1 |\lambda|^{-1}$ and that furthermore we have $|T_s^N(t)| \leq e^{\hat{k}t}$. Thus an estimate of the contour integral for $t \in [1, \infty)$ along with the observation that $|T_s^N(t)| \leq e^{\hat{k}}$ for $t \in [0, 1]$ yields the bound and completes the proof of Lemma 3.3.

We return finally to a discussion of the convergence theory of § 2 as it is applied to the specific parabolic system control problem that is the focus of the present section. We assume (C1) and (C2) hold. Then (POES) along with Theorem 2.1 yields the existence of nonnegative selfadjoint Riccati operators Π and Π^N (for N sufficiently large) associated with the problems (\mathcal{R}) and (\mathcal{R}^N) in $H^0(\Omega)$ and H^N , respectively. Since $D > 0$ and $D^N = P^N D > 0$ on H^N , these Riccati operators are unique and furthermore (H1) obtains.

Turning to Theorem 2.2, we first verify that (2.11) holds.

Recall that

$$(3.24) \quad |\Pi^N|_{H^N} = \sup \{ \langle \Pi^N z, z \rangle \mid z \in H^N, |z| = 1 \} = \sup_{|z|=1} J(z, \bar{u}^N),$$

where \bar{u}^N is the optimal feedback control (2.8) of (\mathcal{R}^N) . Define, for $z^N \in H^N$ with $|z^N| = 1$, the control $u_s^N(t) = K T_s^N(t) z^N$ where $T_s^N(t)$ is the semigroup defined in (POES). Then

$$\begin{aligned} J(z^N, \bar{u}^N) &\leq \int_0^\infty \langle D^N T_s^N(t) z^N, T_s^N(t) z^N \rangle dt + \int_0^\infty \langle Q K T_s^N(t) z^N, K T_s^N(t) z^N \rangle dt \\ &\leq \{|D| + |Q| |K|^2\} M_s^2 \omega_s^{-2} = M_2, \end{aligned}$$

so that from (3.24) we may infer (2.11). To establish (2.10), we first note that $|S^N(t)|_{H^N} \leq K_1 e^{\beta t}$ for some constants K_1 and β independent of N . This follows from (3.15), (3.16), (2.11) and the fact that $S^N(t)$ is generated by $A^N - P^N B Q^{-1} B^* P^N \Pi^N$. Moreover we have

$$\int_0^\infty \langle D S^N(t) z^N, S^N(t) z^N \rangle dt \leq \langle \Pi^N z^N, z^N \rangle \leq M_2 |z^N|^2.$$

Since $D > 0$, a theorem of Datko (see [6], [8, p. 540, Thm. 2.2]) implies existence of

positive constants M_1 and ω , independent of N , such that

$$|S^N(t)|_{H^N} \leq M_1 e^{-\omega t}.$$

Hence (2.10) of Theorem 2.2 holds.

Using the convergence results of (2.12), (2.13) it is easy to argue that the optimal feedback controls for (\mathcal{R}^N) converge to that of (\mathcal{R}) . We summarize our findings for the regulator problems for (3.1) in the following theorem.

THEOREM 3.1. *Assume that the subspace approximation condition (C1) holds for $H^N \subset H_0^1(\Omega)$, that the stabilizability condition (C2) holds for (3.1), and that $Q > 0$, $D > 0$. Then there exist unique Riccati operators Π and Π^N associated with the regulator problems (\mathcal{R}) and (\mathcal{R}^N) on $H^0(\Omega)$ and H^N for (3.1) and*

$$\begin{aligned}\Pi^N P^N z &\rightarrow \Pi z \quad \text{for } z \in H^0(\Omega), \\ S^N(t) P^N z &\rightarrow S(t) z \quad \text{for } z \in H^0(\Omega),\end{aligned}$$

and

$$\bar{u}^N(t) \rightarrow \bar{u}(t),$$

with these last two statements holding uniformly in t on compact subsets of $[0, \infty)$. Here $S^N(t)$ and $S(t)$ are the semigroups generated by $A^N - P^N B Q^{-1} B^* P^N \Pi^N$ and $A - B Q^{-1} B^* \Pi$, and \bar{u}^N and \bar{u} are the optimal feedback controls for (\mathcal{R}^N) and (\mathcal{R}) , respectively. Moreover, $|S(t)| \leq M_1 e^{-\omega t}$ with $\omega > 0$.

Remark 3.1. We note that the techniques of this section lead easily and directly to implementable approximation schemes. Given the approximation subspaces H^N , one does not need to construct the operators A_k^N of (3.9) to consider the resulting approximating systems. Indeed the usual Galerkin type formulation for the approximate system leads to only the need to evaluate $\sigma(B_i^N, B_j^N)$, where $\{B_j^N\}$ is a set of basis elements for H^N .

4. Concluding remarks. The conclusions in Theorem 3.1, especially (3.26) and (3.27), are important since they reveal that the *finite dimensional control laws* when employed in the systems that we can compute (i.e., the *approximate systems*) allow us to anticipate what might happen qualitatively if we used the *infinite dimensional feedback controls* in the *original distributed system*. However of equal importance are findings (which are simple corollaries to the results of section 3) that indicate that use of the *approximate* (readily computed and usually easily implemented) *controls* in the *actual distributed system* can be expected to produce satisfactory performance. More precisely, let $U^N = Q^{-1} B^* \Pi^N P^N$ and consider the sequence $\bar{A}^N = A - B U^N$ of operators in $H^0(\Omega)$. Then the operators \bar{A}^N generate semigroups $\bar{S}^N(t)$ which are uniformly exponentially stable and $\bar{S}^N(t) z \rightarrow S(t) z$, uniformly on compact sets in $[0, \infty)$, $z \in H^0(\Omega)$, provided, of course, that the assumptions of Theorem 3.1 are fulfilled. The uniform exponential stability can be established using arguments similar to those in the proof of Lemma 3.3. The significance of results for such finite dimensional feedback into the original distributed system was noted by Gibson in [9, p. 699].

We note that although the techniques employed in § 3 pertain in a fundamental way to stabilizable equations of parabolic type, the techniques described in that section are not restricted to equations of the form (1.1) with distributed control. Indeed as can be seen from the arguments in § 3, the essential property required for application of these ideas is that the differential equation operator in (2.1) be sectorial or, more precisely, that the systems (including feedback) generate sesquilinear forms with numerical range in some sector (e.g., see the arguments behind Lemmas 3.1, 3.3).

Indeed, even though our treatment here is concerned with the practically important (in view of the applications mentioned in § 1) case of distributed controls, we recognize that there are important applications where boundary control problems for parabolic equations are of primary interest. In some of these applications our treatment and techniques are readily used. (We note that the only restriction on B is that it be bounded and some boundary control problems are readily transformed to the form (2.1)). Furthermore, certain control problems for higher order equations can also successfully be treated with the ideas presented in this paper.

Finally, we note that our approximation approach involves almost no restrictions on the subspaces H^N so that we again can treat a large variety of problems. For example, we specifically do not require that H^N be contained in $\text{dom } A$ (or $\text{dom } A_k$) about which we may have only partial information in some cases. Thus we may readily employ linear spline approximations with second order operators in the framework of our results. Based on our previous efforts with spline based approximations in parameter estimation [1], [2] and control problems [3], we were confident that optimism concerning use of splines in the present framework was justified. In the period since we wrote our first report [Banks and Kunisch, *The linear regulator problem for parabolic systems*, LCDS Rep. 83-18, Brown Univ., May, 1983] on this work, we have, in collaboration with K. Ito, successfully carried out test computations using spline and spectral (tau-Legendre) methods with our expectations fully realized. Details of this and other numerical work presently being pursued will be reported elsewhere. We gratefully acknowledge K. Ito for his efforts on these numerical calculations and for his comments and questions concerning the present manuscript.

Appendix. We give here a proof for Theorem 2.2 using in a fundamental way some of the results of Gibson. As we have already mentioned, Theorem 2.2 in its present form is not given in [8].

To make our arguments, we need to consider regulator problems on the finite intervals $[s, t_f]$, $-\infty < s < t_f$, with a weighting operator G for the final state $y(t_f)$. We assume throughout that A generates the C_0 semigroup $T(t)$ on H , that D , Q , G are selfadjoint with $D \geq 0$, $Q > 0$, $G \geq 0$, and $B \in \mathcal{L}(U, H)$. The finite interval problems are given by:

(\mathcal{R}, t_f) Minimize

$$J(s, y(s), u) = \langle Gy(t_f), y(t_f) \rangle + \int_s^{t_f} \{ \langle Dy(t), y(t) \rangle + \langle Qu(t), u(t) \rangle \} dt$$

subject to

$$y(t) = T(t-s)y(s) + \int_s^t T(t-\sigma)Bu(\sigma) d\sigma \quad \text{for } s \leq t \leq t_f.$$

Under our assumptions a unique nonnegative selfadjoint Riccati operator Π_s can be associated with (\mathcal{R}, t_f) . That is, Π_s is the unique nonnegative selfadjoint solution of the integral Riccati equation for $z \in H$

$$\Pi_s(t)z = T(t_f-t)^*GT(t_f-t)z + \int_t^{t_f} T(\eta-t)^*[D - \Pi_s(\eta)BQ^{-1}B^*\Pi_s(\eta)]T(\eta-t)z d\eta$$

with $\Pi_s(\xi) \in \mathcal{L}(H)$ for $s \leq \xi \leq t_f$, (see [8, Thms., 3.1, 3.2 and eq. (3.28)]). We then have the following limit results.

THEOREM A.1. Assume that the unique nonnegative selfadjoint solution Π of the A.R.E. (2.3) exists. Let Π_s be the unique Riccati operator function associated with the

problem $(\mathcal{R}, 0)$. If $\lim_{t \rightarrow \infty} |S(t)z| = 0$ for all $z \in H$, where $S(t)$ is generated by $A - BQ^{-1}B^*\Pi$, then

$$(A.1) \quad \lim_{s \rightarrow -\infty} \Pi_s(s)z = \Pi z \quad \text{for all } z \in H.$$

If moreover $G \geq \Pi$ and there exist positive constants M and β such that

$$(A.2) \quad |S(t)| \leq M e^{-\beta t}, \quad t \geq 0,$$

then

$$(A.3) \quad \Pi \leq \Pi_s(s) \leq \Pi + M^2 e^{2\beta s} |G| \quad \text{for } s < 0.$$

Proof. If Π is the unique nonnegative, selfadjoint solution of (2.3), then by the calculations in [8, pp. 557–558], it is also the unique solution of the first integral Riccati equation of [8] on the infinite interval and corresponds to the operator P_∞ of that paper. Theorem A.1 then follows directly from [8, Thm. 4.10].

We note that if in addition to the above hypotheses we have $D > 0$, then (A.2) is satisfied (see [8, Thm. 4.8]).

We next recall an approximation result for the finite horizon regulator problem (\mathcal{R}, t_f) in H . Let (H2) hold with operators as in (\mathcal{R}^N) given; in addition assume $G^N \in \mathcal{L}(H^N)$, $G^N \geq 0$, are given. To consider a related finite horizon problem in H , we define $\tilde{G}^N = G^N P^N$ and $\tilde{D}^N = D^N P^N$ on H . Consider for $-\infty < s < t_f$ and $y(s) \in H$ given the problem:

(\mathcal{R}^N, t_f) Minimize

$$J^N(s, y^N(s), u) = \langle \tilde{G}^N y^N(t_f), y^N(t_f) \rangle + \int_s^{t_f} \{ \langle \tilde{D}^N y^N(t), y^N(t) \rangle + \langle Qu(t), u(t) \rangle \} dt$$

subject to

$$y^N(t) = T^N(t-s)P^N y(s) + \int_s^t T^N(t-\sigma)B^N u(\sigma) d\sigma \quad \text{for } s \leq t \leq t_f.$$

The problem (\mathcal{R}^N, t_f) is considered as a problem in H even though we note that $y^N(t) \in H^N$ for each t so that $\tilde{D}^N y^N(t) = D^N y^N(t)$ and $\tilde{G}^N y^N(t_f) = G^N y^N(t_f)$. We denote the unique nonnegative selfadjoint Riccati operator function associated with $(\mathcal{R}^N, 0)$ by Π_s^N (see [8, Thm. 3.2]). The following is a consequence of [8, Thm. 5.1].

THEOREM A.2. Let (H2) hold and assume that $G^N P^N z \rightarrow Gz$ for $z \in H$. Then for $s < 0$ we have

$$\bar{u}^N \rightarrow \bar{u} \text{ uniformly on } [s, 0],$$

$$\bar{y}^N \rightarrow \bar{y} \text{ uniformly on } [s, 0],$$

$$\Pi_s^N(\xi)z \rightarrow \Pi_s(\xi)z \text{ for } z \in H, \text{ uniformly in } \xi \text{ on } [s, 0].$$

Here $\bar{u}^N, \bar{u}, \bar{y}^N, \bar{y}$ denote optimal controls and trajectories of the problems $(\mathcal{R}^N, 0)$ and $(\mathcal{R}, 0)$, respectively.

With these preliminaries, we are now prepared to prove Theorem 2.2.

Proof of Theorem 2.2. Denote by Π_s and Π_s^N , $s \leq 0$, the Riccati operator functions associated with $(\mathcal{R}, 0)$ and $(\mathcal{R}^N, 0)$ in H where we take $G = M_2 I$, $G^N = M_2 P^N$ with M_2 the constant in inequality (2.11). From Theorem A.1 applied to each of the problems $(\mathcal{R}^N, 0)$ on H^N with (2.10) (and hence (A.2) with $M = M_1$, $\beta = \omega$) holding

we conclude that for $s < 0$ one has on H^N

$$\Pi^N \leq \Pi_s^N(s) \leq \Pi^N + M_1^2 e^{2\omega s} M_2.$$

This implies that on H we have for $s < 0$ and each N

$$(A.4) \quad \Pi^N P^N \leq \Pi_s^N(s) P^N \leq \Pi^N P^N + M_1^2 e^{2\omega s} M_2.$$

Since $\Pi^N P^N$ is selfadjoint in H , we conclude from (A.4) and (A.1) that for each $\varepsilon > 0$ and $z \in H$, there exists $\xi = \xi(z, \varepsilon)$ in $(-\infty, 0)$ such that

$$(A.5) \quad |\Pi_\xi^N(\xi) P^N - \Pi^N P^N| < \varepsilon \quad \text{for every } N = 1, 2, \dots,$$

and

$$(A.6) \quad |\Pi z - \Pi_\xi(\xi) z| < \varepsilon.$$

Therefore we have

$$(A.7) \quad \begin{aligned} |\Pi^N P^N z - \Pi z| &\leq |\Pi^N P^N z - \Pi_\xi^N(\xi) P^N z| + |\Pi_\xi^N(\xi) P^N z - \Pi_\xi^N(\xi) z| \\ &\quad + |\Pi_\xi^N(\xi) z - \Pi_\xi(\xi) z| + |\Pi_\xi(\xi) z - \Pi z| \\ &\leq \varepsilon |z| + |\Pi_\xi^N(\xi)| |P^N z - z| + |\Pi_\xi^N(\xi) z - \Pi_\xi(\xi) z| + \varepsilon. \end{aligned}$$

But by Theorem A.2 and the uniform boundedness principle we have $|\Pi_\xi^N(\xi)|$ uniformly bounded in N and $\Pi_\xi^N(\xi) z \rightarrow \Pi_\xi(\xi) z$. Finally from (H2)(ii) we have $P^N z \rightarrow z$ and thus (A.7) implies $\Pi^N P^N z \rightarrow \Pi z$ for every $z \in H$. Hence (2.12) is established.

From (H2)(iii) and (2.11) it follows that $|B^N Q^{-1} B^{N*} \Pi^N|_{H^N}$ is uniformly bounded and moreover $B^N Q^{-1} B^{N*} \Pi^N P^N z \rightarrow B Q^{-1} B^* \Pi z$ for each $z \in H$. Therefore (2.13) follows from use of the variation of parameters representations for $\bar{y}^N(t) = S^N(t) z$ and $\bar{y}(t) = S(t) z$ and the Gronwall inequality along with (2.8), (2.10) and (H2)(i). Finally (2.14) is a consequence of (2.13) and (2.10).

REFERENCES

- [1] H. T. BANKS AND P. KAREIVA, *Parameter estimation techniques for transport equations with application to population dispersal and tissue bulk flow models*, LCDS Tech. Rep. 82-13, Brown Univ., Providence, RI, July 1982; J. Math. Biol., 17 (1983), pp. 253-273.
- [2] H. T. BANKS, P. L. DANIEL AND P. KAREIVA, *Estimation of temporally and spatially varying coefficients in models for insect dispersal*, LCDS Tech. Rep. 83-14, Brown Univ., Providence, RI, June, 1983.
- [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*, ICASE Rep. 82-31, Sept. 1982; SIAM J. Sci. Stat. Comput., 5 (1984), to appear.
- [4] P. L. BUTZER AND H. BERENS, *Semigroups of Operators and Approximation*, Springer, Berlin-Heidelberg-New York, 1967.
- [5] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
- [6] R. DATKO, *Uniform asymptotic stability of evolutionary processes in a Banach space*, SIAM J. Math. Anal., 3 (1972), 428-445.
- [7] H. FUJITA AND A. MIZUTANI, *On the finite element method for parabolic equations, I: Approximation of holomorphic semi-groups*, J. Math. Soc. Japan, 28 (1976), pp. 749-771.
- [8] J. S. GIBSON, *The Riccati integral equations for optimal control problems on Hilbert spaces*, this Journal, 17 (1979), pp. 537-565.
- [9] ———, *An analysis of optimal model regulation: convergence and stability*, this Journal, 19 (1981), pp. 686-707.
- [10] ———, *Linear-quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximations*, this Journal, 21 (1983), pp. 95-139.
- [11] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

- [12] S. G. KREIN, *Linear Differential Equations in Banach Spaces*, Transl. Math. Mono., 29, American Mathematical Society, Providence, RI, 1971.
- [13] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, John Wiley, New York, 1972.
- [14] O. A. LADYZHENSKAYA AND N. N. URAL'TSEVA, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.
- [15] A. PAZY, *Semigroups of linear operators and applications to partial differential equations*, Lecture Notes, 10, Univ. Maryland, College Park, 1974.
- [16] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [17] R. TRIGGIANI, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383–403.
- [18] T. USHIJIMA, *On the finite element approximation of parabolic equations—consistency, boundedness and convergence*, Mem. Numer. Math., 2 (1975), pp. 21–23.

MORE STATES REACHABLE BY BOUNDARY CONTROL OF THE HEAT EQUATION*

N. WECK†

Abstract. Which temperatures can be reached in time T from given initial temperatures y by controlling the boundary temperature? It is known that the answer to this question is independent of T and y . Partial answers have been given using moment methods. In this paper we use a new method (which constructs the controls in the form of a series) and obtain a class of temperatures which can be reached exactly. Generalizations to general parabolic boundary control problems are indicated.

Key words. boundary control, parabolic equations, exact controllability

Introduction. Given $T \in \mathbb{R}_+$ consider the following initial boundary value problem in some bounded region $\Omega \subset \mathbb{R}^N$:

(IBVP) Given $y \in L^2(\Omega)$, $u \in U$ (to be specified later) find $v: [0, T] \times \bar{\Omega} \rightarrow \mathbb{R}$ satisfying

$$(1) \quad (\partial_t - \Delta_x)v(t, x) = 0,$$

$$(2) \quad v|_{\Gamma} = u, \quad \Gamma := (0, T) \times \partial\Omega,$$

$$(3) \quad v(0, \cdot) = y.$$

This has a unique solution if (1)–(3) are interpreted properly (e.g., as in Definition 3 below). Consider a pair of functions $y, z \in L^2(\Omega)$. Which conditions will guarantee that z can be attained from y by some boundary control u ? To make this more precise we introduce the next definition.

DEFINITION 1. A state z is called *reachable* from y in time T if there exists $u \in U$ such that the corresponding solution v of (IBVP) satisfies

$$v(T, \cdot) = z.$$

We shall write this symbolically as

$$y \overset{0}{\underset{u}{\mapsto}}^T z$$

or in short

$$y \overset{u}{\mapsto} z,$$

and define the set of reachable states by

$$R(U, T; y) := \{z \mid y \overset{0}{\underset{u}{\mapsto}}^T z \text{ for some } u \in U\}.$$

Let us review some pertinent results from the literature.

a) *Null-controllability.*

THEOREM 1 ([4], [3]). *We have null-controllability, i.e. $0 \in R(U, T; y)$ for all $t > 0$ and all $y \in L^2(\Omega)$.*

As a direct consequence one obtains the following theorem.

THEOREM 2 ([1], [11]). *The set $R := R(U, T; y)$ does not depend on either T or y .*

b) *D. L. Russell's results.* Put $D(L) := \{w \in H_0^1(\Omega) \mid \Delta w \in L^2(\Omega)\}$ and let $L: D(L) \subset L^2(\Omega) \rightarrow L^2(\Omega)$ be defined by $LW = -\Delta w$. Then L is selfadjoint and positive. Viewing

* Received by the editors August 10, 1981, and in revised form June 25, 1983.

† Universität Essen, Gesamthochschule, Postfach 103764, 4300 Essen 1, West Germany.

the controllability problem as a moment problem D. L. Russell has shown the next theorem.

THEOREM 3 ([4]). *There exists $\gamma \in \mathbb{R}_+$ such that $D := D(e^{\gamma L^{1/2}}) \subset R$.*

Since L has a compact resolvent we have a sequence of eigenpairs (λ_k, w_k) and Russell's reachability criterion $z \in D$ can be written as

$$\sum_k |\langle z, w_k \rangle (L^2(\Omega))|^2 e^{2\gamma\sqrt{\lambda_k}} < \infty.$$

c) *A remark of Seidman* [10]. If some extension \tilde{z} of z defined in a region $\tilde{\Omega} \supset \Omega$ is reachable in $\tilde{\Omega}$ then z is reachable in Ω : Just restrict the process in $[0, T] \times \tilde{\Omega}$ to $[0, T] \times \Omega$!

d) *G. Schmidt's "holdable targets"*.

DEFINITION 2. Consider $w \in H^1(\Omega)$ satisfying $\Delta w = 0$ and put $g := w|_{\partial\Omega}$ ($\in H^{1/2}(\partial\Omega)$ by the trace theorem [12, p. 237]). We say that w is holdable (by g).

The reason for this terminology is the fact that

$$v(t, x) := w(x)$$

defines a solution to (IBVP) with $y := w$ and $u(t, \xi) := g(\xi)$, i.e.

$$w \overset{0}{\underset{u}{\mapsto}}^T w.$$

Thus we have the next theorem.

THEOREM 4 [8]. *Holdable targets are reachable.*

Results c) and d) are rather easily obtained compared to D. L. Russell's rather deep considerations which (as he points out [5, p. 709]) cannot be improved if one insists that the constructed series converge *absolutely*. Nevertheless Seidman's and Schmidt's remarks exhibit reachable states not in D ! This is because the examples in c) *need not* and nontrivial holdable targets *cannot* satisfy homogeneous boundary conditions whereas $z \in D(e^{\gamma L^{1/2}})$ implies $z \in D(L^k)$ hence even $\Delta^k z|_{\partial\Omega} = 0$ for all $k \in \mathbb{N} := \{0, 1, 2, \dots\}$. Thus apparently there are reachable states that cannot be characterized by growth conditions on their Fourier coefficients. Rather the rapidly convergent series of rapidly oscillating functions obtained by moment methods should be complemented by spaces of slowly varying reachable functions. In the present paper we want to describe such spaces.

Notation. We shall follow the notation in [12] with one exception: the Sobolev spaces $H^{k,2}$ will be denoted just by H^k .

Regularity assumption. We assume that $\partial\Omega$ is an $(N-1)$ -dimensional C^∞ -submanifold of \mathbb{R}^N and that Ω "lies on one side of it."

Remark. In the regularity assumption C^∞ could be replaced by some finite degree of smoothness—except in § 5 where we use [7] and hence indirectly a unique continuation result of G. Schmidt and N. Weck [9] which really needs C^∞ .

1. Some facts about parabolic mixed boundary value problems. We now introduce a space of boundary controls:

$$U := \{u = \tilde{u}|_\Gamma \mid \tilde{u} \in \tilde{U}\},$$

$$\|u\|(U) := \inf_{u = \tilde{u}|_\Gamma} \|\tilde{u}\|(\tilde{U}),$$

$$\tilde{U} := \left\{ \tilde{u} \in L^2((0, T), H^1(\Omega)) \mid \frac{d}{dt} \tilde{u} \in L^2((0, T), H^{-1}(\Omega)) \right\},$$

$$\|\tilde{u}\|(\tilde{U}) := \left[\int_0^T \left(\|\tilde{u}(t)\|^2_{H^1(\Omega)} + \left\| \frac{d}{dt} \tilde{u}(t) \right\|^2_{(H^{-1}(\Omega))} \right) dt \right]^{1/2}.$$

DEFINITION 3. $v \in \tilde{U}$ is called a solution of (IBVP) if

$$(1') \quad \frac{d}{dt}v - \Delta_x v = 0 \quad \text{for almost all } t \in [0, T],$$

$$(2') \quad v(t, \cdot)|_{\partial\Omega} = u(t, \cdot)$$

in the sense of the trace theorem for almost all t ,

$$(3') \quad v(t, \cdot) \rightarrow y \quad (\text{in } L^2(\Omega)) \text{ for } t \rightarrow 0.$$

This notion of solution is described in [12, pp. 391–407], e.g. where the following is proved.

LEMMA 1. For any $u \in U$ there exists a unique solution v , and this belongs to $C^0([0, T], L^2(\Omega))$.

Since we shall need Theorem 2 (which is a direct consequence of Theorem 1) we should indicate how Theorem 1 can be proved for our choice of the space U of boundary controls: for any $\varepsilon > 0$ W. Littman [3] has constructed a fundamental solution g of the operator $\partial_t - \Delta_x$ with the following properties.

- (i) $g|_{(0, \infty) \times \mathbb{R}^N} \in C^\infty$,
- (ii) $g|_{(2\varepsilon, \infty) \times \mathbb{R}^N} = 0$,
- (iii) For $t < \varepsilon$ g coincides with the classical fundamental solution g_+ defined by

$$g_+(t, x) := \begin{cases} (4\pi t)^{-N/2} \exp(-|x|^2/4t), & t > 0, \\ 0, & t \leq 0. \end{cases}$$

So in order to prove null-controllability he can use Russell's trick invented for the wave equation: Extend y by zero to the whole of \mathbb{R}^N and put

$$\tilde{v}(t, x) := \int_{\mathbb{R}^N} g(t, x - x') f(x') dx'.$$

Then by ii) $\tilde{v}(t, x) = 0$ for $t > 2\varepsilon$. So by Seidman's remark

$$y \overset{0}{\underset{u}{\mapsto}} 0$$

provided $2\varepsilon < T$ and $u := \tilde{v}|_\Gamma$. So all we have to prove is $u \in U$ i.e. $v := \tilde{v}|_Q \in L^2(0, T; H^1(\Omega))$ and $(d/dt)v \in L^2(0, T; H^{-1}(\Omega))$. But this holds true since

- a) $v \in C^\infty(Q)$ by (i),
- b) in $(0, \varepsilon) \times \mathbb{R}^N$ (by (iii)) \tilde{v} coincides with the solution to the Cauchy problem which belongs to $L^2(0, T; H^1(\mathbb{R}^N))$ and $(d/dt)\tilde{v} \in L^2(0, T; H^{-1}(\mathbb{R}^N))$ (cf. [12, Remark 40.4, p. 398]).

2. Some facts about elliptic boundary value problems. We shall also need some information on the solution theory of elliptic boundary value problems associated with the operator L . The well-known results (formulated to fit our needs) can be read off from [12, pp. 189–209], for example. We consider the inhomogeneous Dirichlet boundary value problem:

(DBVP) Given $f \in L^2(\Omega)$ and $g \in H^{1/2}(\partial\Omega)$ find $w \in H^1(\Omega)$ such that

$$(4) \quad \mathcal{L}w := -\Delta w = f \quad (\text{in the sense of distributions}),$$

$$(5) \quad w|_{\partial\Omega} = g \quad (\text{in the sense of the trace theorem}).$$

LEMMA 2.

- (i) (DBVP) has a unique solution w for all $(f, g) \in L^2(\Omega) \times H^{1/2}(\partial\Omega)$.
- (ii) There exist bounded linear operators

$$G: L^2(\Omega) \rightarrow L^2(\Omega), \quad B: H^{1/2}(\partial\Omega) \rightarrow H^1(\Omega),$$

such that $w = Gf + Bg$.

- (iii) $\|G\| = \lambda_1^{-1}$, where λ_1 is the smallest eigenvalue of L .

For $k = 2, 3, \dots$ consider the following more general elliptic boundary value problem:

(EBVP) Given $f \in L^2(\Omega)$ and $g_j \in H^{1/2}(\partial\Omega)$ ($j \in J := \{0, \dots, k-1\}$) find $w \in W := \{w \in H^1(\Omega) \mid \mathcal{L}^j w \in H^1(\Omega) \text{ for } j \in J\}$ such that

$$(6) \quad \mathcal{L}^k w = f \quad (\text{in the sense of distributions}),$$

$$(7) \quad \mathcal{L}^j w|_{\partial\Omega} = g_j \quad (\text{in the sense of the trace theorem for all } j \in J).$$

LEMMA 3. For all $f \in L^2(\Omega)$ and $g_j \in H^{1/2}(\partial\Omega)$

$$w := G^k f + \sum_{i=0}^{k-1} G^i B g_i$$

is the unique solution of (EBVP).

Proof. We note the identities

$$(8) \quad \mathcal{L}^j G^i = G^{i-j}, \quad j \leq i,$$

$$(9) \quad G^j f|_{\partial\Omega} = 0, \quad j > 0,$$

$$(10) \quad \mathcal{L}^j B = 0, \quad j > 0,$$

$$(11) \quad B g|_{\partial\Omega} = g.$$

These show that w is indeed a solution to (EBVP). On the other hand let w be a solution and $w_j := \mathcal{L}^{k-j} w$. Then the w_j solve the following DBVPs:

$$\begin{aligned} \mathcal{L} w_1 &= f, & \text{and} & & \mathcal{L} w_{j+1} &= w_j, \\ w_1|_{\partial\Omega} &= g_{k-1}, & & & w_{j+1}|_{\partial\Omega} &= g_{k-j-1}. \end{aligned}$$

Hence by the unique solvability of (DBVP) we find $w_j = 0$ and in particular $w = w_k = 0$ if $f = 0$ and $g_j = 0$.

Remark. We might replace W by $H^{2k-1}(\Omega)$. But as we need Lemma 3 for *all* $k \in \mathbb{N}$ this simplification would make our proofs invalid if $\partial\Omega$ has only some finite degree of smoothness.

3. Slowly varying reachable states. We start with a lemma which is both a generalization and a quantitative version of Theorem 4. (In the first version of this paper this was proved by independently establishing a lemma which turned out to be the same result (though in different function spaces and with a different proof) as [6, Thm. 1] which appeared after the submission of this paper.) The following proof is more elementary.

LEMMA 4. For $T \in \mathbb{R}_+$ and $k \in \mathbb{Z}_+$ consider $z \in H^1(\Omega)$ satisfying

- (i) $\mathcal{L}^i z \in H^1(\Omega)$ for $i = 1, \dots, k-1$,

$$(ii) \quad \mathcal{L}^k z = 0,$$

$$(iii) \quad \mathcal{L}^i z|_{\partial\Omega} = \begin{cases} g, & i = k-1, \\ 0, & i < k-1. \end{cases}$$

Then $z \in R$. In particular

$$(iv) \quad y \overset{0}{\underset{u}{\mapsto}}^T z \quad \text{for}$$

$$y := \sum_{i=0}^{k-1} \frac{T^i}{i!} G^{k-1-i} Bg,$$

$$u(t, \xi) := \frac{(T-t)^{k-1}}{(k-1)!} g(\xi), \quad (t, \xi) \in \Gamma.$$

There exist C_1, C_2 (independent of k and z) such that

$$(v) \quad \|y\|_{L^2(\Omega)} \leq C_1 \|G\|^{k-1} \|g\|_{H^{1/2}(\partial\Omega)},$$

$$(vi) \quad \|u\|_U \leq C_2 k^{1/2} \frac{T^{k-1}}{(k-1)!} \|g\|_{H^{1/2}(\partial\Omega)}.$$

Proof. Define

$$w(t, x) := \sum_{i=0}^{k-1} \frac{(T-t)^i}{i!} \mathcal{L}^i z(x).$$

Then by (ii)

$$\mathcal{L}_x w(t, x) = \sum_{i=0}^{k-2} \frac{(T-t)^i}{i!} \mathcal{L}^{i+1} z(x),$$

$$\partial_t w(t, x) = - \sum_{i=1}^{k-1} \frac{(T-t)^{i-1}}{(i-1)!} \mathcal{L}^i z(x).$$

Therefore

$$(\partial_t + \mathcal{L}_x) w(t, x) = 0$$

and hence

$$y \overset{0}{\underset{u}{\mapsto}}^T z$$

where

$$y(x) := w(0, x), \quad x \in \Omega,$$

$$u(t, \xi) := w(t, \xi), \quad (t, \xi) \in \Gamma.$$

For the quantitative results v) and vi) we note that z is a solution to a special EBVP. Hence by Lemma 3

$$z = G^{k-1} Bg.$$

Thus from (8) and (9) we find

$$w(t, x) = \sum_{i=0}^{k-1} \frac{(T-t)^i}{i!} G^{k-1-i} Bg(x),$$

$$u(t, \xi) = \frac{(T-t)^{k-1}}{(k-1)!} g(\xi),$$

$$y = \sum_{i=0}^{k-1} \frac{T^i}{i!} G^{k-1-i} Bg.$$

This proves (iv). Now

$$\|y\| \leq \sum_{i=0}^{k-1} \frac{T^i}{i!} \|G\|^{k-1-i} \|Bg\|_{L^2(\Omega)} \leq e^{T/\|G\|} \|G\|^{k-1} \|Bg\|_{H^1(\Omega)}$$

proves (v) with $C_1 := e^{T/\|G\|} \|B\|$.

Finally for

$$\tilde{u}(t, x) := \frac{(T-t)^{k-1}}{(k-1)!} Bg(x)$$

we have

$$\begin{aligned} \|\tilde{u}\|_{L^2((0, T), H^1(\Omega))} &= \frac{\|Bg\|_{H^1(\Omega)}}{(k-1)!} \left[\int_0^T (T-t)^{2k-2} dt \right]^{1/2} \\ &\leq \frac{\|B\|}{(2k-1)^{1/2}} T^{1/2} \cdot \frac{T^{k-1}}{(k-1)!} \|g\|_{H^{1/2}(\partial\Omega)} \end{aligned}$$

and (for $k > 1$)

$$\begin{aligned} \left\| \frac{d}{dt} \tilde{u} \right\|_{L^2((0, T), H^{-1}(\Omega))} &= \frac{\|Bg\|_{H^{-1}(\Omega)}}{(k-2)!} \left[\int_0^T (T-t)^{2k-4} dt \right]^{1/2} \\ &\leq \frac{\|B\|}{T^{1/2}} \frac{(k-1)}{(2k-3)^{1/2}} \frac{T^{k-1}}{(k-1)!} \|g\|_{H^{1/2}(\partial\Omega)}. \end{aligned}$$

(For $k = 1$ we have $d\tilde{u}/dt = 0$.) Now

$$\begin{aligned} \|u\|(U) &\leq \left[\|\tilde{u}\|_{L^2((0, T), H^1(\Omega))}^2 + \left\| \frac{d}{dt} \tilde{u} \right\|_{L^2((0, T), H^{-1}(\Omega))}^2 \right]^{1/2} \\ &\leq C_2 k^{1/2} \frac{T^{k-1}}{(k-1)!} \|g\|_{H^{1/2}(\partial\Omega)} \end{aligned}$$

with some computable C_2 .

We are now ready to prove the next theorem.

THEOREM 5. $z \in C_\infty(\Omega)$ is reachable if

- (i) $\mathcal{L}^k z \in H^1(\Omega)$ for all $k \in \mathbb{N}$,
- (ii) $\sum_{k=0}^\infty \lambda_1^{-k} \|\mathcal{L}^k z\|_{H^1(\Omega)} < \infty$ (λ_1 : first eigenvalue of L .)

Proof. Define

$$\mathcal{L}^k z := \varphi_k, \quad \psi_k := \varphi_{k-1}|_{\partial\Omega},$$

and solve the EBVP

$$\mathcal{L}^k z_k = 0, \quad \mathcal{L}^i z_k|_{\partial\Omega} = \begin{cases} 0, & i < k-1, \\ \psi_k, & i = k-1. \end{cases}$$

It is clear that

$$(12) \quad \begin{aligned} \mathcal{L}^i z_k &= 0, & i \geq k, \\ \mathcal{L}^i z_k|_{\partial\Omega} &= \begin{cases} \psi_k, & i = k-1, \\ 0, & i \in \mathbb{N} \setminus \{k-1\}. \end{cases} \end{aligned}$$

We define

$$Z_K := z - \sum_{k=1}^K z_k.$$

By Theorem 2 all we have to prove is that there exist *some* y and *some* u such that $y \mapsto_u z$. But this is implied by the following three assertions:

(A) There exist y_k, u_k such that $y_k \xrightarrow{u_k} z_k$.

(B) $\sum y_k, \sum u_k$ are convergent.

(C) $Z_K \rightarrow 0 \quad (K \rightarrow \infty)$.

Proving (A). This follows directly from Lemma 4(iv). We have

$$y_k := \sum_{i=0}^{k-1} \frac{T^i}{i!} G^{k-1-i} B \psi_k,$$

$$u_k(t, \xi) := \frac{(T-t)^{k-1}}{(k-1)!} \psi_k(\xi).$$

Proving (B).

$$\begin{aligned} \|u_k\|(U) &\leq C_2 k^{1/2} \frac{T^{k-1}}{(k-1)!} \|\psi_k\| H^{1/2}(\partial\Omega) \\ &\leq \tilde{C}_2 k^{1/2} \frac{T^{k-1}}{(k-1)!} \|\mathcal{L}^{k-1} z\| H^1(\Omega) \end{aligned}$$

by the trace theorem. Thus we get absolute convergence of $\sum u_k$ from (ii). (Actually at this point it would be enough that $\|\mathcal{L}^k z\| H^1(\Omega) \leq A \cdot B^k$ for some constants A and B .) Similarly

$$\begin{aligned} \|y_k\| L^2(\Omega) &\leq C_1 \|G\|^{k-1} \|\psi_k\| H^{1/2}(\partial\Omega) \\ &\leq \tilde{C}_1 \lambda_1^{-(k-1)} \|\mathcal{L}^{k-1} z\| H^1(\Omega) \end{aligned}$$

and (ii) show that $\sum y_k$ is convergent.

Proving (C). From (12) we find

$$\begin{aligned} \mathcal{L}^K Z_K &= \mathcal{L}^K z, \\ \mathcal{L}^i Z_K|_{\partial\Omega} &= \mathcal{L}^i z|_{\partial\Omega} - \sum_{k=1}^K \mathcal{L}^i z_k|_{\partial\Omega} = \psi_{i+1} - \psi_{i+1} = 0, \quad i \in \{0, \dots, K-1\}. \end{aligned}$$

Furthermore

$$\mathcal{L}^i z_K = \mathcal{L}^i z - \sum_{k=1}^K \mathcal{L}^i z_k \in H^1(\Omega).$$

Thus $Z_K \in W$ is the solution of a special (EBVP) and by Lemma 3 (since $g_j = 0$)

$$Z_K = G^K(\mathcal{L}^K z).$$

We find

$$\|Z_K\| L^2(\Omega) \leq \|G\|^K \|\mathcal{L}^K z\| L^2(\Omega) \leq \lambda_1^{-K} \|\mathcal{L}^K z\| H^1(\Omega) \rightarrow 0.$$

This proves Theorem 5.

Let us introduce

$$X := \left\{ z \in C_\infty(\Omega) \mid \mathcal{L}^k z \in H^1(\Omega) \text{ for all } k \in \mathbb{N} \text{ and } \sum_{k=0}^{\infty} \lambda_1^{-k} \|\mathcal{L}^k z\|_{H^1(\Omega)} < \infty \right\}.$$

Remark 1. Given $y \in L^2(\Omega)$ and $z \in X$ the preceding proof (when combined with W. Littman's constructive proof of null-controllability) actually *constructs* a control $u \in U$ such that $y \mapsto_u z$.

Remark 2. An inspection of the preceding proof shows that Theorem 5(ii) can be replaced by

$$\lambda_1^{-k} \|\mathcal{L}^k z\|_{L^2(\Omega)} \rightarrow 0$$

and

$$\sum_k \lambda_1^{-k} \|\mathcal{L}^k z|_{\partial\Omega}\|_{H^{1/2}(\partial\Omega)} < \infty.$$

Remark 3. If z has an extension \tilde{z} to some region $\tilde{\Omega} \supset \supset \Omega$ then (by interior a priori estimates) we can replace Theorem 5(ii) by

$$\sum_k \lambda_1^{-k} \|\mathcal{L}^k \tilde{z}\|_{L^2(\tilde{\Omega})} < \infty.$$

Theorem 5 has the following simple corollary.

COROLLARY. *All polynomials are in X and hence in the space R of reachable states.*

This result (also obtained in [6]) is not contained in [4]. In fact we have the next theorem.

THEOREM 6. $X \cap D = \{0\}$.

Proof. Let

$$z = \sum_{j=1}^{\infty} \langle z, w_j \rangle_{L^2(\Omega)} w_j \in D \setminus \{0\}.$$

We have $\langle z, w_{j_1} \rangle \neq 0$ for some j_1 and

$$\|\mathcal{L}^k z\|_{L^2(\Omega)}^2 \geq |\langle z, w_{j_1} \rangle|^2 \lambda_{j_1}^{2k}.$$

Therefore

$$\lambda_1^{-k} \|\mathcal{L}^k z\|_{L^2(\Omega)} \geq |\langle z, w_{j_1} \rangle|$$

which implies $z \notin X$. Q.E.D.

Let us introduce

$$D_0 := \{z \mid \sum |\langle z, w_k \rangle|^2 e^{2\gamma \lambda_k} < \infty \text{ for some } \gamma \in \mathbb{R}_+\}.$$

This is the space of states that can be followed backward for some time γ . In still another formulation:

$$D_0 = \{z \mid y \overset{0}{\mapsto}^{\gamma} z \text{ for some } \gamma \in \mathbb{R}_+ \text{ and some } y \in L^2(\Omega)\}.$$

Let us also introduce R_k and R_∞ as the set of those states that are reachable by controls which (considered as maps from $[0, T]$ into $H^{1/2}(\partial\Omega)$) are C_k (resp. C_∞) in *some* interval $[T - \gamma, T]$. We find:

THEOREM 7. *Let $z \in R_{k+1}$ and in particular*

$$y \mapsto_u z$$

where

$$u(t) = \sum_{j=0}^k \frac{(t-T)^j}{j!} u^{(j)}(T) + \rho_k(t),$$

$$\rho_k(t) = O((t-T)^{k+1}) \quad (\text{in } H^{1/2}(\partial\Omega)).$$

Then $z = z_0 + z_1 + z_\rho$ where $z_0 \in D_0$, $z_1 \in X$ and z_ρ can be reached by a control which is

$$O((T-t)^{k+1}) \quad \text{at } t = T.$$

COROLLARY. Let $z \in R_\infty$. Then for arbitrary $k \in \mathbb{N}$ z can be decomposed as follows:

$$z = z_0 + z_1 + z_\rho$$

where $z_0 \in D_0$, $z_1 \in X$ and z_ρ can be reached by a control having a zero of order k at T .

4. Generalizations. Our choice of control space U and notion of a solution is by no means essential in order to carry out the construction of slowly varying reachable states. In fact mutatis mutandis Lemmas 4 and 5 and Theorem 5 can be proved for

- a) other choices of control spaces,
- b) other boundary conditions,
- c) other parabolic differential operators (even of higher order) provided null-controllability is available (cf. the announcements in [3, p. 569]). To illustrate this remark let us write down without proofs the corresponding results in a C^0 -framework: We put $U := C^0(\Gamma) \subset L^\infty(\Gamma)$ and for $y \in L^2(\Omega)$ define the generalized solution of (IBVP) as in [7, Thm. 1], for example.

Let $R(C^0(\Gamma), T; y) (\subset C_0(\bar{\Omega}))$ be the corresponding set of reachable states. Using W. Littman's fundamental solution we can prove null-controllability hence

$$R^0 := R(C^0(\Gamma), T; y)$$

is independent of T and y . Define:

(DBVP⁰) Given $f \in C^0(\bar{\Omega})$ and $g \in C^0(\partial\Omega)$ find $w \in C^0(\bar{\Omega})$ such that

$$(4^0) \quad \mathcal{L}w = f \quad (\text{in the sense of distributions}),$$

$$(5^0) \quad w|_{\partial\Omega} = g.$$

LEMMA 2⁰ (cf. [12; pp. 268 ff]).

- (i) (DBVP⁰) has a unique solution w .
- (ii) There exist bounded linear operators

$$G: C^0(\bar{\Omega}) \rightarrow C^0(\bar{\Omega}), \quad B: C^0(\partial\Omega) \rightarrow C^0(\bar{\Omega}),$$

such that $w = Gf + Bg$.

- (iii) $\|G\| = \mu$ and $\|B\| = 1$ where

$$\mu := \max \{w_0(x) \mid x \in \bar{\Omega}\}, \quad \mathcal{L}w_0 = 1, \quad w_0|_{\partial\Omega} = 0.$$

Remark. There is an elementary explicit estimate for μ if $N \geq 3$: Let $g(x, y) = (1/(N-2))\omega_N^{-1}|x-y|^{-N+2} + h(x, y)$ be Green's function for the above boundary value problem. ($\omega_N := \frac{1}{2}(\pi^{N/2}/\Gamma(N/2))$ is the area of the unit sphere S^{N-1} .) By the maximum principle we have $h \leq 0$. Thus

$$w_0(x) = \int_{\Omega} g(x, y) dy \leq \int_{\Omega} \frac{1}{N-2} \omega_N^{-1} |x-y|^{-N+2} dy.$$

Let $B(0, r)$ be a ball satisfying $\mu(B(0, r)) = \mu(\Omega)$ (Ω : Lebesgue-measure), i.e. $r = (N\mu(\Omega)\omega_N^{-1})^{1/N}$. By symmetrization

$$\int_{\Omega} |x-y|^{-N+2} dy \leq \int_{B(0,r)} |y|^{-N+2} dy = \frac{1}{2}\omega_N r^2.$$

Thus

$$\mu = \max w_0(x) \leq \frac{1}{2(N-2)} (N\mu(\Omega)\omega_N^{-1})^{2/N}.$$

(EBVP⁰) Given $f \in C^0(\bar{\Omega})$ and $g_j \in C^0(\partial\Omega)$ ($j \in J := \{0, \dots, k-1\}$) find $w \in W := \{w \in C^0(\bar{\Omega}) \mid \mathcal{L}^j w \in C_0(\bar{\Omega}) \text{ for } j \in J\}$ such that

$$(6^0) \quad \mathcal{L}^k w = f \quad (\text{in the sense of distributions})$$

$$(7^0) \quad \mathcal{L}^j w|_{\partial\Omega} = g_j \quad (j \in J).$$

LEMMA 3⁰. For all $f \in C^0(\bar{\Omega})$ and $g_j \in C^0(\partial\Omega)$ ($j \in J$)

$$w := G^k f + \sum_{i=0}^{k-1} G^i B g_{k-i-1}$$

is the unique solution of (EBVP⁰).

LEMMA 4⁰: For $T \in \mathbb{R}_+$ and $k \in \mathbb{Z}_+$ consider $z \in C^0(\bar{\Omega})$ satisfying

$$(i) \quad \mathcal{L}^i z \in C^0(\bar{\Omega}) \text{ for } i = 1, \dots, k-1,$$

$$(ii) \quad \mathcal{L}^k z = 0,$$

$$(iii) \quad \mathcal{L}^i z|_{\partial\Omega} = \begin{cases} g, & i = k-1, \\ 0, & i < k-1. \end{cases}$$

Then $z \in R$. In particular

$$(iv) \quad y \overset{0}{\underset{u}{\mapsto}}^T z \text{ for}$$

$$y := \sum_{i=0}^{k-1} \frac{T^i}{i!} G^{k-1-i} B g,$$

$$u(t, \xi) := \frac{(T-t)^{k-1}}{(k-1)!} g(\xi), \quad (t, \xi) \in T.$$

There exists C (independent of k and z) such that

$$(v) \quad \|y\|_{C^0(\bar{\Omega})} \leq C \|G\|^{k-1} \|g\|_{C^0(\partial\Omega)},$$

$$(vi) \quad \|u\|_{C^0(\Gamma)} \leq \frac{T^{k-1}}{(k-1)!} \|g\|_{C^0(\partial\Omega)}.$$

THEOREM 5⁰. $z \in C^\infty(\bar{\Omega})$ is reachable if $z \in X^0$ where

$$X^0 := \left\{ z \mid \sum_{k=0}^{\infty} \mu^k \|\mathcal{L}^k z\|_{C^0(\bar{\Omega})} < \infty \right\}.$$

Similarly the corollary and the remarks of Theorem 5 can be adapted to the present situation.

5. Time-optimal control. One of the advantages of Russell's space D has been its inherent topology which made it easy to prove bang-bang-theorems for time-optimal controls using duality methods. Recently G. Schmidt [7] has proved bang-bang-theorems more directly by introducing to R the topology of U via the control operator. In this final section we would like to show that Schmidt's results may be applied to D as well as to X . Consider the time-optimal problem

$$y \overset{0}{\underset{u}{\mapsto}}^T z, \quad T \stackrel{!}{=} \text{Min}$$

where y, z are given and u is to be chosen optimally from

$$U_{ad} := \{u \in L^\infty((0, \infty) \times \partial\Omega) \mid |u(t, \xi)| \leq M \text{ a.e.}\}.$$

Put

$$R_M(y) := \bigcup_{T>0} R(U_{ad}, T; y).$$

It is well-known that a time-optimal \hat{u} exists provided $z \in R_M(y)$. The problem is to show that \hat{u} is unique and bang-bang, i.e. $u(t, \xi) = \pm M$ almost everywhere.

G. Schmidt defines

$$\|z\|_R := \inf_u \{\|u\|_{L^\infty((0, 1) \times \partial\Omega)} \mid 0 \xrightarrow{u} z\}$$

and proves the next theorem.

THEOREM 8 [7, Thm. 3]. *If $z \in R_M(y)$ and if*

$$\inf \{\|z - z_0\|_R \mid z_0 \in D_0\} < M$$

then there exists a unique time-optimal control \hat{u} and $\hat{u} = \pm M$ almost everywhere.

This result can be applied directly if $z \in D$ because in this case $z \in \overline{D_0}$ (closure with respect to $\|\cdot\|_R$) by Russell's construction. If $z \in X^0$ recall that we constructed u (and some initial state y) such that $y \xrightarrow{u} z$ where, $u = \sum_{k=1}^\infty u_k$ and

$$u_k(t, \xi) := \frac{(T-t)^{k-1}}{(k-1)!} \mathcal{L}^{k-1} z(\xi), \quad (t, \xi) \in [0, T] \times \partial\Omega$$

this series converging $C^0(\Gamma)$. (We now use the results of § 4 which are closer to G. Schmidt's framework.) For $\delta > 0$ define χ_δ by

$$\chi_\delta(t, \xi) := \begin{cases} 1 & \text{for } t < T - \delta, \\ 0 & \text{for } t > T - \delta, \end{cases}$$

and define $z_\delta^I, z_\delta^{II}$ by

$$y \xrightarrow[\chi_\delta u]{0} z_\delta^I, \quad 0 \xrightarrow[(1-\chi_\delta)u]{0} z_\delta^{II}.$$

Then $z_\delta^I \in D_0$, $z_\delta^I + z_\delta^{II} = z$ and

$$\|z_\delta^{II}\|_R \leq \sup_{\substack{t \in [T-\delta, T] \\ \xi \in \partial\Omega}} |u(t, \xi)|.$$

Letting $\delta \rightarrow 0_+$ we obtain

$$\begin{aligned} \inf_{z_0 \in D_0} \|z - z_0\|_R &\leq \inf_\delta \|z - z_\delta^I\|_R \\ &\leq \inf_\delta \sup |u(t, \xi)| = \sup_{\xi \in \partial\Omega} |u(T, \xi)| \\ &= \sup_{\xi \in \partial\Omega} |u_1(T, \xi)| = \sup_{\xi \in \partial\Omega} |z(\xi)|. \end{aligned}$$

We just proved the next theorem.

THEOREM 9. *Consider $z \in R(U_{ad}, T; y)$ belonging to $X^0 \oplus D$. If $\sup_{\xi \in \partial\Omega} |z(\xi)| < M$ then the time-optimal control \hat{u} is unique and satisfies $\hat{u} = \pm M$ almost everywhere.*

REFERENCES

- [1] H. O. FATTORINI, *Reachable states in boundary control of the heat equation are independent in time*, Proc. Roy. Soc. Edinburgh, Sect. A, 81 (1976), pp. 71–77.
- [2] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.
- [3] W. LITTMAN, *Boundary control theory for hyperbolic and parabolic partial differential equations with constant coefficients*, Ann. Scuola Norm. Sup. Pisa, 37 (1978), pp. 567–580.
- [4] D. L. RUSSELL, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Studies in Appl. Math., 52 (1973), pp. 189–211.
- [5] ———, *controllability and stabilization theory for linear partial differential equations: recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.
- [6] E. SACHS AND G. SCHMIDT, *On reachable states in boundary control for the heat equation, and an associated moment problem*, Appl. Math. Opt., 7 (1981), pp. 225–232.
- [7] G. SCHMIDT, *The “bang-bang” principle for the time optimal problem in boundary control of the heat equation*, this Journal, 18 (1980), pp. 101–107.
- [8] ———, *Boundary control for the heat equation with steady state targets*, this Journal, 18 (1980), pp. 145–154.
- [9] G. SCHMIDT AND N. WECK, *On the boundary behaviour of solutions to elliptic and parabolic equations with applications to boundary control for parabolic equations*, this Journal, 16 (1978), pp. 593–598.
- [10] T. SEIDMAN, *Exact boundary control for some evolution equations*, this Journal, 16 (1978), pp. 979–999.
- [11] ———, *Time invariance of the reachable set for linear control problems*, J. Math. Anal. Appl., 72 (1979), pp. 17–20.
- [12] F. TRÈVES, *Basic Linear Partial Differential Equations*, Academic Press, New York, 1975.

CONTROLLABILITE DE SYSTEMES MECANIQUES SUR LES GROUPES DE LIE*

BERNARD BONNARD†

Abstract. In the attitude control problem of a rigid satellite governed by reaction jets, the motion of the free system can be interpreted as the motion on a geodesic of a left invariant Riemannian structure on the group of rotations in \mathbb{R}^3 . We set in this article the general problem of the control of the dynamic system describing the motion on a geodesic of a left invariant Riemannian structure on a Lie group G . When G is compact, we give an algebraic necessary and sufficient controllability condition. Moreover we describe control laws using the property that on a complete Riemannian manifold two points can be joined by a geodesic.

Résumé. Dans le problème de contrôle de l'attitude d'un satellite rigide gouverné par des rétrofusées le déplacement du système libre peut s'interpréter comme le mouvement sur une géodésique d'une structure riemannienne invariante à gauche sur le groupe de rotations de \mathbb{R}^3 . On pose dans cet article le problème plus général de contrôle d'un système dynamique décrivant le mouvement sur une géodésique d'une structure riemannienne invariante à gauche sur un groupe de Lie G . On donne une condition algébrique nécessaire et suffisante de contrôlabilité dans le cas où G est compact. On décrit des politiques de commande exploitant le fait que sur une variété riemannienne complète deux points peuvent être joints par une géodésique.

1. Introduction. Cet article fait suite à un travail, proposé par l'Agence Spatiale Européenne, sur le problème de contrôle de l'attitude d'une satellite réalisé par P. E. Crouch et moi-même et dans lequel j'ai assuré l'étude du problème de contrôlabilité, le rapport final rédigé par P. E. Crouch [28] contenant par ailleurs de nombreux résultats complémentaires obtenus par ce dernier principalement dans le problème de stabilisation.

L'objet de cet article est de présenter et de généraliser à toute une classe de systèmes non linéaires les résultats de contrôlabilité que j'ai obtenus dans le problème de contrôle de l'attitude d'un satellite rigide gouverné par des rétrofusées, en utilisant d'une part les propriétés du système libre et d'autre part les techniques développées ces dernières années pour l'étude de la contrôlabilité des systèmes non linéaires [3], [15], [19], [24].

Dans cet article on étudie donc la contrôlabilité des systèmes décrits par les équations:

$$(C) \quad \frac{dg(t)}{dt} = \sum_{i=1}^n \omega_i(t) X_i(g(t)),$$

$$(D) \quad \frac{d\omega(t)}{dt} = Q(\omega(t)) + \sum_{k=1}^p u^k(t) b^k.$$

$g(t) \in G$ groupe de Lie connexe de dimension n et représente la position (ou attitude) du système à l'instant t , $\{X_i; i=1, \dots, n\}$ est une famille de champs de vecteurs invariants à gauche sur G et linéairement indépendants en tout point, $\omega(t) = (\omega_1(t), \dots, \omega_n(t)) \in \mathbb{R}^n$ et s'appelle la vitesse angulaire du système à l'instant t . Le champ de vecteurs Q est un champ quadratique qui s'écrit (Q_1, \dots, Q_n) où

$$Q_i = \sum_{j,k=1}^n \frac{I_k}{I_i} c_{ji}^k \omega_j \omega_k,$$

* Received by the editors September 15, 1982, and in revised form April 25, 1983.

† Laboratoire d'Automatique de Grenoble, Ecole Nationale Supérieure d'Ingénieurs Electriciens de Grenoble, Saint Martin D'Hères, France.

c_{ji}^k désignant les constantes de structure de l'algèbre de Lie de G et I_i sont des constantes réelles non nulles. Les contrôles u^k sont des applications constantes par morceaux définies sur $[0 + \infty[$, b^k désigne un vecteur constant de \mathbb{R}^n .

Si on pose

$$L\left(\sum_{i=1}^n \omega_i X_i(g)\right) = \frac{1}{2} \sum_{i=1}^n I_i \omega_i^2,$$

la partie libre du système (C), (D) est l'équation différentielle du second ordre sur G décrivant la loi d'évolution d'une géodésique $g(t)$ de la structure pseudo-riemannienne invariante à gauche définie par L [2].

Trois résultats de contrôlabilité sont présentés dans cet article sous forme de trois théorèmes.

Le premier résultat démontré en utilisant une approche non constructive est une caractérisation algébrique de la contrôlabilité dans le cas où G est compact et $u^k(t) \in [-1, +1] \forall k = 1, \dots, n$. Il s'énonce ainsi: si on note $D_{A.L.}$ l'algèbre de Lie engendrée par la famille de champs de vecteurs de \mathbb{R}^n , $D = \{Q, b^1, \dots, b^k\}$, le système (C), (D) est contrôlable si et seulement si le système (D) satisfait la condition suivante appelée condition du rang: la dimension de $D_{A.L.}(x)$ est $n \forall x \in \mathbb{R}^n$.

Ce résultat est prouvé en remarquant d'une part que la partie libre du système (C), (D) est un système hamiltonien sur $G \times \mathbb{R}^n$ [13], d'autre part que la partie libre du système (D) est un système hamiltonien sur chaque orbite d'une représentation équivalente à la représentation coadjointe de G [2]. On utilise alors un résultat de Hopf [13]: si un champ de vecteurs X complet laisse invariant une forme volume alors presque tous les états sont soit fuyants soit stables au sens de Poisson. Le résultat de contrôlabilité est alors une conséquence de [7]: un système asservi dont la partie libre admet un ensemble de points stables au sens de Poisson est contrôlable si et seulement si il satisfait la condition du rang.

Les deux autres résultats de contrôlabilité concernent le cas où G est semi-simple et $I_i > 0$ c'est-à-dire que L définit une structure riemannienne sur G . Le problème de contrôlabilité étant abordé de façon constructive en ce sens que l'on indique des politiques de commande.

Les lois de commande sont fondées sur un résultat classique de géométrie riemannienne [14]: $\forall g, g' \in G$ il existe une géodésique joignant g à g' . En d'autres termes $\forall g, g' \in G \exists \tilde{\omega} \in \mathbb{R}^n$ de sorte que la géodésique $g(t)$ issue en $t=0$ de g avec la vitesse angulaire initiale $\tilde{\omega}$ passe par la position g' avec une vitesse $\tilde{\omega}$. D'où l'idée pour transférer le système de g à g' de chercher à atteindre cette géodésique joignant g à g' puis ensuite d'appliquer simplement le contrôle nul.

On montre que, avec l'hypothèse $u^k(t) \in [-1, +1]$, il est possible de placer le système à partir de l'état $(g, 0)$ sur une géodésique joignant g à n'importe quelle position g' si le système (D) est localement contrôlable en 0, c'est à dire que en un temps arbitrairement petit on peut toujours transférer 0 sur toutes les vitesses suffisamment voisines de 0. D'où le théorème 2 que l'on peut résumer ainsi: le système (C), (D) est contrôlable si le système (D) est localement contrôlable en 0.

Si $u^k(t) \in \mathbb{R}$ tout entier la politique de commande proposée pour transférer l'état (g, ω) en (g', ω') est la suivante. On transfère l'état (g, ω) sur $(g, \tilde{\omega})$, puis on applique le contrôle nul le temps nécessaire pour arriver en $(g', \tilde{\omega})$, ensuite on transfère l'état $(g', \tilde{\omega})$ en l'état final désiré (g', ω') . Dans le théorème 3 on donne des conditions suffisantes sur le système (D), ceci en utilisant un résultat de Kunita [19] qui permettent de décrire une loi de commande assurant le transfert de (g, ω) en $(g, \tilde{\omega})$ et $(g', \tilde{\omega})$ en (g', ω') . Par ailleurs avec ces conditions le système est aussi fortement contrôlable

c'est-à-dire que l'état (g', ω') est accessible à partir de l'état (g, ω) en un temps arbitrairement petit.

L'organisation de l'article est la suivante: Dans la section 2 on rappelle certaines définitions et résultats, concernant les géodésiques, d'une structure pseudo-riemannienne invariante à gauche sur un groupe de Lie, ceci dans le but de faciliter la lecture de cet article, les références utilisées étant [2], [13]. Dans la section 3 on présente les résultats de contrôlabilité. Dans la section 4 ces résultats sont discutés dans le problème de contrôle d'attitude d'un satellite rigide.

2. Géodésiques d'une structure pseudo-riemannienne invariante à gauche sur un groupe de Lie.

DEFINITIONS 2.1. Soit G un groupe de Lie connexe de dimension n , d'élément neutre e . On note TG et T^*G respectivement le fibré tangent et cotangent. Soit X_1, \dots, X_n une famille de champs de vecteurs invariants à gauche sur G et linéairement indépendants en e . On note $[\cdot, \cdot]$ le crochet de Lie et c_{ij}^k les constantes de structure de G définies par

$$[X_i, X_j] = \sum_{k=1}^n c_{ij}^k X_k.$$

Une structure pseudo-riemannienne invariante à gauche sur G est définie en posant

$$L\left(\sum_{i=1}^n \omega_i X_i(g)\right) = \frac{1}{2} \sum_{i=1}^n I_i \omega_i^2, \quad I_i \neq 0.$$

Si $I_i > 0$, L définit une structure riemannienne.

Soit $g_0, g_1 \in G$ et soit Γ l'ensemble des courbes γ de G définies sur $[t_0, t_1]$ et vérifiant $\gamma(t_0) = g_0$, $\gamma(t_1) = g_1$. Considérons la fonctionnelle Φ définie sur Γ par

$$\Phi(\gamma) = \int_{t_0}^{t_1} L\left(\frac{d\gamma(t)}{dt}\right) dt.$$

Une géodésique joignant g_0 à g_1 est par définition une courbe $g(t) \in \Gamma$ telle que $\Phi(g) = \inf_{\gamma \in \Gamma} \Phi(\gamma)$.

PROPOSITION 2.2 [13, p. 171]. Une géodésique $g(t)$ satisfait l'équation différentielle du second ordre sur G appelée équation de Lagrange dont l'expression en termes de coordonnées locales (q, \dot{q}) de TG est:

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{q}} - \frac{\partial L}{\partial q} = 0.$$

PROPOSITION 2.3 [13, p. 170]. L'équation de lagrange est un système hamiltonien sur TG , de hamiltonien L , muni de la forme symplectique $\lambda = d_v L$ où d_v désigne la dérivation verticale.

PROPOSITION 2.4 [25]. L'équation de Lagrange peut s'écrire:

$$(C) \quad \frac{dg(t)}{dt} = \sum_{i=1}^n \omega_i(t) X_i(g(t)),$$

$$(E) \quad I_i \frac{d\omega_i(t)}{dt} = \sum_{j,k=1}^n I_k c_{ji}^k \omega_j(t) \omega_k(t), \quad i = 1, \dots, n.$$

L'équation (E) s'appelle l'équation d'Euler et le vecteur $\omega(t) = (\omega_1(t), \dots, \omega_n(t)) \in \mathbb{R}^n$ le vecteur vitesse angulaire (par rapport au solide généralisé).

Remarque 2.5. Si on note e^i la base canonique de \mathbb{R}^n , on identifie \mathbb{R}^n avec $T_e G$ en identifiant e^i avec $X_i(e)$. L'équation (C) peut s'écrire

$$\omega(t) = dL_{g^{-1}(t)} \left(\frac{dg(t)}{dt} \right) \in T_e G \simeq \mathbb{R}^n,$$

L_g désignant la translation à gauche. L'équation (E) généralise l'équation d'Euler dans le mouvement d'un solide libre autour d'un point fixe et exprime la conservation au cours du mouvement d'une quantité qui s'interprète comme le moment cinétique du solide généralisé.

On peut définir une représentation de G dans $T_e^* G$ appelée représentation coadjointe (cf. [2, p. 320]). Soit V une orbite de cette représentation. Kirillov a montré que l'on peut munir V d'une structure symplectique naturelle notée λ [2, p. 322]. Considérons l'isomorphisme A de $T_e C$ dans $T_e^* G$ représenté dans la base $X_i(e)$ et sa base duale par la matrice diagonale

$$\begin{pmatrix} I_1 & 0 \\ 0 & I_n \end{pmatrix}.$$

L'équation d'Euler (E) sur $T_e G$ peut être transportée à l'aide de A en une équation également dite d'Euler sur $T_e^* G$. Définissons par ailleurs la fonction $H: T_e G \rightarrow \mathbb{R}$ par $H = L \circ A^{-1}$. On rappelle le résultat suivant (cf. [2, p. 328]).

PROPOSITION 2.6. *L'orbite V est invariante par le flot défini par l'équation d'Euler sur $T_e^* G$. Sur V l'équation d'Euler est un système hamiltonien pour la structure symplectique λ , de fonction de Hamilton H .*

3. Contrôlabilité. Rappelons un certain nombre de définitions.

On considère sur une variété M , analytique réelle, paracompacte et connexe un système asservi:

$$(S) \quad \frac{dx(t)}{dt} = X(x(t)) + \sum_{k=1}^p u^k(t) Y^k(x(t))$$

où X, Y^1, \dots, Y^p désignent des champs de vecteurs analytiques réels sur M et u^k une application constante par morceaux définie sur $[0, +\infty[$ et à valeurs dans un sous-ensemble d'intérieur non vide de \mathbb{R} .

Soit $u = (u^1, \dots, u^p)$ un contrôle, on note $x_u(t, x_0)$ la solution du système asservi (S) associée à u et issue en $t=0$ de $x_0 \in M$. On dit que l'état x_1 est accessible (en un temps T) à partir de l'état x_0 s'il existe un contrôle u et $T \geq 0$ non fixé tel que $x_u(T, x_0) = x_1$. On note $A^+(x_0, T)$ l'ensemble des états accessibles à x_0 en un temps T et $A^+(x_0) = \bigcup_{T \geq 0} A^+(x_0, T)$ l'ensemble des états accessibles à x_0 .

Le système (S) est dit contrôlable si $\forall x_0 \in M, A^+(x_0) = M$. Le système est dit fortement contrôlable si $\forall T > 0, \forall x_0 \in M, A^+(x_0, T) = M$. Le système est dit localement contrôlable en x_0 si $\forall T > 0, A^+(x_0, T)$ est un voisinage de x_0 .

Le crochet de Lie de deux champs de vecteurs analytiques réels est le champ de vecteurs noté $[X, Y]$ et défini en coordonnées locales par $[X, Y](x) = \partial X(x)/\partial x Y(x) - \partial Y(x)/\partial x X(x)$. On note $D_{A.L.}$ l'algèbre de Lie engendrée par une famille D de champs de vecteurs. On dit que la famille D satisfait la condition du rang en $x \in M$ si la dimension du sous-espace vectoriel de $T_x M, D_{A.L.}(x) = \{V(x); V \in D_{A.L.}\}$ est égale à la dimension de M . On dit que D satisfait la condition du rang si elle est satisfaite $\forall x \in M$.

Soit X un champ de vecteurs analytique réel sur M et suppose que X est complet c'est-à-dire que ses courbes intégrables sont définies $\forall t \in \mathbb{R}$. On note $(\exp tX)(x)$ la

courbe intégrable de X issue en $t=0$ de $x \in M$. Le point x est dit stable dans le sens de Poisson si $\forall U$ voisinage de x et $\forall T \geq 0 \exists t_1, t_2 \geq T$ tels que $(\exp t_1 X)(x) \in U$ et $(\exp -t_2 X)(x) \in U$.

On rappelle le résultat suivant qui est une conséquence directe d'un théorème de Hopf [13, p. 143].

PROPOSITION 3.1. *Soit X un champ de vecteurs hamiltonien sur une variété symplectique M non forcément compacte. Si toutes les trajectoires de X sont bornées, alors l'ensemble des points stables dans le sens de Poisson de X est dense dans M .*

D'après la proposition 2.6 l'équation d'Euler définit un système hamiltonien de hamiltonien $L(\omega) = \frac{1}{2} \sum_{i=1}^n I_i \omega_i^2$ sur chaque orbite V d'une représentation équivalente à la représentation coadjointe de G . Si une orbite V est ouverte toutes les trajectoires de l'équation d'Euler peuvent être non bornées au sens de la topologie de V . D'où l'intérêt des deux lemmes suivants.

LEMME 3.2 [26, p. 81]. *Si G est compact, toutes les orbites de la représentation coadjointe de G sont des sous-variétés compactes régulières de T_e^*G .*

LEMME 3.3. *Si G est semi-simple, alors pour un sous-ensemble dense de T_e^*G les orbites de la représentation coadjointe de G sont des sous-variétés fermées régulières.*

Preuve. Si G est semi-simple alors la représentation adjointe et coadjointe sont équivalentes. L'orbite d'un élément semi-simple de $T_e G$ sous l'action de la représentation adjointe est fermée [27, p. 106] et donc régulière. Les éléments semi-simples forment un sous-ensemble dense de $T_e G$.

On va maintenant donner des conditions de contrôlabilité pour les systèmes (D) et (C), (D), mais rappelons le résultat suivant.

PROPOSITION 3.4 [7]. *On considère sur M le système:*

$$(S) \quad \frac{dx(t)}{dt} = X(x(t)) + \sum_{k=1}^p u^k(t) Y^k(x(t)),$$

u^k est une application constante par morceaux définie sur $[0, +\infty[$ et à valeur dans l'ensemble $\{-1, 1\}$. On suppose que l'ensemble des points stables dans le sens de Poisson de X est dense dans M . Alors le système est contrôlable si et seulement si la famille $\{X, Y^1, \dots, Y^p\}$ satisfait la condition du rang.

PROPOSITION 3.5. *Si G est compact alors le système*

$$(D) \quad \frac{d\omega(t)}{dt} = Q(\omega(t)) + \sum_{k=1}^p u^k(t) b^k, \quad |u^k(t)| \leq 1$$

est contrôlable si et seulement si la famille de champ de vecteurs de $\mathbb{R}^n \{Q, b^1, \dots, b^p\}$ satisfait la condition du rang.

Preuve. Puisque G est compact, d'après le lemme 3.2 chaque orbite de la représentation coadjointe est compacte. D'après la proposition 3.1 l'ensemble des points stables dans le sens de Poisson de l'équation d'Euler $d\omega(t)/dt = Q(\omega(t))$ est dense sur chaque orbite et donc dans \mathbb{R}^n . Le résultat est donc une conséquence de la proposition 3.4.

PROPOSITION 3.6. *On suppose que G est semi-simple et que L définit une structure riemannienne sur G . Alors le système:*

$$(D) \quad \frac{d\omega(t)}{dt} = Q(\omega(t)) + \sum_{k=1}^p u^k(t) b^k, \quad |u^k(t)| \leq 1$$

est contrôlable si et seulement si la famille de champs de vecteurs de $\mathbb{R}^n \{Q, b^1, \dots, b^p\}$ satisfait la condition du rang.

Preuve. Soit V une orbite fermée invariante par le flot défini par l'équation d'Euler (cf. la proposition 2.6). Soit $\omega \in V$, la trajectoire de l'équation d'Euler issue en $t=0$ de ω est donc contenue dans l'ensemble compact de V , $V \cap L(\omega)$, où $L(\omega) = \frac{1}{2} \sum_{i=1}^n I_i \omega_i^2$ puisque L est une intégrale première de l'équation d'Euler. L'ensemble des points stables au sens de Poisson de l'équation d'Euler est donc dense dans V d'après la proposition 3.1. Puisque l'ensemble des orbites fermées V est dense dans \mathbb{R}^n d'après le lemme 3.3, l'ensemble des points stables au sens de Poisson de l'équation d'Euler est donc aussi dense dans \mathbb{R}^n . L'assertion est une conséquence de la proposition 3.4.

Avant d'énoncer le théorème 1 on va présenter deux résultats concernant la condition du rang. Le premier est un algorithme pour calculer la condition du rang dans le cas du système (D) exploitant la nature quadratique du champ de vecteurs Q et a été démontré dans un cadre plus général par Baillieul [6], cf. aussi [10] pour une interprétation géométrique de cette condition. Le second résultat est la remarque que la condition du rang du système (C), (D) équivaut à la condition du rang du système (D) et généralise [8, 3.5].

LEMME 3.7. *Soit E le plus petit sous-espace vectoriel de \mathbb{R}^n tel que :*

- 1) $\{b^1, \dots, b^p\} \in E$;
- 2) si $v^1, v^2 \in E$ alors $[[Q, v^1], v^2] \in E$.

Alors la famille de champs de vecteurs de $\mathbb{R}^n \{Q, b^1, \dots, b^p\}$ satisfait la condition du rang si et seulement si E est tout \mathbb{R}^n .

LEMME 3.8. *La famille de champs de vecteurs $G \times \mathbb{R}^n$,*

$$H = \left\{ \left(\sum_{i=1}^n \omega_i X_i, Q \right), (0, b^1), \dots, (0, b^p) \right\}$$

satisfait la condition du rang si et seulement si la famille de champs de vecteurs de $\mathbb{R}^n \{Q, b^1, \dots, b^p\}$ satisfait la condition du rang.

Preuve. Notons X le champs de vecteurs $(\sum_{i=1}^n \omega_i X_i, Q)$ et soit $v^1, v^2 \in \mathbb{R}^n$. On remarque que $[[X, (0, v^1)], (0, v^2)] = (0, [[Q, v^1], v^2])$.

L'algèbre de Lie $H_{A.L.}$ contient donc $(0, E)$ où E est défini dans le lemme 3.7. Si la famille $\{Q, b^1, \dots, b^p\}$ satisfait la condition du rang il résulte du lemme 3.7 que $H_{A.L.}$ contient $(0, \mathbb{R}^n)$. Notons e^i la base canonique de \mathbb{R}^n , $H_{A.L.}$ contient donc $[X, (0, e^i)] = (X_i, 0) \forall i = 1, \dots, n$. Donc la famille H satisfait la condition du rang si la famille $\{Q, b^1, \dots, b^p\}$ satisfait la condition du rang. La réciproque est claire.

THÉORÈME 1. *On suppose que G est un groupe de Lie compact. Alors le système :*

$$(C) \quad \frac{dg(t)}{dt} = \sum_{i=1}^n \omega_i(t) X_i(g(t)),$$

$$(D) \quad \frac{d\omega(t)}{dt} = Q(\omega(t)) + \sum_{k=1}^p u^k(t) b^k, \quad |u^k(t)| \leq 1$$

est contrôlable si et seulement si la famille de champs de vecteurs de $\mathbb{R}^n \{Q, b^1, \dots, b^p\}$ satisfait la condition du rang.

Preuve. Puisque G est compact, toutes les orbites de la représentation coadjointe sont donc compactes d'après le lemme 3.2. La partie libre de (C), (D) est donc un système hamiltonien (cf. la proposition 2.3) dont toutes les trajectoires sont bornées. Le système est donc contrôlable d'après les propositions 3.1 et 3.4 si et seulement si H satisfait la condition du rang. Cette condition du rang équivaut d'après le lemme 3.8 à la condition du rang de la famille $\{Q, b^1, \dots, b^p\}$.

Des conditions nécessaires et suffisantes de contrôlabilité ont été obtenues dans le cas où G est compact. Le problème de contrôlabilité va maintenant être abordé, de façon plus constructive, dans le cas où G est semi-simple, mais en faisant l'hypothèse que L définit une structure riemannienne.

THÉORÈME 2. *On suppose que G est un groupe de Lie semi-simple et que L définit une structure riemannienne sur G . On considère le système:*

$$(C) \quad \frac{dg(t)}{dt} = \sum_{i=1}^n \omega_i(t) X_i(g(t)),$$

$$(D) \quad \frac{d\omega(t)}{dt} = Q(\omega(t)) + \sum_{k=1}^p u^k(t) b^k, \quad |u_k(t)| \leq 1.$$

Si le système (D) est localement contrôlable en 0, alors le système (C), (D) est contrôlable.

Preuve. Notons $g_u(t, g_0, \omega_0)$, $\omega_u(t, \omega_0)$, la solution de (C), (D) associée au contrôle $u = (u_1, \dots, u_p)$ issue en $t = 0$ de g_0, ω_0 et $g(t, g_0, \omega_0)$, $\omega(t, \omega_0)$ la solution associée au contrôle $u = 0$, $g(t, \omega_0, g_0)$ est donc la géodésique issue en $t = 0$ de g_0 avec la vitesse angulaire initiale ω_0 .

LEMME 1. $\forall \lambda \in \mathbb{R}, \forall t \in \mathbb{R}, \omega(t, \lambda \omega_0) = \lambda \omega(\lambda t, \omega_0)$ et $g(t, g_0, \lambda \omega_0) = g(\lambda t, g_0, \omega_0)$.

Preuve. C'est un résultat classique qui exprime le fait que la géodésique parcourue ne dépend que de la direction du vecteur vitesse initiale et non de son intensité. C'est clairement une conséquence immédiate de la nature homogène de l'équation différentielle géodésique.

LEMME 2. $\forall g_1, g_2 \in G$ il existe $\omega_1 \in \mathbb{R}^n$ tel que $g(1, g_1, \omega_1) = g_2$.

Preuve. Les géodésiques de la structure riemannienne sont définies $\forall t \in \mathbb{R}$. On applique alors un autre résultat classique de géométrie riemannienne $\forall g_1, g_2 \in G$ il existe une géodésique joignant g_1 à g_2 [14, p. 56]. En d'autres termes $\forall g_1, g_2 \in G \exists \tilde{\omega}_1 \in \mathbb{R}^n, T \in \mathbb{R}$ tel que $g(T, g_1, \tilde{\omega}_1) = g_2$. Posons $\omega_1 = T\tilde{\omega}_1$, alors d'après le lemme 1 $g(1, g_1, \omega_1) = g_2$.

LEMME 3. $\forall g_1, g_2 \in G$ il existe une suite de contrôles v^n définis sur $[0, T^n]$ tels que la suite $g_{v^n}(T^n, g_1, 0)$ converge vers g_2 et $\omega_{v^n}(T^n, 0)$ converge vers 0.

Preuve. Soit $g_1, g_2 \in G$ d'après le lemme 2 il existe $\omega_1 \in \mathbb{R}^n$ tel que $g(1, g_1, \omega_1) = g_2$, posons par ailleurs $\omega_2 = \omega(1, \omega_1)$.

Puisque (D) est localement contrôlable en 0, $\forall n \in \mathbb{N}$ il existe, par définition, $\lambda^n > 0$ et un contrôle u^n défini sur $[0, 1/n]$ tel que $\omega_{u^n}(1/n, 0) = \lambda^n \omega_1$.

Posons $T^n = 1/n + 1/\lambda^n$, le contrôle v^n est défini par

$$v^n = u^n \quad \text{sur} \left[0, \frac{1}{n}\right], \quad v^n = 0 \quad \text{sur} \left[\frac{1}{n}, T^n\right].$$

Posons

$$\begin{aligned} g_1^n &= g_{u^n}\left(\frac{1}{n}, g_1, 0\right), & \omega_1^n &= \omega_{u^n}\left(\frac{1}{n}, 0\right), \\ g_2^n &= g\left(\frac{1}{\lambda^n}, g_1^n, \omega_1^n\right), & u_2^n &= \omega\left(\frac{1}{\lambda^n}, \omega_1^n\right). \end{aligned}$$

Puisque la structure riemannienne est invariante à gauche et puisque $\omega_1^n = \lambda^n \omega_1$ on peut en utilisant le lemme 1 écrire $g_2^n = g_1^n g(1, e, \omega_1)$ où e désigne l'élément neutre de G . Lorsque $n \rightarrow +\infty$, g_1^n converge vers g_1 et donc g_2^n converge vers $g_2 = g_1 g(1, e, \omega_1)$.

Par ailleurs $\omega_1^n = \lambda^n \omega_1$ et donc $\omega_2^n = \lambda^n \omega_2$ d'après le lemme 1. Lorsque $n \rightarrow +\infty$, ω_1^n converge vers 0 et donc ω_2^n converge également vers 0.

On a donc montré que lorsque $n \rightarrow +\infty$, $g_{v^n}(T^n, g_1, 0) = g_2^n \rightarrow g_2$ et $\omega_{v^n}(T^n, 0) \rightarrow 0$, d'où l'assertion du lemme.

LEMME 4. *La famille de champs de vecteurs $H = \{(\sum_{i=1}^n \omega_i X_i, Q), (0, b^1), \dots, (0, b^p)\}$ satisfait la condition du rang.*

Preuve. D'après le lemme 3.8 il suffit de montrer que la famille $\{Q, b^1, \dots, b^p\}$ satisfait la condition du rang. Or (D) est localement contrôlable en 0, donc la condition du rang est satisfaite en 0 d'après [24]. Puisque Q est homogène, il résulte de [10] qu'elle est satisfaite partout.

Démontrons maintenant le théorème.

Soit $(g_I, \omega_I), (g_F, \omega_F) \in G \times \mathbb{R}^n$, montrons que (g_F, ω_F) est accessible à (g_I, ω_I) .

D'après le lemme 4 et [24] il suffit de montrer qu'il existe une suite u^n de contrôles définis sur $[0, \hat{T}^n]$ tels que $g_{u^n}(\hat{T}^n, g_I, \omega_I)$ converge vers g_F et $\omega_{u^n}(\hat{T}^n, \omega_I)$ vers ω_F .

D'après la proposition 3.6 et le lemme 4 le système (D) est contrôlable. Donc $\exists g_1, g_2 \in G$ et des contrôles u_1, u_2 définis respectivement sur $[0, T_1], [0, T_2]$ tels que $g_{u_1}(T_1, g_I, \omega_I) = g_1, \omega_{u_1}(T_1, \omega_I) = 0$ et $g_{u_2}(T_2, g_2, 0) = g_F, \omega_{u_2}(T_2, 0) = \omega_F$.

D'après le lemme 3 il existe une suite v^n de contrôles définis sur $[0, T^n]$ tels que $g_{v^n}(T^n, g_1, 0)$ converge vers g_2 et $\omega_{v^n}(T^n, 0)$ vers 0.

La suite de contrôles u^n recherchée est définie en posant $u^n = u_1$ sur $[0, T_1]$, $u^n = v^n$ sur $[T_1, T_1 + T^n]$, $u^n = u_2$ sur $[T_1 + T^n, \hat{T}_n]$ avec $\hat{T}_n = T_1 + T_2 + T_n$.

On va rappeler une condition pour que le système (D) soit localement contrôlable en 0. Il faut remarquer que puisque est quadratique le comportement local du système ne peut, à moins que $p = n$, être analysé en utilisant la linéarisation classique de [20].

La condition de locale contrôlabilité présentée est extraite de [15] et est obtenue en utilisant les techniques de linéarisation d'ordre supérieur de l'ensemble des états accessibles le long d'une trajectoire, ici un point.

LEMME 3.9. *On note B l'espace vectoriel engendré par $\{b^1, \dots, b^p\}$. Le système (D) est localement contrôlable en 0 si le cône convexe engendré par la famille de vecteurs $\{b, [b, [b, Q]]\}; b \in B\}$ est tout \mathbb{R}^n .*

Remarques 3.10. Si on note $(C^-), (D^-)$ le système décrit par les équations (C), (D) mais avec $t \leq 0$ on peut remarquer que sous les hypothèses sur (D) du lemme 3.9 non seulement (D) est localement contrôlable en 0 mais aussi (D^-) . Par ailleurs il résulte de [15] que $\forall g \in G$ les systèmes (C), (D) et $(C^-), (D^-)$ sont localement contrôlables en $(g, 0)$. Cela permet, par une application de [16], de construire une synthèse locale et une loi de stabilisation locale autour de n'importe quel état d'équilibre du système (C), (D).

Si la condition de locale contrôlabilité du système (D) en 0 est bien plus forte que la condition de contrôlabilité, à moins que la structure riemannienne soit aussi invariante à droite, d'un point de vue pratique la possibilité de stabiliser localement un système autour de l'état final désiré apparaît indispensable.

Le problème de décrire une loi de commande pour mettre le vecteur vitesse angulaire à 0 peut être résolu, sous les hypothèses du lemme 3.9, en s'inspirant de la technique de stabilisation de [17] ou en réalisant une synthèse de l'équation (D), le problème d'atteindre à partir de 0 une vitesse quelconque se résolvant par les mêmes techniques appliquées au système (D^-) .

La loi de commande décrite dans le théorème 2 nécessite la détermination d'une géodésique. En fait ce n'est pas nécessaire, on peut remplacer cette géodésique par une trajectoire sur G qui est par exemple la concaténation de géodésiques qui sont aussi des trajectoires de champs de vecteurs invariants à gauche cf [28, p. 162]. D'un point de vue pratique d'ailleurs il n'est pas recommandé d'utiliser les géodésiques instables, cette instabilité étant liée à la courbure sectionnelle de G [2, p. 300].

On va présenter maintenant un résultat de contrôlabilité forte. C'est la notion de contrôlabilité adaptée au problème de contrôle en temps fixé avec minimisation de la consommation d'énergie.

THÉOREME 3. *On suppose que G est un groupe de Lie semi-simple et que L définit une structure riemannienne sur G . On considère le système*

$$(C) \quad \frac{dg(t)}{dt} = \sum_{i=1}^n \omega_i(t) X_i(g(t)),$$

$$(D) \quad \frac{d\omega(t)}{dt} = Q(\omega(t)) + \sum_{k=1}^p u^k(t) b^k, \quad u^k(t) \in \mathbb{R}.$$

On note B le sous espace vectoriel de \mathbb{R}^n engendré par $\{b^1, \dots, b^p\}$ et on suppose que le cône convexe engendré par $\{b, [b, [b, Q]]; b \in B\}$ est tout \mathbb{R}^n . Alors le système est fortement contrôlable.

Preuve. On rappelle que l'on note $A^+(x, T)$ l'ensemble des états accessibles à $x \in G \times \mathbb{R}^n$ en un temps T . Démontrons le lemme préliminaire:

LEMME. *Soit $(g, \omega) \in G \times \mathbb{R}^n$, alors $\forall T > 0$ l'adhérence de $A^+((g, \omega), T)$ contient (g, \mathbb{R}^n) .*

Preuve. Soit $b \in B$, on remarque que le champ de vecteurs $(ad^k(0, b))$ $(\sum_{i=1}^n \omega_i X_i, Q)$ est nul si $k \geq 3$ et égal au champ de vecteurs constant $(0, [b, [b, Q]])$ si $k = 2$. L'assertion du lemme est alors une simple application de [19, 2.2 et 3.3].

Soit $(g_b, \omega_b), (g_F, \omega_F) \in G \times \mathbb{R}^n$ et $T > 0$ montrons que (g_F, ω_F) appartient à $A^+((g_b, \omega_b), T)$.

Il suffit d'après [24] de montrer que (g_F, ω_F) appartient à l'adhérence de $A^+((g_b, \omega_b), T)$, en effet d'après les lemmes 3.7 et 3.8 (H étant défini en 3.8) l'idéal engendré par $(0, B)$ dans $H_{A.L.}$ coïncide avec $H_{A.L.}$.

D'après le lemme 2 du théorème 2 il existe $\omega \in \mathbb{R}^n$ tel que $g(1, g_b, \omega) = g_F$, posons par ailleurs $\tilde{\omega} = \omega(1, \omega)$.

D'après le lemme préliminaire il existe deux suites u^n, \tilde{u}^n de contrôles définis sur $[0, T/3]$ tels que lorsque $n \rightarrow +\infty$

$$\begin{aligned} g_{u^n} \left(\frac{T}{3}, g_b, \omega_b \right) &\rightarrow g_b, & \omega_{u^n} \left(\frac{T}{3}, \omega_b \right) &\rightarrow \frac{3\omega}{T}, \\ g_{\tilde{u}^n} \left(\frac{T}{3}, g_F, \frac{3\tilde{\omega}}{T} \right) &\rightarrow g_F, & \omega_{\tilde{u}^n} \left(\frac{T}{3}, \frac{3\tilde{\omega}}{T} \right) &\rightarrow \omega_F. \end{aligned}$$

Définissons la suite de contrôles v^n sur $[0, T]$ par $v^n = u^n$ sur $[0, T/3]$, $v^n = 0$ sur $[T/3, 2T/3]$, $v^n = \tilde{u}^n$ sur $[2T/3, T]$.

Clairement $g_{v^n}(T, g_b, \omega_b)$ converge vers g_F et $\omega_{v^n}(T, \omega_b)$ vers ω_F lorsque $n \rightarrow +\infty$, d'où l'assertion du théorème.

Remarques 3.11. La politique de commande utilisant [19] pour à partir d'un état obtenir une vitesse angulaire quelconque sans changer la position est constructive et le vecteur $[b, [b, Q]]$ est colinéaire au champ Q évalué le long de la droite engendré par b [10].

La loi de commande proposée dans le théorème 3 nécessite pour être mise en oeuvre la détermination d'une géodésique soit l'intégration d'une équation différentielle sur $G \times \mathbb{R}^n$ avec condition initiale et finale. D'un point de vue de consommation d'énergie elle est en un sens satisfaisante puisque l'on utilise au maximum le contrôle nul. Cette remarque est importante si l'on considère le fait que la commande optimale donnée par le principe du maximum de Pontriaguine nécessite l'intégration d'une

équation différentielle avec condition initiale et finale sur $T^*(G \times \mathbb{R}^n)$ soit avec deux fois plus de paramètres ce qui interdit par exemple dans le problème de contrôle d'attitude toute mise en oeuvre numérique, les calculs devant être conduits en temps réel [11].

On peut comme dans le théorème 2 ne pas suivre la géodésique mais emprunter un chemin dans G constitué de portions de géodésique qui sont aussi des trajectoires de champs de vecteurs invariants à gauche, ceci notamment pour éviter le calcul de la géodésique et le problème posé par son instabilité. Cependant il est alors nécessaire de fournir plus d'énergie au système car il faut lui donner de l'énergie pour sauter d'une géodésique du chemin choisi à la suivante.

4. Application au problème de contrôle d'attitude. Dans le cas particulier du contrôle de l'attitude d'un satellite rigide gouverné par des rétrofusées l'équation (D) s'écrit:

$$\begin{aligned}\frac{d\omega_1(t)}{dt} &= a_1\omega_2(t)\omega_3(t) + \sum_{k=1}^p u^k(t)b_1^k, \\ \frac{d\omega_2(t)}{dt} &= a_2\omega_1(t)\omega_3(t) + \sum_{k=1}^p u^k(t)b_2^k, \\ \frac{d\omega_3(t)}{dt} &= a_3\omega_1(t)\omega_2(t) + \sum_{k=1}^p u^k(t)b_3^k,\end{aligned}$$

avec

$$a_1 = \frac{I_2 - I_3}{I_1}, \quad a_2 = \frac{I_3 - I_1}{I_2}, \quad a_3 = \frac{I_1 - I_2}{I_3}.$$

$I_1 > I_2 > I_3 > 0$ désignent les moments d'inertie principaux du satellite supposés tous distincts. Le groupe G est $SO(3)$ bien que pour les calculs numériques il est judicieux de représenter le système sur le groupe des quaternions $Sp(1)$, en effet le plongement naturel de $SO(3)$ est dans \mathbb{R}^9 et celui de son revêtement universel dans \mathbb{R}^4 puisque $Sp(1) \simeq S^3$, d'où une économie sensible sur le nombre des paramètres à conserver en mémoire.

Le dispositif de commande est constitué par p couples de rétrofusées, les rétrofusées étant couplées pour émettre du gaz dans deux directions opposées.

Le système libre possède 4 intégrales premières qui expriment la conservation de l'énergie cinétique et du vecteur moment cinétique au cours du mouvement. Les orbites invariantes par le flot défini par les équations d'Euler (cf. le remarque 2.5) sont les surfaces $I_1^2\omega_1^2 + I_2^2\omega_2^2 + I_3^2\omega_3^2 = \text{constante}$ qui expriment la conservation au cours du mouvement de la norme du moment cinétique.

L'équation est complètement intégrable, toutes les trajectoires sont périodiques de période non nulle excepté les trois axes de \mathbb{R}^3 notés e^1, e^2, e^3 qui sont des ensembles de positions d'équilibre et une famille de trajectoires constituées de demi-ellipses et qui sont toutes situées dans deux plans notés H_1 et H_2 définis par l'équation $a_3\omega_1^2 - a_1\omega_3^2 = 0$ (cf. [2] et [8]).

Les trois axes et ces deux plans sont les seuls sous-espaces vectoriels de \mathbb{R}^3 invariants par le flot défini par les équations d'Euler et l'on peut en utilisant le théorème 1 et [10] caractériser géométriquement la contrôlabilité du système. Si on note B l'espace vectoriel engendré par les vecteurs $\{b^k, k=1, \dots, p\}$, le système est contrôlable à moins que B soit contenu dans un des sous-espaces invariants de l'équations d'Euler. En particulier le système est contrôlable avec un seul couple de rétrofusées à moins

que le vecteur $b = (b_1, b_2, b_3)$ soit orienté le long d'un axe de \mathbb{R}^3 ou appartienne à l'un des plans invariants, H_1, H_2 c'est-à-dire $a_3 b_1^2 - a_1 b_3^2 = 0$. Si l'on dispose non pas d'un couple de rétrofusées mais d'une seule rétrofusée et même en supposant que la poussée de celle-ci est aussi grande que l'on veut, i.e., $u(t) \in \mathbb{R}^+$, le système (D) n'est plus contrôlable, en effet on met en évidence l'existence d'une région invariante limitée par un des dièdres formées par les plans H_1 et H_2 .

Si l'on dispose de deux couples de rétrofusées ou plus la condition de locale contrôlabilité de l'équation (D) en 0 du lemme 3.9, qui implique aussi que le système (C), (D) est fortement contrôlable si $u_k(t) \in \mathbb{R}$ tout entier d'après le théorème 3, équivaut à la condition du rang, c'est-à-dire que l'espace vectoriel B doit être distinct de H_1 et H_2 . Cela résulte directement du fait que le champ de vecteurs $[b, [b, Q]]$ est colinéaire à Q évalué le long de la droite engendrée par b [10]. Par ailleurs comme il est noté dans la remarque 3.10 cela permet de construire pour le système une synthèse locale et une loi de stabilisation locale autour de chaque état d'équilibre, ce qui d'un point pratique est fondamental. Il faut remarquer que d'un point de vue physique cette locale contrôlabilité avec seulement deux couples de rétrofusées n'est pas évidente. En effet considérons le dispositif de commande suivant: $p = 2$, $b^1 = e^1$, $b^2 = e^2$, c'est-à-dire que l'on peut opérer à l'aide de ce dispositif des rotations du système autour des deux axes d'inertie principaux e^1 et e^2 du satellite. Supposons que l'on veuille faire tourner le satellite dont la vitesse initiale est nulle autour de e^3 d'un angle ε petit. La loi de commande en termes d'angles d'Euler consiste à réaliser le changement d'attitude en trois rotations successives, une rotation de $\pi/2$ par rapport à e^1 , une rotation de ε par rapport à e^2 , une rotation de $\pi/2$ par rapport à e^1 (cf. [9], [11]). La loi de commande donnée par [16] permet de réaliser le même changement d'attitude mais sans imposer de faire deux fois basculer le satellite d'un angle de $\pi/2$.

REFERENCES

- [1] V. ARNOLD, *Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits*, Annale de l'institut Fourier, XVI, no 1 (1966), pp. 319-361.
- [2] ———, *Méthode mathématiques de la mécanique classique*, Editions Mir, Moscow 1976.
- [3] R. W. BROCKETT, *System theory on group manifold and coset spaces*, this Journal, 10 (1972), pp. 265-284.
- [4] J. BAILLIEUL, *Geometric methods for nonlinear optimal control problems*, J. Optim. Theory Appl., 25 (1978), pp. 519-549.
- [5] ———, *A controllability result with an application to rigid body orientation*, Midwest Conference on Circuits and Systems, 1978.
- [6] ———, *The geometry of homogeneous polynomial dynamical systems*, to appear.
- [7] B. BONNARD, *Contrôlabilité des systèmes nonlinéaires*, C.R. Acad. Sci. Paris, Série I', 292 (1981), pp. 535-537.
- [8] ———, *Contrôle de l'attitude d'un satellite rigide*, RAIRO Automatique, 16 (1982), pp. 85-93.
- [9] ———, *Une loi de commande pour le problème de contrôle de l'attitude d'un satellite*, Université de Bordeaux I, Rapport de Recherche 8020, October 1980.
- [10] ———, *Contrôlabilité et observabilité d'une certaine classe de systèmes non linéaires*, Note Interne Laboratoire d'Automatique de Grenoble no 82-09, January 1982.
- [11] G. M. COUPÉ, *Assessment of the state of the art for simultaneous three-axis large angle attitude manoeuvres*, E.S.T.E.C. working paper 1260, October 1980.
- [12] P. E. CROUCH, *Attitude control of spacecraft*, Proc. Seminar in Mathematical Systems Theory, Banach Math. Centre, Warsaw, September-December 1980, to appear.
- [13] C. GODBILLON, *Géométrie différentielle et mécanique analytique*, Hermann, Paris, 1969.
- [14] S. HEGASON, *Differential Geometry and Symmetric Spaces*, Academic Press, New York, 1962.
- [15] H. HERMES, *Lie algebras of vector fields and local approximation of attainable sets*, this Journal, 16 (1978), pp. 715-727.

- [16] ———, *On the synthesis of a stabilizing feedback control via Lie algebraic methods*, this Journal, 18 (1980), pp. 352–361.
- [17] V. JURDJEVIC AND J. P. QUINN, *Controllability and stability*, J. Differential Equations, 28 (1979), pp. 381–389.
- [18] A. A. KIRILLOV, *Elements of the Theory of Representations*, Springer-Verlag, Berlin, Heidelberg, New York, 1976.
- [19] H. KUNITA, *On the controllability of nonlinear systems with applications to polynomial systems*, Appl. Math. Optim., 5 (1979), pp. 89–99.
- [20] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [21] G. MEYER, *On the use of Euler's theorem on rotations for the synthesis of attitude control systems*, NASA TN D-3643.
- [22] R. MORTENSEN, *A globally stable linear attitude regulator*, Internat. J. Control, 8 (1968), pp. 297–302.
- [23] M. NAIMACK AND A. STERN, *Theorie des représentations des groupes*, Editions Mir, Moscow, 1979.
- [24] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.
- [25] A. M. VINOGRADOV AND B. A. KUPERSHMIT, *The structure of hamiltonian mechanics*, Russian Math. Surveys, 32, 4 (1977), pp. 177–243.
- [26] V. S. VARADARAJAN, *Lie Groups, Lie Algebras and Their Representations*, Prentice-Hall, Englewood Cliffs, NJ.
- [27] G. WARNER, *Harmonic Analysis on Semi-Simple Lie Groups I*, Springer-Verlag, Berlin, Heidelberg, New York, 1972.
- [28] *An appraisal of linear analytic systems theory with applications to attitude control*, Report to E.S.T.E.C., Contract 3771/78/NL/AR (SC), by Applied Systems Studies, Coventry, England, May 1980.
- [29] *Application of linear analytic systems theory to attitude control*, Report to E.S.T.E.C., Contract 3771/78/NL/AK (SC), by Applied Systems Studies, Coventry, England, October 1981.

OPTIMALITY OF PIECEWISE-CONSTANT POLICIES IN SEMI-MARKOV DECISION CHAINS*

LAURENT CANTALUPPI†

Abstract. The control of a finite-state semi-Markov process is investigated. In each state, a finite number of actions is available. Each action determines reward rates and transition rates to the other states. These rates depend on the holding time in the state and the actions can be changed at any point in time—not just at transition times. The goal is to find a policy that maximizes the expected total or discounted reward.

In the infinite-horizon case, necessary and sufficient conditions for the optimality of a stationary policy in the class of nonstationary policies is given. A stationary policy is shown to be optimal in that class, and this policy can be chosen piecewise-constant in the holding time in each state if the rates are piecewise-analytic in the holding time. Several applications are examined in the domains of queueing, inventory and reliability.

In the finite-horizon case, necessary and sufficient conditions for optimality are given.

1. Introduction. We study herein the problem of controlling a finite-state semi-Markov process to maximize the infinite-horizon (expected) present value. Finitely many actions are available in each state, and each such action determines uniformly bounded holding-time-dependent reward and transition rates. Because these rates depend on the holding time, higher present values can be achieved if actions are allowed to be holding-time-dependent than would otherwise be so. For this reason, we allow actions to be changed at *any* values of the holding times, not just at transition times. Our goal is to characterize and establish the existence of stationary optimal policies.

Çınlar [4] gives a definitive work on Markov renewal theory. The control problem has previously been studied under the assumption that actions cannot be changed between transitions by Howard [10], Jewell [11], Denardo and Fox [7] and Osaka and Mine [18]. Chitgopekar [3] considers policies that permit actions to change at a single value of the holding time and Stone [20] allows stationary policies that are piecewise-constant (in the holding time).

Stone [20] shows that in the case of a piecewise-constant stationary policy, the process consisting of the state of the system and the holding time in that state is Markovian and he computes its infinitesimal generator. In §§ 3–5 we generalize these results to the case of measurable nonstationary policies and derive several properties of the associated Markov process. In § 6 some applications are examined in the domains of reliability, queueing, production and inventory.

Stone [20] gave necessary and sufficient conditions for optimality of a piecewise-constant stationary policy among that same class of policies. We generalize this result in § 7 to give necessary and sufficient conditions for optimality of a stationary policy among the class of measurable nonstationary policies. Moreover, we establish the existence of a stationary optimal policy in § 8. When the reward and transition rates are piecewise-analytic in the holding times, we also demonstrate the existence of a piecewise-constant stationary optimal policy. We show in [2] how to approximate piecewise-analytic by piecewise-constant rates. In the last case we also show in [2] that an ε -optimal policy can be computed in polynomial time for each fixed instantaneous interest rate. In § 9 we examine the finite-horizon case and give necessary and sufficient conditions for optimality. Our methods are analytic and follow closely the ideas of

* Received by the editors May 5, 1982, and in revised form August 11, 1983.

† Institut für Operations Research, Federal Institute of Technology, Zürich, Switzerland.

Miller [15], [16] and Veinott [21] in the case of finite-state continuous-time-parameter Markov decision chains.

2. Formulation and definitions. Consider a system which is observed continuously in time. At each point in time the system is in one of n states labeled $1, \dots, n$. The system operates from time 0 to time T , where $T \leq \infty$. When the system is in state i with a *holding time* s , i.e., the system has been in state i for s units of time since the last transition into that state, an action is chosen from a finite set $A(i, s)$ of possible *actions*. The cardinality of the sets $A(i, s)$ is uniformly bounded and $A(i, \cdot)$ is piecewise-constant. (A function $L: [0, \infty) \rightarrow B$ is called *piecewise-constant* if there exists an unbounded sequence $0 = t_0 < t_1 < \dots$ and a sequence $\{b_j: j \geq 1\}$ of elements of B such that

$$L(t) = b_j \quad \text{if } t_{j-1} \leq t < t_j$$

for $j = 1, 2, \dots$.) A *reward rate* $r_a(i, s)$ is received that depends on the state i , the action a and the holding time s in state i . The reward rates are uniformly bounded by a constant R and $r_a(i, s)$ is Lebesgue-measurable in s on each interval where $A(i, s)$ is constant, for all $a \in A(i, s)$, $1 \leq i \leq n$ and $s \geq 0$. The evolution of the system from state to state is described by a probability law determined by the *transition rates* $q_a(j|i, s)$ of the system. The transition rates are assumed to have the following properties:

- (1a) there exists a number $Q < \infty$ such that $0 \leq -q_a(i|i, s) \leq Q$ for all $a \in A(i, s)$, $1 \leq i \leq n$, $s \geq 0$;
- (1b) $0 \leq q_a(j|i, s)$ for all $a \in A(i, s)$, $j \neq i$, $1 \leq i, j \leq n$, $s \geq 0$;
- (1c) $\sum_{j=1}^n q_a(j|i, s) = 0$ for all $a \in A(i, s)$, $1 \leq i \leq n$, $s \geq 0$; and
- (1d) $q_a(j|i, s)$ is Lebesgue-measurable in s on each interval where $A(i, s)$ is constant, for all $a \in A(i, s)$, $1 \leq i \leq n$, $s \geq 0$.

The interpretation of the transition rates is the following. Suppose that $q_a(j|i, s)$ is continuous in s . Then if the system is in state i with a holding time s at clock time t , then the probability of being in state j at time $t + \Delta t$ (knowing that action a was used in the interval $[t, t + \Delta t)$ while in state i) is $q_a(j|i, s)\Delta t + o(\Delta t)$, the probability that the system is still in state i at time $t + \Delta t$ is $1 + q_a(i|i, s)\Delta t + o(\Delta t)$.

We now define the decisions available to control the process. Let $N = \{1, \dots, n\}$ be the set of states of the process, $S = N \times [0, \infty)$ be the set of ordered pairs (i, s) representing the state and the holding time of the process and $A(S) = \bigcup_{i,s} A(i, s)$. The decisions taken at each point in time are given by *policies* $\pi: [0, T) \times N \times [0, \infty) \rightarrow A(S)$, i.e., for each point in time t , state i and holding time s , the policy $\pi = (\pi_t)$ specifies the action used at that time. We consider three classes of policies. A policy π is called *measurable* if $\pi_t(i, s)$ is jointly Lebesgue-measurable in s and t . The policy π is called *stationary* if it is independent of the clock time, i.e., $\pi_t = \pi$ for all $0 \leq t < T$. These policies are not stationary with respect to the process representing the state of the system, but they are stationary with respect to the process consisting of the state and the holding time. Since we will work with the latter process rather than the former, this terminology is appropriate. A *piecewise-constant* stationary policy is a stationary policy that is piecewise-constant in the holding time for each state. Using policy π , the probability of being in state i at time $t + u$, knowing that the process is in state i with a holding time s at time u is given by $\exp \left\{ \int_0^t q_{\pi_{u+v}}(i|i, s+v) dv \right\}$.

3. The Markov transition function. We now show that to each policy corresponds a semi-Markov process and an associated reward. Let $\{X_t: 0 \leq t < \infty\}$ be the stochastic process representing the state of the system at time t . The *holding time* in the state of

the process at time $t \geq 0$ is defined by

$$Y_t \equiv t - \sup \{u: 0 \leq u \leq t \text{ and } X_u \neq X_t\}.$$

Let $Z_t \equiv (X_t, Y_t)$ for $t \geq 0$ and Z_t^π be the process Z_t determined by π .

Stone [20] gives a path-by-path construction of the process Z_t^π in the case of a piecewise-constant stationary policy π and computes its infinitesimal generator G_π . We use his results to show the existence of a right-continuous Markov process Z_t^π in the more general case of a measurable policy π using the reverse approach, i.e., we define a generator and show that it determines the process Z_t^π . This analytic approach has the advantage of giving us immediately some properties of the transition function which are used to derive the optimality equation of the process.

THEOREM 1 (Stone). *If π is a piecewise-constant stationary policy, then $\{X_t^\pi: t \geq 0\}$ is a semi-Markov process and $\{Z_t^\pi: t \geq 0\}$ is a strong Markov process.*

Since Z_t^π is a strong Markov process, it is often easier to work with Z_t^π than X_t^π in order to derive properties of the semi-Markov process. Therefore in the sequel when we talk about the *state* of the process, we will generally mean a value of the process Z_t^π , i.e., an ordered pair (i, s) . This should not cause any confusion. In this *augmented state space*, the various policies we are considering are all *Markovian*, i.e., they depend only on the present state of the process and are independent of its past history.

Suppose now that π is a piecewise-constant stationary policy. Then the process $\{Z_t^\pi: t \geq 0\}$ is a two-dimensional right-continuous Markov process with stationary transition probabilities. Thus we may define G_π , the infinitesimal generator of this process in the following manner, see for example Lamperti [13, p. 138]. For $g: S \rightarrow R$ bounded and measurable, define

$$\|g\| \equiv \sup_{(i,s) \in S} |g(i, s)|$$

and

$$(P_\pi^t g)(i, s) \equiv \sum_{j=1}^n \int_0^\infty g(j, r) P_\pi^t(j, dr|i, s),$$

where $P_\pi^t(j, dr|i, s)$ is the transition function from time 0 to time t , starting in state (i, s) and using policy π . The *infinitesimal generator* G_π is defined for bounded measurable g by

$$G_\pi g = \lim_{t \downarrow 0} t^{-1} (P_\pi^t g - g),$$

wherever the limit of the right-hand side exists pointwise and is bounded. Stone [20] has computed this infinitesimal generator and its domain. Let $Dg(i, s) \equiv (d/d\tau)g(i, s + \tau)|_{\tau=0}$.

LEMMA 2 (Stone). *Suppose that π is a piecewise-constant stationary policy. Then the domain of G_π is the set of all $g: S \rightarrow R$ such that $Dg(i, \cdot)$ exists and is bounded for each $i \in N$ and G_π is given by*

$$(G_\pi g)(i, s) = Dg(i, s) + \sum_{j \neq i} q_\pi(j|i, s)g(j, 0) + q_\pi(i|i, s)g(i, s),$$

with $q_\pi(j|i, s) \equiv q_a(j|i, s)$, when $\pi(i, s) = a$.

We now use this result to define a similar operator in the case of the more general measurable policies. We will show that the operator G_π corresponding to a measurable

policy π generates a nonhomogeneous Markov process having the given transition rates in the sense given in (5) below.

For each real-valued function $g_t(i, s)$ of the clock time t and state (i, s) that is jointly measurable in t and s and *diagonally absolutely continuous*, i.e., $g_{t+\tau}(i, s+\tau)$ is absolutely continuous in τ , let $Dg_t(i, s) \equiv (d/d\tau)g_{t+\tau}(i, s+\tau)|_{\tau=0}$ and define the operator G_π by

$$(1) \quad (G_\pi g)_t(i, s) \equiv Dg_t(i, s) + \sum_{j \neq i} q_{\pi_i}(j|i, s)g_t(j, 0) + q_{\pi_i}(i|i, s)g_t(i, s).$$

It should be noted that G_π is an extension to a larger domain of the operator G_π defined in Lemma 2. We now introduce the operator P^u and the measure $P^u(j, dr|i, s)$ related to it via (2). For each bounded measurable function $g: [0, T) \times S \rightarrow R$ the operator P^u defines a function $(P^u g)_t: S \rightarrow R$ by

$$(2) \quad (P^u g)_t(i, s) = \sum_{j=1}^n \int_0^\infty g_u(j, r) P^u(j, dr|i, s).$$

The operator P^u is called a *complete Markov transition function* (Dynkin [9, p. 96]) if:

$$(3a) \quad (P^u g)_t(i, \cdot) \text{ is Lebesgue-measurable;}$$

$$(3b) \quad P^u g \geq 0 \text{ if } g \geq 0;$$

$$(3c) \quad P^u g = 1 \text{ if } g = 1;$$

$$(3d) \quad (P^u g)_t(i, s) = 0 \text{ if } g_t(i, s) = 0; \text{ and}$$

$$(3e) \quad P^u P^{uv} = P^v.$$

The objective of the remainder of this and the following section is to show that to each measurable policy π there exists a unique solution to

$$(4) \quad G_\pi P^u = 0 \quad \text{and} \quad P^{uu} = I$$

that is a complete Markov transition function and to which corresponds a right-continuous process $\{Z_t^\pi: t \geq 0\}$ satisfying

$$(5) \quad \Pr \{X_{t+\Delta t}^\pi = j | Z_t^\pi = (i, s)\} = q_{\pi_i}(j|i, s)\Delta t + o(\Delta t)$$

for $j \neq i$ in N and almost every s and t , i.e., having the desired transition rates.

Existence of a Markov transition function.

THEOREM 3. *If π is a measurable policy, then there is a unique transition function $P^u = P_\pi^u$ satisfying (4) for all $0 \leq t \leq u$. Moreover, P_π^u is Markov and complete.*

Proof. Consider a particular function $g_t(i, s)$ that is measurable and diagonally absolutely continuous, let $u > 0$ and set $F_t(i, s) = (P^u g)_t(i, s)$. Rewrite (4) in this particular case as

$$(6) \quad (G_\pi F)_t(i, s) = 0 \quad \text{and} \quad F_u(i, s) = g_u(i, s).$$

We now show that if we replace the $F_t(i, 0)$ in (6) by parameters $f_i(t)$, then (6) has a unique solution in terms of those parameters. Moreover, we show that the parameters may be uniquely chosen so that $F_t(i, 0) = f_i(t)$ for all $i \in N$ and $0 \leq t \leq u$, whence (6) has a unique solution. To this end, let $\beta_\tau \equiv \sum_{j \neq i} q_{\pi_{i+\tau}}(j|i, s+\tau)f_j(t+\tau)$, $\alpha_\tau \equiv q_{\pi_{i+\tau}}(i|i, s+\tau)$ and $\phi_\tau \equiv F_{t+\tau}(i, s+\tau)$. Then the first equation in (6) becomes

$$\dot{\phi}_\tau + \alpha_\tau \phi_\tau + \beta_\tau = 0,$$

which has the unique solution

$$(7) \quad \phi_\tau = \left(\exp \left(\int_\tau^{u-t} \alpha_w dw \right) \right) \left(K + \int_\tau^{u-t} \beta_w \exp \left(- \int_w^{u-t} \alpha_v dv \right) dw \right)$$

in terms of a parameter K . The initial condition $\phi_{u-t} = g_u(i, s + u - t)$ determines the parameter as

$$(8) \quad K = g_u(i, s + u - t).$$

Equations (7) and (8) express $F_t(i, s)$ as a function of the parameters $f_i(t)$. It remains to show that those parameters may be chosen uniquely to satisfy $F_t(i, 0) = f_i(t)$. To this end, on setting $s = 0$ we have from (7) and (8) with $\tau = 0$ and using the definition of $f_i(t)$ that

$$(9) \quad f_i(t) = g_i(t) + \sum_{j \neq i} \int_0^{u-t} f_j(t+v) Q_{ij}(t, v) dv,$$

where

$$g_i(t) = g_u(i, u-t) \exp \left(\int_0^{u-t} \alpha_v dv \right), \quad \text{and} \quad Q_{ij}(t, v) = q_{\pi_{t+v}}(j|i, v) \exp \left(\int_v^{u-t} \alpha_w dw \right).$$

Let L^∞ be the space of n -tuples $f = (f_1, \dots, f_n)$ of real-valued Lebesgue-measurable functions defined on $[0, u]$ with finite norm $\|f\| \equiv \sup_{(i,t)} |f_i(t)|$. Define the operator $T: L^\infty \rightarrow L^\infty$ so that $(Tf)_i(t)$ equals the right-hand side of (9). From (1c) and (1a) we have

$$\begin{aligned} \sum_{j \neq i} \int_0^{u-t} Q_{ij}(t, v) dv &= - \int_0^{u-t} \alpha_v \exp \left(\int_v^{u-t} \alpha_w dw \right) dv \\ &= \int_0^{u-t} \frac{d}{dv} \exp \left(\int_v^{u-t} \alpha_w dw \right) dv = 1 - \exp \left(\int_0^{u-t} \alpha_v dv \right) \\ &\leq 1 - e^{-Qu} \equiv C < 1. \end{aligned}$$

Therefore we have $\|Tf - Tf'\| \leq C\|f - f'\|$, so T is a contraction. Thus since L^∞ is complete, it follows from the Banach fixed-point theorem that T has a unique fixed point f^* . Thus f^* is the unique solution to (9), establishing that (6) has a unique solution.

It remains to show that P^{tu} defines a complete Markov transition function, i.e., has the properties (3a)–(3e). Observe first that since $F_t(i, \cdot)$ is Lebesgue-measurable, (3a) holds.

Next we establish (3b), or equivalently that $g \geq 0$ implies $F \geq 0$. It is clear that T is a nonnegative operator, i.e., $f \geq 0$ implies $Tf \geq 0$. Therefore $T^n 0 \geq 0$ and since $T^n 0$ converges to f^* , $f^* \geq 0$. Hence by (7) and the nonnegativity of g we have $F \geq 0$.

In order to verify (3c) we must show that $g = 1$ implies $F = 1$. Using the conditions (1a)–(1c) on the transition rates, it is easy to verify that $F = 1$ is a solution of (6) when $g = 1$. Hence by the uniqueness of the solution, (3c) is proved. Condition (3d) is obvious from (7) and (8), which imply that $F_t(i, s) = 0$, whenever $g_t(i, s) = 0$.

It remains to establish (3e). We have for $0 \leq t \leq u \leq v$ that $0 = G_\pi P^{tu}$ and $P^{uu} = I$, and $0 = G_\pi P^{uv}$ and $P^{vv} = I$. Let $\bar{P}^{tv} \equiv P^{tu} P^{uv}$. We have to show that $\bar{P}^{tv} = P^{tv}$. Now $G_\pi \bar{P}^{tv} = G_\pi P^{tu} P^{uv} = 0$ for all $0 \leq t \leq u \leq v$ and $\bar{P}^{vv} = P^{vv} P^{vv} = I$. Hence by the uniqueness of the solution to (4) we have $\bar{P}^{tv} = P^{tv}$ and (3e) is proved, completing the proof.

4. The Markov process. We know that P_π^{iu} defines a complete Markov transition function and the existence of a corresponding Markov process $\{Z_t^\pi: t \geq 0\}$ is a direct consequence of Dynkin [9, Thm. 4.1, p. 99]. We will show that $\{Z_t^\pi: t \geq 0\}$ has right-continuous paths with left limits and that it satisfies condition (5). To do so we need the following result from Natanson [17, p. 255].

LEMMA 4. *If f is integrable on $[a, b]$, then*

$$\int_x^{x+h} |f(t) - f(x)| dt = o(h) \quad \text{for almost every } x \in (a, b).$$

LEMMA 5. *If g is measurable and diagonally absolutely continuous, and π is a measurable policy, then for almost every t ,*

$$(10) \quad P_\pi^{iu} g = g + (G_\pi g)(u - t) + o(u - t).$$

Proof. Using the notation of Theorem 3 we have

$$\begin{aligned} (P_\pi^{iu} g)_t(i, s) &= \phi_0 = \phi_{u-t} + \int_0^{u-t} (\alpha_\tau \phi_\tau + \beta_\tau) d\tau \\ &= \phi_{u-t} + \int_0^{u-t} [(\alpha_\tau - \alpha_0) \phi_\tau + (\beta_\tau - \beta_0)] d\tau + \int_0^{u-t} (\alpha_0 \phi_\tau + \beta_0) d\tau. \end{aligned}$$

Applying Lemma 4 with $f(t) = \alpha_t \phi_t + \beta_t$ the first integral is $o(u - t)$ for almost every t and

$$\int_0^{u-t} (\alpha_0 \phi_\tau + \beta_0) d\tau = (\alpha_0 \phi_0 + \beta_0)(u - t) + o(u - t),$$

since ϕ_τ is absolutely continuous in τ . Moreover, since g is measurable and diagonally absolutely continuous we have

$$\phi_{u-t} = g_u(i, s + u - t) = g_t(i, s) + (Dg_t(i, s))(u - t) + o(u - t);$$

and this together with Theorem 3 proves the lemma.

The following result shows that for a suitable class of functions g , the Markov transition function P_π^{iu} satisfies the forward equations.

THEOREM 6. *If g is measurable and diagonally absolutely continuous, and π is a measurable policy, then for almost every u ,*

$$\frac{d}{du} P_\pi^{iu} g = P_\pi^{iu} G_\pi g.$$

Proof. From Lemma 5 we have for almost every u ,

$$P_\pi^{u, u+\Delta u} g - g = (G_\pi g) \Delta u + o(\Delta u).$$

Multiplying on the left by the operator P_π^{iu} and dividing by Δu we get

$$\frac{P_\pi^{i, u+\Delta u} g - P_\pi^{iu} g}{\Delta u} = P_\pi^{iu} G_\pi g + \frac{o(\Delta u)}{\Delta u}.$$

Taking the limit as Δu goes to zero gives the theorem.

We now show that the process $\{Z_t^\pi: t \geq 0\}$ satisfies (5).

THEOREM 7. *If π is a measurable policy, then*

$$\Pr \{X_{t+\Delta t}^\pi = j | Z_t^\pi = (i, s)\} = q_{\pi_i}(j|i, s) \Delta t + o(\Delta t)$$

for all $j \neq i$ and almost every s and t , and

$$\Pr \{X_{t+\Delta t}^\pi = i | Z_t^\pi = (i, s)\} = 1 + q_{\pi_i}(i|i, s) \Delta t + o(\Delta t)$$

for almost every s and t .

Proof. In Lemma 5 let $g_{(\cdot)}(j, \cdot) = 1$ and $g = 0$ elsewhere. Then (10) gives the first assertion of Theorem 7 which is exactly (5). The second assertion follows from the first and (1c).

COROLLARY 8. If π is a measurable policy, then

$$\Pr \{Z_{t+u}^\pi = (i, s+t) | Z_u^\pi = (i, s)\} = \exp \left\{ \int_0^t q_{\pi_{u+v}}(i|i, s+v) dv \right\} \geq e^{-Q_t}.$$

Proof. Let $h(t)$ be the left-hand side of the above equality. On letting $g_{u+\tau}(i, s+\tau) = 1$ for $\tau \geq 0$ and $g_{(\cdot)}(\cdot, \cdot) = 0$ otherwise, we have that $h(t) = (P_\pi^{u, u+t} g)_u(i, s)$, so, from Theorem 6,

$$\dot{h}(t) = q_{\pi_{u+t}}(i|i, s+t)h(t)$$

for almost every t . Solving this differential equation with the initial condition $h(0) = 1$ gives

$$h(t) = \exp \left\{ \int_0^t q_{\pi_{u+v}}(i|i, s+v) dv \right\} \geq e^{-Q_t},$$

the inequality being obvious from (1a).

We now show that the process $\{Z_t^\pi: t \geq 0\}$ has paths which are right-continuous and have left limits. Since the probability of a jump during $[t, t+u)$ starting in state (i, s) at time t is bounded by $Qu + o(u)$ uniformly in i, s and t , we can apply Dynkin [9, Thm. 6.3, p. 150] to obtain the result that $\{Z_t^\pi: t \geq 0\}$ has right-continuous paths with left limits. Therefore the process $\{X_t^\pi: t \geq 0\}$ has sample paths which are step-functions. We can take the sample space Ω as the set of all functions $\omega: [0, T) \rightarrow N \times [0, \infty)$ such that the restriction of ω to its first component is a right-continuous step function and the second component is the holding time in the corresponding state. Let Γ be the σ -algebra of sets in the space Ω generated by the sets

$$\{\omega \in \Omega: \omega(t) \in (i, B) \text{ for all } t \in [0, T), i \in N, \text{ Borel sets } B \text{ of } R\}.$$

The process $\{Z_t^\pi: t \geq 0\}$ is defined on (Ω, Γ) .

5. The objective function. It remains to define the *objective function* of the semi-Markov decision chain. In the case of a finite-horizon T , the *value* V_π of the measurable policy π is defined by

$$V_\pi(i, s) \equiv E_{is} \left\{ \int_0^T r_{\pi_t}(Z_t^\pi) dt \right\},$$

where E_{is} is the expectation operator starting in state (i, s) . In the case of an infinite horizon, i.e., $T = \infty$, we discount rewards earned with the instantaneous interest rate $\rho > 0$, i.e., a reward earned at time t is discounted by a factor $e^{-\rho t}$. The *present value* V_π of the policy π is therefore defined by

$$V_\pi(i, s) \equiv E_{is} \left\{ \int_0^\infty e^{-\rho t} r_{\pi_t}(Z_t^\pi) dt \right\}.$$

Our objective is to find a policy, called *optimal*, with maximum value if $T < \infty$ and maximum present value if $T = \infty$.

It is desirable to interchange the integral and the expectation. In order to do so we must show that $r_{\pi_t}(Z_t^\pi)$ satisfies the hypotheses of Fubini's theorem. We first need the measurability of $r_{\pi_t}(Z_t^\pi)$ with respect to $(\Gamma \times \partial)$, where ∂ represents the Lebesgue-measurable sets of $[0, T]$. It follows easily from the definition (Chung [5, p. 143]) that the sample paths of the process are separable with respect to any dense set. This implies (Chung [5, p. 143]) that the process is well separable. Since each sample path has only a finite number of discontinuities, the stochastic process is continuous almost everywhere almost surely. These properties imply (Doob [8, p. 60]) that the process Z_t^π is measurable with respect to $(\Gamma \times \partial)$. But then $r_{\pi_t}(Z_t^\pi)$ is measurable with respect to $(\Gamma \times \partial)$ since for each action $a \in A(i, s)$, state i and Borel subset B of the real line, the set

$$\{(\omega, t): Z_t^\pi(\omega) \in (i, B) \text{ and } \pi_t(Z_t^\pi(\omega)) = a\}$$

is $(\Gamma \times \partial)$ -measurable.

Since the reward rates are uniformly bounded by R , we can apply Fubini's theorem and get the objective function

$$V_\pi = \int_0^T P_\pi^t r_{\pi_t} dt$$

in the finite-horizon case and

$$V_\pi = \int_0^\infty e^{-\rho t} P_\pi^t r_{\pi_t} dt$$

in the infinite-horizon case, where $P_\pi^t \equiv P_\pi^{0t}$.

6. Applications. Semi-Markov decision chains offer an interesting generalization of Markov decision chains. One of the drawbacks of using a Markov decision chain to model a real process (economic, physical, biological, etc.) is that the transition rates have to be constant, i.e., the distributions of the holding times in each state are exponential. In many cases it is much more realistic to consider other types of distributions. In reliability, for example, one can face "increasing failure rate" distributions or "bathtub shaped" distributions. A semi-Markov decision chain allows one to consider such possibilities.

Even when the holding-time distributions are exponential, semi-Markov decision chains allow consideration of reward rates that depend on the holding time in the state of the system. For example, this may occur where the output of a machine diminishes with its age. We now examine several applications of semi-Markov decision chains.

Control of an M/G/1 queueing system. Consider an M/G/1 queueing system with a finite waiting room of n customers (including the one being served). The queue length is observed only at departure times. The state of the process is the number of customers left in the queue just after the last departure time. A customer who finds a full waiting room does not wait. The actions are the available service rates of the system to which correspond different cost rates. There is also a penalty rate depending on the state and the holding time in that state, i.e., on the expected number of arrivals since the last observation of the queue length. The goal is to select state and holding-time dependent service rates to minimize expected costs. This model has previously been studied only in the case where the service rate cannot be changed between departures, see for example Lippman [14].

Maintenance of a deteriorating system. Consider a system whose performance can be represented by classifying it in one of n states labeled $1, \dots, n$. State 1 represents

a new system and state n represents a system that has failed. The intermediate states $2, \dots, n-1$ represent different degrees of deterioration. The system is observed continuously and the transitions always occur from one state to a higher labeled state, reflecting the increasing deterioration of the system. The system is controlled by selecting an output rate depending on its condition, i.e., its state of deterioration, and the sojourn time in that condition. The output rate determines reward rates and transition rates.

At any point in time one can decide to repair the system. After a random time, reflecting the availability of a repair shop, the repair is started. The repair time is random and depends on the condition of the system. When the system is repaired, it is restored to an as-new condition, i.e., the system is in state 1. The goal is to maximize the expected return of the system by choosing the appropriate output rates and repair schedules. This application is a modification of a replacement model proposed by Kao [12].

Maintenance of a two-unit system. Consider a system of two identical units, one in standby and the other operating. Two decisions are available for the operating unit, viz., do nothing or have the unit serviced, the second being available only if the standby unit is not being serviced or repaired, since otherwise the system would go down. After a random time the operating unit fails or is serviced. The standby unit is instantaneously put into operation. We assume that a unit recovers its function perfectly after repair or service. When a unit completes service or repair it is immediately put into operation and the operating unit is serviced. We therefore have the four states:

- (1) one unit begins to be operative and the other is in standby;
- (2) one unit begins to be operative in place of the other failed unit and the failed unit undergoes repair;
- (3) one unit begins to be operative while service of the other unit begins;
- (4) the operating unit fails while the standby unit is being serviced or repaired, bringing the system down.

The goal is to maximize the expected return to the system by scheduling service at appropriate values of the holding time in state 1. This model is described and studied in a special case by Asakura and Shunji [1].

Production and inventory. We have a product in inventory which can be produced at finitely many different rates. The demand for this product is random and the size of an order is correlated with the length of time since the last order. This is so where there is only one customer who buys the product at a constant average rate and thus the longer the time between two consecutive orders, the higher the probability that the second order is large. Orders have to be delivered instantaneously. The state of the system is the number of items left after an order is filled. There is a carrying cost rate when some items are left on hand after an order is filled and there is a shortage cost rate when there are not enough units on hand to fill an order. Backorders are lost. The goal is to minimize the expected production and inventory cost.

7. Infinite-horizon: optimality conditions. In this section we consider the case where the horizon is infinite and the criterion to be maximized is the present value. The optimal policy obviously depends on the instantaneous interest rate ρ , but this dependence will be suppressed in the sequel. If π is a policy, denote by π' its restriction to the interval $[0, t)$ and by ${}^t\pi$ the policy π truncated and shifted to the left by t , i.e., ${}^t\pi$ is the policy π' defined by $\pi'_u = \pi_{t+u}$ for $u \geq 0$. If π and π^* are policies, let $\pi' \pi^*$ denote the policy π' defined by $\pi'_u = \pi_u$ for $0 \leq u < t$ and $\pi'_u = \pi^*_u$ for $u \geq t$, i.e., π replaces π^* during the first t units of time. Let $G_{\pi_t} V_t(i, s) \equiv G_u V_t(i, s) \equiv (G_{\pi} V)_t(i, s)$,

if $\pi_t(i, s) = a$. We have the following comparison lemma which is similar to Veinott [21, Lemma 11] in the case of continuous-time Markov decision chains.

LEMMA 9 (Comparison lemma). *If π and π^* are measurable policies, then*

$$V_\pi - V_{\pi^*} = \int_0^\infty e^{-\rho t} P_\pi^t C_{\pi_t \pi^*}^t dt,$$

where $C_{\pi_t \pi^*}^t \equiv r_{\pi_t} + (G_{\pi_t} - \rho I) V_{t_{\pi^*}}$.

Proof. We have for $t \geq 0$,

$$(11) \quad V_{\pi^*} = \int_0^t e^{-\rho u} P_{\pi^*}^u r_{\pi_u} du + e^{-\rho t} P_{\pi^*}^t V_{t_{\pi^*}}.$$

Since $V_{t_{\pi^*}} = \int_t^\infty e^{-\rho(u-t)} P_{\pi^*}^{u-t} r_{\pi_u} du$, the function $V_{t_{\pi^*}}(i, s)$ is measurable and diagonally absolutely continuous by Theorem 3, and we can use Theorem 6 to differentiate (11) with respect to t yielding

$$(12) \quad \frac{d}{dt} V_{\pi^*} = e^{-\rho t} P_{\pi^*}^t C_{\pi_t \pi^*}^t$$

for almost every $t \geq 0$. Integrating over the nonnegative halfline completes the proof.

We now characterize V_γ , the present value of the stationary measurable policy γ . Equation (13) below is the differential version of the renewal equation of Çinlar [4].

THEOREM 10. *The present value V_γ of the measurable stationary policy γ is the unique bounded absolutely continuous solution V to*

$$(13) \quad r_\gamma + (G_\gamma - \rho I) V = 0.$$

Proof. On replacing π and π^* by γ in (12), we get

$$P_\gamma^t C_{\gamma\gamma} = 0 \quad \text{almost everywhere } t \geq 0,$$

where $C_{\gamma\gamma} \equiv C_{\gamma\gamma}^t$. On letting $t \downarrow 0$ and using Lemma 5 we have $C_{\gamma\gamma} = 0$ and thus, V_γ satisfies (13). We now show that (13) has a unique bounded solution. Let $q(t) = \exp[\int_0^t (q_\gamma(i|i, u) - \rho) du]$ and $V_i = V(i, 0)$. In a manner similar to the proof of Theorem 3, we will solve (13) for $V(i, s)$ as a function of V_j , $1 \leq j \leq n$. We then show that there exists a unique n -tuple $V^0 = (V_1, \dots, V_n)$ such that the solution to (13) is bounded. On integrating (13) we have

$$V(i, s) = \frac{1}{q(s)} \left(V_i - \int_0^s \left[r_\gamma(i, t) + \sum_{j \neq i} V_j q_\gamma(j|i, t) \right] q(t) dt \right).$$

Since $\lim_{s \rightarrow \infty} q(s) = 0$, for $V(i, s)$ to remain bounded as s goes to ∞ , we must have

$$(14) \quad V_i = \int_0^\infty r_\gamma(i, t) q(t) dt + \sum_{j \neq i} V_j \int_0^\infty q_\gamma(j|i, t) q(t) dt.$$

On setting

$$r_i^0 \equiv \int_0^\infty r_\gamma(i, t) q(t) dt, \quad r^0 \equiv (r_i^0), \quad p_{ij} \equiv \int_0^\infty q_\gamma(j|i, t) q(t) dt$$

for $j \neq i$, $p_{ii} \equiv 0$, and $P = (p_{ij})$, (14) can be rewritten as $V^0 = r^0 + P V^0$. By (1b), $p_{ij} \geq 0$.

Also by (1c),

$$\begin{aligned}\sum_{j \neq i} p_{ij} &= - \int_0^\infty q_\gamma(i|i, t) q(t) dt < - \int_0^\infty [q_\gamma(i|i, t) - \rho] q(t) dt \\ &= - \int_0^\infty \frac{d}{dt} q(t) dt = 1.\end{aligned}$$

Thus P is the transition matrix of a transient Markov chain, so (14) has a unique solution, proving the theorem.

It should be noted that (14) is the system whose solution gives the present value of γ in the embedded Markov chain associated with the semi-Markov chain X_t^γ , where p_{ij} is the probability of a transition from state i to state j and r_i^0 is the reward earned in state i .

The following result gives necessary and sufficient conditions for optimality of a stationary policy. It was first proved by Stone [20] in a different manner for the class of piecewise-constant policies.

THEOREM 11. *The stationary measurable policy γ is optimal if and only if*

$$(15) \quad 0 = \max_{a \in A(i, s)} [r_a + (G_a - \rho I) V_\gamma](i, s)$$

almost everywhere.

Proof sufficiency. First note that since γ is stationary, $V_{t_\gamma} = V_\gamma$ and so $C_{\gamma\gamma}^t = C_{\gamma\gamma}$, say, for all $t \geq 0$. Suppose $\gamma(i, s)$ maximizes the right-hand side of (15) for almost every s . Then from Lemma 9 we have $V_\pi - V_\gamma \leq 0$ for every measurable policy π and therefore γ is optimal.

Necessity. Suppose $\gamma(i, s)$ does not maximize the right-hand side of (15) almost everywhere. Let δ be a stationary measurable policy such that $a = \delta(i, s)$ maximizes the right-hand side of (15) almost everywhere. The existence of δ is guaranteed since the expression is measurable and the ties can be broken by an arbitrary enumeration of the actions. But we know from Theorem 3 that $P_\delta^t C_{\delta\gamma} \geq 0$ if $C_{\delta\gamma} \geq 0$. Moreover, for some $i \in N$ and subset B of the nonnegative halfline with positive Lebesgue measure, $C_{\delta\gamma}(i, \cdot)$ is positive on B . Also from Corollary 8 we have

$$\Pr \{Z_t^\delta = (i, t) | Z_0^\delta = (i, 0)\} \geq e^{-\rho t},$$

so $P_\delta^t(i, t | i, 0) > 0$. Thus for each $t \in B$,

$$(P_\delta^t C_{\delta\gamma})(i, 0) \geq P_\delta^t(i, t | i, 0) C_{\delta\gamma}(i, t) > 0.$$

Hence by Lemma 9, $V_\delta(i, 0) > V_\gamma(i, 0)$ and γ is not optimal. The equality in (15) is a direct consequence of Theorem 10, completing the proof.

8. Infinite horizon: existence of stationary measurable and piecewise-constant optimal policies. We now show the existence of an optimal stationary policy γ^* whose present value V_{γ^*} is the unique solution to (15). To do so we look at the embedded Markov decision chain in which the system is observed only at times when the state changes. In this framework the state space $N = \{1, \dots, n\}$ is kept finite but the action space becomes infinite since a decision in state i must specify the action taken for each value of the holding time in state i . To avoid confusion, we will keep the terminology defined earlier. On this embedded Markov decision chain we define the *optimal return operator* $\mathbf{R}: R^n \rightarrow R^n$ in the following way. For $V \in R^n$, let $(\mathbf{R}V)_i = V(i, 0)$, where

$V(i, 0)$ is the value at $s = 0$ of the unique bounded absolutely continuous solution to

$$(16) \quad -\frac{dV(i, s)}{ds} = \max_{a \in A(i, s)} \left\{ r_a(i, s) + \sum_{j \neq i} q_a(j|i, s) V_j + (q_a(i|i, s) - \rho) V(i, s) \right\}.$$

We can interpret $(\mathbf{R}V)_i$ as the maximum present value starting in state $(i, 0)$ if the terminal reward when the process jumps to state $j \neq i$ is V_j . We need the following result.

LEMMA 12. *For each i and $V = (V_j)$, the differential equation (16) has a unique bounded absolutely continuous solution.*

Proof. Consider the following single-state semi-Markov decision chain. The action space in this state i at holding time s is $A(i, s)$; the reward rates are

$$r_a(s) \equiv r_a(i, s) + \sum_{j \neq i} q_a(j|i, s) V_j;$$

and the transition rates to an external “stopped state” are $q_a(s) \equiv q_a(i|i, s)$. Once in the stopped state the process stays there and earns no rewards. From Theorem 11 the optimality equation for this problem is

$$(17) \quad -\dot{V}(s) = \max_{a \in A(i, s)} \{r_a(s) + (q_a(s) - \rho) V(s)\},$$

which is the same equation as (16).

Let $V_\gamma(s)$ be the present value of policy γ starting in state i with a holding time s . Let $V_*(s) = \sup_{\gamma \in \Delta} V_\gamma(s)$ and $s_0 > 0$. Since (17) satisfies a Lipschitz condition, it has a unique solution for each initial value $V(s_0)$. Let \bar{V} be that solution with $\bar{V}(s_0) = V_*(s_0)$. There exists δ such that $V_\delta(s_0) > V_*(s_0) - \varepsilon$. We solve (17) on $0 \leq s \leq s_0$ with the end condition $V(s_0) = V_\delta(s_0)$. Let γ be a policy maximizing the right-hand side of (17) for $0 \leq s \leq s_0$ and equaling δ for $s_0 \leq s$. Since (17) satisfies a Lipschitz condition with Lipschitz factor $Q + \rho$ we have (see, for example, Coddington and Levinson [6, Thm. 2.1, p. 8])

$$(18) \quad V_\gamma(0) > \bar{V}(0) - \varepsilon e^{(Q+\rho)s_0}.$$

Since for every $\varepsilon > 0$ there exists γ such that (18) is satisfied, we have $V_*(0) \geq \bar{V}(0)$. Suppose that $V_*(0) > \bar{V}(0)$. Then there exists θ such that $V_\theta(0) > \bar{V}(0)$. Since from Theorem 10, V_θ satisfies

$$\begin{aligned} -\frac{dV_\theta(s)}{ds} &= r_\theta(s) + (q_\theta(s) - \rho) V_\theta(s) \\ &\leq \max_{a \in A(i, s)} [r_a(s) + (q_a(s) - \rho) V_\theta(s)], \end{aligned}$$

we have that $V_\theta(s_0) > \bar{V}(s_0) = V_*(s_0)$, which is a contradiction. Therefore $\bar{V}(0) = V_*(0)$. By the uniqueness of the solution to (17) with a given initial condition, we also have that $\bar{V}(s) = V_*(s)$ for $0 \leq s \leq s_0$. Since s_0 is arbitrary, $V_*(s)$ satisfies (17) almost everywhere. Also V_* is bounded. Moreover the argmax of (17) determines a measurable policy γ^* and by Theorem 10 V_{γ^*} satisfies (17) and therefore is optimal by Theorem 11.

We next establish the uniqueness of a bounded solution to (17). By Theorem 10 such a solution is the present value of a policy and by Theorem 11 that policy is optimal. Hence, since the optimal return is unique, the claim follows, completing the proof.

LEMMA 13. *The operator \mathbf{R} is an isotone contraction with modulus $Q/(Q + \rho)$.*

Proof. The isotonicity of \mathbf{R} follows from the fact that increasing the terminal reward vector V increases $\mathbf{R}V$. It remains to show that \mathbf{R} is a contraction. To this end let U and V be any vectors in R^n . We must show that

$$\|\mathbf{R}V - \mathbf{R}U\| \leq \frac{Q}{Q + \rho} \|V - U\|,$$

where $\|V\| = \max_{1 \leq i \leq n} |V_i|$. Since the roles of U and V can be interchanged, it is enough to show that

$$\mathbf{R}V - \mathbf{R}U \leq \frac{Q}{Q + \rho} \|V - U\|.$$

Let γ be the policy defined by (16), R_γ the total reward earned in state i before a transition occurs, T_γ the time of the first transition and S_γ the state the process jumps to at the first transition. We have

$$(\mathbf{R}V)_i = E[R_\gamma + e^{-\rho T_\gamma} V_{S_\gamma}] \quad \text{and} \quad (\mathbf{R}U)_i \geq E[R_\gamma + e^{-\rho T_\gamma} U_{S_\gamma}],$$

which gives

$$\mathbf{R}V - \mathbf{R}U \leq E[e^{-\rho T_\gamma} (V_{S_\gamma} - U_{S_\gamma})] \leq \frac{Q}{Q + \rho} \|V - U\|,$$

since $E[e^{-\rho T_\gamma}] \leq \int_0^\infty e^{-\rho t} Q e^{-Q t} dt = Q/(Q + \rho)$, completing the proof.

It follows from Lemma 12 that each $V \in R^n$ has a unique bounded absolutely continuous extension $V(i, s)$ satisfying (16) for each $i \in N$ and almost everywhere.

Existence of an optimal policy.

THEOREM 14. *The operator \mathbf{R} has a unique fixed point $V_* \in R^n$ and its extension $V_*(\cdot, \cdot)$ to S is the unique bounded absolutely continuous function $V(\cdot, \cdot)$ satisfying*

$$0 = \max_{a \in A(i, s)} [r_a + (G_a - \rho I) V](i, s)$$

almost everywhere. Moreover, there is a measurable δ such that $a = \delta(i, s)$ achieves the above maximum almost everywhere, δ is optimal and $V_\delta = V_$.*

Proof. It is well known that since \mathbf{R} is a contraction mapping, it has a unique fixed point V_* say. The vector V_* has a unique extension $V_*(\cdot, \cdot)$ and there is a stationary policy γ^* such that V_{γ^*} satisfies (16). Therefore γ^* is an optimal policy for the semi-Markov decision chain and V_* is its optimal present value.

Optimality of piecewise-constant policies. We have shown the existence of an optimal policy γ^* and therefore the uniqueness of a solution to (15). We now give sufficient conditions for the optimality of a piecewise-constant stationary policy. This result was conjectured by Veinott [22].

Let I be an interval of the real line. We say that the function $L: I \rightarrow R$ is *analytic* at \bar{s} if it has an absolutely convergent power series expansion $L(s) = \sum_{j=1}^\infty a_j s^j$ in some neighborhood of \bar{s} . The function $L(s)$ is analytic on the interval $I_1 \subset I$ if there exists an open interval $I_2 \supset I_1$ such that $L(s)$ is analytic at each $s \in I_2$. The function $L(s)$ is *piecewise-analytic* on I if there exists an unbounded sequence $\{t_i\}$ such that L is analytic on $I \cap (t_i, t_{i+1})$ for all $i \geq 0$. A function $L: S \rightarrow R$ is piecewise-analytic if $L(i, \cdot)$ is piecewise-analytic for all $i \in N$. The following proof is similar to that of Pliska [19] in the case of controlled diffusions.

THEOREM 15. *If the reward rates $r_a(i, s)$ and the transition rates $q_a(j|i, s)$ are piecewise-analytic in $s \geq 0$ for all $a \in A(i, s)$ and $i \in N$, then there exists a piecewise-constant stationary optimal policy.*

Proof. The optimality equation is

$$-\frac{dV(i, s)}{ds} = \max_{a \in A(i, s)} \left\{ r_a(i, s) + \sum_{j \neq i} q_a(j|i, s) V(j, 0) + (q_a(i|i, s) - \rho) V(i, s) \right\}.$$

To prove the existence of a piecewise-constant stationary optimal policy it is sufficient to show that for each $(i, \bar{s}) \in S$ there exists an $\varepsilon > 0$ such that

(19a) an action a_+ is optimal on $\{(i, s) : s \in (\bar{s}, \bar{s} + \varepsilon)\}$

and

(19b) an action a_- is optimal on $\{(i, s) : s \in (\bar{s} - \varepsilon, \bar{s})\}$,

since from (19a) we can find for each state i an increasing sequence of holding times (t_j) and a sequence of actions (a_j) such that the action a_j is optimal in state i on $[t_j, t_{j+1})$, and from (19b) the sequence (t_j) does not accumulate.

We discuss only (19a), the proof of (19b) being similar. Let $V_*(i, s)$ be the optimal present value starting in state (i, s) . For each action $a \in A(i, s)$ let $V_a(i, s)$ be the unique solution to

$$(20) \quad -DV(i, s) = R_a(i, s) + Q_a(i, s) V(i, s)$$

satisfying the initial condition $V(i, \bar{s}) = V_*(i, \bar{s})$, where

$$R_a(i, s) \equiv r_a(i, s) + \sum_{j \neq i} q_a(j|i, s) V_*(j, 0)$$

and

$$Q_a(i, s) = q_a(i|i, s) - \rho.$$

From differential-equation theory, $V_a(i, s)$ is analytic at \bar{s} whenever the coefficients of (20) are analytic at \bar{s} (see for example Coddington and Levinson [6, p. 90]) and thus is piecewise-analytic. Therefore $V_a(i, s)$ is analytic on $(\bar{s}, \bar{s} + \varepsilon)$ for some $\varepsilon > 0$. Hence, for some $\varepsilon > 0$, there exists an action $a_+ \in A(i, s)$ such that

$$(21) \quad V_{a_+}(i, s) \geq V_a(i, s)$$

for all $a \in A(i, \bar{s})$ and all $s \in (\bar{s}, \bar{s} + \varepsilon)$. We now show that the action a_+ is optimal on $\{(i, s) : s \in (\bar{s}, \bar{s} + \varepsilon)\}$. For each $a \in A(i, \bar{s})$ let

$$\Psi_a(i, s) = R_a(i, s) - R_{a_+}(i, s) + [Q_a(i, s) - Q_{a_+}(i, s)] V_{a_+}(i, s)$$

and note that $\Psi_a(i, s)$ is analytic in s on $(\bar{s}, \bar{s} + \varepsilon)$. If we let $\tilde{V}_a = V_a - V_{a_+}$, then \tilde{V}_a is the unique absolutely continuous solution to

$$(22) \quad -D\tilde{V}_a(i, s) = \Psi_a(i, s) + Q_{a_+}(i, s) \tilde{V}_a(i, s)$$

satisfying the initial condition $\tilde{V}_a(i, \bar{s}) = 0$.

Since from (21) $\tilde{V}_a(i, s) \leq 0$ for $s \in (\bar{s}, \bar{s} + \varepsilon)$, we must have $-D\tilde{V}_a(i, s) \geq 0$ for $s \in (\bar{s}, \bar{s} + \varepsilon)$ for some $\varepsilon > 0$ and thus from (22) we have $\Psi_a(i, s) \geq 0$ for $s \in (\bar{s}, \bar{s} + \varepsilon)$ for some $\varepsilon > 0$. Therefore

$$\begin{aligned} -DV_{a_+}(i, s) &= R_{a_+}(i, s) + Q_{a_+}(i, s) V_{a_+}(i, s) - \Psi_{a_+}(i, s) \\ &= \max_{a \in A(i, \bar{s})} \{R_a(i, s) + Q_a(i, s) V_{a_+}(i, s)\} \end{aligned}$$

for $s \in (\bar{s}, \bar{s} + \varepsilon)$ for some $\varepsilon > 0$. Because of the uniqueness of the solution to this equation, we must have that $V_*(i, s) = V_{a_+}(i, s)$ for $s \in (\bar{s}, \bar{s} + \varepsilon)$ and therefore a_+ is optimal for these values of s and the theorem is proved.

In the case of a Markov decision chain with a finite number of states and actions and infinite planning horizon, Miller [15] shows that a policy that is independent of the holding time is optimal. In the semi-Markov case it is easy to give an example where this is not so.

Example. Consider a two-state semi-Markov decision chain with states 1, 2. There are three actions a, b, c in state 1 with transition and reward rates given below:

$$\text{action a. } r_a(1, s) = \begin{cases} 2.41 & \text{for } s < 5, \\ 1 & \text{for } s \geq 5, \end{cases}$$

$$q_a(2|1, s) = \begin{cases} .05 & \text{for } s < 5, \\ .1 & \text{for } s \geq 5; \end{cases}$$

$$\text{action b. } r_b(1, s) = \begin{cases} 1.47 & \text{for } s < 5, \\ -1 & \text{for } s \geq 5, \end{cases}$$

$$q_b(2|1, s) = \begin{cases} .1 & \text{for } s < 5, \\ .15 & \text{for } s \geq 5; \end{cases}$$

$$\text{action c. } r_c(1, s) = \begin{cases} .5 & \text{for } s < 5, \\ 2.5 & \text{for } s \geq 5, \end{cases}$$

$$q_c(2|1, s) = \begin{cases} .15 & \text{for } s < 5, \\ 0 & \text{for } s \geq 5. \end{cases}$$

In state 2 we have a single action a with $r_a(2, s) = 5$ for all $s \geq 0$ and $q_a(2|1, s) = 0$ for all $s \geq 0$. The interest rate ρ is .1.

It is easy to verify that the optimal policy in state 1 uses:

- action 1 for $0 \leq s < 1.47$,
- action 2 for $1.47 \leq s < 3.57$,
- action 3 for $3.57 \leq s < 5$,
- action 1 for $5 \leq s$.

9. Finite horizon: optimality conditions. In this section we consider the finite-horizon problem in which the system operates from time 0 to time T , where $T < \infty$. The total T -period expected reward using policy π is given by

$$V_\pi = \int_0^T P_\pi^t r_{\pi_t} dt.$$

Moreover we define

$$(23) \quad V_\pi^t = \int_t^T P_\pi^{tu} r_{\pi_u} du,$$

i.e., the expected reward earned under policy π during the interval $[t, T)$. The following lemma is a direct consequence of Theorem 3.

LEMMA 16. *For each measurable policy π , $V^t = V_\pi^t$ is the unique measurable diagonally absolutely continuous solution to*

$$(24) \quad r_{\pi_t} + G_{\pi_t} V^t = 0$$

satisfying the end condition $V^T = 0$.

Proof. Using Theorem 3 we differentiate (23) and get (24), proving the existence part of the theorem. Let V_1^t and V_2^t be solutions to (24) and $\bar{V}^t = V_1^t - V_2^t$. Then \bar{V}^t

satisfies $G_{\pi_t} \bar{V}^t = 0$ and $\bar{V}^T = 0$, which is exactly (6). Therefore, from Theorem 3, $\bar{V}^t = 0$ is its unique solution, proving the uniqueness of a solution to (24) and $V^T = 0$.

The following result is a generalization of a theorem of Miller [16].

LEMMA 17 (comparison lemma). *If π and π' are measurable policies, then*

$$(25) \quad V_{\pi} - V_{\pi'} = \int_0^T P_{\pi}^t C_{\pi, \pi'}^t dt,$$

where $C_{\pi, \pi'}^t \equiv r_{\pi_t} + G_{\pi_t} V_{\pi'}^t$.

Proof. We have

$$V_{\pi'}^t = \int_0^t P_{\pi'}^u r_{\pi_u} du + P_{\pi'}^t V_{\pi'}^T.$$

Differentiating with respect to t and using Theorem 6 gives

$$\frac{d}{dt} V_{\pi'}^t = P_{\pi'}^t C_{\pi, \pi'}^t$$

almost everywhere. Integrating over the interval $[0, T]$ completes the proof.

Necessary and sufficient conditions for optimality.

THEOREM 18. *A measurable policy π is optimal for the T -period problem if and only if*

$$(26) \quad 0 = \max_{a \in A(i, s)} [r_a + G_a V_{\pi}^t](i, s)$$

diagonally almost everywhere.

Proof. From (25) it is obvious that if π maximizes (26), then $V_{\pi} - V_{\pi'} \geq 0$ for every measurable policy π' and therefore π is optimal. This proves the sufficiency of the condition.

Now suppose that there exists i such that π does not maximize (26) diagonally almost everywhere, i.e., $a = \pi_t(i, s)$ does not maximize $r_a + G_a V_{\pi}^t(i, s + t)$ for almost every t . Let π' be a policy that maximizes this expression almost everywhere; we can choose π' measurable by breaking the ties according to an arbitrary enumeration of the actions, since (26) is jointly measurable in s and t . Consider a T -period problem starting in state (i, s) . From (25) we have

$$(27) \quad (V_{\pi'} - V_{\pi})(i, s) = \int_0^T (P_{\pi'}^t [r_{\pi_t} + G_{\pi_t} V_{\pi}^t])(i, s) dt.$$

From Corollary 8 we have that $P_{\pi'}^t(i, s + t | i, s) \geq e^{-Q_u}$ and from (24) the expression in brackets in (27) is nonnegative and positive on a subset of $\{(i, s + t): 0 \leq t \leq T\}$ of positive Lebesgue measure. Therefore $V_{\pi}(i, s) < V_{\pi'}(i, s)$ and π is not optimal, completing the proof.

Acknowledgments. The author is indebted to Professor Arthur F. Veinott, Jr. for suggesting and guiding the development of this work and for considerably improving the exposition.

REFERENCES

- [1] T. ASAKURA AND O. SHUNJI (1970), *A two-unit standby redundant system with repair and preventive maintenance*, J. Appl. Probab., 7, pp. 641–648.
- [2] L. J. CANTALUPPI (1981), *Semi-Markov decision chains with holding-time-dependent policies*, Ph.D. Dissertation, Dept. Operations Research, Stanford University, Stanford, CA.

- [3] S. S. CHITGOPEKAR (1969), *Continuous time Markovian sequential control processes*, this Journal, 7, pp. 367–389.
- [4] E. ÇINLAR (1969), *Markov renewal theory*, Adv. Appl. Prob., 1, pp. 123–187.
- [5] K. L. CHUNG (1960), *Markov Chains with Stationary Transition Probabilities*, Springer, New York.
- [6] E. A. CODDINGTON AND N. LEVINSON (1955), *Theory of Ordinary Differential Equations*, McGraw-Hill, New York.
- [7] E. V. DENARDO AND B. L. FOX (1968), *Multichain Markov renewal programs*, SIAM J. Appl. Math., 16, pp. 468–487.
- [8] J. L. DOOB (1961), *Stochastic Processes*, Prentice-Hall, Englewood Cliffs, NJ.
- [9] E. B. DYNKIN (1961), *Theory of Markov Processes*, Prentice-Hall, Englewood Cliffs, NJ.
- [10] R. A. HOWARD (1963), *Semi-Markov decision processes*, Bull. Inst. Internat. Statist., 40, pp. 625–652.
- [11] W. S. JEWELL (1963), *Markov renewal programming, I and II*, Oper. Res., 11, pp. 938–971.
- [12] E. P. KAO (1973), *Optimal replacement rules when changes of state are semi-Markovian*, Oper. Res., 21, pp. 1231–1249.
- [13] J. LAMPERTI (1977), *Stochastic Processes*, Springer, New York.
- [14] S. A. LIPPMAN (1973), *Semi-Markov decision processes with unbounded rewards*, Management Sci., 19, pp. 717–731.
- [15] B. L. MILLER (1968), *Finite state continuous time Markov decision processes with an infinite planning horizon*, J. Math. Anal. Appl., 22, pp. 552–569.
- [16] ——— (1968), *Finite state continuous time Markov decision processes with a finite planning horizon*, this Journal, 6, pp. 266–280.
- [17] I. P. NATANSON (1955), *Theory of Functions of a Real Variable*, Frederick Ungar, New York.
- [18] S. OSAKI AND H. MINE (1968), *Linear programming algorithms for semi-Markovian decision processes*, J. Math. Anal. Appl., 22, pp. 356–381.
- [19] S. R. PLISKA (1973), *Single-person controlled diffusions with discounted costs*, J. Optim. Theory Appl., 12, pp. 248–255.
- [20] L. D. STONE (1973), *Necessary and sufficient conditions for optimal control of semi-Markovian processes*, this Journal, 11, pp. 367–389.
- [21] A. F. VEINOTT, JR. (1969), *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Statist., 40, pp. 1635–1660.
- [22] ——— (1978), *Dynamic programming: some open problems*, in Dynamic Programming and Its Applications, M. Puterman, ed., Academic Press, New York, pp. 397–407.

THE LAGRANGE MULTIPLIER RULE ON MANIFOLDS AND OPTIMAL CONTROL OF NONLINEAR SYSTEMS*

J. C. P. BUS†

Abstract. In this paper we present a differential geometric approach to the Lagrange problem and the fixed time optimal control problem for nonlinear time-invariant control systems. We restrict attention to first order conditions for optimality and present a generalized Lagrange multiplier rule for restricted variational problems. Our treatment of the optimal control problem uses a recently proposed fibre bundle approach for the definition of nonlinear systems.

Key words. Lagrange problem, nonlinear optimal control, differential geometric approach, first order conditions

1. Introduction. The classical problem of the calculus of variations is the Lagrange problem: find a curve $\phi: [0, T] \rightarrow \mathbb{R}^n$ from some class of curves, e.g. piecewise continuous, which satisfies certain end point conditions and minimizes an integral of the form

$$\int_a^b \mathcal{L}(\phi(t), \dot{\phi}(t), t) dt.$$

In addition, one might impose restrictions on the curves of the form

$$F(\phi(t), \dot{\phi}(t), t) = 0.$$

Such problems were already studied by Euler and Lagrange at the end of the eighteenth century. A comprehensive treatment of the calculus of variations and its use to solve the (restricted) Lagrange problem is given by Carathéodory in [8]. It includes references to classical work. An important difficulty using variational techniques for solving the restricted Lagrange problem is caused by the end point conditions. It may occur that restrictions and end point conditions are such that no admissible variations of an admissible curve exist (except for the trivial one). So, such an admissible curve is extremal. Carathéodory studies this phenomenon by introducing the concept “class of the problem;” no problems arise when the class equals zero. In another general reference on the calculus of variations [3], Bliss introduces the concept “abnormality of certain order.” He calls a problem normal (abnormal of order zero), if there exist nontrivial admissible variations. Both Carathéodory and Bliss need the definition and existence of Lagrange multipliers as a prerequisite for defining “class” and “normality.” In this paper, which is based on unpublished course notes of Takens [20], we consider the generalized variational problem on manifolds, restricting attention to first order conditions (we speak of stationarity rather than optimality). We introduce the concept of formal stationarity for restricted problems. This is stationarity with respect to formally admissible (i.e., admissible up to first order in the variation parameter) variations. This concept is stronger than stationarity. We then define restricted variational problems to be “normal” if stationary curves are also formally stationary. Normal as we use it, means not quite the same as for Bliss. In our terminology it might occur that in normal problems there exist neither formally admissible nor admissible variations of a stationary admissible curve. It is the same for those problems which

* Received by the editors January 31, 1983, and in final revised form September 12, 1983.

† Department of Operations Research and System Theory, Mathematical Centre, Amsterdam, the Netherlands.

allow nontrivial formally admissible variations. Our approach to normality does not rely upon the definition of Lagrange multipliers. The Lagrange multiplier rule is given in § 3 expressing that a necessary and sufficient condition for formal stationarity for a restricted variational problem, is the existence of a stationary curve for a related unrestricted problem on a higher dimensional manifold. Then the theory of integral invariants of Cartan [7] can be used to express stationary curves for the latter problem as characteristic curves of a certain differential 2-form. The problem of normality is postponed to §§ 4 and 5 where the unrestricted Lagrange problem and the nonlinear optimal control problem are formulated as restricted variational problems. The former is merely given as an example and normality is proven, as to be expected. In our opinion the latter has value in itself. Moreover, it incorporates a recently introduced formulation of nonlinear control systems on fibre bundles (see [14], [18] and [21]). We shall see that the variational problem associated with an optimal control problem with clamped end points, will not always be normal as was already clear from the results in the books of Carathéodory and Bliss.

Variational problems on manifolds, using differential geometric concepts and Cartan's characterization for unrestricted problems, are also treated in various other papers, e.g. [10], [11], [12], [13], [15] and [17]. The restrictions considered in these references are induced by exterior differential systems or Pfaffian systems. They place more emphasis on the generalized Euler–Lagrange equation as a necessary condition for stationarity, treating the normality problem in about the same way as Bliss, except for their use of modern differential geometric results and formulations. In our approach the multiplier rule plays a natural role and normality is treated differently. Together with the linkage to the fibre bundle approach to nonlinear control systems, we expect that the given formulation of optimal control problems will be useful for studying optimal feedback control laws. It can be extended to infinite horizon problems (see [6]) in which case it might be particularly useful. The given approach is coordinate-free and does not presuppose any regularity conditions on the cost function.

Finally, in this paper we shall use the notation given in [19]. For instance, if M is a smooth manifold, TM is its tangent bundle ($T_x M$ is the tangent space at $x \in M$) and T^*M is the cotangent bundle. If $f: M \rightarrow N$ is a smooth mapping between smooth manifolds M and N then $f_*: TM \rightarrow TN$ is its lift to the tangent bundles and for any k -form ω on N , $f^*\omega$ is a k -form on M which is defined by $(f^*\omega)(v) = \omega(f_*v)$ for all $v \in TM$. Some minor deviations from Spivak's notation occur. The set of smooth vector fields on a smooth manifold is denoted by $\mathcal{X}(M)$. Furthermore, given a k -form ω and a vector field X on M , we define the *contraction* $\iota_X \omega$ of ω with respect to X , to be the $(k-1)$ -form on M defined by

$$\iota_X \omega(X_1, \dots, X_{k-1}) = \omega(X, X_1, \dots, X_{k-1})$$

for

$$X_i \in \mathcal{X}(M) \quad (i = 1, \dots, k-1).$$

Unless stated otherwise all manifolds, mappings, forms and vector fields are assumed to be smooth, i.e. C^∞ .

2. The unrestricted variational problem. Let M be a smooth manifold with $\dim M = m$, α a smooth (differential) 1-form on M and $h: M \rightarrow \mathbb{R}$ a function. Denote $I = [0, T] \subset \mathbb{R}$. Let $x_0 \in M$, the *initial point*, be given and $S \subset M$ be a connected smooth submanifold of M , called the *target set*. Define for smooth curves $\phi: I \rightarrow M$ the action

$$(2.1) \quad \mathcal{J}(\phi) = h(\phi(T)) + \int_I \phi^* \alpha.$$

The *variational problem* w.r.t. this data, denoted by $VP(M, \alpha, x_0, h, S)$ is the problem to find curves with $\phi(0) = x_0, \phi(T) \in S$, which are locally optimal w.r.t. \mathcal{J} , i.e. which produce an optimal value for \mathcal{J} to small variations of the curves. We shall restrict attention to *first order conditions*, hence to *stationarity* rather than optimality.

We distinguish two cases:

1. *Clamped end point (CE) problem.* $S = \{m_T\}$, i.e. just one point $m_T \in M$, and $h \equiv 0$,
2. *Free end point (FE) problem.* S a connected smooth submanifold of M of dimension ≥ 1 .

The following definitions are standard (see [10], [11] and [19]).

DEFINITION 2.1. A mapping $\tilde{\phi}: (-\delta, \delta) \times I \rightarrow M$ (for some $\delta > 0$) is called a *variation of $\phi: I \rightarrow M$* if:

- (i) $\tilde{\phi}$ is C^∞ in each variable;
- (ii) $\tilde{\phi}(0, t) = \phi(t)$ for all $t \in I$;
- (iii) $\tilde{\phi}(\varepsilon, 0) = \phi(0), \tilde{\phi}(\varepsilon, T) \in S$ for all $\varepsilon \in (-\delta, \delta)$.

The set of variations of ϕ is denoted by \mathcal{V}_ϕ and for short we write $\phi_\varepsilon(t) = \tilde{\phi}(\varepsilon, t)$. Depending on S we speak of CE or FE variations.

Stationary curves for the action are curves which make the first variation formula vanish. The following definition makes this precise.

DEFINITION 2.2. A curve $\phi: I \rightarrow M$ is *stationary* for $VP(M, \alpha, x_0, h, S)$ if for all $\phi_\varepsilon \in \mathcal{V}_\phi$:

$$\left. \frac{d}{d\varepsilon} \right|_{\varepsilon=0} \mathcal{J}(\phi_\varepsilon) = 0.$$

For given variation $\phi_\varepsilon \in \mathcal{V}_\phi$ we can choose $\bar{\phi}_\varepsilon$ such that $\bar{\phi}_\varepsilon$ is identically equal to ϕ in some neighbourhood of x_0 and $\int_I (\phi_\varepsilon^* - \bar{\phi}_\varepsilon^*) \alpha$ is arbitrarily close to zero (see [22, §§ 6, 7]). The same holds for the end point in the CE case. Therefore we may assume that variations in \mathcal{V}_ϕ are identically equal to ϕ in neighbourhoods of the initial point and the end point (except of course for the free directions in the FE problem).

From now on we shall assume that the curves we consider are injective immersions. This is a rather natural assumption as curves with double points are usually not optimal, because of occurrence of a loop. In such cases we can formulate the variational problem for piecewise injective curves as a sum of variational problems for each piece (see also [19]).

We can give another, equivalent, definition of stationarity in terms of vector fields along ϕ . By a *vector field along a curve $\phi: I \rightarrow M$* we mean a smooth function $V: I \rightarrow TM$ which satisfies $V(t) \in T_{\phi(t)}M$. Clearly, each variation $\tilde{\phi} \in \mathcal{V}_\phi$ defines a vector field V along ϕ by the formula

$$(2.2) \quad V(t) = \left. \frac{\partial}{\partial \varepsilon} \right|_{\varepsilon=0} \tilde{\phi}(\varepsilon, t), \quad t \in I,$$

with

$$(2.3) \quad V(0) = 0, \quad V(T) = 0 \text{ (CE) or } V(T) \in T_{\phi(T)}S \text{ (FE)}.$$

We shall denote the set of vector fields along ϕ satisfying (2.3) by \mathcal{X}_ϕ . Conversely, given any vector field $V \in \mathcal{X}_\phi$, we can extend it (as ϕ is an injective immersion) to a vector field $X \in \mathcal{X}(M)$ and construct a variation of ϕ by

$$(2.4) \quad \phi_\varepsilon(t) = \gamma_X(\varepsilon)(\phi(t)),$$

where $\gamma_X(\varepsilon)$ denotes the flow of X over ε . Let now ω be an arbitrary 1-form on M and let $L_X\omega$ denote its *Lie derivative* w.r.t. X . Then

$$(2.5) \quad \begin{aligned} \phi^*L_X\omega &= \phi^*\left(\lim_{\varepsilon \rightarrow \infty} \frac{1}{\varepsilon} [(\gamma_X(\varepsilon))^*\omega - (\gamma_X(0))^*\omega]\right) \\ &= \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} [(\gamma_X(\varepsilon) \circ \phi)^*\omega] = \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} (\phi_\varepsilon^*\omega). \end{aligned}$$

We also have the well-known relation

$$(2.6) \quad L_X\omega = \iota_X d\omega + d\iota_X\omega.$$

Given V along ϕ , we have for an arbitrary smooth extension X of V :

$$(2.7) \quad \phi^*L_X\omega \left(\frac{\partial}{\partial t} \right) = d\omega \left(V(t), \phi_* \left(\frac{\partial}{\partial t} \right) \right) + d(\omega(V(t))) \left(\frac{\partial}{\partial t} \right) = \phi^*L_V\omega \left(\frac{\partial}{\partial t} \right).$$

(By $\partial/\partial t$ we mean the tangent vector evaluated at t .) Then, for all extensions X of V and induced variations cf. (2.4) we have the equality

$$(2.8) \quad \phi^*L_V\omega = \frac{d}{d\varepsilon} \bigg|_{\varepsilon=0} \phi_\varepsilon^*\omega.$$

So any $V \in \mathcal{X}_\phi$ defines a class of variations ϕ_ε of ϕ satisfying (2.8). These relations between vector fields along ϕ and variations of ϕ show that we can equivalently define stationarity by:

DEFINITION 2.2'. ϕ is *stationary* for VP (M, α, x_0, h, S) , if for all $V \in \mathcal{X}_\phi$:

$$(2.9) \quad dh(V(T)) + \int_I \phi^*L_V\alpha = 0.$$

This definition easily leads to a useful and well-known characterization of stationary curves.

PROPOSITION 2.3. ϕ is *stationary* for VP (M, α, x_0, h, S) if and only if

$$(2.10) \quad \phi_* \left(\frac{\partial}{\partial t} \bigg|_t \right) \in \ker d\alpha \quad \forall t \in I,$$

with $\ker d\alpha = \{v \in TM \mid d\alpha(v, w) = 0, \forall w \in T_{\pi(v)}M\}$, and

$$(2.11) \quad (dh + \alpha)|_S(\phi(T)) = 0,$$

where $|_S$ denotes restriction to S .

Proof. Using Stokes' theorem we have for $V \in \mathcal{X}_\phi$

$$\int_I \phi^*L_V\alpha = \int_I \phi^*\iota_V d\alpha + \alpha(V(T)).$$

So sufficiency is trivial.

Now suppose ϕ is stationary and (2.10) is not satisfied for some $t \in I$. Then by the smoothness we can construct a V along ϕ with $V(0) = 0$, $V(T) = 0$ (hence $\in T_{\phi(T)}S$) and

$$\int_I \phi^*\iota_V d\alpha \neq 0.$$

However, this contradicts the stationarity of ϕ . Hence (2.10), and therefore (2.11), is satisfied. \square

Condition (2.10) expresses that $d\alpha$ is an *integral invariant* for stationary curves (see [7]). ϕ is called a *characteristic curve* of $d\theta$. Another way of looking to (2.10) is to say that ϕ is an integral curve of the Cartan system $C(d\alpha)$ (i.e. the Pfaffian system generated by all 1-forms $\iota_X d\alpha$, $X \in \mathcal{X}(M)$ arbitrary) (see [11]). Condition (2.11) is the so-called *transversality condition* at the end point. It is interpreted as to disappear, or be trivially satisfied, in the CE problem ($h \equiv 0$ and S consists of one point).

3. The restricted variational problem. We may introduce restrictions on curves in M via smooth codistributions on M . In §§ 4 and 5 it is shown that the classical Lagrange problem and the nonlinear optimal control problem can be formulated as variational problems with such restrictions. Let E be a given codistribution on M . Denote the variational problem $VP(M, \alpha, x_0, h, S)$ with restriction E by $VP(M, \alpha, x_0, h, S, E)$. We call a curve $\phi: I \rightarrow M$ *admissible* for this problem if

$$(3.1) \quad \phi^* \beta \equiv 0 \quad \forall \beta \in E.$$

We shall assume throughout that E is smooth and of fixed dimension p , spanned locally by 1-forms β_1, \dots, β_p . So locally (3.1) has to be satisfied for $\beta = \beta_i$ ($i = 1, \dots, p$) only. We denote the class of admissible variations of ϕ by

$$(3.2) \quad \mathcal{V}_\phi^E = \{\xi_\epsilon \in \mathcal{V}_\phi \mid \xi_\epsilon^* \beta \equiv 0, \forall \beta \in E\}.$$

In the vector field terminology we consider the set of admissible vector fields $V \in \mathcal{X}_\phi$:

$$(3.3) \quad \mathcal{X}_\phi^E = \{V \in \mathcal{X}_\phi \mid \beta(V) = 0, \forall \beta \in E\}.$$

The following definition is then natural.

DEFINITION 3.1. An admissible curve $\phi: I \rightarrow M$ is *stationary* for $VP(M, \alpha, x_0, h, S, E)$ if one of the following two equivalent conditions is satisfied:

- (i) $d/d\epsilon|_{\epsilon=0} \mathcal{J}(\phi_\epsilon) = 0$, for all $\phi_\epsilon \in \mathcal{V}_\phi^E$,
- (ii) $dh(V(T)) + \int_I \phi^* L_V \alpha = 0$, for all $V \in \mathcal{X}_\phi^E$.

Note that this definition implies that isolated admissible curves, i.e. admissible curves for which there exist no admissible variations, are stationary. Such situations may occur as shows the following example.

Example 3.2. We consider on $M = T\mathbb{R}^2 \times \mathbb{R}$ the restricted variational problem $VP(M, \alpha, m_0, 0, \{m_T\}, E)$ with $m_0, m_T \in M$ and E spanned by:

$$\beta_1 = dx - \sqrt{1 + \dot{y}^2} dt, \quad \beta_2 = dy - \dot{y} dt,$$

where $(x, y, \dot{x}, \dot{y}, t)$ are coordinates for $T\mathbb{R}^2 \times \mathbb{R}$. Now let $\phi: [0, T] \rightarrow T\mathbb{R}^2 \times \mathbb{R}$, given by

$$\phi(t) = (\phi_x, \phi_y, \phi_{\dot{x}}, \phi_{\dot{y}}, t)$$

be admissible. Then

$$\dot{\phi}_x = \sqrt{1 + \phi_{\dot{y}}^2}, \quad \dot{\phi}_y = \phi_{\dot{y}}.$$

So $\phi_x(t)$ is the length of the curve ϕ_y from 0 to t . Hence, any variation of ϕ_y with fixed end point yields a change of the x -coordinate of the end point. Therefore there are no nontrivial admissible variations of ϕ .

Clearly, the situation of isolated admissible curves requires careful attention and its occurrence depends on both the restrictions and the end point conditions. The way we shall handle this difficulty is suggested by Takens [20]. First observe that admissible variations satisfy the codistribution constraints for all small $|\epsilon|$. We may consider variations of the unrestricted problem which satisfy the restrictions to first order only.

To do so denote

$$(3.4) \quad \begin{aligned} \mathcal{W}_\phi^E &= \left\{ \xi_\varepsilon \in \mathcal{V}_\phi \left| \frac{d}{d\varepsilon} \right|_{\varepsilon=0} \xi_\varepsilon^* \beta = 0, \forall \beta \in E \right\}, \\ \mathcal{Y}_\phi^E &= \{ V \in \mathcal{X}_\phi \mid \phi^* L_V \beta = 0, \forall \beta \in E \} \end{aligned}$$

and define:

DEFINITION 3.3. An admissible curve $\phi: I \rightarrow M$ is *formally stationary* for VP $(M, \alpha, x_0, h, S, E)$, if one of the conditions of Definition 3.1 is satisfied with \mathcal{V}_ϕ^E replaced by \mathcal{W}_ϕ^E and \mathcal{X}_ϕ^E by \mathcal{Y}_ϕ^E .

We call elements of \mathcal{W}_ϕ^E *formal variations* of ϕ . Note that $\mathcal{V}_\phi^E \subsetneq \mathcal{W}_\phi^E$, so that formal stationarity implies stationarity, but not necessarily the converse. Example 3.2 can be used to show that. We define

DEFINITION 3.4. VP $(M, \alpha, x_0, h, S, E)$ is *normal* for an admissible curve ϕ , if stationarity of ϕ implies formal stationarity of ϕ .

We defer the problem of normality to § 5, where it is studied for the special classes of variational problems which are of concern to us here. For historic reasons we use the terminology of [3]. However, our notion is weaker in the sense that it also allows the situation that neither formally admissible nor admissible variations exist.

Before giving the main result of this section we shall dwell some time upon the global character of the results to be obtained. In fact, a global problem can easily be broken up in finitely many local problems.

PROPOSITION 3.5. Let $\phi: I \rightarrow M$ be an injective curve. Let $\{I^\mu\}$ ($I^\mu = [a^\mu, b^\mu]$) be a finite collection of closed subintervals of I such that $\{\text{int } I^\mu\}$ is an open covering of $\text{int } I$. Define $\phi^\mu = \phi|_{I^\mu}$, the restriction of ϕ to I^μ . Then, ϕ is (formally) stationary for the CE problem VP $(M, \alpha, x_0, 0, \{x_T\})$ (or its restricted variant) if and only if ϕ^μ is (formally) stationary for VP $(M, \alpha, \phi(a^\mu), 0, \{\phi(b^\mu)\})$ (or restricted), for all μ . Similarly for the FE problem VP (M, α, x_0, h, S) with local problems VP $(M, \alpha, \phi(a^\mu), h^\mu, S^\mu)$ where $h^\mu = 0$, $S^\mu = M$ if $b^\mu \neq T$ and $h^\mu = h$, $S^\mu = S$, otherwise.

Proof. Recall that variations are identically equal to ϕ at some neighbourhood of the clamped begin (and end) point, according to the remark after Definition 2.2.

First consider stationarity for CE problems. If ϕ is stationary then any variation of a subproblem on I^μ can be considered to be a variation of ϕ on I (equal to ϕ outside I^μ). So stationarity holds for the subproblem. To prove the converse choose a partition of unity $\{f^\mu\}$ ($f^\mu: I \rightarrow \mathbb{R}$) and write

$$(3.5) \quad \phi_\varepsilon = \phi + \eta_\varepsilon = \phi + \sum_\mu f^\mu \eta_\varepsilon.$$

As $\phi_\varepsilon^\mu = \phi^\mu + f^\mu \eta_\varepsilon$ is a variation of ϕ^μ the result follows immediately. For formal stationarity we need the additional observation that

$$(3.6) \quad \frac{d}{d\varepsilon} \Big|_{\varepsilon=0} (\phi_\varepsilon^{\mu*} \beta) = f^{\mu*} \frac{d}{d\varepsilon} \Big|_{\varepsilon=0} (\eta^* \beta) = f^{\mu*} \frac{d}{d\varepsilon} \Big|_{\varepsilon=0} (\phi_\varepsilon^* \beta),$$

so that global formal variations yields local formal variations and vice versa. In the case of an FE problem we note that a variation of a subproblem (both for $b^\mu = T$ or $b^\mu \neq T$) can be approximated arbitrarily close by an FE variation of ϕ on I which equals ϕ outside I^μ . Hence stationarity of ϕ yields stationarity of ϕ^μ . Conversely, note that if ϕ^μ is FE stationary ($b^\mu \neq T$) then ϕ^μ is also CE stationary. Hence we can again use a partition of unity argument as above. \square

After this intermezzo we return to the development of the main theorem. Let $\pi: T^*M \rightarrow M$ denote natural projection and recall from [11] the definition of the canonical 1-form θ on T^*M :

$$(3.7) \quad \theta(\xi)(v) = \xi(\pi_*(v)) \quad \text{for all } \xi \in T^*M, v \in T_\xi T^*M.$$

We need one more important 1-form.

DEFINITION 3.6. Let M be a manifold with 1-form α and codistribution E of fixed dimension. Let $\pi_E: E \rightarrow M$ denote the natural projection and let θ_E be the canonical 1-form on T^*M restricted to E . Then the *Cartan form* θ_α on E associated with α is defined by

$$(3.8) \quad \theta_\alpha = \pi_E^* \alpha + \theta_E.$$

Now we are ready to formulate the generalized Lagrange multiplier rule.

THEOREM 3.7. An injective curve $\phi: I \rightarrow M$ is formally stationary for $VP(M, \alpha, x_0, h, S, E)$ if and only if there exists an injective curve $\eta: I \rightarrow E$ with $\pi_E \circ \eta = \phi$ which is stationary for $VP(E, \theta_\alpha, e_0, h \circ \pi_E, \chi(S))$, for some $e_0 \in \pi_E^{-1}(x_0)$ and some section $\chi: M \rightarrow E$.

Proof. We first give the proof for the CE problem. Let $\eta: I \rightarrow E$ be given with $\pi_E \circ \eta = \phi$ and η stationary for the problem on E . By Proposition 3.5 we can restrict attention to curves in a coordinate neighbourhood such that E is spanned by forms β_1, \dots, β_p on this neighbourhood. Furthermore, note that an arbitrary vector field along η yields a projected vector field along ϕ as ϕ and η are injective immersions.

To prove that ϕ is formally stationary we first have to prove that ϕ is admissible. Therefore choose local coordinates x for M and let β_1, \dots, β_p be a local basis for E . Then we can give coordinates (x, y) for E ; that is, an element $(x, \sum_{i=1}^p y_i \beta_i(x)) \in E$ has coordinates (x, y) ($y = (y_1, \dots, y_p)$). By definition of the canonical form on $E \subset T^*M$ we have for $v \in T_{(x,y)}E$:

$$(3.9) \quad \theta_E(x, y)(v) = \left(\sum_{i=1}^p y_i \beta_i(x) \right) (\pi_E^* v) = \sum_{i=1}^p y_i (\pi_E^* \beta_i)(v).$$

Therefore, given an arbitrary vector field X on E ,

$$(3.10) \quad X = \sum_{i=1}^n X_i \frac{\partial}{\partial x_i} + \sum_{j=1}^p Y_j \frac{\partial}{\partial y_j},$$

we obtain

$$(3.11) \quad \begin{aligned} (L_X \theta_E)(x, y) &= \sum_{i=1}^p L_X(y_i \pi_E^* \beta_i)(x, y) \\ &= \sum_{i=1}^p Y_i (\pi_E^* \beta_i)(x, y) + \sum_{i=1}^p y_i L_X(\pi_E^* \beta_i)(x, y). \end{aligned}$$

Now let in these coordinates η be given by

$$(3.12) \quad \eta(t) = (\phi(t), \lambda(t))$$

(ϕ and λ are x and y coordinates, respectively) and define

$$W_i(t) = w_i(t) \frac{\partial}{\partial y_i} \Big|_{\eta(t)}, \quad i = 1, \dots, p,$$

where w_i is arbitrary on I with $w_i(0) = w_i(T) = 0$. Clearly W_i ($i = 1, \dots, p$) are vector

fields along η with projection $\pi_{E*} W_i = 0$. Then with use of (3.11)

$$(3.13) \quad \int_I \eta^* L_{W_i} \theta_\alpha = \int_I w_i(t) \eta^* \pi_E^* \beta_i + \sum_{i=1}^p \int_I \lambda_i(t) \eta^* L_{W_i}(\pi_E^* \beta_i).$$

Moreover, use of (2.7) shows that the last term equals zero. As the stationarity of η makes the left-hand side of (3.13) equal to zero, we have

$$0 = \int_I w_i(t) \eta^* \pi_E^* \beta_i = \int_I w_i(t) \phi^* \beta_i,$$

for arbitrary w_i . This proves that $\phi^* \beta_i = 0$ ($i = 1, \dots, p$), hence ϕ is admissible.

To prove the formal stationarity of ϕ let a vector field V along ϕ with $V(0) = V(T) = 0$ be given in coordinates:

$$V(t) = \sum_{i=1}^n V_i(t) \frac{\partial}{\partial x_i} \Big|_{\phi(t)}.$$

Define a vector field W along η by

$$W(t) = \sum_{i=1}^n V_i(t) \frac{\partial}{\partial x_i} \Big|_{\eta(t)}.$$

Then $\pi_{E*} W = V$ and the $\partial/\partial y_i$ -components of W are zero. So use of (3.11) yields:

$$(3.14) \quad \begin{aligned} \int_I \eta^* L_W \theta_\alpha &= \int_I \eta^* L_W(\pi_E^* \alpha) + \int_I \eta^* L_W \theta_E \\ &= \int_I \eta^* L_W(\pi_E^* \alpha) + \sum_{i=1}^p \int_I \lambda_i(t) \eta^* L_W(\pi_E^* \beta_i). \end{aligned}$$

Moreover, use of (2.7) shows that

$$\eta^* L_W(\pi_E^* \beta_i) \left(\frac{\partial}{\partial t} \Big|_{\eta(t)} \right) = \phi^* L_V \beta_i \left(\frac{\partial}{\partial t} \Big|_{\phi(t)} \right).$$

Substituting this in (3.14) yields

$$\int_I \eta^* L_W \theta_\alpha = \int_I \phi^* L_V \alpha + \sum_{i=1}^p \int_I \lambda_i(t) \phi^* L_V \beta_i.$$

Stationarity of η makes the left-hand side zero. So $\phi^* L_V \beta_i = 0$ ($i = 1, \dots, p$) yields $\int_I \phi^* L_V \alpha = 0$. This implies formal stationarity.

To prove the converse, let ϕ be formally stationary. Given any vector field W along η with $W(0) = W(T) = 0$ we obtain, using (3.11),

$$\int_I \eta^* L_W \theta_\alpha = \int_I \eta^* L_W(\pi_E^* \alpha) + \sum_{i=1}^p \int_I W_{y_i} \phi^* \beta_i + \sum_{i=1}^p \int_I \lambda_i \eta^* L_W(\pi_E^* \beta_i),$$

with W_{y_i} the $\partial/\partial y_i$ -component of W . As ϕ is admissible ($\phi^* \beta_i = 0$) we obtain, with $V = \pi_{E*} W$:

$$(3.15) \quad \int_I \eta^* L_W \theta_\alpha = \int_I \phi^* L_V \alpha + \sum_{i=1}^p \int_I \lambda_i \phi^* L_V \beta_i.$$

Hence, we have to prove that we can find $\lambda_i: I \rightarrow \mathbb{R}$ ($i = 1, \dots, p$) such that for all

$V \in \mathcal{X}_\phi$ the following equality is satisfied

$$(3.16) \quad \int_I \phi^* L_V \alpha = - \sum_{i=1}^p \int_I \lambda_i \phi^* L_V \beta_i.$$

Note that we then have $\eta(t) = \sum_{i=1}^p \lambda_i(t) \beta_i(\phi(t))$ satisfying the conditions of the theorem. For simplicity we assume that $p=1$, i.e. E is spanned by one 1-form. We omit the subscripts for λ and β . To find an appropriate λ in this case define a vector field Z along ϕ such that $\beta(Z) = 1$ along ϕ . Let

$$\mathcal{F}_1 = \{V \mid V \text{ vector field along } \phi, \phi^* L_V \beta = 0, V(0) = 0\},$$

$$\mathcal{F}_2 = \{V \mid V \text{ vector field along } \phi, V = \psi Z, \psi(0) = 0\}.$$

Then, any vector field $V \in \mathcal{X}_\phi$ can be written uniquely as the sum

$$V = V_1 + V_2, \quad V_1 \in \mathcal{F}_1, \quad V_2 \in \mathcal{F}_2.$$

This is shown by the following argument. Given V , the differential equation:

$$(3.17) \quad \begin{aligned} \phi^* L_V \beta \left(\frac{\partial}{\partial t} \right) &= \psi(t) d\beta \left(Z(t), \phi_* \left(\frac{\partial}{\partial t} \right) \right) + d\psi \left(\frac{\partial}{\partial t} \right), \\ \psi(0) &= 0 \end{aligned}$$

defines $\psi: I \rightarrow \mathbb{R}$ uniquely. Now define

$$V_2 = \psi Z; \quad V_1 = -\psi Z.$$

Then we have the appropriate splitting as $V_2 \in \mathcal{F}_2$ by choice and $V_1 \in \mathcal{F}_1$ because (use (2.7) and (3.17))

$$\phi^* L_{V_1} \beta \left(\frac{\partial}{\partial t} \right) = \phi^* L_V \beta \left(\frac{\partial}{\partial t} \right) - \psi(t) d\beta \left(Z(t), \phi_* \left(\frac{\partial}{\partial t} \right) \right) + d\psi \left(\frac{\partial}{\partial t} \right) = 0.$$

Note that $V_1(T) = -V_2(T) = -\psi(T)Z(T)$ is not necessarily equal to zero. Now let V be arbitrary with $V(0) = V(T) = 0$ and $V = V_1 + V_2 = V_1 + \psi Z$ its unique splitting. Then (2.7) and Stokes' theorem yield

$$\int_I \phi^* L_V \alpha = \int_I \phi^* \iota_V \alpha + \int_I d(\alpha(V)) = \int_I \phi^* \iota_{V_1} \alpha + \int_I \phi^* \iota_{V_2} \alpha,$$

where $\phi^* \iota_V \alpha (\partial/\partial t) = d\alpha(V(t), \phi_*(\partial/\partial t))$, by definition.

If $\psi(T) \neq 0$ ($V_1(T) \neq 0$) we define a constant C_0 such that

$$(3.18) \quad \int_I \phi^* L_V \alpha = \int_I \phi^* \iota_{V_2} \alpha - C_0 \psi(b).$$

If $\psi(T) = 0$ then $\int_I \phi^* L_V \alpha = \int_I \phi^* \iota_{V_2} \alpha$ by the formal stationarity of ϕ , so that we can choose C_0 arbitrarily and (3.18) still holds. Then define $\Psi_1, \Psi_2: I \rightarrow \mathbb{R}$ by

$$(3.19) \quad \Psi_1 dt = \phi^* \iota_Z d\beta, \quad \Psi_2 dt = \phi^* \iota_Z d\alpha$$

and $\lambda: I \rightarrow \mathbb{R}$ by

$$(3.20) \quad \dot{\lambda} = \Psi_2 + \Psi_1 \lambda, \quad \lambda(T) = C_0.$$

Then we have with use of (3.18)–(3.20)

$$- \int_I \lambda \phi^* L_V \beta = \int_I \psi \Psi_2 dt - C_0 \psi(T) = \int_I \phi^* L_V \alpha.$$

So the chosen λ satisfies (3.16) for $p=1$. Hence η , given by $\eta(t) = \lambda(t)\beta(\phi(t))$, is stationary w.r.t. θ_α and $\pi_E\eta = \phi$. For $p>1$ the proof is similar. For the FE problem there is only a slight difference where we use Stokes' theorem in the definition of C_0 . Here we choose C_0 such that

$$(3.21) \quad dh(V(T)) + \int_I \phi^* L_V \alpha = \int_I \phi^* \iota_{V_2} d\alpha + C_0 \psi(T),$$

which is fine for $\psi(T) \neq 0$. If $\psi(T) = 0$, then $V_2(T) = \psi(T)Z(T) = 0 \in T_{\phi(T)}S$ and as $V = V_2 + V_1 \in T_{\phi(T)}S$ we also have $V_1(T) \in T_{\phi(T)}S$. Then formal stationarity with $V_2(T) = 0$ shows that (3.21) holds for arbitrary choice of C_0 if $\psi(T) = 0$. Then the proof is valid for the FE case. Note that the section χ defining the target set in the problem on E is given locally by $\chi(x) = (x, C_0)$ with C_0 as in (3.18) or (3.21). \square

Note that the Lagrange multipliers are hidden in the formulation of Theorem 3.7. They appear in the coordinate representation as $\lambda_i(t)$ ($i=1, \dots, n$). Theorem 3.7 forms the heart of this paper. It enables us to formulate the Lagrange problem and the optimal control problem as a problem of finding characteristic curves of the differential of a certain Cartan form (recall Proposition 2.3), provided the associated restricted variational problem is normal for admissible curves. The most significant examples of the use of Theorem 3.7 are the unrestricted Lagrange problem and the optimal control problem. We discuss these in the next sections.

4. The Lagrange problem. Consider a smooth manifold Q (the *configuration space*) with $\dim Q = n$, together with its 1-jet manifold $J^1(I, Q)$ (see [11]), we should in fact write $J^1(\mathbb{R}, Q)$ but to express that t is restricted to I we use the above notation). Note that a point in $J^1(I, Q)$ consists of a point $t \in I$ together with a point $(q, v) \in TQ$. Thus

$$(4.1) \quad J^1(I, Q) \cong TQ \times I,$$

and moreover, given a curve $\psi: I \rightarrow Q$ there exists a naturally associated curve $\phi: I \rightarrow J^1(I, Q)$ defined by

$$(4.2) \quad \phi(t) = \left(\psi_* \left(\frac{\partial}{\partial t} \right) \Big|_t, t \right), \quad t \in I.$$

We denote $\phi = \psi^l$. Now suppose we have been given:

$$(4.3) \quad \mathcal{L}: J^1(I, Q) \rightarrow \mathbb{R}, \quad h: Q \rightarrow \mathbb{R},$$

called the *Lagrangian* and the *end cost*, respectively. Then the (unrestricted) *Lagrange problem* is to find curves $\psi: I \rightarrow Q$, with $\psi(0) = q_0, \psi(T) \in S \subset Q$, which minimize the action

$$(4.4) \quad \mathcal{J}(\psi) = h(\psi(T)) + \int_I \mathcal{L}(\psi^l(t)) dt.$$

We can formulate this as a variational problem on $M = J^1(I, Q)$ with restriction on curves in M to be naturally associated cf. (4.2) with curves in Q . Using [11, § 0.b], this restriction is defined by a codistribution E on M , which is a canonical subbundle of $T^*M = T^*(J^1(I, Q))$. Moreover, in local coordinates $(q_1, \dots, q_n, \dot{q}_1, \dots, \dot{q}_n, t)$ for M , this subbundle E is spanned by 1-forms:

$$(4.5) \quad \beta_i = dq_i - \dot{q}_i dt, \quad i = 1, \dots, n.$$

We shall call E the *canonical (restriction) codistribution* of the Lagrange problem and

the variational problem so obtained the *Lagrange variational problem*. We already noted that we may restrict attention to variations which are identical to ϕ on a neighbourhood of clamped end points. So, CE conditions for $\psi: I \rightarrow Q$ yield CE conditions for $\phi: I \rightarrow J^1(I, Q)$. Moreover, a target set $S \subset Q$ gives rise to a target set $TS \times \{T\}$ in $J^1(I, Q)$ with end cost $h \circ \pi$ ($\pi: J^1(I, Q) \rightarrow Q$ natural projection). The following result is important.

PROPOSITION 4.1. *The Lagrange variational problem is a normal restricted variational problem for every admissible curve.*

Proof. Given any admissible curve $\phi = \psi^l$ ($\psi: I \rightarrow Q$). We have to prove that if ϕ is stationary, then ϕ is formally stationary. We may restrict attention to vector fields $V \in \mathcal{X}_\phi$ which can be given in canonical coordinates by

$$(4.6) \quad V(t) = V^q(t) \frac{\partial}{\partial q} \Big|_{\phi(t)} + V^{\dot{q}}(t) \frac{\partial}{\partial \dot{q}} \Big|_{\phi(t)}.$$

Suppose such a vector field satisfies:

$$(4.7) \quad \phi^* L_V \beta = 0 \quad \forall \beta \in \text{span} \{dq_i - \dot{q}_i dt\}.$$

(Note that we may work locally, by Proposition 3.5.) We first assume that Q is 1-dimensional, so E is spanned by the form $\beta = dq - \dot{q} dt$. Thus (4.7) implies, using (2.7):

$$0 = \phi^* L_V \beta \left(\frac{\partial}{\partial t} \right) = -d\dot{q} \wedge dt \left(V(t), \phi_* \left(\frac{\partial}{\partial t} \right) \right) + d(V^q(t)) \left(\frac{\partial}{\partial t} \right).$$

So

$$(4.8) \quad \frac{dV^q(t)}{dt} = V^{\dot{q}}(t).$$

Now choose ϕ_ε by

$$\phi_\varepsilon(t) = (\psi(t) + \varepsilon V^q(t), \dot{\psi}(t) + \varepsilon V^{\dot{q}}(t), t).$$

Then ϕ_ε is a CE variation of ϕ according to Definition 2.1 with

$$(4.9) \quad \frac{\partial}{\partial \varepsilon} \Big|_{\varepsilon=0} \phi_\varepsilon(t) = (V^q(t), V^{\dot{q}}(t), 0),$$

and

$$\phi_\varepsilon^* \beta \left(\frac{\partial}{\partial t} \right) = \dot{\psi}(t) + \varepsilon \dot{V}^q(t) - (\dot{\psi}(t) + \varepsilon V^{\dot{q}}(t)) = 0,$$

using (4.8). So ϕ_ε is an admissible CE variation of ϕ , so that by stationarity and (4.9)

$$0 = \frac{d}{d\varepsilon} \Big|_{\varepsilon=0} \int_I \phi_\varepsilon^* \alpha = \int_I \phi^* L_V \alpha.$$

This proves the theorem for $\dim Q = 1$. For $\dim Q > 1$ the proof is similar. \square

A direct consequence of Proposition 4.1 and Theorem 3.7 is the following corollary.

COROLLARY 4.2. *An injective curve $\psi: I \rightarrow Q$ is a stationary curve for the Lagrange problem, if and only if there exists an injective curve $\eta: I \rightarrow E$, with E the canonical codistribution, such that $\pi_E \circ \eta = \psi^l$ and η stationary for an unrestricted variational problem $VP(E, \theta_{\mathcal{L}}, \psi^l(0), \tilde{h}, \tilde{S})$ with Cartan form*

$$\theta_{\mathcal{L}} = \pi_E^*(\mathcal{L} dt) + \theta_E$$

and

$$\tilde{h} = h \circ \pi \circ \pi_E, \quad \tilde{S} = \chi(TS \times \{T\}),$$

where $\pi_E: E \rightarrow M$, $\pi: M \rightarrow Q$ are natural projections, θ_E is the canonical 1-form restricted to E and $\chi: M \rightarrow E$ is some section.

Using Proposition 2.3 the stationary curves for the given unrestricted variational problem are characteristic curves of $d\theta_{\mathcal{L}}$ satisfying the transversality conditions. If we choose canonical coordinates q, \dot{q}, t for $J^1(Q, I)$ and λ for the fibres of E ($\beta \in E: \beta = \sum_{i=1}^n \lambda_i \beta_i$, β_i given by (4.5)), then

$$\theta_{\mathcal{L}} = \sum_{i=1}^n \lambda_i \beta_i + \mathcal{L} dt.$$

Then $\eta: t \rightarrow (q(t), \dot{q}(t), \lambda(t), t)$ is a characteristic curve of $d\theta_{\mathcal{L}}$ if

$$\frac{d}{dt} \lambda(t) = \frac{\partial \mathcal{L}}{\partial q}(q, \dot{q}, t), \quad \lambda(t) = \frac{\partial \mathcal{L}}{\partial \dot{q}}(q, \dot{q}, t), \quad \frac{d}{dt} q(t) = \dot{q}(t),$$

with given initial and end point conditions for q, \dot{q} . This yields the Euler–Lagrange equation:

$$\frac{d}{dt} \left(\frac{\partial \mathcal{L}}{\partial \dot{q}}(q, \dot{q}, t) \right) - \frac{\partial \mathcal{L}}{\partial q}(q, \dot{q}, t) = 0$$

as a necessary condition on optimal curves. The transversality condition yields

$$\lambda_i(T) = \frac{\partial h}{\partial q_i}(q(T)), \quad i = 1, \dots, n.$$

Remark 4.3. It is easily seen that we may also choose $\alpha = \mathcal{L} dt + \beta$ for any $\beta \in E$ in the formulation of the Lagrange variational problem. Indeed, we then also have $\phi^* \alpha = \phi^*(\mathcal{L} dt)$, for all admissible ϕ . Such a choice does not change the solution of the Lagrange problem but only induces a translation of the canonical coordinates λ in E .

5. The nonlinear optimal control problem. We shall first recall the notion of a general nonlinear control system as given in [4] and [21] and worked out in [18].

DEFINITION 5.1. A *nonlinear* (time-invariant) *control system* $\Sigma = \Sigma(Q, B, f)$ is defined by a smooth manifold Q , a fibre bundle $\tau: B \rightarrow Q$ and a smooth map $f: B \rightarrow TQ$ such that the following diagram commutes

$$(5.1) \quad \begin{array}{ccc} B & \xrightarrow{f} & TQ \\ & \searrow \tau & \swarrow \pi_Q \\ & Q & \end{array}$$

We call Σ *affine* if B is a vector bundle and f restricted to the fibres of B is an affine map into the fibres of TQ .

Σ is called *analytic* if B and Q are analytic manifolds and f is an analytic map.

We say that $\psi: I \rightarrow Q$ is a *trajectory* of Σ if ψ is absolutely continuous and

$$\psi_* \left(\frac{\partial}{\partial t} \Big|_t \right) \in f(\tau^{-1}(\psi(t))),$$

almost everywhere on I . With each trajectory ψ we can associate a *state-input trajectory*

$\zeta: I \rightarrow B$ such that

$$(5.2) \quad \tau(\zeta(t)) = \psi(t), \quad \psi_* \left(\frac{\partial}{\partial t} \Big|_t \right) = f(\zeta(t)), \quad t \in I.$$

Q is called the *configuration space* cf. the Lagrange context. The fibres of B represent the (state dependent) input spaces. In local coordinates q for Q and u for the fibres $\tau^{-1}(q)$ we obtain the familiar equation $\dot{q} = f(q, u)$ (with abuse of notation $f: (q, u) \rightarrow (q, f(q, u))$). A state-input trajectory ζ will in such coordinates be denoted by: $\zeta(t) = (\psi(t), \nu(t))$, ψ and ν denoting the q and u coordinates respectively. In the sequel we will use f in both ways, how it is used will be clear from the context. If Σ is affine then, in coordinates, f has the form

$$(5.3) \quad f(q, u) = f_0(q) + \sum_{i=1}^m u_i f_i(q),$$

with $u_i \in \mathbb{R}$, f_0 and f_i vector fields on Q ($i = 1, \dots, m$).

We shall assume in the rest of this paper that f is an injective immersion.

Now, an optimal control problem can be interpreted as a certain variational problem on the space of states and inputs, i.e. B , under certain restrictions, one of these being the restriction to curves in B which are state-input trajectories of the system. In fact, the approach to the Lagrangian problem for curves in Q can be followed here with respect to curves in B . Therefore, let us first assume to be given a function $\mathcal{G}: J^1(I, B) \rightarrow \mathbb{R}$, in analogy with the Lagrangian \mathcal{L} in § 4 and an end cost function $h: Q \rightarrow \mathbb{R}$. We restrict attention to two cases:

CE optimal control problem: $h \equiv 0$, clamped end point;

FE optimal control problem: $S = Q$.

The *optimal control problem* OP $(\Sigma, \mathcal{G}, q_0, h, S)$ is to find $\zeta: I \rightarrow B$ of Σ with $\tau \circ \zeta(0) = q_0$, $\tau \circ \zeta(T) \in S$ and which are optimal w.r.t. the action

$$(5.4) \quad \mathcal{J}(\zeta) = h(\tau \circ \zeta(T)) + \int_I \mathcal{G}(\zeta^l(t)) dt.$$

As before we restrict attention to stationarity rather than optimality. The optimal control problem can be defined as a variational problem on $J^1(I, B)$ where the curves are restricted to be naturally associated (cf. (4.2)) with curves in B which are state-input trajectories of Σ . This implies restriction to a submanifold $M \subset J^1(I, B)$ defined by

$$(5.5) \quad M = \{(w, t) \in J^1(I, B) | f \circ \pi(w, t) = \tau_*(w)\},$$

with $\pi: J^1(I, B) \rightarrow B$ natural projection, together with restriction to the canonical restriction codistribution on $J^1(I, B)$, similar as in the Lagrange problem. Therefore the given optimal control problem can be defined as a restricted variational problem VP $(M, \alpha, x_0, \tilde{h}, \tilde{S}, E)$, with E the canonical codistribution on $J^1(I, B)$ restricted to M , $\alpha = \mathcal{G}|_M dt$, $\tau \circ \pi(x_0) = q_0$ and $\tilde{h} \equiv 0$, \tilde{S} one point in the CE case, or $\tilde{h} = \tau \circ \pi \circ h$ and $\tilde{S} = (TQ \times \{T\}) \cap M$ in the FE case. If we choose local coordinates q on Q , u on the fibres $\tau^{-1}(q)$, then canonical coordinates on $J^1(I, B)$ are given by $(q, u, \dot{q}, \dot{u}, t)$. Elements of M are then given by $(q, u, f(q, u), \dot{u}, t)$, so that as f is an injective immersion, natural coordinates on M are given by (q, u, \dot{u}, t) . Then E , the canonical codistribution restricted to M , is locally spanned by the 1-forms

$$(5.6) \quad \begin{aligned} \beta_i &= dq_i + f_i(q, u) dt, & i &= 1, \dots, n, \\ \beta_{n+j} &= du_j - \dot{u}_j dt, & j &= 1, \dots, m, \end{aligned}$$

where $f_i(q, u)$ denotes the i th coordinate of $f(q, u) \in T_q Q$.

The question arises whether this variational problem is normal. The answer to this question appears to be relatively easy for the important class of affine analytic control systems, if we use some recent geometric techniques (see e.g. [14] and [15]). Let the system be given by (5.3), let $\mathfrak{L}(\Sigma)$ denote the Lie algebra generated by f_i ($i=0, 1, \dots, m$) and $\mathfrak{S}(q_0)$ the maximal integral submanifold of $\mathfrak{L}(\Sigma)$ containing the trajectory under consideration which initiates at q_0 . Define $\text{ad}^0(f_0, f_i) = f_i$, $\text{ad}^{k+1}(f_0, f_i) = [f_0, \text{ad}^k(f_0, f_i)]$ for $k=0, 1, \dots$, and $\mathfrak{S}^1 = \{\text{ad}^k(f_0, f_i); k=0, 1, \dots, i=1, \dots, m\}$. Then we can give the following proposition.

PROPOSITION 5.2. *Let Σ be analytic and affine. Then:*

The FE variational problems associated with OP $(\Sigma, \mathcal{G}, q_0, h, Q)$ are normal.

If rank $\mathfrak{S}^1(q_0) = \dim \mathfrak{S}(q_0)$, then the CE variational problems associated with OP $(\Sigma, \mathcal{G}, q_0, 0, \{q_{1j}\})$ are normal.

Remark 5.3. The condition in the CE case implies that the system restricted to $\mathfrak{S}(q_0)$ has a controllable linear variational equation along the trajectory initiating at q_0 , or this restricted system is locally controllable of first order along this trajectory (see [2], [15] and [16]).

Proof. We assume that $\mathfrak{S}(q_0) = n$. The other cases are proved similarly by restricting the system to the lower dimensional manifold $\mathfrak{S}(q_0)$. The manifold appearing in the variational problem may be given coordinates such that $\phi(t) = (\phi^q(t), 0, 0, t)$ (trajectory ϕ^q in Q for input $u \equiv 0$). By breaking up the global problem in a series of local problems we may assume that $\phi(t)$ belongs to this coordinate neighbourhood for all $t \in I$. Let a formal variation in these coordinates be given by

$$(5.7) \quad \xi(\varepsilon, t) = (\xi^q(\varepsilon, t), \xi^u(\varepsilon, t), \xi^{\dot{u}}(\varepsilon, t), t).$$

We shall prove that under the given conditions we can find an admissible variation $\bar{\xi}_\varepsilon$ of ϕ which is an order ε^2 perturbation of ξ . As stationarity involves first order conditions in ε only, the proof then follows immediately. Working out the conditions for formal variations we see that

$$(5.8) \quad \begin{aligned} \xi^q(\varepsilon, t) &= \phi^q(t) + \varepsilon \eta(t) + \bar{C}^q(\varepsilon, t), \\ \xi^u(\varepsilon, t) &= \varepsilon \mu(t) + \bar{C}^u(\varepsilon, t), \\ \xi^{\dot{u}}(\varepsilon, t) &= \varepsilon \dot{\mu}(t) + \bar{C}^{\dot{u}}(\varepsilon, t), \end{aligned}$$

where $\bar{C}^q(\varepsilon, t)$, $\bar{C}^u(\varepsilon, t)$ and $\bar{C}^{\dot{u}}(\varepsilon, t)$ are all of order ε^2 and equal to zero for $t=0$ and arbitrary ε . Moreover, $\eta(t)$ satisfies the linear equation of variations with input $\mu(t) = (\mu_1(t), \dots, \mu_m(t))^T$:

$$(5.9) \quad \begin{aligned} \dot{\eta}(t) &= \frac{df_0}{dq}(\phi^q(t)) \cdot \eta(t) + \sum_{i=1}^m f_i(\phi^q(t)) \cdot \mu_i(t), \\ \eta(0) &= 0. \end{aligned}$$

Now consider for arbitrary $C^u(\varepsilon, t) = (C_1^u(\varepsilon, t), \dots, C_m^u(\varepsilon, t))^T$ of order ε^2 and satisfying $C^u(\varepsilon, 0) = \dot{C}^u(\varepsilon, 0) = 0$ (arbitrary ε) the equation

$$(5.10) \quad \begin{aligned} \dot{q}(t) &= f_0(q(t)) + \sum_{i=1}^m f_i(q(t))(\varepsilon \mu_i(t) + C_i^u(\varepsilon, t)), \\ q(0) &= q_0. \end{aligned}$$

Then from [5, Thm. 6] and [9], we know that for ε small enough any solution of (5.17) can be written as a unique convergent Volterra series. Working out this series and

using the facts that ϕ^q is a trajectory of (5.10) for $\mu = 0$ and η satisfies (5.9), we see that any solution of (5.10) can be written as $\zeta^q(\varepsilon, t) = \phi^q(t) + \varepsilon\eta(t) + C^q(\varepsilon, t)$, with $C^q(\varepsilon, 0) = 0$, $C^q(\varepsilon, t) = O(\varepsilon^2)$. Hence,

$$(5.11) \quad \zeta^q(\varepsilon, t) = (\zeta^q(\varepsilon, t), \varepsilon\mu(t) + C^u(\varepsilon, t), \varepsilon\dot{u}(t) + \dot{C}^u(\varepsilon, t), t)$$

is an admissible variation and an ε^2 perturbation of a formal variation. This proves the assertion for FE problems. For CE problems we still have to show that we can choose $C^u(\varepsilon, t)$ such that the solution $\zeta^q(\varepsilon, t)$ satisfies $\zeta^q(\varepsilon, 1) = q_1 = \phi^q(1)$ (i.e. $C^q(\varepsilon, 1) = 0$). By the local controllability of the system we can find $\tilde{C}^u(\varepsilon, t)$ such that $\varepsilon\mu_i(t) + C^u(\varepsilon, t) + \tilde{C}^u(\varepsilon, t)$ yields a trajectory of (5.10) terminating in q_1 . Moreover, the fact that the linear equation of variations (5.9) is controllable assures that we can choose $\tilde{C}^u(\varepsilon, t)$ of the same order in ε as $C^q(\varepsilon, 1)$, i.e. $O(\varepsilon^2)$. This completes the proof of the proposition. \square

The restriction to affine systems does not seem to be essential. The results of Brockett and Crouch yielding the Volterra series solution for (5.10) can also be given for nonaffine systems.

As controllable linear systems are first order controllable, the condition $\text{rank } \mathfrak{S}^1(q_0) = \dim \mathfrak{S}(q_0)$ is satisfied for all (also noncontrollable) linear systems.

We give one example of a system which is controllable but not first order locally controllable and which may give nonnormal variational problems.

Example 5.4. Let $I = [0, 1]$, $q^0 = (0, 0)^T$ and

$$\dot{q}_1 = u, \quad \dot{q}_2 = q_1^2 + 1.$$

Then

$$f_0 = \begin{pmatrix} 0 \\ q_1^2 + 1 \end{pmatrix}, \quad f_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad [f_0, f_1] = \begin{pmatrix} 0 \\ 2q_1 \end{pmatrix}, \quad \text{ad}^k(f_0, f_1) = 0 \quad \text{for } k \geq 2.$$

However

$$[f_1, [f_0, f_1]] = -\begin{pmatrix} 0 \\ 2 \end{pmatrix},$$

so that the system is locally controllable, but

$$\mathfrak{S}^1(q_0) = \text{span} \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\}.$$

For $u = 0$ we have the trajectory $\phi(t) = (0, t)^T$ and $q^1 = (0, 1)^T$. Any formal variation $\phi(t) + \varepsilon\eta(t)$ for control $\varepsilon\mu(t)$ satisfies the linear equation of variations

$$\dot{\eta}_1(t) = \mu(t), \quad \dot{\eta}_2(t) = 2\phi_1(t)\eta_1(t) = 0,$$

with $\eta(0) = \eta(1) = 0$; $\mu(0) = \mu(1) = \dot{\mu}(0) = \dot{\mu}(1) = 0$. Hence

$$\eta_1(t) = \int_0^1 \mu(\sigma) d\sigma, \quad \eta_2(t) = 0.$$

Therefore, any formal variation is of the form:

$$\left(\varepsilon \int_0^1 \mu(\sigma) d\sigma, t \right)^T + O(\varepsilon^2).$$

We consider the formal variations with order ε terms only. If these are to be ε^2 perturbations of an admissible variation $\xi(\varepsilon, t)$ we should have, for some control $\varepsilon\mu(t) + C^u(\varepsilon, t)(C^u(\varepsilon, t) = O(\varepsilon^2))$,

$$\dot{\xi}_1(\varepsilon, t) = \varepsilon\mu(t) + C^u(\varepsilon, t), \quad \dot{\xi}_2(\varepsilon, t) = \xi_1^2(\varepsilon, t) + 1,$$

with $\xi(\varepsilon, 0) = q^0$, $\xi(\varepsilon, 1) = q^1$. Hence

$$\xi_1(\varepsilon, t) = \varepsilon\eta_1(t) + \int_0^t C^u(\varepsilon, \sigma) d\sigma \stackrel{\nabla}{=} \varepsilon\eta_1(t) + C_1^q(\varepsilon, t),$$

$$\xi_2(\varepsilon, t) = t + \int_0^t (\varepsilon\eta_1(\sigma) + C_1^q(\varepsilon, \sigma))^2 d\sigma,$$

with the conditions $\xi_1(\varepsilon, 1) = 0$, $\xi_2(\varepsilon, 1) = 1$, for all $|\varepsilon|$ small. As $\eta_1(1) = 0$ the first condition is satisfied for all $C_1^q(\varepsilon, t)$ such that $C_1^q(\varepsilon, 1) = 0$, which can be obtained by an appropriate choice of $C^u(\varepsilon, t)$. The second condition implies that

$$\varepsilon^2 \int_0^1 (\eta_1(\sigma))^2 d\sigma + O(\varepsilon^3) = 0.$$

So the choice of C^u (i.e. C^q) does not influence the ε^2 term. Therefore if $\eta_1(t)$ is such that $\int_0^1 (\eta_1(\sigma))^2 d\sigma \neq 0$, then $\xi_2(\varepsilon, 1) \neq 1$. Such a choice can be made. With $\mu(t) = \eta_1(t)$ we have a formal variation which is no ε^2 perturbation of an admissible variation.

The final conclusion of this section is a consequence of Theorem 3.7 and the given formulation of an optimal control problem as a variational problem.

COROLLARY 5.5. *Let $\text{OP}(\Sigma, \mathcal{G}, q_0, h, S)$ be a given optimal control problem. Assume that the associated variational problem is normal for a given trajectory-input $\zeta: I \rightarrow B$. Then ζ is stationary if and only if there exists an injective curve $\eta: I \rightarrow E$ (E the canonical codistribution) such that $\pi_M \circ \pi_E \circ \eta = \zeta$ ($\pi_M: M \rightarrow B$, $\pi_E: E \rightarrow M$ natural projections),*

$$(5.12) \quad \eta_* \left(\frac{\partial}{\partial t} \Big|_t \right) \in \ker d\theta_{\mathcal{G}}$$

and, in case of a FE problem,

$$(5.13) \quad (d\tilde{h} + \theta_{\mathcal{G}})|_{\zeta(\eta(T))} = 0,$$

where $\theta_{\mathcal{G}} = \pi_E^*(\mathcal{G}|_M dt) + \theta_E$ is the Cartan form, $\tilde{h} = \tau \circ \pi_M \circ \pi_E \circ h$ and $\tilde{S} = \chi((TQ \times \{T\}) \cap M)$ for some section $\chi: M \rightarrow E$.

Formula (5.12) defines a Cartan system. In fact one can use the intrinsic reduction procedure given in [11, § I.e.1.] to study existence and uniqueness of solutions (see [6]).

In many practical optimal control problems \mathcal{G} depends on q and u only. We can work out such a situation in coordinates as we did for the Lagrange problem. Choose natural coordinates (q, u, \dot{u}, t) on M and on $E: (q, u, \dot{u}, \lambda, \mu, t)$ ($\beta \in E \Rightarrow \beta = \sum_{i=1}^n \lambda_i \beta_i + \sum_{j=1}^m \mu_j \beta_{n+j}$ with β_k , cf. (5.6)). Then

$$(5.14) \quad \theta_{\mathcal{G}} = \mathcal{G}(q, u) dt + \sum_{i=1}^n \lambda_i (dq_i - f_i(q, u) dt) + \sum_{j=1}^m \mu_j (du_j - \dot{u}_j dt).$$

Some computation shows that condition (5.12) on a stationary curve $\eta(t) =$

$(q(t), u(t), \dot{u}(t), \lambda(t), \mu(t), t)$ yields the equations

$$\begin{aligned}\frac{d}{dt}q &= f(q, u), \\ \frac{d}{dt}\lambda &= \frac{\partial \mathcal{G}}{\partial q}(q, u) - \left(\frac{\partial f}{\partial q}(q, u) \right)^T \lambda, \\ \frac{\partial \mathcal{G}}{\partial u}(q, u) - \left(\frac{\partial f}{\partial u}(q, u) \right)^T \lambda &= 0, \\ \mu &= 0, \quad \frac{d}{dt}u = \dot{u}.\end{aligned}$$

With the definition

$$\mathcal{H}(q, \lambda, u) = \mathcal{G}(q, u) - \lambda^T f(q, u)$$

we obtain the familiar equations of Pontryagin's maximum principle (smooth case):

$$(5.15) \quad \dot{q} = -\frac{\partial \mathcal{H}}{\partial \lambda}(q, \lambda, u), \quad \dot{\lambda} = \frac{\partial \mathcal{H}}{\partial q}(q, \lambda, u), \quad \frac{\partial \mathcal{H}}{\partial u}(q, \lambda, u) = 0.$$

The transversality condition (5.13) yields:

$$(5.16) \quad \lambda(T) = \frac{dh}{dq}(q(T)).$$

Remark 5.6. The Lagrange multiplier theorem 3.7 gives a necessary and sufficient condition for formal stationarity. Therefore, these conditions are necessary for optimality. If the problem is not normal, then the conditions of Corollary 5.5 only are sufficient for stationarity. Hence, for nonnormal problems ϕ may be stationary (although not necessarily optimal) without being a projection of a stationary η in E . Higher order theory provides better insight here. That means that we consider k th order conditions:

$$\left. \frac{d^j}{d\epsilon^j} \right|_{\epsilon=0} \mathcal{J}(\phi_\epsilon) = 0 \quad (j = 1, \dots, k)$$

and “ k -formal variations” (variations which satisfy restrictions up to k th order). A Lagrange multiplier theorem similar to 3.7 should then be formulated for “ k -formal optimality.” Obviously we speak of “ k -normality” in that case and in the CE case k th order local controllability will actually ensure k -normality cf. Proposition 5.2 for $k = 1$. Other approaches to higher order conditions can be found in the literature. [17] in particular is closely related to the approach suggested here.

Examples of practical optimal control problems to illustrate the given set up would easily lead to complicated calculations in coordinates, which do not essentially differ from the normal approach on \mathbb{R}^n except for possible lower dimension and less constraints (e.g. we can use the two-dimensional sphere instead of \mathbb{R}^3 with a restriction to the sphere). We did not search for examples where the coordinate-free approach might be profitable. Some may be found in [11]. The given approach can be generalized to the infinite horizon optimal control problem (see [6]). We expect that our approach might be particularly profitable there, for instance to obtain methods for computation of optimal feedback controls in infinite horizon problems.

Acknowledgments. Special gratitude is owed to H. Nijmeijer for numerous discussions about the subject of this paper. Furthermore, I would like to thank

Prof. J. C. Willems for introducing me to this research area and Dr. J. H. van Schuppen, Prof. F. Takens and an anonymous referee for valuable suggestions.

REFERENCES

- [1] V. I. ARNOLD (1973), *Ordinary Differential Equations*, MIT Press, Cambridge, MA.
- [2] R. M. BIANCHINI AND G. STEFANI (1982), *Local controllability for analytic families of vector fields*, Rep. Università degli Studi di Firenze.
- [3] G. A. BLISS (1946), *Lectures on the Calculus of Variations*, Univ. Chicago Press, Chicago.
- [4] R. W. BROCKETT (1977); *Control theory and analytical mechanics*, in 1976 Ames Research Center Conference on Geometric Control Theory, C. Martin and R. Hermann, eds., Math. Sci. Press.
- [5] ——— (1976), *Nonlinear systems and differential geometry*, Proc. IEEE, 64, pp. 61–72.
- [6] J. C. P. BUS (1983), *Infinite horizon optimal control on manifolds*, Paper presented at Berkeley-Ames Conference on Nonlinear Problems in Control and Fluid Dynamics.
- [7] E. CARTAN (1922), *Leçons sur les invariants intégraux*, Hermann, Paris.
- [8] C. CARATHÉODORY (1935), *Variationsrechnung und partielle Differentialgleichungen erster Ordnung*, Teubner, Leipzig.
- [9] P. CROUCH (1981), *Dynamical realizations of finite Volterra series*, this Journal, 19, pp. 177–202.
- [10] R. B. GARDNER (1983), *Differential geometric methods interfacing control theory*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman and H. J. Sussman, eds., Birkhäuser, Boston.
- [11] P. A. GRIFFITHS (1983), *Exterior Differential Systems and the Calculus of Variations*, Birkhäuser, Boston.
- [12] R. HERMANN (1962), *Some differential-geometric aspects of the Lagrange variational problem*, Illinois J. Math., 6, pp. 634–673.
- [13] ——— (1977), *Differential Geometry and the Calculus of Variations*, 2nd ed., Interdisc. Math. XVII, Math. Sci. Press.,
- [14] R. HERMANN AND A. KRENER (1977), *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, AC-22, pp. 728–740.
- [15] H. HERMES (1974), *On local and global controllability*, this Journal, 12, pp. 252–261.
- [16] ——— (1982), *On local controllability*, this Journal, 20, pp. 211–220.
- [17] A. J. KRENER (1977), *The high order maximal principle and its application to singular extremals*, this Journal, 15, pp. 256–293.
- [18] H. NIJMEIJER AND A. VAN DER SCHAFT (1982), *Controlled invariance for nonlinear systems*, IEEE Trans. Automat. Control. AC-27, pp. 904–914.
- [19] M. SPIVAK (1979), *A Comprehensive Introduction to Differential Geometry*, Vols. I and II, Publish or Perish Inc., Berkeley, CA.
- [20] F. TAKENS (1978), Unpublished course notes.
- [21] J. C. WILLEMS (1979), *System theoretic models for the analysis of physical systems*, Ric. di Autom., Spec. Issue on Systems theory and Physics.
- [22] L. C. YOUNG (1969), *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia.

RECURSIVE SYSTEM IDENTIFICATION AND ADAPTIVE CONTROL BY USE OF THE MODIFIED LEAST SQUARES ALGORITHM*

H. F. CHEN†

Abstract. In this paper we first use a weaker than persistent excitation condition to establish the strong consistency of MLS for MIMO stochastic systems without monitoring, and give a comparison between various conditions for consistency. Then we show the global convergence for adaptive tracking based on MLS and finally the adaptive tracking and strong consistency results are combined.

Key words. system identification, strong consistency, least squares, adaptive tracking

1. Introduction. Because of its potential for applications and its theoretical interest, recursive system identification and adaptive control of stochastic linear systems has been studied for many years. For a tracking problem for stochastic systems, Goodwin, Ramadge and Caines [12] have proved the global convergence and the asymptotic optimality of adaptive control by using a stochastic gradient (SG) algorithm. This is the stochastic analogue of their results for deterministic systems (see Goodwin et al [11]). Instead of the stochastic gradient algorithm, Sin and Goodwin [20] introduced an estimation algorithm called modified least squares (MLS) and obtained results similar to those in [12].

Concerning the strong consistency of recursive parameter estimation when the system noise is uncorrelated, most work has concentrated on least squares (LS) methods (Ljung [16], Moore [18], Chen [3, 4], Lai and Wei [15]). As far as the author knows, for strong consistency of LS, the conditions in Chen [4] and Lai and Wei [15] may be the weakest, although they are different from each other in both conditions and assertions. When the system noise is correlated, Solo [21] and Sin [19] have given conditions to guarantee the strong consistency of the approximate maximum likelihood (AML) algorithm under persistent excitation condition. By using conditions weaker than the persistent excitation condition, the author has obtained the strong consistency for LS (Chen [4]), for SG (Chen [6]), for MLS (Chen [5]) and for the stochastic approximation type algorithm (Chen [7]).

Probably, the weakest conditions are given in Chen [5], where we have introduced a new method for consistency proofs, namely, a method which combines the probabilistic method with the ordinary differential equation treatment (Ljung [17], Kushner and Clark [14]). It is worth noting that this combined method is also very successful in dealing with stochastic approximation problems (Chen [8], [10]).

There are some problems involved with the use of the persistent excitation conditions on the system input-output variables in consistency proofs, these include the difficulty of verification and the issue of the incompatibility of the conditions with control laws designed for a specific purpose. Caines [1] proposed to introduce a “dither” in controls to guarantee the persistent excitation condition and he called such controls “continually disturbed controls (CDC)”. This plan, later on, has been realized by Caines and Lafortune [2]. They used the SG algorithm to obtain the value of CDC, then they apply the AML algorithm to give the strong consistency estimates for parameters.

* Received by the editors December 27, 1982, and in revised form September 26, 1983.

† Institute of Systems Science, Academia Sinica and Department of Electrical Engineering, McGill University, Montreal, Quebec, Canada H3A 2A7.

In Chen [9] we have used only one estimation algorithm to give both the adaptive control and the strongly consistent estimate of the parameters. But the conditions used there are still not easy to verify.

In this paper we first prove a strong consistency theorem for MLS without monitoring, this is done by use of a weaker than persistent condition (numbered (28) below) which is compared with various persistent excitation-like conditions. Then we establish results similar to Goodwin, Ramadge and Caines [12] and Sin and Goodwin [20] for stochastic adaptive control. Finally, following Caines and Lafortune [2] we apply their suboptimal adaptive tracking control to guarantee the validity of (28) and hence the strong consistency. In contrast to [2], which uses two algorithms, we use only the MLS algorithm for both tracking and estimation purposes and this is done with generally weaker conditions.

We consider MIMO systems of the form

$$(1) \quad y_n + A_1 y_{n-1} + \cdots + A_p y_{n-p} = B_1 u_{n-1} + \cdots + B_q u_{n-q} + w_n + C_1 w_{n-1} + \cdots + C_r w_{n-r},$$

where y_n , u_n , w_n are m -, l - and m -dimensional respectively.

Set

$$(2) \quad A(z^{-1}) = I + A_1 z^{-1} + \cdots + A_p z^{-p},$$

$$(3) \quad B(z^{-1}) = B_1 + B_2 z^{-1} + \cdots + B_q z^{-q+1},$$

$$(4) \quad C(z^{-1}) = I + C_1 z^{-1} + \cdots + C_r z^{-r},$$

where z^{-1} is the shift-back operator and A_i , B_j , C_k are unknown matrices.

Let \mathcal{F}_n be the σ -field generated by the random variables $\{w_i, i \leq n\}$ which we write as

$$\mathcal{F}_n \triangleq \sigma\{w_i, i \leq n\}$$

and assume that

$$(5) \quad w_n = 0, \quad n < 0, \quad E(w_n | \mathcal{F}_{n-1}) = 0, \quad E(w_n^T w_n | \mathcal{F}_{n-1}) \leq k_0 r_{n-1}^\varepsilon, \quad 0 \leq \varepsilon < 1,$$

where and henceforth k_i denotes a constant independent of n and r_n defined later by (9) can increase unboundedly as $n \rightarrow \infty$.

Assume that u_n is \mathcal{F}_n -measurable, an assumption which includes the case of feedback, i.e., y -dependent controls.

We write

$$(6) \quad \theta^T = [-A_1 \dots -A_p B_1 \dots B_q C_1 \dots C_r],$$

$$(7) \quad \phi_n^T = [y_n^T y_{n-1}^T \dots y_{n-p+1}^T u_n^T \dots u_{n-q+1}^T y_n^T - \phi_{n-1}^T \theta_n \dots y_{n-r+1}^T - \phi_{n-r}^T \theta_{n-r+1}],$$

$$(8) \quad \phi_n^{0T} = [y_n^T y_{n-1}^T \dots y_{n-p+1}^T u_n^T \dots u_{n-q+1}^T w_n^T \dots w_{n-r+1}^T].$$

$\phi_{-1} = \phi_{-1}^0$ is any deterministic vector, where θ_n is the estimate for θ and is given by the MLS algorithm as follows:

Define r_n and R_n recursively by

$$(9) \quad r_n = r_{n-1} + \|\phi_n\|^2, \quad r_0 = 1, \quad \text{i.e. } r_n = 1 + \sum_{i=1}^n \|\phi_i\|^2,$$

and

$$(10) \quad R_n = dI \text{ with } d = mp + lq + mr.$$

If R_{n-1} has been defined, then calculate the matrix R'_n ,

$$(11) \quad R'_n = R_{n-1} - (1 + \phi_n^T R_{n-1} \phi_n)^{-1} R_{n-1} \phi_n \phi_n^T R_{n-1},$$

and its maximum and minimum eigenvalues $\mu_{\max}^n, \mu_{\min}^n$.

If

$$(12) \quad \mu_{\max}^n (\mu_{\min}^n)^{-1} \leq k_1, \quad \phi_n^T R'_n \phi_n \leq k_2 < 1$$

then define

$$(13) \quad R_n = R'_n, \quad a_n = 1,$$

but if the inequalities in (12) are not satisfied simultaneously, then define

$$(14) \quad R_n = \frac{r_{n-1}}{r_n} R_{n-1}, \quad a_n = (1 + \phi_n^T R_n \phi_n)^{-1}.$$

Finally the estimate θ_n for θ is given by

$$(15) \quad \theta_{n+1} = \theta_n + a_n R_n \phi_n (y_{n+1}^T - \phi_n^T \theta_n)$$

with any deterministic θ_0 .

We first establish some properties of the algorithm.

1) We note at once that

$$(16) \quad \lambda_{\max}^n (\lambda_{\min}^n)^{-1} \leq k_1,$$

where λ_{\max}^n and λ_{\min}^n are maximum and minimum eigenvalues of R_n respectively since if (14) takes place the ratio of eigenvalues for R_n and R_{n-1} is the same.

2) Set

$$(17) \quad \beta_i = \frac{\|\phi_i\|^2}{r_i}.$$

By use of Theorem 162 in Hardy et al. [13] it is immediate that

$$(18) \quad r_n \rightarrow \infty \Leftrightarrow \sum_{i=0}^{\infty} \beta_i = \infty.$$

We have to distinguish $r_n \rightarrow \infty$ from $\lim_{n \rightarrow \infty} r_n < \infty$. If $p \geq 1$ for the latter case, $y_n \rightarrow 0$, $u_n \rightarrow 0$, $C(z^{-1})w_n \rightarrow 0$ and θ_n converges to a limit which may be shown to depend upon θ_0 . Besides, if

$$\sum_{i=0}^{\infty} \|u_i\|^2 = \infty \quad \text{or} \quad \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \|w_i\|^2 > 0,$$

and the zeros of $\det C(z)$ are outside the closed unit disk, then $r_n \rightarrow \infty$. In Theorem 4, where controls are designed with a prescribed purpose, this property will be proved.

3)

$$(19) \quad r_n = \text{tr } R_n^{-1}.$$

We prove this fact inductively. From (9), (10) we have $r_0 = \text{tr } R_0^{-1}$. Assume (19) is true for $n-1$. If R_n is defined by (13) then by the matrix inversion formula

$$R'_n = [R_{n-1}^{-1} + \phi_n \phi_n^T]^{-1}$$

it follows that

$$\text{tr } R_n^{-1} = \text{tr } R'_n = \text{tr } (R_{n-1}^{-1} + \phi_n \phi_n^T) = \text{tr } R_{n-1}^{-1} + \|\phi_n\|^2 = r_n.$$

If R_n is defined by (14), then

$$\text{tr } R_n^{-1} = \frac{r_n}{r_{n-1}} \text{tr } R_{n-1}^{-1} = r_n.$$

Therefore (19) holds for any n .

4) There exists a constant $k_3 > 0$ such that

$$(20) \quad 1 - a_n \phi_n^T R_n \phi_n \geq k_3,$$

since, if (14) holds, then

$$\begin{aligned} 1 - a_n \phi_n^T R_n \phi_n &= (1 + \phi_n^T R_n \phi_n)^{-1} \geq (1 + \lambda_{\max}^n \|\phi_n\|^2)^{-1} \\ &\geq (1 + k_1 \lambda_{\min}^n \|\phi_n\|^2)^{-1} = (1 + k_1 \|R_n^{-1}\|^{-1} \|\phi_n\|^2)^{-1} \\ &\geq \left[1 + k_1 \frac{d \|\phi_n\|^2}{\text{tr } R_n^{-1}} \right]^{-1} \geq [1 + k_1 d]^{-1} > 0, \end{aligned}$$

while if (12) holds then (20) is immediate.

5) Set

$$(21) \quad \tilde{\theta}_n = \theta - \theta_n$$

and

$$(22) \quad \xi_n = y_n - w_n - \theta_n^T \phi_{n-1}.$$

Then since

$$\begin{aligned} C(z^{-1})(y_n - w_n - \theta_n^T \phi_{n-1}) &= \{(y_n - C(z^{-1})w_n) + (C(z^{-1}) - I)(y_n - \theta_n^T \phi_{n-1})\} - \theta_n^T \phi_{n-1} \\ &= \theta^T \phi_{n-1} - \theta_n^T \phi_{n-1} = \tilde{\theta}_n^T \phi_{n-1} \end{aligned}$$

we obtain the important relation

$$(23) \quad C(z^{-1})\xi_n = \tilde{\theta}_n^T \phi_{n-1}.$$

6) From (15) (21)–(23) we have

$$(24) \quad y_n - \theta_n^T \phi_{n-1} = (1 - a_{n-1} \phi_{n-1}^T R_{n-1} \phi_{n-1})(y_n - \theta_{n-1}^T \phi_{n-1}),$$

$$(25) \quad \tilde{\theta}_{n+1} = \tilde{\theta}_n - a_n (1 - a_n \phi_n^T R_n \phi_n)^{-1} R_n \phi_n (\xi_{n+1} + w_{n+1})^T.$$

7) The last property we should mention is the following:

$$(26) \quad E \|\theta_n\|^2 < \infty, \quad E \|\tilde{\theta}_n\|^2 < \infty \quad \forall n.$$

We shall prove it inductively. Since θ_0 is deterministic (26) holds for $n = 0$. Assume, it is true for n , i.e. $E \|\theta_n\|^2 < \infty$.

Noticing that $0 < a_n \leq 1$ and

$$(27) \quad R_n \leq k_1 \lambda_{\min}^n I \leq k_1 \frac{d}{\text{tr } R_n^{-1}} I = \frac{k_1 d}{r_n} I,$$

(15) yields

$$E \|\theta_{n+1}\|^2 \leq 2E \|\theta_n\|^2 + 4E \frac{k_1^2 d^2 \|\phi_n\|^2 (\|y_{n+1}\|^2 + \|\phi_n\|^2 \|\theta_n\|^2)}{r_n^2}.$$

By (9) $\|\phi_n\|^2 < r_n$, so if we can prove $E\|y_{n+1}\|^2/r_n < \infty$ we will have

$$E\|\theta_{n+1}\|^2 \leq 2E\|\theta_n\|^2 + 4k_1^2 d^2 \left(E \frac{\|y_{n+1}\|^2}{r_n} + E\|\theta_n\|^2 \right) < \infty$$

by use of the induction assumption.

Expressing y_{n+1} via its driving terms given by (1) we find that

$$\begin{aligned} E \frac{\|y_{n+1}\|^2}{r_n} &\leq E (\|A_1 y_n\| + \cdots + \|A_p\| \|y_{n-p+1}\| + \|B_1\| \|u_n\| + \cdots \\ &\quad + \|B_q\| \|u_{n-q+1}\| + \|w_{n+1}\| + \cdots + \|C_r\| \|w_{n-r+1}\|)^2 / r_n < \infty \end{aligned}$$

since

$$\frac{\|y_i\|^2}{r_n} < 1, \quad \frac{\|u_j\|^2}{r_n} < 1, \quad i = n, \dots, n-p+1, \quad j = n, \dots, n-q+1$$

and

$$E \frac{\|w_k\|^2}{r_n} \leq E \frac{\|w_k\|^2}{r_{k-1}} \leq E \left[\frac{1}{r_{k-1}} E(\|w_k\|^2 / \mathcal{F}_{k-1}) \right] \leq k_0 E \frac{1}{r_{k-1}^{1-\varepsilon}} \leq k_0$$

by (5), $k = n+1, \dots, n-r+1$.

2. Strong consistency. The proof of Theorem 1 below is similar to that in Chen [5] but here we have weakened the conditions and the definitions for ϕ_n and θ are modified. Because of its importance in our analysis, and the difficulty in obtaining [5], we present the complete proof of the following theorem in the interests of a self-contained exposition:

THEOREM 1. *For the system and algorithm (1)–(15) let conditions a) and b) be satisfied.*

a) *Either $r = 0$, or $r \geq 1$ and the transfer matrix $C^{-1}(z^{-1}) - \frac{1}{2}I$ is strictly positive real (i.e. zeros of $\det C(z)$ are outside the closed unit disk and $C^{-1}(e^{i\theta}) + C^{-1}(e^{-i\theta}) - I > 0$, $\theta \in [0, 2\pi]$).*

b) *There exist random variables $0 < \alpha < \infty$, $0 < \beta < \infty$ and $T > 0$ such that*

$$(28) \quad \sum_{i=m(t)}^{m(t+\alpha)} \frac{\phi_i \phi_i^T}{r_i} \geq \beta I \quad \forall t \geq T, \quad \forall \omega \in [\omega : r_n \rightarrow \infty]$$

where

$$(29) \quad m(t) = \max[n; t_n \leq t], \quad t \geq 0, \quad t_n = \sum_{i=0}^{n-1} \beta_i, \quad t_0 = 0.$$

Then for almost all $\omega \in [\omega : r_n \rightarrow \infty]$

$$(30) \quad \theta_n \xrightarrow[n \rightarrow \infty]{} \theta.$$

Proof. The proof divides into two parts. First we prove the boundedness of θ_n and the convergence of the stochastic Lyapunov function V_n

$$(31) \quad V_n = \frac{\text{tr } \tilde{\theta}_n^T R_{n-1}^{-1} \tilde{\theta}_n}{r_{n-1}}$$

by a probabilistic method, and then, by a treatment using the ODE method, we establish the strong consistency of $\{\theta_n; n \geq 1\}$.

Part 1. By (19), (26) we have $EV_n < \infty, \forall n$. If R_n is calculated according to (13), then

$$1 - \phi_n^T R_n \phi_n = (1 + \phi_n^T R_{n-1} \phi_n)^{-1},$$

$$(1 + \phi_n^T R_{n-1} \phi_n) R_n \phi_n = R_{n-1} \phi_n,$$

and hence

$$(32) \quad \begin{aligned} \tilde{\theta}_{n+1} &= \tilde{\theta}_n - R_{n-1} \phi_n (\xi_{n+1} + w_{n+1})^T, \\ \text{tr } \tilde{\theta}_{n+1}^T R_n^{-1} \tilde{\theta}_{n+1} &= \text{tr } \tilde{\theta}_n^T R_{n-1}^{-1} \tilde{\theta}_n - 2(\xi_{n+1} + w_{n+1})^T \tilde{\theta}_{n+1}^T \phi_n \\ &\quad - \phi_n^T R_{n-1} \phi_n \|\xi_{n+1} + w_{n+1}\|^2 + \|\tilde{\theta}_{n+1}^T \phi_n\|^2 \\ &\leq \text{tr } \tilde{\theta}_n^T R_{n-1}^{-1} \tilde{\theta}_n - 2\phi_n^T \tilde{\theta}_{n+1} (\xi_{n+1} + w_{n+1}) + \|\tilde{\theta}_{n+1}^T \phi_n\|^2. \end{aligned}$$

From this we obtain immediately

$$(33) \quad V_{n+1} \leq V_n - \frac{2}{r_n} \phi_n^T \tilde{\theta}_{n+1} (\xi_{n+1} + w_{n+1}) + \frac{\|\tilde{\theta}_{n+1}^T \phi_n\|^2}{r_n}.$$

If R_n is calculated according to (14), then

$$(34) \quad a_n(1 - a_n \phi_n^T R_n \phi_n)^{-1} = 1, \quad \tilde{\theta}_{n+1} = \tilde{\theta}_n - R_n \phi_n (\xi_{n+1} + w_{n+1})^T$$

and

$$\begin{aligned} \text{tr } \tilde{\theta}_{n+1}^T R_n^{-1} \tilde{\theta}_{n+1} &= \text{tr } \tilde{\theta}_n^T R_n^{-1} \tilde{\theta}_n - 2 \text{tr } (\xi_{n+1} + w_{n+1}) \phi_n^T [\tilde{\theta}_{n+1} + R_n \phi_n (\xi_{n+1} + w_{n+1})^T] \\ &\quad + (\xi_{n+1} + w_{n+1})^T \phi_n^T R_n \phi_n (\xi_{n+1} + w_{n+1}) \\ &\leq \text{tr } \tilde{\theta}_n^T R_n^{-1} \tilde{\theta}_n - 2 \text{tr } (\xi_{n+1} + w_{n+1}) \phi_n^T \tilde{\theta}_{n+1} \\ &\leq \text{tr } \tilde{\theta}_n^T R_n^{-1} \tilde{\theta}_n - 2\phi_n^T \tilde{\theta}_{n+1} (\xi_{n+1} + w_{n+1}) + \|\tilde{\theta}_{n+1}^T \phi_n\|^2. \end{aligned}$$

By using (14) from here we again obtain (33).

All $y_{n+1} - w_{n+1}$, R_n , $\tilde{\theta}_n$, ϕ_n and r_n are \mathcal{F}_n -measurable and so

$$\begin{aligned} E\left(\frac{w_{n+1}^T \tilde{\theta}_{n+1}^T \phi_n}{r_n} \middle/ \mathcal{F}_n\right) &= E\left[\frac{w_{n+1}^T (\tilde{\theta}_n^T - (y_{n+1} - w_{n+1} + w_{n+1} - \theta_n^T \phi_n) \phi_n^T R_n a_n) \phi_n}{r_n} \middle/ \mathcal{F}_n\right] \\ &= \frac{-a_n \phi_n^T R_n \phi_n}{r_n} E(\|w_{n+1}\|^2 / \mathcal{F}_n). \end{aligned}$$

From this and (5), (33) we find that

$$(35) \quad E(V_{n+1} | \mathcal{F}_n) \leq V_n - E\left\{\frac{2}{r_n} [\phi_n^T \tilde{\theta}_{n+1} (\xi_{n+1} - \frac{1}{2} \tilde{\theta}_{n+1}^T \phi_n)] \middle/ \mathcal{F}_n\right\} + 2k_0 \frac{\phi_n^T R_n \phi_n}{r_n^{1-\varepsilon}}.$$

Since $C^{-1}(z^{-1}) - \frac{1}{2}I$ is strictly positive real there exist k_4, k_5 such that

$$(36) \quad S_n \triangleq 2 \sum_{i=1}^n \phi_{i-1}^T \tilde{\theta}_i \left(\xi_i - \frac{1+k_4}{2} \tilde{\theta}_i^T \phi_{i-1} \right) + k_5 \geq 0 \quad \forall n,$$

where ξ_i is given by (23). Introduce

$$(37) \quad M_n = V_n + \frac{S_n}{r_{n-1}} + E\left(\sum_{i=1}^{\infty} \frac{2k_0}{r_i^{1-\varepsilon}} \phi_i^T R_i \phi_i \middle/ \mathcal{F}_n\right) - \sum_{i=1}^{n-1} \frac{2k_0}{r_i^{1-\varepsilon}} \phi_i^T R_i \phi_i + k_4 \sum_{i=1}^n \frac{\|\tilde{\theta}_i^T \phi_{i-1}\|^2}{r_{i-1}},$$

where

$$\sum_{i=1}^{\infty} \frac{\phi_i^T R_i \phi_i}{r_i^{1-\varepsilon}} \leq k_1 d \sum_{i=1}^{\infty} \frac{\|\phi_i\|^2}{r_i^{2-\varepsilon}} \leq \frac{k_1 d}{1-\varepsilon}$$

by (27) and Hardy et al. [13].

By using (35), (36) it is easy to see that (M_n, \mathcal{F}_n) is a nonnegative super martingale which converges to a finite limit a.s., hence

$$(38) \quad \sum_{i=1}^{\infty} \frac{\|\tilde{\theta}_i^T \phi_{i-1}\|^2}{r_{i-1}} < \infty \quad \text{a.s.},$$

and for any fixed ω , V_n is bounded. But by (27) and (31)

$$V_n = \frac{1}{r_{n-1}} \text{tr } \tilde{\theta}_n^T R_{n-1}^{-1} \tilde{\theta}_n \geq \frac{\text{tr } \tilde{\theta}_n^T \tilde{\theta}_n}{k_1 d},$$

hence for any ω $\|\tilde{\theta}_n\|$ is uniformly bounded in n . We now prove that V_n is convergent a.s. on $[\omega : r_n \rightarrow \infty]$. Set

$$(39) \quad F = \begin{bmatrix} -C_1 & I & 0 & \cdots & 0 \\ -C_2 & 0 & I & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ \vdots & \vdots & & \ddots & I \\ -C_r & 0 & & \cdots & 0 \end{bmatrix}, \quad r > 0,$$

$$(40) \quad F = \underbrace{[0]}_m, \quad F^0 = I, \quad r = 0, \\ G = \begin{cases} [I \underbrace{0 \cdots 0}_{mr}]_m, & r > 0, \\ I & r = 0. \end{cases}$$

It is easy to see that there exists a random mr -dimensional vector η_0 depending on initial values of $\{\xi_i\}$ such that

$$(41) \quad \xi_n = G\eta_n, \quad \eta_{n+1} = F\eta_n + G^T \tilde{\theta}_{n+1}^T \phi_n,$$

hence

$$(42) \quad \xi_{n+1} = G \sum_{i=0}^n F^{n-i} G^T \tilde{\theta}_{i+1}^T \phi_i + G F^{n+1} \eta_0.$$

By using the expression of a determinant by blocks it is directly verified that the eigenvalues of F coincide with the reciprocals of the zeros of $\det C(z)$. Hence there exists a $\rho \in (0, 1)$ such that

$$(43) \quad \|F^i\| \leq k_6 \rho^i.$$

Assume $\omega \in [\omega : r_n \rightarrow \infty]$.

From (37) it is easy to see that for the convergence of V_n we only need to prove

$$(44) \quad \frac{S_n}{r_{n-1}} \xrightarrow{n \rightarrow \infty} 0.$$

By (38) and the Kronecker lemma it is sufficient to show that

$$(45) \quad \frac{1}{r_{n-1}} \sum_{i=1}^n \phi_{i-1}^T \tilde{\theta}_i \xi_i \xrightarrow{n \rightarrow \infty} 0$$

for this by using (38) and the Schwarz inequality it is sufficient to prove that

$$(46) \quad \frac{1}{r_{n-1}} \sum_{i=1}^n \|\xi_i\|^2 \xrightarrow{n \rightarrow \infty} 0.$$

Taking notice of (42), (43) we have

$$\begin{aligned} \sum_{i=1}^{\infty} \frac{\|\xi_i\|^2}{r_{i-1}} &\leq 2k_6^2 \|\eta_0\|^2 \sum_{i=1}^{\infty} \frac{\rho^{2i}}{r_{i-1}} + 2k_6^2 \sum_{i=1}^{\infty} \frac{1}{r_i} \left(\sum_{j=1}^i \rho^{i-j} \|\tilde{\theta}_j^T \phi_{j-1}\| \right)^2 \\ &\leq \frac{2k_6^2 \|\eta_0\|^2}{1-\rho^2} + 2k_6^2 \sum_{i=1}^{\infty} \frac{1}{r_i} \left(\sum_{j=1}^i \rho^{i-j} \right) \left(\sum_{j=1}^i \rho^{j-1} \|\tilde{\theta}_{i-j+1}^T \phi_{i-j}\|^2 \right) \\ &\leq \frac{2k_6^2 \|\eta_0\|^2}{1-\rho^2} + \frac{2k_6^2}{1-\rho} \sum_{j=1}^{\infty} \left(\sum_{i=j}^{\infty} \frac{\|\tilde{\theta}_{i-j+1}^T \phi_{i-j}\|^2}{r_{i-j}} \right) \rho^{j-1} < \infty. \end{aligned}$$

From here, by the Kronecker lemma and (46) the convergence of V_n follows immediately.

Part 2. By using a probabilistic method we have succeeded in proving that

$$(47) \quad V_n \xrightarrow{n \rightarrow \infty} V < \infty \quad \text{a.s. on } [\omega : r_n \rightarrow \infty],$$

but we do not know whether V vanishes. Now we use the ODE method to show that on $[\omega : r_n \rightarrow \infty]$ $V = 0$ and from here it will follow that $\theta_n \rightarrow \theta$.

From (25), (42) we have

$$(48) \quad \tilde{\theta}_{n+1} = \tilde{\theta}_0 - \sum_{j=0}^n \alpha_j R_j \phi_j \left[\sum_{i=0}^j \phi_i^T \tilde{\theta}_{i+1} G F^{T(j-i)} G^T + \eta_0^T F^{T(j+1)} G^T + w_{j+1}^T \right],$$

where

$$(49) \quad 1 \leq \alpha_i \triangleq a_i (1 - a_i \phi_i^T R_i \phi_i)^{-1} \leq \frac{1}{k_3}$$

by (20). Assume $\omega \in [\omega : r_n \rightarrow \infty]$.

Let X_i be matrices and denote by $X(t)$ and $\bar{X}(t)$ the linear and the constant interpolations of $\{X_i\}$ respectively with interpolating length equal to β_i given by (17), i.e.

$$\begin{aligned} X(t_n) &= X_n, \\ (50) \quad X(t) &= \frac{t_{n+1}-t}{\beta_n} X_n + \frac{t-t_n}{\beta_n} X_{n+1}, \quad t \in [t_n, t_{n+1}], \\ \bar{X}(t) &= X_n, \quad t \in [t_n, t_{n+1}]. \end{aligned}$$

Denote the last two terms in (48) by J_{n+1} , H_{n+1} respectively and write

$$(51) \quad G_{n+1,i} = -\frac{1}{\beta_i} \sum_{j=i}^n \alpha_j R_j \phi_j \phi_i^T \tilde{\theta}_{i+1} G F^{T(j-i)} G^T, \quad G_{n,n} = 0.$$

Then from (48) it follows that

$$(52) \quad \tilde{\theta}_{n+1} = \tilde{\theta}_0 + \sum_{i=0}^n \beta_i G_{n+1,i} + J_{n+1} + H_{n+1}.$$

Let $G_{t,i}^0$ denote the linear interpolation of $\{G_{n,i}\}$ with the interpolating length $\{\beta_n\}$ for $t \geq t_i$ if i is fixed. When $t = t_k$, then $G_{t_k,i}^0 = G_{k,i}$. Now for fixed t denote by $\bar{G}_{t,s}^0$ the

constant interpolation for $\{G_{t,i}^0\}$ on $[0, t]$ with interpolating length $\{\beta_i\}$. Thus for $t = t_{n+1}$

$$(53) \quad \int_0^t \bar{G}_{t,s}^0 ds = - \sum_{i=0}^n \sum_{j=i}^n \alpha_j R_j \phi_j \phi_i^T \tilde{\theta}_{i+1} G F^{T(j-i)} G^T.$$

Define an interpolating function $\tilde{\theta}(t)$ for $t \geq 0$ as follows:

$$(54) \quad \tilde{\theta}(t) = \tilde{\theta}_0 + \int_0^t \bar{G}_{t,s}^0 ds + J(t) + H(t),$$

where $J(t)$ and $H(t)$ are defined according to (50). From (18) we know that $\tilde{\theta}(t)$ is defined for all $t \geq 0$. Clearly,

$$(55) \quad \tilde{\theta}(t_n) = \tilde{\theta}_n.$$

Define a family of matrix functions $\{\tilde{\theta}_n(t)\}$ by shifting the argument of $\tilde{\theta}(t)$ to the left:

$$(56) \quad \tilde{\theta}_n(t) = \tilde{\theta}(t+n), \quad t \geq 0.$$

In order to use the Arzela-Ascoli theorem we prove the following.

LEMMA. For any fixed $\omega \in [\omega : r_n \rightarrow \infty]$ $\{\tilde{\theta}_n(t)\}$ is uniformly bounded and equicontinuous.

Proof. Let $t_n \leq t < t_{n+1}$. Then for $s \in [t_n, t]$

$$\bar{G}_{t,s}^0 = G_{t,n}^0 = \frac{t-t_n}{\beta_n} G_{n+1,n} + \frac{t_{n+1}-t}{\beta_n} G_{n,n} = -\frac{(t-t_n)}{\beta_n} \alpha_n R_n \phi_n \phi_n^T \tilde{\theta}_{n+1}.$$

Hence

$$(57) \quad \begin{aligned} & \left\| \int_0^t \bar{G}_{t,s}^0 ds - \int_0^{t_n} \bar{G}_{t_n,s}^0 ds \right\| \\ & \leq \left\| \int_0^{t_n} (\bar{G}_{t,s}^0 - \bar{G}_{t_n,s}^0) ds \right\| + \left\| \int_{t_n}^t \bar{G}_{t,s}^0 ds \right\| \\ & \leq k_7 \frac{\|\phi_n\| \|\phi_n^T \tilde{\theta}_{n+1}\|}{r_n} + \sum_{i=0}^{n-1} \beta_i \|G_{t,i}^0 - G_{t_n,i}^0\| \\ & = k_7 \frac{\|\phi_n\| \|\phi_n^T \tilde{\theta}_{n+1}\|}{r_n} + \sum_{i=0}^{n-1} \frac{t-t_n}{\beta_n} \|\alpha_n R_n \phi_n \phi_i^T \tilde{\theta}_{i+1} G F^{T(n-i)} G^T\| \\ & \leq k_7 \frac{\|\phi_n\| \|\phi_n^T \tilde{\theta}_{n+1}\|}{r_n} + k_6 k_7 \sum_{i=0}^{n-1} \frac{\|\phi_n\|}{r_n^{1/2}} \frac{\|\phi_i^T \tilde{\theta}_{i+1}\|}{r_i^{1/2}} \rho^{(n-i)/2} \\ & \leq k_7 \frac{\|\tilde{\theta}_{n+1}^T \phi_n\|}{r_n^{1/2}} + \frac{k_6 k_7 \sqrt{\rho}}{1 - \sqrt{\rho}} \left[\sum_{i=0}^N \frac{\|\tilde{\theta}_{i+1}^T \phi_i\|^2}{r_i} \rho \frac{n-i}{2} + \sum_{i=N+1}^{n-1} \frac{\|\tilde{\theta}_{i+1}^T \phi_i\|^2}{r_i} \right]^{1/2} \end{aligned}$$

where $k_7 = k_1 d / k_3$. Letting $n \rightarrow \infty$ and then $N \rightarrow \infty$, and noticing (38) we conclude that for $t \in [t_n, t_{n+1})$

$$(58) \quad \left\| \int_0^t \bar{G}_{t,s}^0 ds - \int_0^{t_n} \bar{G}_{t_n,s}^0 ds \right\| \xrightarrow{n \rightarrow \infty} 0.$$

J_n is clearly convergent a.s. while the convergence of H_n follows from the martingale convergence theorem since

$$\sum_{j=0}^{\infty} E(\|\alpha_j R_j \phi_j w_{j+1}^T\|^2 | \mathcal{F}_j) \leq \frac{k_0 k_1^2 d^2}{k_3^2} \sum_{j=0}^{\infty} \frac{\|\phi_j\|^2}{r_j^{2-\varepsilon}} < \infty.$$

Hence both $J(t)$ and $H(t)$ are bounded uniformly continuous functions on $[0, \infty)$ which, in addition, converge to limits as $t \rightarrow \infty$. Thus

$$(59) \quad \sup_{t \geq s} \|H(t) - H(s)\| \xrightarrow{s \rightarrow \infty} 0, \quad \sup_{t \geq s} \|J(t) - J(s)\| \xrightarrow{s \rightarrow \infty} 0.$$

For any $t \in [0, \infty)$ we can find n satisfying $t_n \leq t < t_{n+1}$ since $t_n \rightarrow \infty$ as $n \rightarrow \infty$. Hence

$$(60) \quad \|\tilde{\theta}(t) - \tilde{\theta}(t_n)\| \leq \left\| \int_0^t \bar{G}_{t,s}^0 ds - \int_0^{t_n} \bar{G}_{t_n,s}^0 ds \right\| + \|J(t) - J(t_n)\| + \|H(t) - H(t_n)\|.$$

By taking account of $\|\tilde{\theta}(t_n)\| = \|\tilde{\theta}_n\|$ and (58)–(60) it follows that $\tilde{\theta}(t)$ is uniformly bounded for $t \in [0, \infty)$ and hence $\tilde{\theta}_n(t)$ is uniformly bounded for $n \geq 0$ and $t \geq 0$. For any n , $\tilde{\theta}_n(t)$ is continuous, hence in order to prove equicontinuity of $\{\tilde{\theta}_n(t)\}$ we only need to prove it for $n \geq N$ with sufficiently large N .

Let $\Delta > 0$, $t \in [0, \infty)$.

$$(61) \quad \begin{aligned} \|\tilde{\theta}_n(t+\Delta) - \tilde{\theta}_n(t)\| \leq & \left\| \int_0^{t_{m(t+n+\Delta)}} \bar{G}_{t_{m(t+n+\Delta)},s}^0 ds - \int_0^{t_{m(t+n)}} \bar{G}_{t_{m(t+n)},s}^0 ds \right\| \\ & + \left\| \int_0^{t+n+\Delta} \bar{G}_{t+n+\Delta,s}^0 ds - \int_0^{t_{m(t+n+\Delta)}} \bar{G}_{t_{m(t+n+\Delta)},s}^0 ds \right\| \\ & + \left\| \int_0^{t+n} \bar{G}_{t+n,s}^0 ds - \int_0^{t_{m(t+n)}} \bar{G}_{t_{m(t+n)},s}^0 ds \right\| \\ & + \|J(t+n+\Delta) - J(t+n)\| + \|H(t+n+\Delta) - H(t+n)\|. \end{aligned}$$

By (58), (59) all terms except the first one on the right-hand side of (61) tend to zero as $n \rightarrow \infty$.

By use of (53)

$$(62) \quad \begin{aligned} & \left\| \int_0^{t_{m(t+n+\Delta)}} \bar{G}_{t_{m(t+n+\Delta)},s}^0 ds - \int_0^{t_{m(t+n)}} \bar{G}_{t_{m(t+n)},s}^0 ds \right\| \\ & \leq \left\| \sum_{j=m(t+n)}^{m(t+n+\Delta)-1} \sum_{i=0}^j \alpha_j R_j \phi_j \phi_i^T \tilde{\theta}_{i+1} G F^{T(j-i)} G^T \right\| \\ & \leq k_6 k_7 \sum_{j=m(t+n)}^{m(t+n+\Delta)-1} \sum_{i=0}^j \beta_j^{1/2} \frac{\|\tilde{\theta}_{i+1}^T \phi_i\|}{r_i^{1/2}} \rho^{(j-i)/2} \\ & \leq k_6 k_7 \left(\sum_{j=m(t+n)}^{m(t+n+\Delta)-1} \sum_{i=0}^j \beta_j \rho^{(j-i)/2} \right)^{1/2} \left(\sum_{j=m(t+n)}^{m(t+n+\Delta)-1} \sum_{i=0}^j \frac{\|\tilde{\theta}_{i+1}^T \phi_i\|^2}{r_i} \rho^{(j-i)/2} \right)^{1/2} \\ & \leq k_6 k_7 \left(\frac{1+\Delta}{1-\sqrt{\rho}} \right)^{1/2} \left[\sum_{i=m(t+n)}^{m(t+n+\Delta)-1} \sum_{j=i}^{m(t+n+\Delta)-1} \frac{\|\tilde{\theta}_{i+1}^T \phi_i\|^2}{r_i} \rho^{(j-i)/2} \right. \\ & \quad \left. + \sum_{i=0}^{m(t+n)-1} \sum_{j=m(t+n)}^{m(t+n+\Delta)-1} \frac{\|\tilde{\theta}_{i+1}^T \phi_i\|^2}{r_i} \rho^{(j-i)/2} \right]^{1/2} \\ & \leq k_6 k_7 \frac{\sqrt{1+\Delta}}{1-\sqrt{\rho}} \left[\sum_{i=m(t+n)}^{m(t+n+\Delta)-1} \frac{\|\tilde{\theta}_{i+1}^T \phi_i\|^2}{r_i} + \sum_{i=0}^N \frac{\|\tilde{\theta}_{i+1}^T \phi_i\|^2}{r_i} \rho^{(m(t+n)-i)/2} \right. \\ & \quad \left. + \sum_{i=N+1}^{m(t+n)-1} \frac{\|\tilde{\theta}_{i+1}^T \phi_i\|^2}{r_i} \right]^{1/2} \xrightarrow[n \rightarrow \infty]{N \rightarrow \infty} 0 \end{aligned}$$

since (38) and $0 < \rho < 1$.

A similar result is true for $\Delta < 0$. Thus the equicontinuity of $\{\tilde{\theta}_n(t)\}$ has been established. \square

According to the Arzela-Ascoli theorem for any fixed $\omega \in [\omega : r_n \rightarrow \infty]$ there exists a subsequence $\tilde{\theta}_{n_k}(t)$ of $\{\tilde{\theta}_n(t)\}$ and a continuous matrix $\theta(t)$ which is the uniform limit of $\tilde{\theta}_{n_k}(t)$ in any finite interval.

From (58), (59), (61), (62) it follows that

$$\|\theta(t+\Delta) - \theta(t)\| = \lim_{k \rightarrow \infty} \|\tilde{\theta}_{n_k}(t+\Delta) - \tilde{\theta}_{n_k}(t)\| = 0.$$

Thus $\theta(t)$ is a constant matrix:

$$(63) \quad \theta(t) \equiv \theta^0.$$

Now we show that $\theta^0 = 0$; for a fixed $t \in [0, \infty)$ we have

$$(64) \quad \lim_{k \rightarrow \infty} \left\| \sum_{i=m(t+n_k)}^{m(t+n_k+\alpha)} \frac{\phi_i \phi_i^T}{r_i} \tilde{\theta}_{i+1} \right\| \leq \sqrt{2+\alpha} \lim_{k \rightarrow \infty} \left(\sum_{i=m(t+n_k)}^{m(t+n_k+\alpha)} \frac{\|\tilde{\theta}_{i+1}^T \phi_i\|^2}{r_i} \right)^{1/2} = 0.$$

Notice that for all

$$(65) \quad \begin{aligned} i \in [0, 1, \dots, m(t+n_k+\alpha) - m(t+n_k)] &\triangleq S, \\ t < t_{m(t+n_k)+1} - n_k &\leq t_{m(t+n_k)+1+i} - n_k = t_{m(t+n_k)} + \sum_{j=m(t+n_k)}^{m(t+n_k)+i} \beta_j - n_k \\ &\leq t_{m(t+n_k)} + \sum_{j=m(t+n_k)}^{m(t+n_k+\alpha)} \beta_j - n_k \leq t + \alpha + 2, \end{aligned}$$

in other words,

$$(66) \quad t_{m(t+n_k)+i+1} - n_k \in [t, t + \alpha + 2] \quad \forall i \in S.$$

Hence

$$\tilde{\theta}_{n_k}(t_{m(t+n_k)+i+1} - n_k) \xrightarrow[k \rightarrow \infty]{} \theta^0$$

and by (55), (56)

$$(67) \quad \tilde{\theta}_{m(t+n_k)+i+1} \xrightarrow[k \rightarrow \infty]{} \theta^0$$

uniformly in $i \in S$. Consequently, we assert

$$(68) \quad \lim_{k \rightarrow \infty} \sum_{i=m(t+n_k)}^{m(t+n_k+\alpha)} \beta_i \|\tilde{\theta}_{i+1} - \theta^0\| = 0.$$

From (64), (67) it is easy to conclude that

$$(69) \quad \theta^{0T} \lim_{k \rightarrow \infty} \sum_{i=m(t+n_k)}^{m(t+n_k+\alpha)} \frac{\phi_i \phi_i^T}{r_i} \theta^0 = 0.$$

It is worth remarking that until this point all results have been obtained without invoking condition b).

By using (28) from (69) it follows immediately that

$$(70) \quad \theta^0 = 0,$$

and $\tilde{\theta}(t+n_k) \rightarrow 0$ uniformly in $t \in [a, b]$, where $[a, b]$ is any finite interval. By paying attention to $\tilde{\theta}(t_n) = \tilde{\theta}_n$ it is easy to see that for any fixed $\omega \in [\omega : r_n \rightarrow \infty]$ there exists a

subsequence

$$(71) \quad \tilde{\theta}_{m_k} \xrightarrow[k \rightarrow \infty]{} 0.$$

From here by (27), (31), (47) we conclude that

$$\begin{aligned} \overline{\lim}_{n \rightarrow \infty} \text{tr } \tilde{\theta}_n^T \tilde{\theta}_n &\leq k_1 d \lim_{n \rightarrow \infty} V_n = k_1 d \lim_{k \rightarrow \infty} \frac{1}{r_{m_k-1}} \text{tr } \tilde{\theta}_{m_k}^T R_{m_k-1}^{-1} \tilde{\theta}_{m_k} \\ &\leq k_1 d \lim_{k \rightarrow \infty} \text{tr } \tilde{\theta}_{m_k}^T \tilde{\theta}_{m_k} = 0. \end{aligned}$$

Hence

$$\theta_n \xrightarrow[n \rightarrow \infty]{} \theta \quad \forall \omega \in [\omega : r_n \rightarrow \infty]. \quad \square$$

At first glance condition b) is rather complicated, but indeed it is weaker than the persistent excitation condition, which means that

$$(72) \quad \frac{1}{n} \sum_{i=1}^n \phi_i^0 \phi_i^{0T} \xrightarrow[n \rightarrow \infty]{} R^0 > 0 \quad \text{a.s.}$$

or

$$(73) \quad \frac{1}{n} \sum_{i=1}^n \phi_i \phi_i^T \xrightarrow[n \rightarrow \infty]{} R > 0 \quad \text{a.s.}$$

Denote the maximum and minimum eigenvalues of

$$\sum_{i=1}^n \phi_i^0 \phi_i^{0T} + d^{-1} I \quad \text{and} \quad \sum_{i=1}^n \phi_i \phi_i^T + d^{-1} I \quad \text{by } \nu_{\max}^{0n},$$

ν_{\min}^{0n} , ν_{\max}^n and ν_{\min}^n respectively, we introduce conditions c) and c⁰).

c⁰) There exists a random variable $0 < \gamma < \infty$, such that

$$\nu_{\max}^{0n} / \nu_{\min}^{0n} \leq \gamma \quad \forall n \geq 0.$$

c) There exists a random variable $0 < \gamma < \infty$ such that

$$\nu_{\max}^n / \nu_{\min}^n \leq \gamma \quad \forall n \geq 0.$$

It is clear that c⁰) is weaker than (72) and c) weaker than (73).

THEOREM 2. *For the MLS algorithm under condition a) of Theorem 1, Conditions c⁰) and c) are equivalent on $[\omega : r_n \rightarrow \infty]$ and they imply condition b).*

Proof. Let ϕ_n^ξ be of the same dimension as ϕ_n :

$$\phi_n^\xi = [0 \cdots 0 \quad \xi_n^T \quad \xi_{n-1}^T \cdots \xi_{n-r+1}^T]^T.$$

From (7), (8), (22) it is seen that

$$(74) \quad \phi_n^0 = \phi_n - \phi_n^\xi.$$

Assume $\omega \in [\omega : r_n \rightarrow \infty]$. Set

$$(75) \quad r_n^0 = 1 + \sum_{i=1}^n \|\phi_i^0\|^2.$$

By (19), (46), (74) we have

$$(76) \quad \frac{r_n^0}{r_n} = \frac{r_n - 2 \sum_{i=1}^n \phi_i^T \phi_i^\xi + \sum_{i=1}^n \|\phi_i^\xi\|^2}{r_n} \xrightarrow[n \rightarrow \infty]{} 1.$$

Thus

$$(77) \quad \left\| \frac{1}{r_n^0} \sum_{i=1}^n \phi_i^0 \phi_i^{0T} - \frac{1}{r_n} \sum_{i=1}^n \phi_i \phi_i^T \right\| \\ \leq \left\| \frac{1}{r_n} \sum_{i=1}^n \phi_i^0 \phi_i^{0T} - \frac{1}{r_n} \sum_{i=1}^n \phi_i \phi_i^T \right\| + \left\| \frac{1}{r_n} \sum_{i=1}^n \phi_i^0 \phi_i^{0T} - \frac{1}{r_n^0} \sum_{i=1}^n \phi_i^0 \phi_i^{0T} \right\| \xrightarrow{n \rightarrow \infty} 0$$

since by (46) and the Schwarz inequality

$$\left\| \frac{1}{r_n} \left(\sum_{i=1}^n \phi_i^0 \phi_i^{0T} - \sum_{i=1}^n \phi_i \phi_i^T \right) \right\| \leq \frac{1}{r_n} \left\| \sum_{i=1}^n \phi_i^\xi \phi_i^{\xi T} - \phi_i \phi_i^{\xi T} - \phi_i^\xi \phi_i^T \right\| \xrightarrow{n \rightarrow \infty} 0$$

and

$$\left\| \frac{1}{r_n} \sum_{i=1}^n \phi_i^0 \phi_i^{0T} - \frac{1}{r_n^0} \sum_{i=1}^n \phi_i^0 \phi_i^{0T} \right\| \leq \left(\frac{r_n^0}{r_n} - 1 \right) \left\| \frac{1}{r_n^0} \sum_{i=1}^n \phi_i^0 \phi_i^{0T} \right\| \xrightarrow{n \rightarrow \infty} 0.$$

Now suppose that c^0 is true. Then

$$(78) \quad I \geq \frac{1}{r_n^0} \left(\sum_{i=1}^n \phi_i^0 \phi_i^{0T} + \frac{1}{d} I \right) \geq \frac{d}{\nu_{\max}^{0n}} \left(\sum_{i=1}^n \phi_i^0 \phi_i^{0T} + \frac{1}{d} I \right) \\ \geq \frac{d}{\gamma \nu_{\min}^{0n}} \left(\sum_{i=1}^n \phi_i^0 \phi_i^{0T} + \frac{1}{d} I \right) \geq \frac{d}{\gamma} I.$$

From here and (76), (77) we conclude that for any fixed $\omega \in [\omega : rn \rightarrow \infty]$ we can find a positive number $\delta > 0$ such that

$$(79) \quad \delta I \leq \frac{1}{r_n} \left(\sum_{i=1}^n \phi_i \phi_i^T + \frac{1}{d} I \right) \leq I \quad \forall n.$$

Hence we have

$$\nu_{\max}^n / \nu_{\min}^n \leq \nu_{\max}^n / \delta r_n \leq 1 / \delta \quad \forall n,$$

and this shows that c^0 implies c). The proof for the converse inclusion is completely similar.

Now assume that condition c) or c^0 is fulfilled, then, by summation by parts

$$(80) \quad \sum_{i=m(t)}^{m(t+\alpha)} \frac{\phi_i \phi_i^T}{r_i} = \sum_{i=m(t)}^{m(t+\alpha)} \frac{1}{r_i} \left(\sum_{j=1}^i \phi_j \phi_j^T - \sum_{j=1}^{i-1} \phi_j \phi_j^T \right) \\ = \frac{1}{r_{m(t+\alpha)}} \sum_{j=1}^{m(t+\alpha)} \phi_j \phi_j^T - \frac{1}{r_{m(t)}} \sum_{j=1}^{m(t)-1} \phi_j \phi_j^T \\ + \sum_{i=m(t)+1}^{m(t+\alpha)} \sum_{j=1}^{i-1} \phi_j \phi_j^T \left(\frac{1}{r_{i-1}} - \frac{1}{r_i} \right) \\ \geq \sum_{i=m(t)+1}^{m(t+\alpha)} \sum_{j=1}^{i-1} \phi_j \phi_j^T \frac{\|\phi_i\|^2}{r_{i-1} r_i} - I \\ \geq \sum_{i=m(t)+1}^{m(t+\alpha)} \left(\delta - \frac{d}{r_{i-1}} \right) \frac{\|\phi_i\|^2}{r_i} I - I,$$

where c) or c^0 yields the last inequality via (79).

Since $r_n \rightarrow \infty$ there exists $T > 0$, such that

$$(81) \quad \left(\delta - \frac{d}{r_{i-1}} \right) \geq \frac{\delta}{2} \quad \forall i \geq m(T).$$

Then we take

$$(82) \quad \alpha > \frac{2}{\delta} + 1, \quad \beta \triangleq \frac{\delta}{2}(\alpha - 1) - 1 > 0$$

and from (80)–(82) we obtain condition b):

$$\sum_{i=m(t)}^{m(t+\alpha)} \frac{\phi_i \phi_i^T}{r_i} \geq \left(\frac{\delta}{2} \sum_{i=m(t)+1}^{m(t+\alpha)} \frac{\|\phi_i\|^2}{r_i} - 1 \right) I \geq \left[\frac{\delta}{2}(\alpha - 1) - 1 \right] I = \beta I \quad \forall t \geq T. \quad \square$$

3. Adaptive tracking. We still consider the system and algorithm described by (1)–(15) but with $l \leq m$ and $p \geq 1$. Let $\{y_i^*\}$ be a bounded deterministic reference sequence. We need the following condition d).

d) B_1 is of full rank, the zeros of $B_1^+ B(z)$ are outside the closed unit disk and the \mathcal{F}_n -measurable control u_n is selected such that

$$(83) \quad \theta_n^T \phi_n = y_{n+1}^*.$$

THEOREM 3. Assume that for system and algorithm (1)–(15) conditions a) and d) are satisfied, and that w satisfies

$$(84) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n w_i w_i^T = R > 0.$$

Then $r_n \rightarrow \infty$ and

$$(85) \quad \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \|u_i\|^2 < \infty, \quad \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \|y_i\|^2 < \infty,$$

$$(86) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (y_i - y_i^*)(y_i - y_i^*)^T = R.$$

If, in addition, condition b) (or c) or c⁰)) is satisfied, then

$$(87) \quad \theta_n \xrightarrow[n \rightarrow \infty]{} \theta \quad \text{a.s.}$$

Proof. First of all we note that $r_n \rightarrow \infty$ a.s. follows from (84) immediately as mentioned in the Introduction. Since $C(z)$ and $B_1^+ B(z)$ are asymptotically stable, then from (1), (22), (23), (84) it can be seen that there exist constants k_8, k_9, k_{10}, k_{11} (which may depend on ω if R so does) such that

$$(88) \quad \frac{1}{n} \sum_{i=1}^n \|u_i\|^2 \leq \frac{k_8}{n} \sum_{i=1}^n \|y_{i+1}\|^2 + k_9,$$

$$(89) \quad \frac{1}{n} \sum_{i=1}^n \|y_i - \theta_i^T \phi_{i-1}\|^2 \leq \frac{k_{10}}{n} \sum_{i=1}^n \|\tilde{\theta}_i^T \phi_{i-1}\|^2 + k_{11}.$$

By (83), (24) we have

$$(90) \quad (y_{n+1} - y_{n+1}^*)(y_{n+1} - y_{n+1}^*)^T = \frac{1}{(1 - a_n \phi_n^T R_n \phi_n)^2} (y_{n+1} - \theta_{n+1}^T \phi_n)(y_{n+1} - \theta_{n+1}^T \phi_n)^T.$$

By using the boundedness of $\{y_n^*\}$ and (20), (89), (90) we conclude that there are constants k_{12} , k_{13} such that

$$(91) \quad \frac{1}{n} \sum_{i=1}^n \|y_{i+1}\|^2 \leq \frac{k_{12}}{n} \sum_{i=1}^n \|\tilde{\theta}_{i+1}^T \phi_i\|^2 + k_{13},$$

combining (88), (89), (91) we have

$$\frac{r_n}{n} \leq \frac{k_{14}}{n} \sum_{i=0}^n \|\tilde{\theta}_{i+1}^T \phi_i\|^2 + k_{15} = k_{14} \frac{r_n}{n} \frac{1}{r_n} \sum_{i=0}^n \|\tilde{\theta}_{i+1}^T \phi_i\|^2 + k_{15},$$

and by (38) have

$$\frac{r_n}{n} \leq k_{15} / \left(1 - k_{14} \frac{1}{r_n} \sum_{i=0}^n \|\tilde{\theta}_{i+1}^T \phi_i\|^2 \right) \xrightarrow{n \rightarrow \infty} k_{15}.$$

Hence

$$(92) \quad \overline{\lim}_{n \rightarrow \infty} \frac{r_n}{n} < \infty \quad \text{a.s.},$$

and from (46)

$$(93) \quad \frac{1}{n} \sum_{i=1}^n \|\xi_i\|^2 \xrightarrow{n \rightarrow \infty} 0 \quad \text{a.s.}$$

By (84), (93), (22) we have

$$(94) \quad \overline{\lim}_{n \rightarrow \infty} \frac{n}{r_n} \leq \overline{\lim}_{n \rightarrow \infty} \frac{n}{\sum_{i=1}^n \|y_i - \theta_i^T \phi_{i-1}\|^2} = \overline{\lim}_{n \rightarrow \infty} \frac{n}{\sum_{i=1}^n \|\xi_i + w_i\|^2} = \frac{1}{\text{tr } R} < \infty,$$

and by (84)

$$(95) \quad \frac{1}{n} \|w_i\|^2 = \frac{1}{n} \sum_{i=1}^n \|w_i\|^2 - \frac{n-1}{n} \frac{1}{n-1} \sum_{i=1}^{n-1} \|w_i\|^2 \rightarrow 0.$$

Similarly, from (93) we have

$$(96) \quad \frac{1}{n} \|\xi_n\|^2 \xrightarrow{n \rightarrow \infty} 0 \quad \text{a.s.}$$

Hence from (22) we see

$$(97) \quad \frac{1}{n} \|y_n - \theta_n^T \phi_{n-1}\|^2 \xrightarrow{n \rightarrow \infty} 0 \quad \text{a.s.}$$

By (20) and the boundedness of $\{y_i^*\}$ from (90) it follows that

$$(98) \quad \frac{\|y_n\|^2}{n} \xrightarrow{n \rightarrow \infty} 0 \quad \text{a.s.},$$

and hence

$$(99) \quad \frac{\|u_n\|^2}{n} \xrightarrow{n \rightarrow \infty} 0 \quad \text{a.s.}$$

From (97)–(99) we obtain

$$(100) \quad \frac{\|\phi_n\|^2}{n} \xrightarrow{n \rightarrow \infty} 0 \quad \text{a.s.}$$

Denoting

$$b_n \triangleq a_n \phi_n^T R_n \phi_n$$

and using (27), (94), (100), we have

$$(101) \quad |b_n| \leq \|R_n\| \|\phi_n\|^2 \leq \frac{k_1 d \|\phi_n\|^2 n}{r_n n} \xrightarrow{n \rightarrow \infty} 0.$$

From (90), (22) we have

$$(102) \quad \begin{aligned} & \frac{1}{n} \sum_{i=1}^n (y_i - y_i^*)(y_i - y_i^*)^T \\ &= \frac{1}{n} \sum_{i=1}^n \frac{(\xi_i + w_i)(\xi_i + w_i)^T}{(1 - b_{i-1})} \\ &= \frac{1}{n} \sum_{i=1}^n \left(\frac{\xi_i}{1 - b_{i-1}} + \frac{b_{i-1}}{1 - b_{i-1}} w_i + w_i \right) \left(\frac{\xi_i}{1 - b_{i-1}} + \frac{b_{i-1}}{1 - b_{i-1}} w_i + w_i \right)^T. \end{aligned}$$

By (84), (96), (101) it is easy to see

$$(103) \quad \frac{1}{n} \sum_{i=1}^n \left(\frac{\|\xi_i\|}{1 - b_{i-1}} \right)^2 \xrightarrow{n \rightarrow \infty} 0, \quad \frac{1}{n} \sum_{i=1}^n \left(\frac{b_{i-1}}{1 - b_{i-1}} \right)^2 \|w_i\|^2 \rightarrow 0;$$

thus (86) immediately follows from (102), and (85) follows from (86) since the boundedness of $\{y_i^*\}$ and the asymptotical stability of $B_1^+ B(z)$.

Strong consistency follows from Theorems 1 and 2. \square

Remark. Theorem 3 in a certain sense is a multidimensional version of the result in Sin and Goodwin [20], but here we have stronger results; namely in (86) the predicted error is unconditioned. The conditions imposed on $\{w_i\}$ differ slightly from those in Sin and Goodwin [20], namely, we use condition (84) rather than

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n w_i w_i^T < \infty,$$

but the sequence of conditional covariances of w_i need not be bounded (see condition (5)).

4. Adaptive tracking and parameter estimation. For the strong consistency of MLS in Theorems 1 and 3 we imposed condition b), which is weaker than the persistent excitation condition as shown in Theorem 2, but it still is inconvenient to verify. Now for adaptive tracking and consistency of estimates we apply the CDC used in Caines [1] and Caines and Lafortune [2], to obtain conditions imposed on coefficients of the system only.

Let $\{w_i\}, \{v_i\} - \infty < i < \infty$ be two m -dimensional iid random processes which are mutually independent with

$$(104) \quad E v_i = E w_i = 0, \quad E w_i w_i^T = R_1, \quad E v_i v_i^T = R_2, \quad \forall i$$

and let $\mathcal{F}_n = \sigma\{w_i, v_i, i \leq n\}$. With such a strengthening of our hypotheses let us consider (1)–(15).

The essentials of CDC are that \mathcal{F}_n -measurable controls $\{u_n\}$ are defined by

$$(105) \quad \theta_n^T \phi_n = y_{n+1}^* + v_n$$

rather than (83). We need conditions e), f) on the coefficients of the system.

e). The same condition as d)—with (83) replaced by (105)—is assumed to hold.

f). $B_1^+ A(z)$ and $B_1^+ B(z)$ are left coprime and $B_1^+ A_p$ and $B_1^+ B_q$ are of full rank.

THEOREM 4. Let $\{y_i^*\}$ be a bounded deterministic sequence. If for the system and algorithm (1)–(15), with $\{w_i\}, \{v_i\}$ being iid and (104) holding, conditions a), e) and f) are satisfied, then

$$(106) \quad \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \|u_i\|^2 < \infty, \quad \overline{\lim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \|y_i\|^2 < \infty,$$

$$(107) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (y_i - y_i^*)(y_i - y_i^*)^T = R_1 + R_2$$

and

$$\theta_n \xrightarrow[n \rightarrow \infty]{} \theta \quad \text{a.s.}$$

Proof. Everything up to and including (89) is still valid for the present case. Now at the right-hand side of (90) instead of $(y_{n+1} - \theta_{n+1}^T \phi_n)$ will be $y_{n+1} - \theta_{n+1}^T \phi_n + v_n$. But v_n is ergodic, so for fixed ω (91)–(101) remain valid.

Denote

$$(108) \quad z_n = \frac{\xi_n}{1 - b_{n-1}} + \frac{b_{n-1}}{1 - b_{n-1}} w_n.$$

Then instead of (102) we have, by (22), (24), (105),

$$(109) \quad y_n = z_n + y_n^* + w_n + v_{n-1}.$$

By (103) it is clear that

$$(110) \quad \frac{1}{n} \sum_{i=1}^n \|z_i\|^2 \xrightarrow[n \rightarrow \infty]{} 0 \quad \text{a.s.},$$

and from here, (109) and the independence of $\{w_i\}$ and $\{v_i\}$ we obtain (107) and also (106).

To complete the proof of the theorem we proceed to show that condition c^0) is satisfied for the present case.

Notice

$$(111) \quad u_n = [B_1^+ B(z^{-1})]^{-1} B_1^+ A(z^{-1}) y_{n+1} - [B_1^+ B(z^{-1})]^{-1} B_1^+ C(z^{-1}) w_{n+1}.$$

Denoting

$$(112) \quad \begin{aligned} H_1(z^{-1}) &= [B_1^+ B(z^{-1})]^{-1} B_1^+ A(z^{-1}), \\ H_2(z^{-1}) &= H_1(z^{-1}) - [B_1^+ B(z^{-1})]^{-1} B_1^+ C(z^{-1}), \end{aligned}$$

from (8), (109), (111) we obtain

$$(113) \quad \begin{aligned} \phi_n^0 &= [y_n^{*T} \cdots y_{n-p+1}^{*T} H_1(z^{-1}) y_{n+1}^{*T} \cdots H_1(z^{-1}) y_{n-q+2}^{*T} 0 \cdots 0]^T \\ &\quad + [z_n^T \cdots z_{n-p+1}^T H_1(z^{-1}) z_{n+1}^T \cdots H_1(z^{-1}) z_{n-q+2}^T 0 \cdots 0]^T \\ &\quad + [w_n^T \cdots w_{n-p+1}^T H_2(z^{-1}) w_{n+1}^T \cdots H_2(z^{-1}) w_{n-q+2}^T w_n^T \cdots w_{n-r+1}^T]^T \\ &\quad + [v_{n-1}^T \cdots v_{n-p}^T H_1(z^{-1}) v_n^T \cdots H_1(z^{-1}) v_{n-q+1}^T 0 \cdots 0]^T \\ &\triangleq \phi_n^1 + \phi_n^2 + \phi_n^3 + \phi_n^4. \end{aligned}$$

Under our assumptions it is easy to prove that for any $\rho \in (0, 1)$

$$\sum_{j=0}^n \rho^j |v_{n-j}|, \quad \sum_{j=0}^n \rho^j |w_{n-j}|,$$

$$\sum_{j=0}^{\infty} \rho^j |w_{n-j}|, \quad \sum_{j=0}^{\infty} \rho^j |v_{n-j}|$$

are all ergodic processes, by ergodicity of which in Caines and Lafortune [2] it is shown that

$$(114) \quad \frac{1}{n} \sum_{i=1}^n (\phi_i^3 + \phi_i^4)(\phi_i^3 + \phi_i^4)^T \xrightarrow[n \rightarrow \infty]{} R > 0 \quad \text{a.s.}$$

By using (110) and the boundedness of $\{y_i^*\}$ in a completely similar way as in Caines and Lafortune [2] we can prove that

$$(115) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \phi_i^1 (\phi_i^3 + \phi_i^4)^T = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \phi_i^1 \phi_i^{2T}$$

$$= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \phi_i^2 (\phi_i^3 + \phi_i^4)^T = 0.$$

Since $\{y_i^*\}$ is a bounded sequence there exists a constant k_{16} such that

$$(116) \quad \|\phi_n^1\|^2 \leq k_{16}$$

and hence

$$(117) \quad 0 \leq \frac{1}{n} \sum_{i=1}^n \phi_i^1 \phi_i^{1T} \leq k_{16} I.$$

From (110), (114)–(117) it is known that for any ω there exist $\delta_2 \geq \delta_1 > 0$ and $N > 0$ such that

$$\delta_1 I \leq \frac{1}{n} \sum_{i=1}^n \phi_i^0 \phi_i^{0T} \leq \delta_2 I \quad \forall n \geq N,$$

and hence for all n

$$\left(\delta_1 \Lambda \frac{1}{Nd} \right) I \leq \frac{1}{n} \left(\sum_{i=1}^n \phi_i^0 \phi_i^{0T} + \frac{1}{d} I \right) \leq \left[\delta_2 + \frac{1}{d} \right] I,$$

which means condition c^0) is fulfilled, and the strong consistency of θ_n follows from Theorems 1 and 2. \square

Remark. In comparison with Caines and Lafortune [2] in Theorem 4 we only apply one algorithm for both adaptive control and parameter identification. In addition, no further condition on y_i^* beyond boundedness is required and (107) is given in an unconditional form by using the ergodicity only.

Acknowledgment. The author would like to thank Professor P. E. Caines for his helpful discussions.

REFERENCES

- [1] P. E. CAINES, *Stochastic adaptive control: randomly varying parameters and continually disturbed controls*, Control Science and Technology for the Progress of Society, H. Akashi, ed., Pergamon, New York, 1981, pp. 925–930.

- [2] P. E. CAINES, AND S. LAFORTUNE, *Adaptive control with recursive identification for stochastic linear systems*, IEEE Trans. Automat., Control, to appear). Conference version presented at Conference on Decision and Control, Orlando, Florida, December 1982.
- [3] H. F. CHEN, *Least squares identification for continuous-time systems*, Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Control and Information Science 28, Springer, Berlin, 1980, pp. 264–277.
- [4] ———, *Strong consistency and convergence rate of least squares identification*, Scientia Sinica Ser. A, 25 (1982), pp. 771–784.
- [5] ———, *Strong consistency of recursive identification under correlated noise*, J. Systems Sci. Math. Sci., 1 (1981), pp. 34–52.
- [6] ———, *Quasi-least-squares identification and its strong consistency*, Internat. J. Control, 34 (1981), pp. 921–936.
- [7] ———, *Strong consistency in system identification under correlated noise*, Proc. 6th IFAC Symposium on Identification & System Parameter Estimation, 1982, pp. 964–969.
- [8] ———, *Stochastic approximation with ARMA measurement errors*, J. Systems Sci. Math. Sci., 2 (1982), pp. 227–239.
- [9] ———, *Self-tuning controller and its convergence under correlated noise*, Internat. J. Control, 35 (1982), pp. 1051–1059.
- [10] ———, *On continuous-time stochastic approximation*, Proc. 5th International Conference on Analysis and Optimization of Systems, to appear.
- [11] G. C. GOODWIN, P. T. RAMADGE AND P. E. CAINES, *Discrete-time multi-variable adaptive control*, IEEE Trans. Autom. Control, AC-25 (1980), pp. 449–456.
- [12] ———, *Discrete-time stochastic adaptive control*, this Journal, 19 (1981), pp. 829–853.
- [13] G. H. HARDY, J. E. LITTLEWOOD AND G. POLYA, *Inequalities*, Cambridge Univ. Press, Cambridge, 1934.
- [14] H. J. KUSHNER, AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer, New York, 1978.
- [15] T. L. LAI AND C. Z. WEI, *Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems*, Ann. Statist., 10 (1982), pp. 154–166.
- [16] L. LJUNG, *Consistency of the least squares identification method*, IEEE Trans. Autom. Control, AC-21 (1976), pp. 779–781.
- [17] ———, *Analysis of recursive stochastic algorithms*, IEEE Trans. Autom. Control, AC-22, (1977), pp. 551–575.
- [18] J. B. MOORE, *On strong consistency of least squares identification algorithms*, Automatica, 14, (1978), pp. 505–509.
- [19] K. S. SIN, *Adaptive filtering, prediction and control*, Ph.D. thesis, Univ. Newcastle, Newcastle NSW 2308, Australia, 1981.
- [20] K. S. SIN AND G. C. GOODWIN, *Stochastic adaptive control using a modified least squares algorithm*, Automatica, 18 (1982), pp. 315–321.
- [21] V. SOLO, *The convergence of AML*, IEEE Trans. Autom. Control, AC-24, 6 (1979), pp. 958–962.

THE CANONICAL DIOPHANTINE EQUATIONS WITH APPLICATIONS*

W. A. WOLOVICH† AND P. J. ANTSAKLIS‡

Abstract. A fundamental relationship between appropriate pairs of polynomial matrices is presented. This relationship, termed canonical Diophantine equations, can be used to resolve a number of standard polynomial matrix problems. Here, the general Diophantine equation is constructively resolved in a unique minimal way; in addition, prime canonical factorizations of a system transfer matrix are derived from knowledge of any dual factorization.

Key words. multivariable control systems, linear systems, algebraic system theory, polynomial matrix algebra

1. Introduction. Polynomial matrices play an important role in many different aspects of linear system theory, especially when one describes the dynamical behavior of a given system in terms of either a right or left polynomial matrix factorization of the transfer matrix which defines the system; i.e. $T(s) = R(s)P_R^{-1}(s) = P_L^{-1}(s)Q(s)$. Questions such as obtaining state-space realizations of $T(s)$, or state observers associated with $T(s)$, or stabilizing compensators, which perform one or several simultaneous control functions, have been constructively resolved through the manipulation of polynomial matrices, and such results are cited in various texts and papers too numerous to delineate.

Generally speaking, there are certain “standard problems” (SP) involving polynomial matrix pairs, such as $\{R(s), P_R(s)\}$ or $\{P_L(s), Q(s)\}$, which underlie most of the polynomial matrix manipulations required to obtain solutions to questions such as those posed above. Some of these are the following:

(SP1) Solve the general Diophantine equation, $H(s)R(s) + K(s)P_R(s) = F(s)$, for an appropriate polynomial matrix pair, $\{H(s), K(s)\}$ given any arbitrary $F(s)$. The Bezout equation, when $F(s) = I$, would represent a special case of the general Diophantine equation.

(SP2) Obtain a dual, prime factorization, $P_L^{-1}(s)Q(s)$, of $T(s)$ from any given (not necessarily prime) factorization, $R(s)P_R^{-1}(s)$.

(SP3) Divide one polynomial matrix by another nonsingular one to obtain the unique strictly proper part and quotient.

(SP4) Determine a greatest common right or left divisor of a given pair of polynomial matrices.

Clearly, all of these “standard problems” are interrelated, and various solutions to all of them have been documented in numerous references, and this report will not even attempt to judge the merits of one solution relative to another.

It is important to note however that there is a fundamental, underlying relationship common to all of these standard polynomial matrix problems, which can be used to solve them all. In this paper, we will develop such a relationship which we will term “canonical Diophantine equations” because the solutions to such equations can be uniquely determined from canonical state-space representations.

* Received by the editors March 15, 1983, and in revised form August 22, 1983. This work was supported in part by the National Science Foundation under grant ECS79-16584 and by the Air Force Office of Scientific Research under grant AFOSR-82-0034.

† Division of Engineering and the Lefschetz Center for Dynamical Systems, Brown University, Providence, Rhode Island 02912.

‡ Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana 46556.

In § 2, we formally establish both types of canonical Diophantine equations. The general Diophantine equation (SP1) is then constructively resolved in § 3 in a unique, minimal way. In § 4, we again employ the canonical Diophantine equation, along with the algorithm of § 3, to solve (SP2); i.e. to obtain canonical, dual prime factorizations of a given transfer matrix from knowledge of any matrix fraction description, and we conclude with some final remarks in § 5. It might be noted that solutions to (SP3) and (SP4), as well as other "standard polynomial matrix problems," will be addressed in subsequent reports using the canonical Diophantine formulation developed here.

2. The canonical Diophantine equations. Consider a pair $\{R(s), P_R(s)\}$ of polynomial matrices in the Laplace operator s with $R(s)$ $p \times m$ and $P_R(s)$ $m \times m$ and column proper; i.e. the $m \times m$ constant matrix, $\Gamma_c[P_R(s)]$, consisting of the coefficients of the highest degree terms in each column of $P_R(s)$ is nonsingular. If¹ μ_i denotes the degree of each (i th) column of $P_R(s)$, a relation we denote as

$$(1) \quad \partial_{ci}[P_R(s)] = \mu_i$$

it follows [1] that $|P_R(s)|$, the determinant of $P_R(s)$, will be a polynomial of degree n , where

$$(2) \quad n = \sum_{i=1}^m \mu_i.$$

If the pair $\{R(s), P_R(s)\}$ is used to denote a right transfer matrix factorization of some multivariable system, so that the transfer matrix of the system

$$(3) \quad T(s) = R(s)P_R^{-1}(s),$$

then a state-space realization $\{A, B, C, E(s)\}$ of $T(s)$ can readily be determined by the well-known "structure theorem" for linear multivariable systems [1], [2]. In particular, if we apply the polynomial matrix division algorithm to (3) to separate $T(s)$ into its strictly proper part, $\bar{R}(s)P_R^{-1}(s)$, and quotient, $E(s)$, i.e.

$$(4) \quad T(s) = \bar{R}(s)P_R^{-1}(s) + E(s),$$

then a real triple $\{A, B, C\}$ of dimensions $n \times n$, $n \times m$, and $p \times n$, respectively, can be found such that

$$(5) \quad C(sI - A)^{-1}B = \bar{R}(s)P_R^{-1}(s).$$

More specifically, following the development in [1], if the $(n \times m)$ polynomial matrix $S_R^\mu(s)$ is defined by the relation

$$(6) \quad S_R^\mu(s) = \text{block diagonal} \begin{bmatrix} 1 \\ s \\ \vdots \\ s^{\mu_i-1} \end{bmatrix},$$

then there exists a real pair of matrices $\{A, B\}$ in multi-input controllable companion form [1] such that

$$(7) \quad (sI - A)S_R^\mu(s) = BP_R(s).$$

¹ We will assume throughout for convenience, and without much loss of generality, that each $\mu_i \geq 1$ and, later, that each $\nu_j \geq 1$.

Furthermore, the μ_i defined by (1) will represent the controllability indices of $\{A, B\}$ in the sense that the following $n \times n$ "column ordered controllability matrix" associated with the pair, namely

$$(8) \quad \bar{L} = [b_1, Ab_1, \dots, A^{\mu_1-1}b_1, b_2, Ab_2, \dots, A^{\mu_2-1}b_2, \dots, A^{\mu_m-1}b_m],$$

will be nonsingular. In (8), b_i denotes the i th column of B .

Since $\bar{R}(s)P_R^{-1}(s)$ is strictly proper, it can be shown [1] that

$$(9) \quad \bar{R}(s) = CS_R^\mu(s)$$

for some constant $(p \times n)$ matrix C , so that (7) and (9) together imply that

$$(10) \quad CS_R^\mu(s)P_R^{-1}(s) = \bar{R}(s)P_R^{-1}(s) = C(sI - A)^{-1}B,$$

thus verifying (5).

Now consider the "total observability matrix", \mathcal{O} , associated with the pair $\{C, A\}$, i.e., the $np \times n$ real matrix

$$(11) \quad \mathcal{O} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}.$$

Let \bar{n} denote the rank of \mathcal{O} , a relation we represent as

$$(12) \quad \rho[\mathcal{O}] = \bar{n} \leq n,$$

and select from top to bottom the first \bar{n} linearly independent rows of \mathcal{O} . If these \bar{n} rows are then reordered so that all rows containing the k th row of C , c_k , precede those containing c_{k+1} , in increasing powers of A , we will obtain a set of observability indices, ν_j , associated with the pair $\{C, A\}$ as well as an $\bar{n} \times n$ real matrix \bar{M} , analogous to the \bar{L} of (8), which we will call a "row ordered observability matrix" of $\{C, A\}$. In particular,

$$(13) \quad \bar{M} = \begin{bmatrix} c_1 \\ c_1 A \\ \vdots \\ c_1 A^{\nu_1-1} \\ c_2 \\ c_2 A \\ \vdots \\ c_2 A^{\nu_2-1} \\ \vdots \\ c_p A^{\nu_{p-1}} \end{bmatrix},$$

a real matrix of full rank \bar{n} , where

$$(14) \quad \bar{n} = \sum_{j=1}^p \nu_j.$$

In view of this observation, a $(\bar{n} \times p)$ polynomial matrix, $S_R^\nu(s)$, can be defined by the relation

$$(15) \quad S_R^\nu(s) = \text{block diagonal} \begin{bmatrix} 1 \\ s \\ \vdots \\ s^{\nu_j-1} \end{bmatrix}.$$

In view of these preliminaries, we can now state the main result of this section.

THEOREM 1. *Consider a polynomial matrix pair, $\{R(s), P_R(s)\}$, with $R(s)$ $p \times m$ and $P_R(s)$ $m \times m$ and column proper. Let $S_R^\mu(s)$ be given by (6), \bar{M} by (13), and $S_R^\nu(s)$ by (15). There exists another unique $(\bar{n} \times m)$ polynomial matrix, $\hat{M}(s)$, such that*

$$(*) \quad S_R^\nu(s)R(s) - \hat{M}(s)P_R(s) = \bar{M}S_R^\mu(s).$$

Proof. Using the results in [1], we first note that the given polynomial matrix pair $\{R(s), P_R(s)\}$ directly defines a Laplace transformed differential operator representation of a system which is equivalent to the state-space realization of the $T(s)$ given by (3), namely

$$(16a) \quad P_R(s)z(s) = u(s),$$

$$(16b) \quad y(s) = R(s)z(s),$$

where the relationship between the partial state, $z(s)$, of (16) and the state, $x(s)$, of the state-space realization is given by

$$(17) \quad S_R^\mu(s)z(s) = x(s).$$

By repeated “differentiation” of the (Laplace transformed) state-space output equation

$$(18a) \quad y(s) = Cx(s) + E(s)u(s),$$

while substituting the state equation,

$$(18b) \quad sx(s) = Ax(s) + Bu(s),$$

we obtain the relation

$$(19) \quad \begin{bmatrix} y(s) \\ sy(s) \\ \vdots \\ s^{n-1}y(s) \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} x(s) + \begin{bmatrix} E(s) & 0 & 0 & \cdots & 0 \\ CB & E(s) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ CA^{n-2}B & CA^{n-3}B & \cdots & & E(s) \end{bmatrix} \begin{bmatrix} I \\ sI \\ \vdots \\ s^{n-1}I \end{bmatrix} u(s).$$

We next observe that the \bar{M} of (13) can be directly obtained from the \mathcal{O} of (11) by premultiplying \mathcal{O} by a real $(\bar{n} \times np)$ “row selector matrix” which contains only 0’s and (\bar{n}) 1’s. If we now premultiply (19) by exactly the same row selector matrix, we obtain the relation

$$(20) \quad S_R^\nu(s)y(s) = \bar{M}x(s) + \hat{M}(s)u(s),$$

where

$$(21) \quad \hat{M}(s) = \begin{bmatrix} \hat{M}_1(s) \\ \vdots \\ \hat{M}_p(s) \end{bmatrix}$$

with each

$$(22) \quad \hat{M}_j(s) = \begin{bmatrix} E_j(s) & 0 & 0 & \cdots & 0 \\ C_j B & E_j(s) & 0 & \cdots & 0 \\ \vdots & & & & \vdots \\ C_j A^{\nu_j-2} B & \cdots & & & E_j(s) \end{bmatrix} \begin{bmatrix} I \\ sI \\ \vdots \\ s^{\nu_j-1} I \end{bmatrix}.$$

In (22), $E_j(s)$ and C_j denote the j th rows of $E(s)$ and C , respectively.

If we now use (16) and (17) to substitute into (20), we obtain

$$(23) \quad S_R^\nu(s)R(s)z(s) = \bar{M}S_R^\mu(s)z(s) + \hat{M}(s)P_R(s)z(s),$$

or since (23) must hold for arbitrary $z(s)$, that (*) must hold.

Finally, by right dividing (*) by $P_R(s)$, we note that

$$(24) \quad S_R^\nu(s)R(s)P_R^{-1}(s) = \bar{M}S_R^\mu(s)P_R^{-1}(s) + \hat{M}(s).$$

Since $\bar{M}S_R^\mu(s)P_R^{-1}(s)$ is strictly proper, it thus follows that the polynomial matrix $\hat{M}(s)$ represents the unique quotient of $S_R^\nu(s)R(s)P_R^{-1}(s)$; i.e. given the pair $\{R(s), P_R(s)\}$, $\hat{M}(s)$ is uniquely specified via (24) by the choice for $S_R^\nu(s)$. Theorem 1 is therefore established. We call (*) a *right canonical Diophantine equation of the pair* $\{R(s), P_R(s)\}$.

As we have now shown, (*) can be solved by first determining $S_R^\mu(s)$ by inspection of the column degrees of the column proper $P_R(s)$. The structure theorem of [1] can then be used to determine a state-space realization $\{A, B, C, E(s)\}$ of $T(s) = R(s)P_R^{-1}(s)$. The pair $\{C, A\}$ then defines the total observability matrix via (11), from which \bar{M} and $S_R^\nu(s)$ are derived. Finally, $\hat{M}(s)$ can be obtained directly via (21) and (22).

We next note that nothing has been said, thus far, regarding the “primeness” of lack thereof of $R(s)$ and $P_R(s)$. In particular, (*) holds whether or not $R(s)$ and $P_R(s)$ are relatively right prime (*rrp*). We observe, moreover, that in light of the results given in [1], the zeros of the determinant of any greatest common right divisor, $\tilde{G}_R(s)$, of $R(s)$ and $P_R(s)$ will represent all of the unobservable modes of the system defined by (10). In view of this observation, it follows that

$$(25) \quad \bar{n} = \rho[\mathcal{O}] = n - \partial|\tilde{G}_R(s)|,$$

and, as a result, that $\bar{n} = n$ if and only if $R(s)$ and $P_R(s)$ are *rrp*. In such cases, the \bar{M} given by (13) will be $n \times n$ and nonsingular.

It is of interest to note that once the full rank matrix \bar{M} has been determined, premultiplication of $R(s)P_R^{-1}(s)$ by any polynomial matrix $H(s)$, followed by a separation of the resulting rational matrix into its *spp* and quotient, will always yield a constant premultiplier of $S_R^\mu(s)P_R^{-1}(s)$ in the span of \bar{M} , i.e. for any $q \times m$ polynomial matrix, $H(s)$,

$$(26) \quad H(s)R(s)P_R^{-1}(s) = \bar{H}\bar{M}S_R^\mu(s)P_R^{-1}(s) + \tilde{R}(s)$$

for some constant matrix \bar{H} . If this were not true, \bar{M} would not represent the full rank row ordered observability matrix it is. The relationship represented by (26) is a useful one, as we will later show.

A dual result, analogous to (*), can now be readily established by considering any left matrix factorization $P_L^{-1}(s)Q(s)$ of a $\tilde{T}(s)$; i.e. consider a pair of polynomial matrices $\{P_L(s), Q(s)\}$, with $P_L(s)$ $p \times p$ and row proper, and $Q(s)$ $p \times m$, which define the $(p \times m)$ rational transfer matrix,

$$(27) \quad \tilde{T}(s) = P_L^{-1}(s)Q(s).$$

Define

$$(28) \quad \tilde{\nu}_j = \partial_{\eta}[P_L(s)]$$

for $j = 1, 2, \dots, p$,

$$(29) \quad d = \sum_1^p \tilde{\nu}_j,$$

and the $(p \times d)$ polynomial matrix

$$(30) \quad S_L^{\nu}(s) = \text{block diagonal} [1 \quad s \quad \dots \quad s^{\tilde{\nu}_j-1}].$$

Let $\tilde{\mu}_i$ denote an ordered set of controllability indices of any minimal state-space realization of the system defined by (27). Define

$$(31) \quad \bar{d} = \sum_1^m \tilde{\mu}_i$$

and the $(m \times \bar{d})$ polynomial matrix

$$(32) \quad S_L^{\mu}(s) = \text{block diagonal} [1 \quad s \quad \dots \quad s^{\tilde{\mu}_i-1}].$$

In view of (30) and (32), a *left canonical Diophantine equation of the polynomial matrix pair* $\{P_L(s), Q(s)\}$ could then be written as

$$(*)^T \quad Q(s)S_L^{\mu}(s) - P_L(s)\hat{L}(s) = S_L^{\nu}(s)\bar{L}$$

where $\hat{L}(s)$ is the quotient of $P_L^{-1}(s)Q(s)S_L^{\mu}(s)$ and $(d \times \bar{d})$ \bar{L} is a real, full rank (\bar{d}) , column ordered controllability matrix of the system defined by (27).

It is of interest to note that if $\tilde{T}(s)$ given by (27) is equal to the $T(s)$ given by (3), then $R(s)P_R^{-1}(s)$ and $P_L^{-1}(s)Q(s)$ will represent "dual factorizations" of the same transfer matrix which thus satisfy the relation:

$$(33) \quad Q(s)P_R(s) = P_L(s)R(s).$$

Furthermore, if the $\tilde{\nu}_j$ given by (28) correspond to the ν_j defined by (13), so that

$$(34) \quad d = \sum_1^p \tilde{\nu}_j = \sum_1^p \nu_j = \bar{n},$$

then $P_L(s)$ and $Q(s)$ will be relatively left prime, and $P_L^{-1}(s)Q(s)$ will represent a minimal order, left prime factorization of $T(s) = R(s)P_R^{-1}(s)$. It is well known [3] that such a factorization of $T(s)$ can always be found such that $P_L(s)$ is both row proper and column proper with

$$(35) \quad \partial_{\eta}[P_L(s)] = \partial_{\eta}[P_L(s)] = \nu_p,$$

and

$$(36) \quad \Gamma_c[P_L(s)] = I_p.$$

A new constructive procedure for obtaining such canonical, dual, prime factorizations of $T(s)$ via (*) will be presented in § 4.

3. General Diophantine equations. Polynomial matrix Diophantine equations of the form

$$(37) \quad H(s)R(s) + K(s)P_R(s) = F(s)$$

play an important role in many aspects of linear system theory, and numerous references can be cited to substantiate this observation; e.g. see [4]–[12], a nonexhaustive list. It is well known that if, in general, $R(s)$, $P_R(s)$, and $F(s)$ are of dimensions $p \times m$, $k \times m$, and $q \times m$, respectively, and if the rank (over the field of rational functions in s) of the composite $(p+k) \times m$ matrix

$$\begin{bmatrix} R(s) \\ P_R(s) \end{bmatrix}$$

is m , i.e. if

$$(38) \quad \rho \begin{bmatrix} R(s) \\ P_R(s) \end{bmatrix} = m,$$

then (37) has a solution $\{H(s), K(s)\}$ if and only if any gcd, $\tilde{G}_R(s)$, of $R(s)$ and $P_R(s)$ is a right divisor of $F(s)$. A variety of different techniques have been devised to solve (37) when solutions do exist. In this section, a new constructive proof of sufficiency will be given, based on (*), which directly yields a unique, solution $\{H(s), K(s)\}$ to (37) with $H(s)$ of minimum column degree modulo the choice for $S_R^\nu(s)$.

To obtain this general result, first consider (37) when $P_R(s)$ is $m \times m$ and column proper, so that (*) holds, and right divide $F(s)$ by $P_R(s)$, i.e.,

$$(39) \quad F(s)P_R^{-1}(s) = \bar{F}S_R^\mu(s)P_R^{-1}(s) + \hat{F}(s)$$

to obtain the $(q \times n)$ real matrix \bar{F} and the polynomial matrix quotient, $\hat{F}(s)$.

THEOREM 2. Consider the Diophantine equation (37) with $P_R(s)$ $m \times m$ and column proper. If (37) is solvable, there exists a real $(q \times \bar{n})$ matrix \bar{H} such that

$$(40) \quad \bar{H}\bar{M} = \bar{F}$$

where \bar{F} is given by (39). Furthermore, the polynomial matrix pair

$$(41) \quad H(s) = \bar{H}S_R^\nu(s)$$

and

$$(42) \quad K(s) = \hat{F}(s) - \bar{H}\hat{M}(s)$$

solves (37) with $H(s)$, as given by (41), a unique $(q \times p)$ polynomial matrix of minimum column degree in the sense that

$$(43) \quad \partial_{c_j}[H(s)] < \nu_j$$

for $j = 1, 2, \dots, p$.

Proof. First, recall that (26) holds for any polynomial matrix, $H(s)$; i.e.

$$(44) \quad H(s)R(s) = \bar{H}\bar{M}S_R^\mu(s) + \tilde{R}(s)P_R(s)$$

for some real matrix, \bar{H} . Since

$$(45) \quad F(s) = \bar{F}S_R^\mu(s) + \hat{F}(s)P_R(s)$$

in light of (39), the solvability of (37) now implies that

$$(46) \quad [\bar{H}\bar{M} - \bar{F}]S_R^\mu(s) = [\hat{F}(s) - K(s) - \tilde{R}(s)]P_R(s).$$

or that

$$(47) \quad [\bar{H}\bar{M} - \bar{F}]S_R^\mu(s)P_R^{-1}(s) = \hat{F}(s) - K(s) - \tilde{R}(s).$$

Since the left side of (47) is either a strictly proper rational matrix or zero, while the right side is either a polynomial matrix or zero, both sides must be zero, thus establishing (40).

If the right canonical Diophantine equation (*) is now premultiplied by \bar{H} and $F(s) - \hat{F}(s)P_R(s)$ is substituted for $\bar{H}\bar{M}S_R^\mu(s) = \bar{F}S_R^\mu(s)$ in light of (45), the general Diophantine equation, (37), is satisfied with $H(s)$ given by (41) and $K(s)$ given by (42).

To establish Theorem 2, we must finally show that the particular $H(s)$ given by (41) uniquely satisfies (43). To do this, consider any dual, relatively left prime factorization, $P_L^{-1}(s)Q(s)$, of $T(s) = R(s)P_R^{-1}(s)$, such that (33) through (35) hold. The composite $p \times (p+m)$ matrix

$$[P_L(s) \mid -Q(s)]$$

represents a prime basis for the null space of the composite $(p+m) \times m$ matrix [4]

$$\begin{bmatrix} R(s) \\ P_R(s) \end{bmatrix}.$$

Therefore, if $\{H(s), K(s)\}$ represents any particular solution to (37), any other solution to (37), $\{\tilde{H}(s), \tilde{K}(s)\}$, can be written as

$$(48) \quad \tilde{H}(s) = H(s) + J(s)P_L(s)$$

and

$$(49) \quad \tilde{K}(s) = K(s) - J(s)Q(s),$$

for some polynomial matrix $J(s)$. In particular,

$$(50) \quad \begin{aligned} & [H(s) + J(s)P_L(s)]R(s) + [K(s) - J(s)Q(s)]P_R(s) \\ & = H(s)R(s) + K(s)P_R(s) + J(s)[P_L(s)R(s) - Q(s)P_R(s)] = F(s) \end{aligned}$$

in light of (33). In light of (41), (35), and (36), however,

$$(51) \quad \partial_{cj}[H(s)] < \nu_j = \partial_{cj}[P_L(s)],$$

with $\Gamma_c[P_L(s)] = I_p$. It therefore follows from (48) that the unique

$$(52) \quad \text{spp}\{\tilde{H}(s)P_L^{-1}(s)\} = H(s)P_L^{-1}(s),$$

or that the $H(s)$ given by (41) represents a unique solution to (37) of minimum column degree ν_j for $j=1, 2, \dots, p$. Theorem 2 is thus established.

It might be noted that if $P_R(s)$ of (37) is nonsingular but not column proper, (37) can be postmultiplied by any unimodular matrix, $U_R(s)$, which reduces $P_R(s)U_R(s)$ to column proper form. The results of this section can then be directly employed to obtain the unique solution $\{H(s), K(s)\}$ to (37) with $H(s)$ of minimum column degree in the sense of (43).

We finally note that $P_R(s)$ need not be square or nonsingular in order to utilize (*) to solve (37). In particular, note that (37) can be written in composite form as

$$(53) \quad [H(s) \mid K(s)] \begin{bmatrix} R(s) \\ P_R(s) \end{bmatrix} = F(s).$$

Therefore, if (38) holds, (37) has a solution if and only if any gcd , $\tilde{G}_R(s)$, of (the

rows of)

$$\begin{bmatrix} R(s) \\ \overline{P_R(s)} \end{bmatrix}$$

is a right divisor of $F(s)$. If this condition holds, a new “ $P_R(s)$ ” can be defined as any m linearly independent rows of

$$\begin{bmatrix} R(s) \\ \overline{P_R(s)} \end{bmatrix},$$

with the new “ $R(s)$ ” given by the remaining rows. With these new definitions, (53) can then be solved for “ $H(s)$ ” and “ $K(s)$ ” using the results presented in this section. Finally, the actual $H(s)$ and $K(s)$, consistent with their original definitions, can be obtained by repositioning the columns of “ $H(s)$ ” and “ $K(s)$ ”.

4. Canonical dual prime factorizations. At the conclusion of § 2, we noted that given a right matrix factorization, $R(s)P_R^{-1}(s)$, of $T(s)$, a dual, left matrix factorization, $P_L^{-1}(s)Q(s)$, can always be found which satisfies properties (35) and (36). In this section, we will present a new algorithm for obtaining such dual, canonical factorizations via the right canonical Diophantine equation (*), of § 2. In particular we will now constructively establish the following known result [3] in a new, direct way using (*).

THEOREM 3. *Consider any polynomial matrix pair $\{R(s), P_R(s)\}$ with $R(s)$ $p \times m$ and $P_R(s)$ $m \times m$ and column proper. There exists a “dual” relatively left prime pair $\{P_L(s), Q(s)\}$ of polynomial matrices with*

$$(54) \quad \partial_{ej}[P_L(s)] = \partial_{ej}[P_R(s)] = \nu_j,$$

for $j = 1, 2, \dots, p$,

$$(55) \quad \Gamma_c[P_L(s)] = I_p,$$

and $\Gamma_r[P_L(s)]$ nonsingular with 1's along the diagonal such that $T(s) = R(s)P_R^{-1}(s) = P_L^{-1}(s)Q(s)$, or

$$(56) \quad P_L(s)R(s) = Q(s)P_R(s).$$

Proof. In light of (*) define the $p \times p$ diagonal polynomial matrix

$$(57) \quad D''(s) = \text{diagonal}[s^{\nu_j}],$$

and set

$$(58) \quad F_L(s) = D''(s)R(s)$$

in the general Diophantine equation (37), so that (39) implies that

$$(59) \quad F_L(s) = D''(s)R(s) = \bar{F}_L S_R''(s) + \hat{F}_L(s)P_R(s).$$

Equation (37) is clearly solvable in that any gcd of $R(s)$ and $P_R(s)$ will be a right divisor of the particular $F_L(s) = D''(s)R(s)$ given by (59). In light of Theorem 2, therefore, a constant \bar{H}_L exists such that

$$(60) \quad \bar{H}_L \bar{M} = \bar{F}_L.$$

It thus follows that the polynomial matrix pair

$$(61) \quad H_L(s) = \bar{H}_L S_R''(s),$$

and

$$(62) \quad K_L(s) = \hat{F}_L(s) - \bar{H}_L \hat{M}(s)$$

solves (37) i.e., that

$$(63) \quad \bar{H}_L S_R^\nu(s) R(s) + [\hat{F}_L(s) - \bar{H}_L \hat{M}(s)] P_R(s) = D^\nu(s) R(s),$$

or that

$$(64) \quad [D^\nu(s) - \bar{H}_L S_R^\nu(s)] R(s) = [\hat{F}_L(s) - \bar{H}_L \hat{M}(s)] P_R(s),$$

so that (56) holds with

$$(65) \quad Q(s) = \hat{F}_L(s) - \bar{H}_L \hat{M}(s)$$

and

$$(66) \quad P_L(s) = D^\nu(s) - \bar{H}_L S_R^\nu(s).$$

Note that the $P_L(s)$ thus defined will be both column proper and row proper. In particular, in light of (66), it is clear that

$$(67) \quad \partial_{cj}[P_L(s)] = \nu_j$$

for $j = 1, 2, \dots, p$, and that

$$(68) \quad \Gamma_c[P_L(s)] = I_p.$$

Since the ν_j represent an appropriately ordered set of observability indices of the system, $P_L(s)$ will be row proper as well [3] with 1's along the diagonal of $\Gamma_r[P_L(s)]$. We finally observe that since $\bar{n} = \sum_1^p \nu_j$ represents the order of a minimal realization of $T(s) = R(s)P_R^{-1}(s) = P_L^{-1}(s)Q(s)$, and $\partial|P_L(s)| = \bar{n}$, $P_L(s)$ and $Q(s)$ will be relatively left prime. Theorem 3 is thus established.

5. Concluding remarks. A new, fundamental relationship between appropriate pairs of polynomial matrices has now been presented and employed to resolve some "standard (polynomial matrix) problems" in a new and direct manner. In particular, the utility of the dual, canonical Diophantine equations (*) and $(*^T)$, has now been thoroughly demonstrated with respect to (SP1) obtaining unique minimal degree solutions to general Diophantine equations, and (SP2) determining canonical, prime, transfer matrix factorizations of a given system from knowledge of any dual factorization.

It is of interest to note that the dimension of the largest matrix, M , which need be inverted in order to solve a general matrix Diophantine equation of the form (37) is n , the system order, unlike earlier algorithms which require the inversion of a matrix of generic dimension $2n$. This could significantly reduce the computations necessary to implement a variety of adaptive control algorithms.

Finally, it should be noted that additional implications of (*) and $(*^T)$ do exist with respect to other polynomial matrix problems, and that some of these will be addressed in subsequent reports.

REFERENCES

- [1] W. A. WOLOVICH, *Linear Multivariable Systems*, Springer-Verlag, New York, 1974.
- [2] W. A. WOLOVICH AND P. L. FALB, *On the structure of multivariable systems*, this Journal, 7 (1969), pp. 437-451.

- [3] R. GUIDORZI, *Canonical structures in the identification of multivariable systems*, Automatica, 11 (1975), pp. 361–374.
- [4] G. D. FORNEY, JR., *Minimal bases of rational vector spaces with applications to multivariable linear systems*, this Journal, 13 (1975), pp. 493–520.
- [5] W. A. WOLOVICH, P. ANTSAKLIS AND H. ELLIOTT, *On the stability of solutions to minimal and nonminimal design problems*, IEEE Trans. Automat. Control, 22 (1977), pp. 88–94.
- [6] R. R. BITMEAD, S.-Y. KUNG, B. D. O. ANDERSON AND T. KAILATH, *Greatest common divisors via generalized Sylvester and Bezout matrices*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 1043–1047.
- [7] L. CHENG AND J. B. PEARSON, JR., *Frequency domain synthesis of multivariable linear regulators*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 3–15.
- [8] P. J. ANTSAKLIS, *Some relations satisfied by prime polynomial matrices and their role in linear multivariable system theory*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 611–616.
- [9] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
- [10] L. CHENG AND J. B. PEARSON, JR., *Synthesis of linear multivariable regulators*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 194–202.
- [11] V. KUCERA, *Discrete Linear Control*, John Wiley, New York, 1979.
- [12] E. EMRE, *The polynomial equation $AA_c + RP_c = \Phi$ with application to dynamic feedback*, this Journal, 18 (1980), pp. 611–620.

COMPLETE CONTROLLABILITY OF ONE-DIMENSIONAL VIBRATING SYSTEMS WITH BANG-BANG CONTROLS*

KIMIYAKI NARUKAWA†

Abstract. This paper deals with the problem of controllability for one-dimensional vibrating systems with controls which are scalar valued functions of time only. We give some constraint sets of controls and prove exact controllability of vibrations in those constraint sets by studying the moment problem attached to this control problem precisely. Using these results, and Lyapunov's convexity theorem for vector measures, we obtain complete controllability only with bang-bang controls.

Key words. vibrating systems, constraint set of controls, complete controllability, bang-bang controls

1. Introduction. In our previous paper [11], we discussed admissible controllability of vibrations of a vibrating string by constrained controls. The control functions which were considered there are functions both in space and time variables (x, t) , and are therefore difficult to realize practically. In [13], Russell considered the case where a control is exercised by means of a force whose spacial distribution is fixed but whose sign and amplitude are variable with time t . And he obtained exact controllability by studying the moment problem attached to this control problem.

The control system considered by Russell is described by the equation

$$(1.1) \quad \rho(x) \frac{\partial^2 u}{\partial t^2}(x, t) - \frac{\partial}{\partial x} \left(c(x) \frac{\partial u}{\partial x}(x, t) \right) = g(x)f(t), \quad 0 < x < 1, \quad t > 0,$$

with the Dirichlet boundary condition

$$(1.2) \quad u(0, t) = u(1, t) = 0, \quad t > 0,$$

which describes forced motion of a string with density $\rho(x)$ and modulus of elasticity $c(x)$. Let us assume $\rho(x)$ and $c(x)$ are sufficiently smooth, bounded and strictly positive and $g(x)$ belongs to $L^2(0, 1)$. The function $g(x)$ is called a force distribution function. For this control system a function $f(t)$ is considered as a control. In this paper, we will give admissible controllability of a higher order Sobolev space by continuous controls $f(t)$ constrained as $0 \leq f(t) \leq 1$.

Next we will study controllability of (1.1), (1.2) with bang-bang controls, that is, controls assuming only the values 0 or 1. The control problem with bang-bang controls is closely related to the optimal time control problem. For finite-dimensional control systems, various constraint sets of controls and admissible controllability in those constraint sets have been considered by many authors (see e.g. [1], [17], [18]). Furthermore, assuming a certain normality condition is satisfied, they have shown that time optimal controls are necessarily bang-bang. Consequently the attainable set in the constraint set is controllable with bang-bang controls. In the infinite-dimensional case, the relation between the optimal time control problem and the bang-bang principle has been studied for some control systems (see e.g. [3], [4], [5], [8], [19]), but little attention appears to have been given to the problem of controllability using bang-bang controls.

* Received by the editors March 8, 1983, and in revised form September 26, 1983.

† Department of Mathematics, Faculty of Integrated Arts and Sciences, Hiroshima University, Hiroshima 730, Japan.

This problem is closely related to the study of Lyapunov vector measures. Knowles [7], [8] investigated Lyapunov vector measures and the relation to the bang-bang control problems for both finite- and infinite-dimensional control systems. But here we use only Lyapunov's convexity theorem for finite-dimensional vector measures and obtain complete controllability of the control system (1.1) with (1.2).

2. Definitions and fundamental notions. For a real number m and interval (a, b) , we denote by $H^m(a, b)$ the usual Sobolev space of order m on (a, b) . The usual norm of $H^m(a, b)$ and the inner product of $L^2(a, b)$ are denoted by $\|\cdot\|_m$ and (\cdot, \cdot) respectively. Further we denote by $H_0^1(a, b)$ and $H_p^1(a, b)$ the spaces of all functions $f(t)$ in $H^1(a, b)$ satisfying $f(a) = f(b) = 0$ and $f(a) = f(b)$ respectively.

By putting $V(t) = [u(t), (\partial u / \partial t)(t)]$,

$$A = \begin{pmatrix} 0 & I \\ L & 0 \end{pmatrix}, \quad Lu = \rho^{-1}(x) \frac{d}{dx} \left[c(x) \frac{du}{dx} \right]$$

and

$$(Bf)(x, t) = \begin{pmatrix} 0 \\ \rho^{-1}(x)g(x)f(t) \end{pmatrix},$$

the equation (1.1) is reduced to the first order equation

$$(2.1) \quad \frac{dV}{dt}(t) = AV(t) + Bf(t).$$

As in [11], we put $[u, v] = \begin{pmatrix} u \\ v \end{pmatrix}$ and introduce the inner products which define the norms equivalent to those of $H_0^1(0, 1)$ and $L^2(0, 1)$ respectively as

$$((u_1, u_2))_c = \int_0^1 c(x) \frac{du_1}{dx} \frac{du_2}{dx} dx \quad \text{for } u_1, u_2 \in H_0^1(0, 1)$$

and

$$(v_1, v_2)_\rho = \int_0^1 \rho(x) v_1(x) v_2(x) dx \quad \text{for } v_1, v_2 \in L^2(0, 1).$$

Further let \mathcal{H} be the Hilbert space $H_0^1(0, 1) \times L^2(0, 1)$ endowed with the inner product

$$([u_1, v_1], [u_2, v_2])_{\mathcal{H}} = ((u_1, u_2))_c + (v_1, v_2)_\rho.$$

Attached to the boundary condition (1.2), we define the domain of A as $(H^2(0, 1) \cap H_0^1(0, 1)) \times H_0^1(0, 1)$. Then A generates the unitary group $U(t)$ on \mathcal{H} . For any $V_0 \in \mathcal{H}$ and locally summable function $f(t)$, we define

$$(2.2) \quad V(t) = U(t)V_0 + \int_0^t U(t-s)Bf(s) ds$$

to be a mild solution of (2.1) with the initial state $V(0) = V_0$. It is well known that if $f(t)$ is continuously differentiable in $t > 0$ and $V_0 = [u_0, v_0]$ is in the domain of A , then the mild solution $V(t) = [u(t), (\partial u / \partial t)(t)]$ is a strong solution of (1.1) with (1.2) and $[u(0), (\partial u / \partial t)(0)] = [u_0, v_0]$. When, for the mild solution $V(t)$ of (2.2), $V(0) = [u_0, v_0]$ and $V(T) = [u_1, v_1]$ hold, we say that the control $f(t)$ steers $[u_0, v_0]$ to $[u_1, v_1]$ at T .

For a subset \mathcal{F} of $L_{loc}^1[0, \infty)$, we define the attainable set $R_T(\mathcal{F})$ at T of the control system (2.1) (or (1.1) with (1.2)) in the constraint set of controls \mathcal{F} as

$$R_T(\mathcal{F}) = \left\{ \int_0^T U(T-s)Bf(s) ds \mid f \in \mathcal{F} \right\}.$$

We now recall definitions of various types of controllability.

DEFINITION 2.1. Let \mathcal{F} be a subset of $L^1_{\text{loc}}[0, \infty)$.

- 1) A subset D of \mathcal{H} is said to be *exactly controllable* at T in \mathcal{F} if $D \subset R_T(\mathcal{F})$.
- 2) A subset D of \mathcal{H} is said to be *exactly controllable* in \mathcal{F} if $D \subset \bigcup_{T>0} R_T(\mathcal{F})$.
- 3) The control system (2.1) (or (1.1) with (1.2)) is said to be *completely controllable* in \mathcal{F} if $\mathcal{H} = \overline{\bigcup_{T>0} R_T(\mathcal{F})}$.

In the following we consider several constraint sets of controls and investigate controllability of the control system (1.1) with (1.2) in those constraint sets.

3. Exact controllability of $D(A^2)$ in some constraint sets. In [13] Russell obtained exact controllability at some time $T > 0$ of the space $D(A)$ ($\equiv (H^2(0, 1) \cap H^1_0(0, 1)) \times H^1_0(0, 1)$) in the set $L^2(0, T)$ by solving the moment problem attached to this control problem. This moment problem is equivalent to the problem of constructing the biorthogonal system $\{p_k(t), q_k(t)\}$ in $L^2(0, T)$ of the system $\{\sin(\omega_k t), \cos(\omega_k t)\}$ with the square roots $\{\omega_k\}$ of the eigenvalues of the Sturm–Liouville operator which is stated below precisely, namely the system satisfying the equations

$$\begin{aligned} \int_0^T \sin(\omega_j t) p_k(t) dt &= \delta_{jk}, & \int_0^T \cos(\omega_j t) p_k(t) dt &= 0, \\ \int_0^T \sin(\omega_j t) q_k(t) dt &= 0 & \text{and} & \int_0^T \cos(\omega_j t) q_k(t) dt = \delta_{jk}. \end{aligned}$$

Taking $T > 0$ appropriately, we can show that there exists a bounded linear operator F on $L^2(0, T)$ such that $F[\sin(\omega_k t)] = \sin[2(k+1)\pi t/T]$ and $F[\cos(\omega_k t)] = \cos[2(k+1)\pi t/T]$, $k = 0, 1, 2, \dots$. Taking $p_k = F^*[\sin[2(k+1)\pi t/T]]$ and $q_k = F^*[\cos[2(k+1)\pi t/T]]$ with the adjoint operator F^* of F with respect to the $L^2(0, T)$ -inner product, we obtain the biorthogonal system $\{p_k(t), q_k(t)\}$. For details see [13], [16]. Although the operator F is also bounded on $H^1(0, T)$, it is not clear that F^* is bounded on $H^1(0, T)$. Hence it is not obvious that the domain of A^2 is exactly controllable in the set $H^1(0, T)$.

Now we show exact controllability of $D(A^2)$ in the set $H^1_p(0, T)$ for some $T > 0$ under a certain assumption on $g(x)$.

By Liouville's transformation, that is,

$$(3.1) \quad u^* = [c(x)\rho(x)]^{1/4}u, \quad x^* = \int_0^x \left[\frac{\rho(\xi)}{c(\xi)} \right]^{1/2} d\xi,$$

we obtain the simple control system in u^* involving derivatives with respect to x^* and t . Reverting to the use of u and x rather than u^* and x^* , we have

$$(3.2) \quad \frac{\partial^2 u}{\partial t^2}(x, t) - \frac{\partial^2 u}{\partial x^2}(x, t) - r(x)u(x, t) = \gamma(x)f(t),$$

$$(3.3) \quad u(0, t) = u(l, t) = 0,$$

where

$$l = \int_0^1 \left[\frac{\rho(x)}{c(x)} \right]^{1/2} dx,$$

$r(x)$ is a continuous function on $[0, l]$ and $\gamma(x) \in L^2(0, l)$.

Put

$$\tilde{A} = \begin{pmatrix} 0 & I \\ P & 0 \end{pmatrix}, \quad P = \frac{d^2}{dx^2} + r(x),$$

$$D(\tilde{A}) = (H^2(0, l) \cap H_0^1(0, l)) \times H_0^1(0, l) \subset H_0^1(0, l) \times L^2(0, l),$$

$$D(P) = H^2(0, l) \cap H_0^1(0, l) \subset L^2(0, l),$$

and $\{-\lambda_k\}$ and $\{\phi_k\}$, $k=0, 1, 2, \dots$, be the eigenvalues and eigenfunctions of P respectively. Further let $\{\phi_k\}$ form an orthonormal basis in $L^2(0, l)$.

In his proof of exact controllability, Russell [13] made the following basic assumption on $\gamma(x)$.

Assumption (A). The coefficients γ_k , which are defined as

$$\gamma_k = \int_0^l \gamma(x) \phi_k(x) dx, \quad k=0, 1, 2, \dots,$$

satisfy $\liminf_{k \rightarrow \infty} k|\gamma_k| > 0$, and $\gamma_k \neq 0$, $k=0, 1, 2, \dots$.

Then we have the next theorem.

THEOREM 3.1. *Let $\gamma(x)$, which is the function obtained by Liouville's transformation (3.1) of the distributed function $g(x)$, satisfy Assumption (A). Then $D(A^2)$ (\equiv the domain of A^2) is exactly controllable at $t=2l$ in the control set $H_p^1(0, 2l)$ for the control system (1.1) with (1.2).*

Proof. The control system (3.2) with (3.3) is equivalent to the system (1.1) with (1.2). Hence it is sufficient to give controllability for (3.2) with (3.3). Let us set $\omega_k = \lambda_k^{1/2}$. Russell showed that there exists a control function $f(t) \in L^2(0, 2l)$ satisfying

$$(3.4) \quad \mu_k = \int_0^{2l} \frac{\gamma_k}{\omega_k} \sin(\omega_k(2l-s)) f(s) ds,$$

$$(3.5) \quad \nu_k = \int_0^{2l} \gamma_k \cos(\omega_k(2l-s)) f(s) ds, \quad k=0, 1, 2, \dots,$$

for any given $[u_1, v_1] \in D(A)$ with the expansions

$$u_1(x) = \sum_{k=0}^{\infty} \mu_k \phi_k(x), \quad v_1(x) = \sum_{k=0}^{\infty} \nu_k \phi_k(x),$$

and he obtained exact controllability of the space $D(A)$ in $L^2(0, 2l)$. For the details of the proof and further results see [13].

Now we construct a function $f(t)$ in $H_p^1(0, 2l)$ satisfying (3.4) and (3.5) for any given $[u_1, v_1] \in D(A^2)$. Then, by the same way as in [13], we have the result.

For any nonnegative real number m , let us define the Hilbert space \mathfrak{h}^m as the space of all sequences $c = (c_k)$ with

$$\|c\|_{\mathfrak{h}^m} = \left(c_0^2 + \sum_{k=1}^{\infty} k^{2m} c_k^2 \right)^{1/2} < \infty$$

endowed with the norm $\|\cdot\|_{\mathfrak{h}^m}$. Then, noting $\omega_k = (k+1)\pi/l + O(1/k)$, we easily see that the mapping which maps $u = \sum_{k=0}^{\infty} c_k \phi_k$ to $c = (c_k)$ is an isomorphism from $D(P^{m/2})$ to \mathfrak{h}^m . And hence the spaces $D(A^m)$ and $\mathfrak{h}^{m+1} \times \mathfrak{h}^m$ are isomorphic. Now let us rearrange an element $(\alpha, c, d) = (\alpha, (c_k), (d_k)) \in R \times \mathfrak{h}^m \times \mathfrak{h}^m$ (respectively $(c, d) = ((c_k), (d_k)) \in \mathfrak{h}^m \times \mathfrak{h}^m$) as $\langle \alpha, c, d \rangle = (\alpha, c_0, d_0, c_1, d_1, \dots)$ (respectively $\langle c, d \rangle = (c_0, d_0, c_1, d_1, \dots)$) and denote this space by $R \times \overline{\mathfrak{h}}^m$ (respectively $\overline{\mathfrak{h}}^m$). From now we consider an element in $R \times \overline{\mathfrak{h}}^m$ or $\overline{\mathfrak{h}}^m$ as a column vector and represent a linear operator on $R \times \overline{\mathfrak{h}}^m$ or $\overline{\mathfrak{h}}^m$

by a matrix with infinite components. By [13], [16], there exists an isomorphism F on $L^2(0, 2l)$ which maps $1, \sin [(k+1)\pi t/l]$ and $\cos [(k+1)\pi t/l]$ to $2l, 2l \sin (\omega_k t)$ and $2l \cos (\omega_k t)$, $k = 0, 1, 2, \dots$, respectively. Put

$$(3.6) \quad \begin{aligned} \xi_{-1}(t) &= 1, & \eta_{-1}(t) &= 1, \\ \xi_{2k}(t) &= \sin \left[\frac{(k+1)\pi t}{l} \right], & \xi_{2k+1}(t) &= \cos \left[\frac{(k+1)\pi t}{l} \right], \\ \eta_{2k}(t) &= \sin (\omega_k t), & \eta_{2k+1}(t) &= \cos (\omega_k t), \quad k = 0, 1, 2, \dots \end{aligned}$$

Then this implies that the matrix

$$F = ((\xi_i, \eta_j))_{\substack{i \geq -1 \\ j \geq -1}}$$

defines an isomorphism on $R \times \bar{\mathfrak{h}}^0$. Now for a fixed integer N , let us define a linear operator H_N on $R \times \bar{\mathfrak{h}}^0$ as $H_N = ((h_{ij}^N))$ where

$$h_{ij}^N = (\eta_i, \xi_j) \quad \text{when } -1 \leq i \leq 2N$$

and

$$h_{ij}^N = 2l\delta_{ij} \quad \text{when } i \geq 2N+1$$

with the functions ξ_i, η_j defined by (3.6). Put $G_N = F^* - H_N$, where F^* means the adjoint operator of F in the space $R \times \bar{\mathfrak{h}}^0$, i.e., the transposed matrix of F . Let the operator G_N maps $\langle \alpha, c, d \rangle = \langle \alpha, (c_k), (d_k) \rangle$ to $\langle \tilde{\alpha}, \tilde{c}, \tilde{d} \rangle = \langle \tilde{\alpha}, (\tilde{c}_k), (\tilde{d}_k) \rangle$, then we have

$$(3.7) \quad \begin{aligned} \tilde{\alpha} &= \tilde{c}_k = \tilde{d}_k = 0, & 0 \leq k \leq N, \\ \tilde{c}_k &= 2l\alpha(\eta_{2k}, \xi_{-1}) + \sum_{\substack{j=0 \\ j \neq k}}^{\infty} (\eta_{2k}, \xi_{2j})c_j + \sum_{j=0}^{\infty} (\eta_{2k}, \xi_{2j+1})d_j + [(\eta_{2k}, \xi_{2k}) - 1]c_k, \\ \tilde{d}_k &= 2l\alpha(\eta_{2k+1}, \xi_{-1}) + \sum_{j=0}^{\infty} (\eta_{2k+1}, \xi_{2j})c_j \\ &\quad + \sum_{\substack{j=0 \\ j \neq k}}^{\infty} (\eta_{2k+1}, \xi_{2j+1})d_j + [(\eta_{2k+1}, \xi_{2k+1}) - 1]d_k, \quad k \geq N+1. \end{aligned}$$

Noting $\omega_k = (k+1)\pi/l + O(1/k)$, we have

$$\begin{aligned} &\left| \int_0^{2l} \sin(\omega_k t) \sin \left[\frac{(j+1)\pi t}{l} \right] dt \right| \\ &= \frac{1}{2} \left| \left[\frac{1}{\omega_k + (j+1)\pi/l} - \frac{1}{\omega_k - (j+1)\pi/l} \right] \sin(2l\omega_k) \right| \\ &\leq (M_0/k) \left(\frac{1}{j} + \frac{1}{|j-k|} \right) \end{aligned}$$

for $k \geq N+1$, $k \neq j$. Here and hereafter we denote by M_n , $n = 0, 1, 2, \dots$, constants independent of j, k and N . Hence we have

$$|(\eta_{2k}, \xi_{2j})| \leq (M_1/k) \left(\frac{1}{j} + \frac{1}{|j-k|} \right) \quad \text{for } k \geq N+1, \quad k \neq j.$$

Similarly we have

$$\begin{aligned} |(\eta_{2k}, \xi_{2j+1})| &\leq (M_1/k) \left(\frac{1}{j} + \frac{1}{|j-k|} \right), \\ |(\eta_{2k+1}, \xi_{2j})| &\leq (M_1/k) \left(\frac{1}{j} + \frac{1}{|j-k|} \right), \\ |(\eta_{2k+1}, \xi_{2k+1})| &\leq (M_1/k) \left(\frac{1}{j} + \frac{1}{|j-k|} \right) \end{aligned}$$

for $k \geq N+1$, $k \neq j$. Further we have

$$\begin{aligned} |(\eta_{2k}, \xi_{-1})| &\leq M_1/k^2, & |(\eta_{2k+1}, \xi_{-1})| &\leq M_1/k^2, \\ |(\eta_{2k}, \xi_{2k}) - 1| &\leq M_1/k^2, & |(\eta_{2k+1}, \xi_{2k+1}) - 1| &\leq M_1/k^2, \\ |(\eta_{2k}, \xi_{2k+1})| &\leq M_1/k, & |(\eta_{2k+1}, \xi_{2k})| &\leq M_1/k \end{aligned}$$

for $k \geq N+1$. Hence, by Schwarz's inequality, we have

$$\begin{aligned} \tilde{c}_k^2 &\leq 2M_1^2 \left\{ 1/k^4 + (2/k^2) \sum_{j \neq k} (1/j^2 + 1/|j-k|^2) + 1/k^2 \right\} \\ &\quad \times \left(\alpha^2 + \sum_{j=0}^{\infty} c_j^2 + \sum_{j=0}^{\infty} d_j^2 \right) \\ &\leq (M_2/k^2) \|\langle \alpha, c, d \rangle\|_{R \times \mathfrak{h}^0}^2 \end{aligned}$$

for $k \geq N+1$. Similarly we have

$$\tilde{d}_k^2 \leq (M_2/k^2) \|\langle \alpha, c, d \rangle\|_{R \times \mathfrak{h}^0}^2 \quad \text{for } k \geq N+1.$$

Hence we have

$$\begin{aligned} \|\langle \tilde{\alpha}, \tilde{c}, \tilde{d} \rangle\|_{R \times \mathfrak{h}^0}^2 &= \tilde{\alpha}^2 + \sum_{k=0}^{\infty} \tilde{c}_k^2 + \sum_{k=0}^{\infty} \tilde{d}_k^2 \\ (3.8) \quad &= \sum_{k=N+1}^{\infty} \tilde{c}_k^2 + \sum_{k=N+1}^{\infty} \tilde{d}_k^2 \leq 2M_2 \sum_{k=N+1}^{\infty} (1/k^2) \|\langle \alpha, c, d \rangle\|_{R \times \mathfrak{h}^0}^2 \\ &\leq (M_3/N) \|\langle \alpha, c, d \rangle\|_{R \times \mathfrak{h}^0}^2 \end{aligned}$$

for any $\langle \alpha, c, d \rangle \in R \times \overline{\mathfrak{h}^0}$. This inequality (3.8) means

$$\|G_N\|^2 \leq M_3/N.$$

Taking N_0 so large that

$$M_3/N_0 \leq 1/(4\|F^{*-1}\|^2),$$

we see that there exists $(I - F^{*-1}G_N)^{-1}$ and the estimate

$$\|(I - F^{*-1}G_N)^{-1}\| \leq 2$$

holds for $N \geq N_0$. Since $H_N = F^* - G_N = F^*(I - F^{*-1}G_N)$, there exists $H_N^{-1} = (I - F^{*-1}G_N)^{-1}F^{*-1}$ and

$$(3.9) \quad \|H_N^{-1}\| \leq 2\|F^{*-1}\| \quad (\equiv M_4)$$

holds for $N \geq N_0$.

Now we define bounded linear operators $\underline{K}_N, \tilde{K}_N$ from $\mathfrak{h}^2 \times \mathfrak{h}^1$ to $R \times \overline{\mathfrak{h}^0}$, J from $R \times \overline{\mathfrak{h}^0}$ to $L^2(0, 2l)$, R_N from $L^2(0, 2l)$ to $R \times \overline{\mathfrak{h}^0}$ and \bar{R} from $L^2(0, 2l)$ to $\mathfrak{h}^2 \times \mathfrak{h}^1$ as

$$K_N(\mu, \nu) = \langle 0, (c_k), (d_k) \rangle$$

where

$$\begin{aligned} c_k &= \mu_k \omega_k / \gamma_k, & d_k &= \nu_k / \gamma_k, & 0 \leq k \leq N, \\ c_k &= d_k = 0, & k &\geq N+1, \\ \tilde{K}_N(\mu, \nu) &= \langle 0, (\mu_k \omega_k / \gamma_k), (\nu_k / \gamma_k) \rangle - K_N(\mu, \nu) \end{aligned}$$

for $(\mu, \nu) \equiv ((\mu_k), (\nu_k)) \in \mathfrak{h}^2 \times \mathfrak{h}^1$,

$$(J\langle \alpha, c, d \rangle)(t) = \alpha + \sum_{k=0}^{\infty} c_k \sin \left[\frac{(k+1)\pi t}{l} \right] + \sum_{k=0}^{\infty} d_k \cos \left[\frac{(k+1)\pi t}{l} \right]$$

for $\langle \alpha, c, d \rangle \equiv \langle \alpha, (c_k), (d_k) \rangle \in R \times \overline{\mathfrak{h}^0}$,

$$R_N f = \langle 0, (c_k), (d_k) \rangle$$

where

$$\begin{aligned} c_k &= (f(t), \sin(\omega_k t)), & d_k &= (f(t), \cos(\omega_k t)), & 0 \leq k \leq N, \\ c_k &= d_k = 0, & k &\geq N+1, \end{aligned}$$

and

$$\bar{R}f = \left(\left(\frac{\gamma_k}{\omega_k} (f(t), \sin(\omega_k t)) \right), \left(\gamma_k (f(t), \cos(\omega_k t)) \right) \right)$$

for $f \in L^2(0, 2l)$, respectively. And consider the operator

$$T_N \equiv \bar{R}[J\tilde{K}_N + JH_N^{-1}(K_N - R_N J\tilde{K}_N)] - I.$$

By Assumption (A) and $\omega_k = (k+1)\pi/l + 0(1/k)$, we have

$$\begin{aligned} &(\mu_0 \omega_0 / \gamma_0)^2 + \sum_{k=1}^{\infty} k^2 (\mu_k / \gamma_k)^2 + (\nu_0 / \gamma_0)^2 + \sum_{k=1}^{\infty} k^2 (\nu_k / \gamma_k)^2 \\ (3.10) \quad &\cong M_5 \left\{ \mu_0^2 + \sum_{k=1}^{\infty} k^6 \mu_k^2 + \nu_0^2 + \sum_{k=1}^{\infty} k^4 \nu_k^2 \right\} \\ &= M_5 \|(\mu, \nu)\|_{\mathfrak{h}^3 \times \mathfrak{h}^2}^2 \end{aligned}$$

with a constant M_5 independent of N for any $(\mu, \nu) \in \mathfrak{h}^3 \times \mathfrak{h}^2$. This implies that the operators K_N and \tilde{K}_N are bounded from $\mathfrak{h}^3 \times \mathfrak{h}^2$ to $R \times \overline{\mathfrak{h}^1}$ and the operator norms, which are denoted by $\|K_N\|$ and $\|\tilde{K}_N\|$ respectively, are estimated by a constant independent of N . Next, integrating by parts, we have

$$\begin{aligned} \int_0^{2l} f(t) \sin(\omega_k t) dt &= -\frac{1}{\omega_k} \cos(2l\omega_k) f(2l) + \frac{1}{\omega_k} f(0) + \frac{1}{\omega_k} \int_0^{2l} f'(t) \cos(\omega_k t) dt, \\ \int_0^{2l} f(t) \cos(\omega_k t) dt &= \frac{1}{\omega_k} \sin(2l\omega_k) f(2l) - \frac{1}{\omega_k} \int_0^{2l} f'(t) \sin(\omega_k t) dt \end{aligned}$$

for any $f(t) \in H_p^1(0, 2l)$. Noting $\omega_k = (k+1)\pi/l + 0(1/k)$, $f(0) = f(2l)$ and $\sup_{0 < t < 2l} |f(t)| \leq \text{const.} \|f\|_1$, we see that there exists a constant M_6 independent of k

such that

$$|(f(t), \sin(\omega_k t))| \leq (M_6/k^2) \|f\|_1 + \frac{1}{\omega_k} |(f'(t), \cos(\omega_k t))|,$$

$$|(f(t), \cos(\omega_k t))| \leq (M_6/k^2) \|f\|_1 + \frac{1}{\omega_k} |(f'(t), \sin(\omega_k t))|,$$

$k = 0, 1, 2, \dots$. If a function $f'(t) \in L^2(0, 2l)$ is represented as

$$f'(t) = \alpha + \sum_{k=0}^{\infty} c_k \sin \left[\frac{(k+1)\pi t}{l} \right] + \sum_{k=0}^{\infty} d_k \cos \left[\frac{(k+1)\pi t}{l} \right],$$

then we easily see that F^* maps $\langle \alpha, (c_k), (d_k) \rangle$ to $\langle 2l\alpha, ((f'(t), \sin(\omega_k t))), ((f'(t), \cos(\omega_k t))) \rangle$. Noting $L^2(0, 2l)$ and $R \times \mathfrak{h}^0$ are isomorphic and F^* is bounded on $R \times \mathfrak{h}^0$, we have

$$\begin{aligned} & \sum_{k=0}^{\infty} (f'(t), \sin(\omega_k t))^2 + \sum_{k=0}^{\infty} (f'(t), \cos(\omega_k t))^2 \\ & \leq \|\langle 2l\alpha, ((f'(t), \sin(\omega_k t))), ((f'(t), \cos(\omega_k t))) \rangle\|_{R \times \mathfrak{h}^0}^2 \\ & \leq \|F^*\|^2 \|\langle \alpha, c, d \rangle\|_{R \times \mathfrak{h}^0}^2 \leq M_7 \|f'\|_0^2, \end{aligned}$$

Hence we have

$$\begin{aligned} \|R_N f\|_{R \times \mathfrak{h}^1}^2 & \leq |(f(t), \sin(\omega_0 t))|^2 + |(f(t), \cos(\omega_0 t))|^2 \\ & \quad + \sum_{k=1}^{\infty} k^2 |(f(t), \sin(\omega_k t))|^2 + \sum_{k=1}^{\infty} k^2 |(f(t), \cos(\omega_k t))|^2 \\ (3.11) \quad & \leq M_8 \left\{ \|f\|^2 + \|f\|_1^2 \sum_{k=1}^{\infty} (1/k^2) \right. \\ & \quad \left. + \sum_{k=1}^{\infty} (|(f'(t), \sin(\omega_k t))|^2 + |(f'(t), \cos(\omega_k t))|^2) \right\} \\ & \leq M_9 \|f\|_1^2. \end{aligned}$$

Therefore R_N is a bounded operator from $H_p^1(0, \infty)$ to $R \times \mathfrak{h}^1$ with the operator norm estimated by a constant independent of N . Further it is clear to see that J and \bar{R} are bounded operators from $R \times \mathfrak{h}^1$ to $H_p^1(0, 2l)$ and $H_p^1(0, 2l)$ to $\mathfrak{h}^3 \times \mathfrak{h}^2$ respectively. Hence if $H_N^{-1}(K_N - R_N J \bar{K}_N)$ is bounded from $\mathfrak{h}^3 \times \mathfrak{h}^2$ to $R \times \mathfrak{h}^1$, then T_N is a bounded operator on $\mathfrak{h}^3 \times \mathfrak{h}^2$. Now we show the boundedness of $H_N^{-1}(K_N - R_N J \bar{K}_N)$ and give the estimate of the operator norm. For $(\mu, \nu) \equiv ((\mu_k), (\nu_k)) \in \mathfrak{h}^3 \times \mathfrak{h}^2$, put $\langle \hat{\alpha}, \hat{c}, \hat{d} \rangle (\equiv \langle \hat{\alpha}, (\hat{c}_k), (\hat{d}_k) \rangle) = H_N^{-1}(K_N - R_N J \bar{K}_N)(\mu, \nu)$. If we put $\langle \tilde{\alpha}, \tilde{c}, \tilde{d} \rangle (\equiv \langle \tilde{\alpha}, (\tilde{c}_k), (\tilde{d}_k) \rangle) = (K_N - R_N J \bar{K}_N)(\mu, \nu)$, then

$$\begin{aligned} \tilde{\alpha} &= \tilde{c}_k = \tilde{d}_k = 0, \quad k \geq N+1, \\ (3.12) \quad \tilde{c}_k &= \mu_k \omega_k / \gamma_k - (h_N(t) \sin(\omega_k t)), \\ \tilde{d}_k &= \nu_k / \gamma_k - (h_N(t), \cos(\omega_k t)), \quad 0 \leq k \leq N, \end{aligned}$$

where $h_N(t) = (J \bar{K}_N(\mu, \nu))(t)$. By the definition of H_N , we easily see

$$(3.13) \quad \hat{\alpha} = \tilde{\alpha} = 0, \quad \hat{c}_k = \tilde{c}_k = 0, \quad \hat{d}_k = \tilde{d}_k = 0, \quad k \geq N+1.$$

Hence we have

$$\begin{aligned}
 \|\langle \hat{\alpha}, \hat{c}, \hat{d} \rangle\|_{R \times \mathfrak{h}^1}^2 &= \hat{\alpha}^2 + \hat{c}_0^2 + \hat{d}_0^2 + \sum_{k=1}^{\infty} k^2 \hat{c}_k^2 + \sum_{k=1}^{\infty} k^2 \hat{d}_k^2 \\
 &= \hat{c}_0^2 + \hat{d}_0^2 + \sum_{k=1}^N k^2 \hat{c}_k^2 + \sum_{k=1}^N k^2 \hat{d}_k^2 \\
 (3.14) \quad &\leq N^2 \|\langle \hat{\alpha}, \hat{c}, \hat{d} \rangle\|_{R \times \mathfrak{h}^0}^2 \leq N^2 \|H_N^{-1}\|^2 \|\langle \tilde{\alpha}, \tilde{c}, \tilde{d} \rangle\|_{R \times \mathfrak{h}^1}^2 \\
 &\leq M_4^2 N^2 \|\langle \tilde{\alpha}, \tilde{c}, \tilde{d} \rangle\|_{R \times \mathfrak{h}^1}^2 \\
 &\leq M_4^2 N^2 \|K_N - R_N J \tilde{K}_N\|^2 \|(\mu, \nu)\|_{\mathfrak{h}^3 \times \mathfrak{h}^2}^2 \leq M_{10} N^2 \|(\mu, \nu)\|_{\mathfrak{h}^3 \times \mathfrak{h}^2}^2,
 \end{aligned}$$

where $\|K_N - R_N J \tilde{K}_N\|$ denotes the operator norm from $\mathfrak{h}^3 \times \mathfrak{h}^2$ to $R \times \mathfrak{h}^1$, which is independent of N by (3.11). Thus T_N is bounded on $\mathfrak{h}^3 \times \mathfrak{h}^2$.

Finally we give the estimate of the norm of T_N . Let $(\mu, \nu) \equiv ((\mu_k), (\nu_k))$ be any element in $\mathfrak{h}^3 \times \mathfrak{h}^2$ and $(\bar{\mu}, \bar{\nu}) \equiv ((\bar{\mu}_k), (\bar{\nu}_k)) = \bar{R}[J \tilde{K}_N + J H_N^{-1}(K_N - R_N J \tilde{K}_N)](\mu, \nu)$. Further let us put

$$\begin{aligned}
 h_N(t) &= (J \tilde{K}_N(\mu, \nu))(t) \\
 &= \sum_{k=N+1}^{\infty} (\mu_k \omega_k / \gamma_k) \sin \left[\frac{(k+1)\pi t}{l} \right] + \sum_{k=N+1}^{\infty} (\nu_k / \gamma_k) \cos \left[\frac{(k+1)\pi t}{l} \right], \\
 g_N(t) &= [J H_N^{-1}(K_N - R_N J \tilde{K}_N)(\mu, \nu)](t) \\
 &= \sum_{k=0}^N \hat{c}_k \sin \left[\frac{(k+1)\pi t}{l} \right] + \sum_{k=0}^N \hat{d}_k \cos \left[\frac{(k+1)\pi t}{l} \right].
 \end{aligned}$$

Then

$$\begin{aligned}
 (3.15) \quad &(g_N(t), \sin(\omega_k t)) \\
 &= \left(\sum_{j=0}^N \hat{c}_j \sin \left[\frac{(j+1)\pi t}{l} \right] + \sum_{j=0}^N \hat{d}_j \cos \left[\frac{(j+1)\pi t}{l} \right], \sin(\omega_k t) \right) \\
 &= \sum_{j=0}^N \left(\sin(\omega_k t), \sin \left[\frac{(j+1)\pi t}{l} \right] \right) \hat{c}_j + \sum_{j=0}^N \left(\sin(\omega_k t), \cos \left[\frac{(j+1)\pi t}{l} \right] \right) \hat{d}_j
 \end{aligned}$$

and

$$\begin{aligned}
 (3.16) \quad &(g_N(t), \cos(\omega_k t)) = \sum_{j=0}^N \left(\cos(\omega_k t), \sin \left[\frac{(j+1)\pi t}{l} \right] \right) \hat{c}_j \\
 &\quad + \sum_{j=0}^N \left(\cos(\omega_k t), \cos \left[\frac{(j+1)\pi t}{l} \right] \right) \hat{d}_j.
 \end{aligned}$$

Noting the definition of H_N and putting

$$\begin{aligned}
 \bar{a} = \bar{c}_k = \bar{d}_k &= 0, \quad k \geq N+1, \\
 \bar{c}_k &= (g_N(t), \sin(\omega_k t)), \quad \bar{d}_k = (g_N(t), \cos(\omega_k t)), \quad 0 \leq k \leq N,
 \end{aligned}$$

we see that (3.15) and (3.16) imply

$$\langle \bar{a}, \bar{c}, \bar{d} \rangle = H_N \langle \hat{\alpha}, \hat{c}, \hat{d} \rangle = H_N H_N^{-1} \langle \tilde{\alpha}, \tilde{c}, \tilde{d} \rangle = \langle \tilde{\alpha}, \tilde{c}, \tilde{d} \rangle,$$

where $\langle \tilde{\alpha}, \tilde{c}, \tilde{d} \rangle$ is the element defined by (3.12). Hence, for $0 \leq k \leq N$, we have

$$\begin{aligned} \bar{\mu}_k &= \frac{\gamma_k}{\omega_k} (h_N(t) + g_N(t), \sin(\omega_k t)) \\ &= \frac{\gamma_k}{\omega_k} [(h_N(t), \sin(\omega_k t)) + \tilde{c}_k] = \mu_k, \end{aligned} \quad (3.17)$$

$$\bar{\nu}_k = \gamma_k [(h_N(t), \cos(\omega_k t)) + \tilde{d}_k] = \nu_k. \quad (3.18)$$

Put $f_N(t) = g_N(t) + h_N(t)$. Then, for $k \geq N+1$,

$$\mu_k = (\gamma_k / \omega_k) \left(f_N(t), \sin \left[\frac{(k+1)\pi t}{l} \right] \right)$$

and

$$\bar{\mu}_k = (\gamma_k / \omega_k) (f_N(t), \sin(\omega_k t)).$$

Hence

$$|\mu_k - \bar{\mu}_k| = \left| (\gamma_k / \omega_k) \left(f_N(t), \sin(\omega_k t) - \sin \left[\frac{(k+1)\pi t}{l} \right] \right) \right|$$

for $k \geq N+1$. Integrating by parts and noting $f_N(0) = f_N(2l)$, we have

$$\begin{aligned} & \int_0^{2l} f_N(t) \left\{ \sin(\omega_k t) - \sin \left[\frac{(k+1)\pi t}{l} \right] \right\} dt \\ &= \frac{1}{\omega_k} \{1 - \cos(2l\omega_k)\} f_N(0) \\ &+ \int_0^{2l} f'_N(t) \left\{ \frac{1}{\omega_k} \cos(\omega_k t) - \frac{l}{(k+1)\pi} \cos \left[\frac{(k+1)\pi t}{l} \right] \right\} dt. \end{aligned}$$

Since

$$\begin{aligned} \frac{1}{\omega_k} |1 - \cos(2l\omega_k)| &= \frac{1}{\omega_k} |\cos[2(k+1)\pi] - \cos(2l\omega_k)| \leq M_{11}/k^3, \\ \left| \frac{1}{\omega_k} \cos(\omega_k t) - \frac{l}{(k+1)\pi} \cos \left[\frac{(k+1)\pi t}{l} \right] \right| \\ &\leq \left| \frac{1}{\omega_k} \cos(\omega_k t) - \frac{1}{\omega_k} \cos \left[\frac{(k+1)\pi t}{l} \right] \right| + \left| \frac{1}{\omega_k} - \frac{l}{(k+1)\pi} \right| \left| \cos \left[\frac{(k+1)\pi t}{l} \right] \right| \\ &\leq M_{11}/k^3 \end{aligned}$$

and

$$|f_N(0)| + \|f'_N\| \leq M_{12} \|f_N\|_1,$$

we have

$$\left| \int_0^{2l} f_N(t) \left\{ \sin(\omega_k t) - \sin \left[\frac{(k+1)\pi t}{l} \right] \right\} dt \right| \leq M_{13} \|f_N\|_1 / k^3.$$

Hence

$$\begin{aligned} |\mu_k - \bar{\mu}_k| &= \left| \frac{\gamma_k}{\omega_k} \left(f_N(t), \sin(\omega_k t) - \sin \left[\frac{(k+1)\pi t}{l} \right] \right) \right| \\ &\leq M_{14} \|f_N\|_1 / k^5 \quad \text{for } k \geq N+1. \end{aligned} \quad (3.19)$$

Similarly we have

$$(3.20) \quad \begin{aligned} |\nu_k - \bar{\nu}_k| &= \left| \gamma_k \left(f_N(t), \cos(\omega_k t) - \cos \left[\frac{(k+1)\pi t}{l} \right] \right) \right| \\ &\leq M_{14} \|f_N\|_1 / k^4 \quad \text{for } k \geq N+1. \end{aligned}$$

By (3.17), (3.18), (3.19) and (3.20), the inequality

$$\begin{aligned} \|T_N(\mu, \nu)\|_{\mathfrak{h}^3 \times \mathfrak{h}^2}^2 &= \|(\mu, \nu) - (\bar{\mu}, \bar{\nu})\|_{\mathfrak{h}^3 \times \mathfrak{h}^2}^2 \\ &= \sum_{k=N+1}^{\infty} (k^6 |\mu_k - \bar{\mu}_k|^2 + k^4 |\nu_k - \bar{\nu}_k|^2) \\ &\leq M_{15} \|f_N\|_1^2 \sum_{k=N+1}^{\infty} (1/k^4) \leq M_{16} \|f_N\|_1^2 / N^3 \end{aligned}$$

holds. Further we have, by (3.10) and (3.14),

$$\begin{aligned} \|f_N\|_1^2 &= \|h_N\|_1^2 + \|g_N\|_1^2 \\ &\leq \|J\|^2 \{ \|\tilde{K}_N(\mu, \nu)\|_{R \times \mathfrak{h}^1}^2 + \|H_N^{-1}(K_N - R_N J \tilde{K}_N)(\mu, \nu)\|_{R \times \mathfrak{h}^1}^2 \} \\ &\leq \|J\|^2 \{ \|\tilde{K}_N\|^2 + \|H_N^{-1}(K_N - R_N J \tilde{K}_N)\|^2 \} \|(\mu, \nu)\|_{\mathfrak{h}^3 \times \mathfrak{h}^2}^2 \\ &\leq M_{17} N^2 \|(\mu, \nu)\|_{\mathfrak{h}^3 \times \mathfrak{h}^2}^2, \end{aligned}$$

where $\|J\|$ and $\|\tilde{K}_N\|$ denote the operator norm of J and K_N from $R \times \mathfrak{h}^1$ to $H_p^1(0, 2l)$ and $\mathfrak{h}^3 \times \mathfrak{h}^2$ to $R \times \mathfrak{h}^1$ respectively. Therefore

$$\|T_N(\mu, \nu)\|_{\mathfrak{h}^3 \times \mathfrak{h}^2}^2 \leq (M_{18}/N) \|(\mu, \nu)\|_{\mathfrak{h}^3 \times \mathfrak{h}^2}^2 \quad \text{for any } (\mu, \nu) \in \mathfrak{h}^3 \times \mathfrak{h}^2.$$

Thus the operator norm of T_N on $\mathfrak{h}^3 \times \mathfrak{h}^2$ is estimated as

$$\|T_N\| \leq (M_{18}/N)^{1/2}.$$

Take N large so that $\|T_N\| \leq 1$. Then there exists $(I - T_N)^{-1}$. For any given $[u_1, v_1] \in D(\tilde{A}^2)$ with the expansions

$$u_1(x) = \sum_{k=0}^{\infty} \mu_k \phi_k(x), \quad v_1(x) = \sum_{k=0}^{\infty} \nu_k \phi_k(x),$$

let us define $\bar{f}(t) \in H_p^1(0, 2l)$ as

$$\bar{f}(t) = [(J\tilde{K}_N + JH_N^{-1}(K_N - R_N J\tilde{K}_N))(I + T_N)^{-1}(\mu, \nu)](t).$$

Then, by the definition of T_N , we have

$$(3.21) \quad \bar{R}\bar{f} = (\mu, \nu).$$

If we put $f(t) = \bar{f}(2l - t)$, then (3.21) implies

$$\int_0^{2l} \frac{\gamma_k}{\omega_k} \sin[\omega_k(2l - s)] f(s) ds = \mu_k, \quad \int_0^{2l} \gamma_k \cos[\omega_k(2l - s)] f(s) ds = \nu_k,$$

$k = 0, 1, 2, \dots$. This completes the proof.

We denote by $\tilde{H}_p^1(0, \infty)$ the space of all real valued functions defined on $t \geq 0$ such that $f(t)$, $2kl \leq t < 2(k+1)l$, belong to $H_p^1(2kl, 2(k+1)l)$, $k = 0, 1, 2, \dots$, and

$$\|f\|_{\tilde{H}_p^1(0, \infty)} \equiv \left(\sum_{k=0}^{\infty} \|f\|_{H_p^1(2kl, 2(k+1)l)}^2 \right)^{1/2} < \infty.$$

For any $\eta > 0$ let us take a constraint set of controls as

$$\{f \in \tilde{H}_p^1(0, \infty) \mid \|f\|_{\tilde{H}_p^1(0, \infty)} \leq \eta\}.$$

Then we have

THEOREM 3.2. *Let Assumption (A) be satisfied. Then, for any $\eta > 0$, the space $D(A^2)$ is exactly controllable in the constraint set $\{f \in \tilde{H}_p^1(0, \infty) \mid \|f\|_{\tilde{H}_p^1(0, \infty)} \leq \eta\}$.*

Proof. The proof is similar to that of [10, Thm. 1]. Hence we omit the details of the proof. Note, by Theorem 3.1, $D(A^2)$ is exactly controllable in $H_p^1(0, 2l)$. Further as is mentioned in § 3, A generates a unitary group on \mathcal{H} , and hence on $D(A^2)$. And consider the space $H_p^1(0, 2l)$ and the constraint set $\{f \in \tilde{H}_p^1(0, \infty) \mid \|f\|_{\tilde{H}_p^1(0, \infty)} \leq \eta\}$ in place of $L^p(0, T_0; Y)$ and \mathcal{F}_η^p in [10]. Then we obtain null controllability of $D(A^2)$ in $\{f \in \tilde{H}_p^1(0, \infty) \mid \|f\|_{\tilde{H}_p^1(0, \infty)} \leq \eta\}$. Since the control system (1.1) with (1.2) is invariant under time reversal, we obtain the result.

Next we define the space $\tilde{C}^{0,1/2}(0, \infty)$ as the space of all functions which are Hölder continuous of order $1/2$ on each interval $(2kl, 2(k+1)l)$, $k = 0, 1, 2, \dots$, endowed with the norm

$$\|f\|_{\tilde{C}^{0,1/2}(0, \infty)} \equiv \sup_k \|f\|_{\tilde{C}^{0,1/2}(2kl, 2(k+1)l)},$$

where

$$\|f\|_{\tilde{C}^{0,1/2}(2kl, 2(k+1)l)} = \sup_{t \in (2kl, 2(k+1)l)} |f(t)| + \sup_{\substack{t \neq \tau \\ t, \tau \in (2kl, 2(k+1)l)}} \frac{|f(t) - f(\tau)|}{|t - \tau|^{1/2}}.$$

Then, by Sobolev's inequality, $\tilde{H}_p^1(0, \infty) \subset \tilde{C}^{0,1/2}(0, \infty)$ and

$$\|f\|_{\tilde{C}^{0,1/2}(2kl, 2(k+1)l)} \leq \bar{M} \|f\|_{H_p^1(2kl, 2(k+1)l)}$$

for any $f \in H_p^1(2kl, 2(k+1)l)$, with a constant \bar{M} independent of f and k . Hence, by Theorem 3.2, we easily have the following corollary.

COROLLARY 3.1. *Let Assumption (A) be satisfied. Then, for any $\eta > 0$, the space $D(A^2)$ is exactly controllable in the constraint set $\{f \in \tilde{C}^{0,1/2}(0, \infty) \mid \|f\|_{\tilde{C}^{0,1/2}(0, \infty)} \leq \eta\}$.*

4. Complete controllability with bang-bang controls. In this section we are concerned with the control problem with bang-bang controls, i.e., the control problem in the constraint set

$$\{f(t) = \chi_E(t)|E; \text{ measurable set in } (0, \infty)\},$$

where $\chi_E(t)$ is a characteristic function of E .

First we begin with this lemma.

LEMMA 4.1. *Let Assumption (A) be satisfied. Then the control system (1.1) with (1.2) is completely controllable in the constraint set*

$$\{f \in \tilde{C}^{0,1/2}(0, \infty) \mid \|f\|_{\tilde{C}^{0,1/2}(0, \infty)} \leq 1 \text{ and } 0 \leq f(t) \leq 1 \text{ for all } t \geq 0\}.$$

Proof. Let $w(x, t)$ be the mild solution of (2.1) with $f(t) = \frac{1}{2}$ and initial data $w(x, 0) = (\partial w / \partial t)(x, 0) = 0$. Then, by Lemma 4.1 in [11], $w(x, t) \equiv w(t)$ satisfies the inequality

$$\|[w(t), (\partial w / \partial t)(t)]\|_{D(A)} \leq r_0$$

with a constant r_0 independent of t . For the proof, see [11]. If we denote by \mathcal{R}_T the attainable set at T in $\{f \in \tilde{C}^{0,1/2}(0, \infty) \mid \|f\|_{\tilde{C}^{0,1/2}(0, \infty)} \leq \frac{1}{2}\}$, then the attainable set at T in $\{f + \frac{1}{2}f \in \tilde{C}^{0,1/2}(0, \infty), \|f\|_{\tilde{C}^{0,1/2}(0, \infty)} \leq \frac{1}{2}\}$ is equal to $[w(T), (\partial w / \partial t)(T)] + \mathcal{R}_T$. Taking

$\eta = \frac{1}{2}$ in Corollary 3.1 and noting that $D(A^2)$ is dense in \mathcal{H} , we have $\overline{\bigcup_{T>0} \mathcal{R}_T} = \mathcal{H}$. Since $\| [w(T), (\partial w / \partial t)(T)] \| \leq r_0$ for any $T > 0$, we have

$$\overline{\bigcup_{T>0} \{ [w(T), (\partial w / \partial t)(T)] \} + \mathcal{R}_T} = \mathcal{H}.$$

Clearly

$$\begin{aligned} & \{ f + \tfrac{1}{2} | f \in \tilde{C}^{0,1/2}(0, \infty), \| f \|_{\tilde{C}^{0,1/2}(0, \infty)} \leq \tfrac{1}{2} \} \\ & \subset \{ f \in \tilde{C}^{0,1/2}(0, \infty) | \| f \|_{\tilde{C}^{0,1/2}(0, \infty)} \leq 1 \text{ and } 0 \leq f(t) \leq 1 \text{ for all } t \geq 0 \}. \end{aligned}$$

Hence the attainable set in

$$\{ f \in \tilde{C}^{0,1/2}(0, \infty) | \| f \|_{\tilde{C}^{0,1/2}(0, \infty)} \leq 1 \text{ and } 0 \leq f(t) \leq 1 \text{ for all } t \geq 0 \}$$

contains $\bigcup_{T>0} \{ [w(T), (\partial w / \partial t)(T)] + \mathcal{R}_T \}$. Thus we complete the proof.

LEMMA 4.2. *Let $y(t)$ be an integrable m -dimensional vector valued function on a compact interval J . For each measurable subset $E \subset J$ consider the m -dimensional vector*

$$x_E = \int_E y(t) dt.$$

Let K be the set of all such points x_E as E varies over the collection of all measurable subsets of J . Then K is convex in R^m .

This lemma is a corollary of the following lemma which is usually called Lyapunov's convexity theorem.

LEMMA 4.3. *Let Σ be a σ -field of subsets of Ω , X be a finite-dimensional Banach space and $G: \Sigma \rightarrow X$ be a countably additive vector measure. If G is nonatomic, then the range of G is a compact convex subset of X .*

For the proof, see e.g. Diestel and Uhl [2].

Now we have complete controllability with bang-bang controls.

THEOREM 4.1. *Let Assumption (A) be satisfied. Then the control system (1.1) with (1.2) is completely controllable in the constraint set*

$$\{ f(t) = \chi_E(t) | E; \text{ measurable set in } (0, \infty) \}.$$

Proof. By Lemma 4.1, for any $[u_1, v_1] \in \mathcal{H}$ and $\varepsilon > 0$, there exist a state $[\bar{u}_1, \bar{v}_1]$ in \mathcal{H} , a control $f(t)$ in $\tilde{C}^{0,1/2}(0, \infty)$ satisfying $\| f \|_{\tilde{C}^{0,1/2}(0, \infty)} \leq 1$ and $0 \leq f(t) \leq 1$ and a positive integer N such that $\| [u_1, v_1] - [\bar{u}_1, \bar{v}_1] \| < \varepsilon$ and

$$\begin{pmatrix} \bar{u}_1 \\ \bar{v}_1 \end{pmatrix} = \int_0^{2Nl} U(2Nl - t) B f(t) dt.$$

Here $U(t)$ and B are the operators defined in § 2. For any $\varepsilon > 0$, we subdivide the intervals $[2kl, 2(k+1)l]$, $k = 0, 1, \dots, N-1$, into finite union of semi-open intervals as $[2kl, 2(k+1)l] = \sum_{j=1}^{m_k} K_{kj}$ with $|K_{kj}| < \varepsilon^2 / (4\|B\|Nl)^2$, where $|K_{kj}|$ is the length of the interval K_{kj} . We rearrange the suffix and denote these intervals by $\{J_m\}_{m=1,2,\dots,M} \equiv \{[a_m, b_m)\}_{m=1,2,\dots,M}$. Since $\| f \|_{\tilde{C}^{0,1/2}(0, \infty)} \leq 1$,

$$|f(t) - f(a_m)| \leq |t - a_m|^{1/2} \leq |J_m|^{1/2} \quad \text{for any } t \in J_m.$$

Hence putting

$$\bar{f}(t) = \sum_{m=1}^M \xi_m \chi_{J_m}(t) \quad \text{with} \quad \xi_m = f(a_m),$$

we have

$$|f(t) - \bar{f}(t)| \leq |J_m|^{1/2} < \varepsilon / (4\|B\|Nl) \quad \text{on } [0, 2Nl].$$

Thus, noting that $U(t)$ is unitary, we have

$$\begin{aligned}
 (4.1) \quad & \left\| \int_0^{2Nl} U(2Nl-t)Bf(t) dt - \int_0^{2Nl} U(2Nl-t)B\bar{f}(t) dt \right\| \\
 & \leq \int_0^{2Nl} \|B\| |f(t) - \bar{f}(t)| dt \\
 & \leq \|B\| \sum_{m=1}^M \int_{J_m} |f(t) - \bar{f}(t)| dt \\
 & \leq \varepsilon/2.
 \end{aligned}$$

Next, for any $\eta > 0$, let us take an integer K such that the inequality $(\sum_{k=K+1}^{\infty} \gamma_k^2)^{1/2} < \eta$ holds. We denote by Q_K the orthogonal projection from \mathcal{H} onto the closed subspace spanned by $\{[\phi_k, 0], [0, \phi_k]\}_{0 \leq k \leq K}$, where $\{\phi_k\}_{k=0,1,2,\dots}$ are eigenfunctions of the operator L which form an orthonormal basis in $L^2(0, l)$. Namely

$$Q_K \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \sum_{k=0}^K c_k \phi_k \\ \sum_{k=0}^K d_k \phi_k \end{pmatrix}$$

for an element $[u, v] = [\sum_{k=0}^{\infty} c_k \phi_k, \sum_{k=0}^{\infty} d_k \phi_k] \in \mathcal{H}$. Then, since $Q_K U(2Nl-t)Bf(t)$ is a summable K -dimensional vector valued function on each J_m , by Lemma 4.2, there exists a measurable subset $E_m (\subset J_m)$ such that

$$\begin{aligned}
 (4.2) \quad & \int_{J_m} Q_K U(2Nl-t)B\bar{f}(t) dt = \int_{J_m} Q_K U(2Nl-t)B\xi_m dt \\
 & = \int_{E_m} Q_K U(2Nl-t)B1 dt \\
 & = \int_{J_m} Q_K U(2Nl-t)B\chi_{E_m}(t) dt
 \end{aligned}$$

holds.

For any $\alpha \in R$ and $t \geq 0$, putting $[u(t), v(t)] = U(t)B\alpha$, we easily see

$$(4.3) \quad u(t) = \sum_{k=0}^{\infty} \frac{\alpha \gamma_k}{\omega_k} \sin(\omega_k t) \phi_k(x), \quad v(t) = \sum_{k=0}^{\infty} \alpha \gamma_k \cos(\omega_k t) \phi_k(x).$$

Further

$$\begin{aligned}
 (4.4) \quad & (I - Q_K)U(t)B\alpha = (I - Q_K) \begin{pmatrix} u(t) \\ v(t) \end{pmatrix} \\
 & = \begin{pmatrix} \sum_{k=K+1}^{\infty} \frac{\alpha \gamma_k}{\omega_k} \sin(\omega_k t) \phi_k \\ \sum_{k=K+1}^{\infty} \alpha \gamma_k \cos(\omega_k t) \phi_k \end{pmatrix} = U(t)(I - Q_K)B\alpha \quad \text{for any } \alpha \in R.
 \end{aligned}$$

Therefore we have, by (4.3) and (4.4),

$$\begin{aligned}
 & \left\| \int_{J_m} (I - Q_K) U(2Nl - t) B \bar{f}(t) dt - \int_{J_m} (I - Q_K) U(2Nl - t) B \chi_{E_m}(t) dt \right\| \\
 &= \left\| \int_{J_m} U(2Nl - t) (I - Q_K) B [\bar{f}(t) - \chi_{E_m}(t)] dt \right\| \\
 (4.5) \quad & \leq \int_{J_m} \|(I - Q_K) B [\bar{f}(t) - \chi_{E_m}(t)]\| dt \\
 & \leq 2 \left(\sum_{k=N+1}^{\infty} |\gamma_k|^2 \right)^{1/2} \cdot |J_m| \leq 2\eta |J_m|.
 \end{aligned}$$

Putting $\tilde{f}(t) = \sum_{m=1}^M \chi_{E_m}(t)$, we have, by (4.2) and (4.5),

$$\begin{aligned}
 & \left\| \int_0^{2Nl} U(2Nl - t) B \tilde{f}(t) dt - \int_0^{2Nl} U(2Nl - t) B \tilde{f}(t) dt \right\| \\
 & \leq \sum_{m=1}^M \left\| \int_{J_m} U(2Nl - t) B \bar{f}(t) dt - \int_{J_m} U(2Nl - t) B \chi_{E_m}(t) dt \right\| \\
 (4.6) \quad & \leq \sum_{m=1}^M \left\{ \left\| \int_{J_m} Q_K U(2Nl - t) B \bar{f}(t) dt - \int_{J_m} Q_K U(2Nl - t) B \chi_{E_m}(t) dt \right\| \right. \\
 & \quad \left. + \left\| \int_{J_m} (I - Q_K) U(2Nl - t) B \bar{f}(t) dt \right. \right. \\
 & \quad \left. \left. - \int_{J_m} (I - Q_K) U(2Nl - t) B \chi_{E_m}(t) dt \right\| \right\} \\
 & \leq 2\eta \sum_{m=1}^M |J_m| = 4\eta Nl.
 \end{aligned}$$

Taking K so large that $4\eta Nl < \varepsilon/2$, we have, by (4.1) and (4.6),

$$\left\| \int_0^{2Nl} U(2Nl - t) B f(t) dt - \int_0^{2Nl} U(2Nl - t) B \tilde{f}(t) dt \right\| < \varepsilon.$$

Consequently, for any $[u_1, v_1] \in \mathcal{H}$ and $\varepsilon > 0$, we have obtained a control

$$\tilde{f}(t) = \sum_{m=1}^M \chi_{E_m}(t) = \chi_E(t) \quad \left(E = \bigcup_{m=1}^M E_m \right)$$

and have shown the inequality

$$\left\| [u_1, v_1] - \int_0^{2Nl} U(2Nl - t) B \tilde{f}(t) dt \right\| < 2\varepsilon.$$

This shows complete controllability with bang-bang controls.

Remark 4.1. For simplicity we have considered the Dirichlet boundary condition. But we can obtain similar results for the control systems with more general boundary conditions, as were considered by Russell [13],

$$A_0 u(0, t) + B_0 \frac{\partial u}{\partial x}(0, t) = 0, \quad A_1 u(1, t) + B_1 \frac{\partial u}{\partial x}(1, t) = 0,$$

where A_0, B_0, A_1 and B_1 are real constants with $A_0^2 + B_0^2 \neq 0, A_1^2 + B_1^2 \neq 0$.

Remark 4.2. Fattorini and Russell [6], Krabs [9] considered some one-dimensional vibrating systems with boundary controls, and obtained exact controllability of those systems. We can obtain also complete controllability with bang-bang controls of those systems in the same manner.

Remark 4.3. Fattorini [5] considered a control system described by the wave equation as

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2}(x, t) - \sum_{i=1}^n \frac{\partial^2 u}{\partial x_i^2}(x, t) &= f(x, t) \quad \text{in } \Omega \times (0, \infty), \\ u(x, t) &= 0 \quad \text{on } \partial\Omega \times (0, \infty), \end{aligned}$$

where a function $f(x, t)$ is regarded as a control. There he showed that exact controllability holds for this system and that optimal time controls $f_0(x, t)$ under the constraint

$$\int_{\Omega} |f(x, t)|^2 dx \leq 1 \quad \text{almost all } t$$

satisfy the bang-bang principle, i.e.,

$$\int_{\Omega} |f_0(x, t)|^2 dx = 1 \quad \text{almost all } t.$$

Hence, for this control system, complete (even exact) controllability with bang-bang controls holds.

For control systems described by the wave equation with boundary controls, or more generally, elastodynamic systems, exact controllability has been obtained (see [14], [15], [12]), but the question of controllability with bang-bang controls appears to be open.

Remark 4.4. For the control system (1.1) with (1.2), it is well known that an optimal time control does not generally satisfy the bang-bang principle. Hence Theorem 4.1 is not trivial. Here we showed only complete controllability, i.e., approximate controllability, with bang-bang controls. We do not know whether $D(A^2)$ is exactly controllable or not with bang-bang controls. For the control system described by the heat equation, there exist temperature distributions attainable by constrained controls but not attainable by bang-bang controls. But Knowles [8] defined normal systems and showed that, for the problem of approximate controllability, the optimal time control for the normal system is unique and bang-bang. Namely, in any neighborhood of an element w_0 steered by constrained controls, there exists an element w_1 which is steered by a bang-bang control. For details see [8]. Thus for normal systems the statement in Theorem 4.1 is obvious as far as the result in Lemma 4.1 is valid. Knowles showed that the control system described by the parabolic equation is a normal system, under appropriate conditions, using the fact that it generates the analytic semigroup. But it seems to us that the control system (1.1) with (1.2) is not normal.

In [8], it is proven that the above-mentioned element w_1 can be steered by means of a bang-bang control in a time no larger than the optimal time for w_0 . Considering the optimal time in place of $2Nl$ in the proof of Theorem 4.1, we obtain a similar result also for the control system (1.1) with (1.2).

Acknowledgments. The author would like to express his hearty thanks to Professors A. Inoue, F-Y. Maeda and T. Miyakawa for their constant encouragement and kind discussions. Thanks are also due to the referees for their helpful comments and suggestions.

REFERENCES

- [1] R. BRAMMER, *Controllability in linear autonomous systems with positive controllers*, this Journal, 10 (1972), pp. 339–353.
- [2] J. DIESTEL AND J. J. UHL, JR., *Vector Measures*, Math. Surveys no. 15, American Mathematical Society, Providence, RI, 1977.
- [3] H. O. FATTORINI, *The time optimal control problem in Banach spaces*, Appl. Math. Optim., 1 (1974), pp. 163–188.
- [4] ———, *The time-optimal problem for boundary control of the heat equation* in Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976, pp. 305–320.
- [5] ———, *The time optimal problem for distributed control of systems described by the wave equation* in Control Theory of Systems Governed by Partial Differential Equations, A. K. Aziz, J. W. Wingate and M. J. Balas, eds., Academic Press, New York, 1977, pp. 151–175.
- [6] H. O. FATTORINI AND D. L. RUSSELL, *Exact controllability theorems for linear parabolic equations in one space dimension*, Arch. Rational Mech. Anal., 42 (1971), pp. 272–292.
- [7] G. KNOWLES, *Lyapunov vector measures*, this Journal, 13 (1975), pp. 294–303.
- [8] ———, *Time optimal control of infinite dimensional systems*, this Journal, 14 (1976), pp. 919–933.
- [9] W. KRABS, *On boundary controllability of one-dimensional vibrating systems*, Math. Mech. in the Appl. Sci., 1 (1979), pp. 322–345.
- [10] K. NARUKAWA, *Admissible null controllability and optimal time control*, Hiroshima Math. J., 11 (1981), pp. 533–551.
- [11] ———, *Admissible controllability of vibrating systems with constrained controls*, this Journal, 20 (1982), pp. 770–782.
- [12] ———, *Boundary value control of isotropic elastodynamic systems with constrained controls*, J. Math. Anal. Appl., 93 (1983), pp. 250–272.
- [13] D. L. RUSSELL, *Nonharmonic Fourier series in the control theory of distributed parameter systems*, J. Math. Anal. Appl., 18 (1967), pp. 542–560.
- [14] ———, *Control theory of hyperbolic equations related to certain questions in harmonic analysis and spectral theory*, J. Math. Anal. Appl., 40 (1972), pp. 336–368.
- [15] ———, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions* in Differential Games and Control Theory, Roxin, Liu and Sternberg, eds., Marcel Dekker, New York, 1974.
- [16] ———, *Controllability and stabilizability theory for linear partial differential equations: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.
- [17] S. H. SAPERSTONE, *Global controllability of linear systems with positive controls*, this Journal, 11 (1973), pp. 417–423.
- [18] S. H. SAPERSTONE AND J. YORKE, *Controllability of linear oscillatory systems using positive controls*, this Journal, 9 (1971), pp. 253–262.
- [19] G. SCHMIDT, *The “bang-bang” principle for the time optimal problem in boundary control of the heat equation*, this Journal, 18 (1980), pp. 101–107.

NONLINEAR OPTIMAL CONTROL PROBLEMS FOR PARABOLIC EQUATIONS*

AVNER FRIEDMAN†

Abstract. Let u be the solution of the heat equation with diffusion coefficient $k(x)$ ($-1 < x < 1$), initial values $h(x)$ and boundary values 0 on $x = \pm 1$. The function $k(x)$ is a control variable to be chosen from a suitable class so as to minimize $\int_{-1}^1 u^2(x, T) dx$ for some given $T > 0$. An explicit characterization of the optimal k is given. Other related problems are considered.

Key words. parabolic equations, optimal control, heat equation, Green's function

Introduction. The mathematical theory of optimal control for partial differential equations has been dealing almost exclusively with linear models, that is, with controls appearing either in the inhomogeneous term of the equation or in the initial or boundary data; see [1]–[4], [8].

In this paper we deal with control problems for which the control $k(x)$ appears as a coefficient in a parabolic equation for an unknown function $u(x, t)$. This is then a nonlinear control problem, and the objective is to find a control k which minimizes a certain given functional of $u(x, t)$ at time $t = T$. The functionals we consider are either

$$(0.1) \quad \int h(x)u(x, T) dx$$

or

$$(0.2) \quad \int u^2(x, T) dx.$$

The parabolic equations we shall work with are either

$$(0.3) \quad u_t = (ku_x)_x$$

or

$$(0.4) \quad u_t = au_{xx} + bu_x - ku,$$

with suitable initial and boundary conditions.

In a recent paper Friedman and Jiang [5] considered a nonlinear control problem for the Stefan problem as well as a special case of (0.4) with a functional (0.1) and characterized the optimal control. Their method is based on “linearization”, i.e., on small perturbations about the optimal control. This method requires quite a bit of regularity and therefore does not extend to models like (0.3) in which the control appears in higher derivatives of u . The method of the present paper is based on direct comparison of functionals corresponding to two controls.

The physical motivation for the control problem corresponding to (0.3) is as follows. Suppose an inhomogeneous rod of length 1 is to be constructed with some specifications for the conductivity coefficient $k(x)$ ($0 \leq x \leq 1$). We wish to design the rod in such a way that, with prescribed initial temperature $u = \phi(x)$ and boundary conditions $u_x(0, t) = 0$, $u(1, t) = 0$, the rod “cools off” as much as possible after T units

* Received by the editors January 14, 1983, and in revised form September 15, 1983. This paper was partially supported by the National Science Foundation under grant MCS-8300293.

† Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

of time. The "cooling off" is measured either by (0.1) or by (0.2). The specifications on the conductivity are

$$(0.5) \quad \alpha \leq k(x) \leq \beta, \quad \int_0^1 k(x) dx = \gamma, \quad kh' \downarrow \text{ if } x \uparrow$$

in case (0.1) (it is assumed that $h > 0$, $h' \leq 0$); for case (0.2) we replace the last condition in (0.5) by $k \uparrow$ if $x \uparrow$.

In the model corresponding to (0.4) the control k appears as a dissipation coefficient and the physical motivation is similar.

In this paper we shall obtain explicit description of the optimal control $k(x)$. For instance, for the problems associated with (0.3) and (0.1) or (0.2) the optimal control is given by

$$k(x) = \begin{cases} \alpha & \text{if } 0 \leq x < \theta, \\ \beta & \text{if } \theta < x \leq 1, \end{cases}$$

for suitable constant θ .

In § 1 we state the main results for (0.3); proofs are given in § 2 (for the functional (0.1)) and § 3 (for the functional (0.2)). In § 4 we study the optimal control problem for (0.4); we also consider problems in which the control appears as a coefficient of u in the boundary condition.

1. Control problems for $u_t = (ku_x)_x$. Consider the parabolic problem

$$(1.1) \quad \begin{aligned} u_t &= (k(x)u_x)_x \quad \text{in } Q_T = \{(x, t); 0 < x < 1, 0 < t < T\}, \\ u(x, 0) &= \phi(x) \quad \text{if } 0 < x < 1, \\ u_x(0, t) &= 0 \quad \text{if } 0 < t \leq T, \\ u(1, t) &= 0 \quad \text{if } 0 < t \leq T. \end{aligned}$$

The initial values $\phi(x)$ are assumed to satisfy:

$$(1.2) \quad \begin{aligned} &\text{either } \phi \equiv 1, \\ &\text{or } \phi \in C^2[0, 1], \quad \phi \geq 0, \\ &\quad \phi' \leq 0, \quad \phi'' \leq 0, \\ &\quad \phi(0) > 0, \quad \phi'(0) = 0, \quad \phi(1) = 0. \end{aligned}$$

The conductivity coefficient $k(x)$ will belong to the class

$$K = \left\{ k(x) \text{ measurable, } \alpha \leq k(x) \leq \beta, \int_0^1 k(x) dx = \gamma \right\}$$

where α, β, γ are positive and $\alpha < \gamma < \beta$. [If $\gamma \notin [\alpha, \beta]$ then K is empty, and if $\gamma = \alpha$ or $\gamma = \beta$ then K consists of the single function $k \equiv \gamma$.]

The solution of (1.1) is understood in the usual weak sense [7]; recall that

$$u_x \in L^2(Q_T) \quad \text{and} \quad u \in C^\delta(\bar{Q}_T) \quad \text{for some } \delta > 0.$$

Let $h(x)$ be a function satisfying:

$$(1.3) \quad \begin{aligned} &\text{either } h \equiv \text{const.} > 0 \\ &\text{or } h \in C^2[0, 1], \quad h \geq 0, \\ &\quad h'' \leq 0 \text{ in } [0, 1], \quad h'(x) < 0 \text{ in } (0, 1], \\ &\quad h(0) > 0, \quad h'(0) = 0, \quad h(1) = 0. \end{aligned}$$

We define

$$K_h = \{k \in K; (kh')' \leq 0 \text{ in } \mathcal{D}'(0, 1)\}.$$

Notice that if $h \equiv \text{const.}$ then $K_h = K$. Notice also that if $k \in K$ and $k(x)$ is increasing then $k \in K_h$ since, for smooth k ,

$$(kh')' = k'h' + kh'' \leq kh'' \leq 0;$$

for nonsmooth k , we approximate by mollification k_m .

Similarly $(k\phi')' \leq 0$ if $k \in K$ and $k(x)$ are increasing.

For any $k \in K_h$ we denote by u_k the unique solution of (1.1) and consider the functional

$$(1.4) \quad J_h(k) = \int_0^1 h(x) u_k(x, T) dx.$$

Problem (J_h). Find k_0 such that

$$(1.5) \quad k_0 \in K_h, \quad J_h(k_0) = \min_{k \in K_h} J_h(k).$$

Since $\alpha < \gamma < \beta$, there is a unique number $\theta \in (0, 1)$ such that $\theta\alpha + (1 - \theta)\beta = \gamma$. The function

$$(1.6) \quad k_0(x) = \begin{cases} \alpha & \text{if } 0 \leq x \leq \theta, \\ \beta & \text{if } \theta < x \leq 1, \end{cases}$$

is monotone increasing and in K ; hence it belongs to K_h .

THEOREM 1.1. *There exists a unique solution of problem (1.5), given by (1.6).*

For any control function $l \in K$ set

$$(1.7) \quad K_{h,l} = \{k \in K_h; k \text{ is a rearrangement of } l\};$$

by rearrangement k of l we mean that

$$\text{meas}\{k \in A\} = \text{meas}\{l \in A\} \quad \forall \text{ Borel set } A.$$

We introduce the following problem.

Problem ($J_{h,l}$). Find k_1 such that

$$(1.8) \quad k_1 \in K_{h,l}, \quad J_h(k_1) = \min_{k \in K_{h,l}} J_h(k).$$

THEOREM 1.2. *There exists a solution of problem (1.8), given by l^* , the monotone increasing rearrangement of l .*

Observe that the parabolic problem (1.1) is equivalent to the problem

$$(1.9) \quad \begin{aligned} u_t &= (k(x)u_x)_x \quad \text{in } \{(x, t); -1 < x < 1, 0 < t < T\}, \\ u(x, 0) &= \phi(x) \quad \text{if } -1 < x < 1, \\ u(\pm 1, t) &= 0 \quad \text{if } 0 < t < T \end{aligned}$$

provided $k(x) = k(-x)$, $\phi(x) = \phi(-x)$. The optimal control problem we considered is that of optimizing the cooling of a rod $-1 \leq x \leq 1$, with symmetric conductivity and initial temperature about its center, when its endpoints are kept at zero temperature. Theorems 1.1, 1.2 show that the best design is to take the conductivity coefficient "as large as possible" near the endpoints and "as small as possible" near the center. The amount of cooling at time $t = T$ was measured by (1.4).

Let us now measure the amount of cooling by the functional

$$(1.10) \quad \tilde{J}(k) = \int_0^1 u_k^2(x, T) dx$$

and take the control set to be

$$\tilde{K} = \{k \in K; k(x) \text{ is increasing}\}.$$

Problem (\tilde{J}). Find k_0 such that

$$(1.11) \quad k_0 \in \tilde{K}, \quad \tilde{J}(k_0) = \min_{k \in \tilde{K}} \tilde{J}(k).$$

THEOREM 1.3. *There exists a unique solution of problem (1.11), given by the function in (1.6).*

Theorems 1.1, 1.2 are proved in § 2 and Theorem 1.3 is proved in § 3.

Joel Friedman [6] considered problem (J_h) and proved that $u_{k_0} \leq u_{k^0}$ where

$$k^0(x) = \beta \quad \text{if } 0 \leq x \leq 1 - \theta, \quad k^0(x) = \alpha \quad \text{if } 1 - \theta < x \leq 1;$$

this implies that

$$J_h(k_0) \leq J_h(k^0) \quad \forall h, h \geq 0.$$

His proof works for any two piecewise-constant functions k_0, k^0 such that k_0 is an increasing rearrangement of k^0 and k^0 is monotone decreasing.

2. Proof of Theorems 1.1, 1.2.

LEMMA 2.1. *If $k \in K$, $(k\phi')' \leq 0$ in $\mathcal{D}'(0, 1)$ and $u = u_k$, then*

$$(2.1) \quad u_x \leq 0 \quad a.e.,$$

$$(2.2) \quad u_t \leq 0 \quad \text{in } \mathcal{D}'(Q_T).$$

Proof. Consider first the case $\phi \neq 1$ and $k(x)$ is smooth. Applying $k(x)(\partial/\partial x)$ to the first equation in (1.1), we get

$$(ku_x)_t = k(ku_x)_{xx}.$$

Since $u(1, t) = 0$, $u(x, t) \geq 0$, we have $u_x(1, t) \leq 0$. Also $u_x(0, t) = 0$ and $ku_x = k\phi' \leq 0$ at $t = 0$. Hence, by the maximum principle, $u_x < 0$ in Q_T (Notice that due to the consistency conditions $\phi'(0) = 0$, $\phi(1) = 0$, u_x is continuous in \bar{Q}_T .)

To prove that $u_t \leq 0$, we represent u by means of Green's function $G(x, t; y, s)$ of (0.1):

$$u(x, t) = \int_0^1 G(x, t; y, 0) \phi(y) dy.$$

Then

$$\begin{aligned} u_t(x, t) &= \int_0^1 G_t(x, t; y, 0) \phi(y) dy \\ (2.3) \quad &= - \int_0^1 G_s(x, t; y, 0) \phi(y) dy \\ &= \int_0^1 (k(y)G_y(x, t; y, 0))_y \phi(y) dy \end{aligned}$$

$$\begin{aligned}
&= - \int_0^1 k(y) G_y(x, t; y, 0) \phi'(y) dy \\
&\quad \text{(since } G_y(x, t; 0, s) = 0, \phi(1) = 0) \\
&= \int_0^1 G(x, t; y, 0) (k\phi')' dy \\
&\quad \text{(since } G(x, t; 1, s) = 0, \phi'(0) = 0) \\
&\leq 0 \quad \text{(since } (k\phi')' \leq 0).
\end{aligned}$$

We next consider the case where k is not assumed to be smooth.

We shall need the fact that

$$(2.4) \quad \int_0^1 k\phi'\zeta' \geq 0 \quad \forall \zeta \in C^{0,1}[0, 1], \quad \zeta \geq 0, \quad \zeta(1) = 0.$$

Since $(k\phi')' \leq 0$ in $\mathcal{D}'(0, 1)$, (2.4) is satisfied for any $\zeta \in C_0^\infty(0, 1)$ and, by approximation, for any $\zeta \in C^{0,1}[0, 1]$ with $\zeta(0) = 0, \zeta(1) = 0$. In case $\zeta(0) \neq 0$, we apply (2.4) to $\zeta(x) \min\{mx, 1\}$ and take $m \rightarrow \infty$; since $\phi'(0) = 0$, the assertion (2.4) follows.

To prove (2.1), (2.2) for general k , we use mollifications

$$k_m(x) = \int \rho_m(x-y)k(y) dy,$$

where $\text{supp } \rho_m = [0, 1/m]$. For any $\zeta \in C^{0,1}[0, 1]$, $\zeta \geq 0$ with $\zeta(1) = 0$, we have (taking $\zeta(x) = 0$ if $x > 1$)

$$\begin{aligned}
(2.5) \quad & - \int k_m \phi' \zeta' dx = - \int_0^{1/m} dy \int_0^1 \rho_m(y) k(x-y) \phi'(x) \zeta'(x) dx \\
& \leq - \int_0^{1/m} \rho_m(y) \int_0^1 k(x-y) \phi'(x-y) \zeta'(x) dx + \frac{C}{m} \int |\zeta'(x)| \\
& \leq - \int_0^{1/m} \rho_m(y) \int_0^1 k(z) \phi'(z) \zeta'(z+y) + C \int_0^{1/m} |\zeta'(x)| + \frac{C}{m} \int |\zeta'(x)| \\
& \leq C \int_0^{1/m} |\zeta'(x)| + \frac{C}{m} \int |\zeta'(x)| \quad \text{by (2.4).}
\end{aligned}$$

Denote by u_m the solution u corresponding to k_m .

By (2.3) (applied to u_m) and (2.5),

$$\frac{\partial}{\partial t} u_m(x, t) \leq C \int_0^{1/m} \left| \frac{\partial}{\partial y} G_m(x, t; y, 0) dy \right| + \frac{C}{m} \int_0^1 \left| \frac{\partial}{\partial y} G_m(x, t; y, 0) \right| dy$$

where G_m is Green's function corresponding to k_m . Replacing 0 by any s , $0 \leq s \leq \delta$ (with $0 < \delta < t$) and integrating with respect to s , we get

$$\begin{aligned}
(2.6) \quad & \frac{\partial}{\partial t} u_m(x, t) \leq \frac{C}{\delta} \int_0^\delta \int_0^{1/m} \left| \frac{\partial}{\partial y} G_m(x, t; y, s) \right| dy ds \\
& + \frac{C}{m\delta} \int_0^\delta \int_0^1 \left| \frac{\partial}{\partial y} G_m(x, t; y, s) \right| dy ds \leq \frac{C}{\sqrt{m}}
\end{aligned}$$

since $\iint |\partial G_m / \partial y|^2 dy ds \leq C$, C independent of m . Since $u_m \rightarrow u$ uniformly in Q_T as

$m \rightarrow \infty$, we conclude from (2.6) that (2.2) holds. Next, the proof that $\partial u_m / \partial x \leq 0$ is the same as before, and (2.1) follows by letting $m \rightarrow \infty$.

In case $\phi \equiv 1$, the proof of (2.1) is obtained by approximating ϕ by functions ϕ_m as in (1.2) with $\phi_m(1) = 0$. The proof of (2.2) can be obtained by showing (using the maximum principle) that $u(x, t + \varepsilon) - u(x, t) < 0$ for any $\varepsilon > 0$.

We now proceed to prove Theorems 1.1, 1.2. We take any smooth control functions k, \tilde{k} in K_h with \tilde{k} monotone increasing and compare $J_h(\tilde{k})$ with $J_h(k)$. Denoting by u and \tilde{u} the solutions of (1.1) corresponding to k and \tilde{k} , respectively, we can write

$$\tilde{u}_t = (k\tilde{u}_x)_x + ((\tilde{k} - k)\tilde{u}_x)_x.$$

Then, in terms of Green's function $G(y, s; x, t)$ of (1.1) ($t < s$), we have

$$\begin{aligned} u(y, T) &= \int_0^1 G(y, T; x, 0) \phi(x) dx, \\ \tilde{u}(y, T) &= \int_0^1 G(y, T; x, 0) \phi(x) dx + \int_0^T \int_0^1 G(y, T; x, t) ((\tilde{k} - k)\tilde{u}_x)_x(x, t) dx dt. \end{aligned}$$

It follows that

$$\begin{aligned} (2.7) \quad J_h(\tilde{k}) - J_h(k) &= \int_0^1 h(y) \tilde{u}(y, T) dy - \int_0^1 h(y) u(y, T) dy \\ &= \int_0^1 h(y) dy \int_0^T \int_0^1 G(y, T; x, t) ((\tilde{k} - k)\tilde{u}_x)_x(x, t) dx dt \equiv \Phi. \end{aligned}$$

By integration by parts,

$$\Phi = - \int_0^1 h(y) dy \int_0^T \int_0^1 G_x(y, T; x, t) (\tilde{k} - k)(x) \tilde{u}_x(x, t) dx dt$$

since $G(1, T; x, t) = 0$, $\tilde{u}_x(0, t) = 0$. Set

$$\begin{aligned} \psi_1(x, t) &= \tilde{k}(x) \tilde{u}_x(x, t), \\ \psi_2(x, t) &= \int_0^1 [k(x) G_x(y, T; x, t)] h(y) dy. \end{aligned}$$

Then

$$(2.8) \quad \Phi = - \int_0^1 \left[\frac{1}{k(x)} - \frac{1}{\tilde{k}(x)} \right] \left[\int_0^T \psi_1(x, t) \psi_2(x, t) dt \right] dx.$$

Since \tilde{k} is monotone increasing, $(\tilde{k}\phi')' \geq 0$ and thus, by Lemma 2.1,

$$(2.9) \quad \psi_1 \leq 0, \quad \frac{\partial \psi_1}{\partial x} = \frac{\partial}{\partial x} (\tilde{k} \tilde{u}_x) = \tilde{u}_t \leq 0.$$

Consider the function

$$V(x, t) = \int_0^1 G(y, T; x, t) h(y) dy.$$

It satisfies:

$$\begin{aligned} V_t + (k(x) V_x)_x &= 0 && \text{in } Q_T, \\ V(x, T) &= h(x) && \text{if } 0 < x < 1, \\ V_x(0, t) &= 0, \quad V(1, t) = 0 && \text{if } 0 < t < T. \end{aligned}$$

Since $h' \leq 0$ and $(kh')' \leq 0$, the proof of Lemma 2.1 can be applied; it yields

$$V_x \leq 0, \quad (kV_x)_x = -V_t \leq 0.$$

Consequently

$$(2.10) \quad \psi_2 \leq 0, \quad \frac{\partial \psi_2}{\partial x} \leq 0.$$

From (2.7)–(2.10) we see that

$$(2.11) \quad J_h(\tilde{k}) - J_h(k) = \int_0^1 \left(\frac{1}{k(x)} - \frac{1}{\tilde{k}(x)} \right) \psi(x) dx,$$

where

$$(2.12) \quad \psi(x) = - \int_0^T \psi_1(x, t) \psi_2(x, t) dt$$

satisfies

$$(2.13) \quad \psi(x) \leq 0, \quad \psi'(x) \leq 0.$$

Approximating general k, \tilde{k} by smooth functions k_m, \tilde{k}_m we deduce that (2.11)–(2.13) hold for general k, \tilde{k} in K_h with \tilde{k} monotone increasing; more precisely, $\psi \in L^1(0, 1)$, and

$$\psi \leq 0 \quad \text{a.e.}, \quad \psi' \leq 0 \quad \text{in } \mathcal{D}'(0, 1).$$

It follows that if \tilde{k} is a monotone increasing rearrangement of k then

$$J_h(\tilde{k}) \leq J_h(k),$$

and Theorem 1.2 follows. Similarly

$$(2.14) \quad J_h(k_0) \leq J_h(k) \quad \forall k \in K_h.$$

Thus, in order to complete the proof of Theorem 1.1, it remains to prove uniqueness of the minimizer.

Suppose equality holds in (2.14) for some $k \neq k_0$. It is easy to see that if $\tilde{k} = k_0$ then the strong maximum principle holds for $\tilde{k}\tilde{u}_x, \tilde{u}_t$. Hence

$$(2.15) \quad \psi_1 < 0, \quad \psi_{1,x} < 0 \quad \text{in } Q_T.$$

Next, approximating ψ by mollifications we see that ψ has a version which is a monotone increasing function. Let

$$b = \inf \{x; \psi(x) \text{ is strictly increasing in } (x, 1)\}.$$

Since equality holds in (2.14), b must be strictly positive. Further, by (2.9)–(2.13),

$$(2.16) \quad k_0(x) = k(x) \quad \text{if } b < x < 1.$$

In view of (2.15) we then have

$$\psi_2(x, t) = 0 \quad \text{in rectangles } (a_i, b_i) \times (0, T)$$

with $a_i > b_{i-1}$, $b_i \uparrow b$; since $\psi_{2,x} \leq 0$ we also have

$$k(x)V_x(x, t) \equiv \psi_2(x, t) = 0 \quad \text{if } a_1 < x < b, \quad 0 < t < T.$$

Hence $k(x)h'(x) = 0$ if $a_1 < x < b$ and, in view of (1.3), $h \equiv \text{const.} = c > 0$.

Since $V_t = 0$ if $a_i < x < b_i$, $0 < t < T$, we deduce that

$$V(b, t) = c$$

and, by uniqueness,

$$(2.17) \quad V(x, t) = c \quad \text{if } 0 < x < b, \quad 0 < t < T.$$

If $b \leq \gamma$ then (2.16) and $\int k = \int k_0$ imply that $k(x) \equiv k_0(x)$. Thus we may suppose that $b > \gamma$ and, then, V is a smooth function in $\{b \leq x < 1\}$; in particular, V_t is bounded. We conclude that

$$(2.18) \quad KV_x \text{ is Lipschitz continuous in a neighborhood of } \{x = b\}.$$

Further, by the strong maximum principle (in $\{b < x < 1\}$) $V_x(b+0, t) < 0$. Since, by (2.17), $V_x(b-0, t) = 0$, we get a contradiction to (2.18).

Remark 2.1. If $l^* \in C^{2,0}[0, 1]$ and $\tilde{k} = l^*$ then the strong maximum principle can be applied to conclude that $\psi_1 < 0$ in Q_T . But then the proof of uniqueness for Theorem 1.1 applies also to Theorem 1.2; thus l^* is the unique minimizer.

Remark 2.2. Theorems 1.1, 1.2 remain valid also in case $h \in C^1[0, 1]$ (instead of $h \in C^2[0, 1]$) provided h is concave. Indeed, we simply approximate h by smooth functions h_m for which (2.9)–(2.13) hold, and thus derive (as $m \rightarrow \infty$) these formulas for h . The rest of the proof is the same as before.

Remark 2.3. Theorems 1.1, 1.2 extend to the case of

$$(2.19) \quad u_x(1, t) + \mu u(1, t) = 0 \quad \text{if } 0 < t < T \quad (\mu > 0)$$

instead of $u(1, t) = 0$. Indeed by the maximum principle we easily deduce that $u > 0$, $u_x < 0$ in Q_T . The proof that $u_t \leq 0$ is the same, using the appropriate Green function. Theorem 1.3 (to be proved in § 3) also extends to the boundary condition (2.19).

3. Proof of Theorem 1.3. Consider first the case where \tilde{k} is an increasing step function, and set $h(x) = \tilde{u}(x, T)$. By Lemma 2.1

$$h' \leq 0 \quad \text{a.e.,} \quad (\tilde{k}h')' \leq 0 \quad \text{in } \mathcal{D}'(0, 1).$$

Since $\tilde{k}' = 0$ everywhere except at the points of discontinuity x_i of $\tilde{k}(x)$, and since $\tilde{k} \geq \alpha > 0$, we get $h''(x) \leq 0$ if $x \neq x_i$. Further, by the strong maximum principle we can deduce that $h'(x) < 0$ if $0 < x < 1$. Also, $h(0) > 0$, $h(1) = 0$, $h'(0) = 0$. Hence, by Remark 2.2, Theorem 1.1 is valid for this function h . Since $k \in \tilde{K} \subset K_h$, we conclude that, if $\tilde{k} = k_0$,

$$(3.1) \quad \int_0^1 \tilde{u}(x, T) \tilde{u}(x, T) dx \leq \int_0^1 \tilde{u}(x, T) u(x, T) dx$$

with equality if and only if $k = k_0$.

Similarly, if k is an increasing step function then Theorem 1.1 is valid for $h(x) = u(x, T)$. Thus, if $\tilde{k} = k_0$,

$$(3.2) \quad \int_0^1 u(x, T) \tilde{u}(x, T) dx \leq \int_0^1 u(x, T) u(x, T) dx.$$

By approximation we find that (3.2) is valid for any $k \in \tilde{K}$. Combining (3.1) with (3.2) we deduce that

$$(3.3) \quad \tilde{J}(k_0) \leq \tilde{J}(k) \quad \forall k \in \tilde{K}.$$

Finally, if equality holds in (3.3) for some $k \neq k_0$, then equality must also hold in (3.1), a contradiction.

4. Other control problems. Consider the parabolic problem

$$\begin{aligned}
 (4.1) \quad & u_t = a(x)u_{xx} + b(x)u_x - k(x)u \quad \text{in } Q_T, \\
 & u(x, 0) = \phi(x) \quad \text{if } 0 < x < 1, \\
 & u_x(0, t) = 0 \quad \text{if } 0 < t < T, \\
 & u(1, t) = 0 \quad \text{if } 0 < t < T
 \end{aligned}$$

where a, b are $C^{1+\delta}$ functions in $0 \leq x \leq 1$, $a(x) \geq a_0 > 0$, and $k(x)$ is a control variable belonging to the set

$$K_* = \left\{ k(x) \text{ monotone increasing, } 0 \leq k(x) \leq M, \int_0^1 k(x) dx = \theta \right\};$$

M and θ are given positive numbers, and $M > \theta$.

We define the functional $J_h(k)$ as in (1.4) and set

$$(4.2) \quad \tilde{J}_m(k) = \int_0^1 u_k^m(x, T) dx, \quad m \text{ positive integer.}$$

We replace the conditions (1.2), (1.3) by the conditions:

$$(4.3) \quad \phi \in C^1[0, 1], \quad \phi(x) > 0, \quad \phi'(x) < 0 \quad \text{if } 0 < x < 1,$$

$$(4.4) \quad h \in C^1[0, 1], \quad h(0) > 0, \quad h(x) \geq 0 \quad \text{and} \quad h'(x) \leq 0 \quad \text{if } 0 < x < 1.$$

Consider the following problems.

Problem (J_h^).* Find k_* such that

$$(4.5) \quad k_* \in K_*, \quad J_h(k_*) = \min_{k \in K_*} J_h(k).$$

Problem (J_m^).* Find k_* such that

$$(4.6) \quad k_* \in K_*, \quad \tilde{J}_m(k_*) = \min_{k \in K_*} \tilde{J}_m(k).$$

THEOREM 4.1. *There exists a unique solution of problem (4.5), given by $k_*(x) \equiv \theta$.*

THEOREM 4.2. *There exists a unique solution of problem (4.6), given by $k_*(x) \equiv \theta$.*

Proof of Theorem 4.1. For any k, \tilde{k} in K_* we compute (cf. (2.7))

$$\begin{aligned}
 (4.7) \quad J_h(\tilde{k}) - J_h(k) &= \int_0^1 k(y) dy \int_0^T \int_0^1 G(y, T; x, t)(k(x) - \tilde{k}(x))\tilde{u}(x, t) dx dt \\
 &= \int_0^1 \psi(x)(k(x) - \tilde{k}(x)) dx
 \end{aligned}$$

where

$$\psi(x) = \int_0^T \psi_1(x, t)\psi_2(x, t) dt,$$

$$\psi_1(x, t) = \tilde{u}(x, t).$$

$$\psi_2(x, t) = \int_0^1 G(y, T; x, t)h(y) dy.$$

Using (4.3), (4.4) and $k' \geq 0, \tilde{k}' \geq 0$, we deduce, by the proof of Lemma 2.1, that

$$\frac{\partial \psi_1}{\partial x} \leq 0, \quad \frac{\partial \psi_2}{\partial x} \leq 0.$$

Hence,

$$\psi > 0, \quad \psi' \leq 0.$$

Taking $\tilde{k} = k_*$ and recalling (4.7) we see that

$$J_h(k_*) \leq J(k).$$

The proof of uniqueness is similar to the proof in Theorem 1.1. Indeed, since $\psi_2 > 0$ in Q_T , if k is another minimizer then $\partial\psi_1/\partial x = 0$ for $a < x < b$, $0 < t < T$ and some $0 < a < b < 0$. This implies that $\phi'(x) = 0$ if $a < x < b$, a contradiction to (4.3).

The proof of Theorem 4.2 is similar to the proof of Theorem 1.3. In fact, it follows by repeated application of Theorem 4.1 with

$$\begin{aligned} h(x) &= \tilde{u}^{m-1}(x, T), & h(x) &= \tilde{u}^{m-2}(x, T)u(x, T), \dots, \\ h &= u^{m-1}(x, T). \end{aligned}$$

Remark 4.1. Theorems 4.1, 4.2 extend to the case of the boundary condition (2.19).

Remark 4.2. If we consider the problem of maximizing $J_h(k)$ or $\tilde{J}_m(k)$, then the same method yields the unique solution

$$k^*(x) = \begin{cases} M & \text{if } 1 - \frac{\theta}{M} < x \leq 1, \\ 0 & \text{if } 0 \leq x \leq 1 - \frac{\theta}{M}. \end{cases}$$

We shall now consider another parabolic problem

$$\begin{aligned} (4.8) \quad & u_t = u_{xx} && \text{in } Q_T, \\ & u(x, 0) = \phi(x) && \text{if } 0 < x < 1, \\ & u(0, t) = 1 && \text{if } 0 < t < T, \\ & u_x(1, t) + k(t)u(1, t) = 0 && \text{if } 0 < t < T \end{aligned}$$

where the control $k(t)$ appears in one boundary condition as a coefficient of u . We take the cost functional to be $J_h(k)$, as in (1.4), and assume that h and ϕ satisfy:

$$\begin{aligned} (4.9) \quad & \phi(0) = h(0) = 1, \quad \phi''(0) = h''(0) = 0, \\ & \phi > 0, \quad \phi'' \geq 0, \quad h > 0, \quad h'' \leq 0. \end{aligned}$$

We also assume that

$$(4.10) \quad \beta \equiv -\frac{\phi'(1)}{\phi(1)} > -\frac{h'(1)}{h(1)} \equiv \alpha > 0$$

and take the control set

$$\hat{K} = \left\{ k(t) \in L^\infty(0, T), k(0) = \beta, k(T) = \alpha, -M \leq k'(t) \leq 0, \int_0^T k(t) dt = \gamma T \right\},$$

where $\alpha < \gamma < \beta$. If M is large enough then $\hat{K} \neq \emptyset$ and, in fact, there is a unique

$t_0 \in (0, T)$ such that the function

$$(4.11) \quad \hat{k}(t) = \begin{cases} \beta & \text{if } 0 \leq t < t_0, \\ \beta - M(t - t_0) & \text{if } t_0 \leq t < t_1 \quad \left(t_1 = t_0 + \frac{\beta - \alpha}{M} \right), \\ \alpha & \text{if } t_1 < t < 1, \end{cases}$$

belongs to \hat{K} .

Consider the problem: Find \hat{k} such that

$$(4.12) \quad \hat{k} \in \hat{K}, \quad J_h(\hat{k}) = \max_{k \in \hat{K}} J_h(k).$$

THEOREM 4.3. *There exists a unique solution of problem (4.12), given by (4.11).*

Proof. Proceeding as in previous theorems, we write, for a control \tilde{k} ,

$$\tilde{u}_x(1, t) + k(t)\tilde{u}(1, t) = (k(t) - \tilde{k}(t))\tilde{u}(1, t)$$

and then compute that

$$J_h(\tilde{k}) - J_h(k) = \int_0^1 dy \int_0^T h(y)G(y, T; 1, t)(k(t) - \tilde{k}(t))\tilde{u}(1, t) dt$$

where G is Green's function for the parabolic problem (4.8). Setting

$$V(x, t) = \int_0^t h(y)G(y, T; x, t) dy$$

and using (4.9) and the facts that $k' \leq 0$, $\tilde{k}' \leq 0$, we find that

$$\tilde{u} > 0, \quad \tilde{u}_t \geq 0, \quad V > 0, \quad V_t \geq 0.$$

Thus

$$J_h(\tilde{k}) - J_h(k) = \int_0^T (k(t) - \tilde{k}(t))\psi(t) dt$$

with $\psi > 0$, $\psi' \geq 0$, and the proof is easily completed by familiar arguments.

Remark 4.3. One can establish a similar result for the problem of minimizing $J_h(k)$, $k \in \hat{K}$.

Remark 4.4. The results of this paper extend to some elliptic control problems. Consider, for instance,

$$\begin{aligned} u_{xx} + u_{yy} - k(x)u &= 0 & \text{in } Q = \{0 < x < 1, 0 < y < 1\}, \\ u(x, 0) &= \phi(x) & \text{if } 0 < x < 1, \\ u_y(x, 1) &= 0 & \text{if } 0 < x < 1, \\ u_x(0, y) = u(1, y) &= 0 & \text{if } 0 < y < 1, \end{aligned}$$

with control $k(x)$ and cost functional

$$J_h(k) = \int_0^1 h(x)u_k(x, 1) dx,$$

or

$$\tilde{J}(k) = \int_0^1 u_k^m(x, 1) dx \quad (m \text{ positive integer}).$$

Then

$$J_h(\tilde{k}) - J_h(k) = \int_0^1 (k(x) - \tilde{k}(x))\psi(x) dx$$

with $\psi > 0$, $\psi' \leq 0$ (provided $k' \geq 0$, $\tilde{k}' \geq 0$). We can thus obtain results analogous to Theorems 4.1, 4.2.

REFERENCES

- [1] H. D. FATTORINI, *Time optimal control of solutions of differential equations*, this Journal, 2 (1964), pp. 54–59.
- [2] A. FRIEDMAN, *Optimal control for parabolic equations*, J. Math. Anal. Appl., 18 (1967), pp. 479–491.
- [3] ———, *Optimal control in Banach spaces*, J. Math. Anal. Appl., 19 (1967), pp. 35–55.
- [4] ———, *Optimal control in Banach space with fixed end-points*, J. Math. Anal. Appl., 24 (1968), pp. 161–181.
- [5] A. FRIEDMAN AND L. S. JIANG, *Nonlinear optimal control in heat conduction*, this Journal, 21 (1983), pp. 940–952.
- [6] J. FRIEDMAN, *Optimal control for cooling problems* (1980, unpublished).
- [7] O. A. LADYZHENSKAYA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasilinear Equations of Parabolic Type*, Trans. Amer. Math. Soc., Vol. 23, Providence, RI, 1968.
- [8] J. L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

ALGEBRAIC AND GEOMETRIC METHODS IN NONLINEAR FILTERING*

STEVEN I. MARCUS†

Abstract. The purpose of this paper is to show how the structure of the recursive nonlinear filtering problem leads naturally to the use of methods from nonlinear system theory and the theory of Lie algebras, and to illustrate the application of these methods to a number of specific nonlinear filtering problems. The paper is expository in nature and provides sufficient review of the requisite background in nonlinear systems and Lie algebras and enough examples of the application of tools from these fields to nonlinear filtering problems, so that the value of these methods in nonlinear filtering can be understood by researchers in all of these fields. The application of these methods to nonlinear filtering problems has led to a number of new results concerning finite dimensional filters and to a deeper understanding of the structure of nonlinear filtering problems in general. In particular, new finite dimensional and lower dimensional filters have been obtained; it has been shown that certain problems are inherently infinite dimensional; the understanding of some known filters has been enhanced; and these methods, along with asymptotic methods, have led to new interesting suboptimal filters. We conclude by outlining a procedure for using these tools and by posing some open problems.

Key words. nonlinear filtering, Lie algebras, nonlinear system theory, finite dimensional filters

1. Introduction. The nonlinear filtering problem involves the estimation of a stochastic process $x = \{x_t\}$ (called the *signal* or *state* process) which cannot be observed directly. Information concerning x is obtained from observations of a related process $y = \{y_t\}$ (the *observation* process). The goal is the computation, for each t , of least-squares estimates of functions of the signal x_t given the observation history $\{y_s, 0 \leq s \leq t\}$ —i.e., the computation of conditional expectations of the form $E[\phi(x_t)|y_s, 0 \leq s \leq t]$, or perhaps even the computation of the entire conditional distribution of x_t given the observation history. When the observations are received sequentially, as in many engineering applications, it is preferable that this computation be performed *recursively* in terms of a statistic $\theta = \{\theta_t\}$ which can be updated using only the latest observations:

$$(1) \quad \theta_{t+\tau} = \alpha(t, \tau, \theta_t, \{y_s, t \leq s \leq t + \tau\}),$$

and from which estimates can be calculated in a “pointwise” or “memoryless” manner:

$$(2) \quad E[\phi(x_t)|y_s, 0 \leq s \leq t] = \beta(t, y_t, \theta_t).$$

In general, θ_t is related to the conditional distribution of x_t given $\{y_s, 0 \leq s \leq t\}$, but in certain special cases θ_t is computable with a finite set of stochastic differential equations driven by y . In these special cases, the practical implication of recursiveness is the possible implementation of the filter (1)–(2) in real time: α and β can be thought of as the description of the filter in terms of hardware or a program that does not depend upon that data; θ_t , the state of the filter, is the only quantity which must be stored in memory; and $\{y_s, t \leq s \leq t + \tau\}$ or dy_t is the new observation that is fed into the filter at each time increment.

The purpose of this paper is to show how the structure of the recursive nonlinear filtering problem leads naturally to the use of methods from nonlinear system theory

* Received by the editors September 14, 1983. This is a special expository paper written at the invitation of the editors. This research was supported in part by the National Science Foundation under grant ECS-8022033, in part by the Air Force Office of Scientific Research under grant AFOSR-79-0025, and in part by the Joint Services Electronics Program under contract F49620-82-C-0033.

† Department of Electrical Engineering, University of Texas at Austin, Austin, Texas 78712.

and the theory of Lie algebras, and to illustrate the application of these methods to a number of specific nonlinear filtering problems. The paper is expository in nature and is intended to provide sufficient review of the requisite background in nonlinear systems and Lie algebras and enough examples of the application of tools from these fields to nonlinear filtering problems, so that the value of these methods in nonlinear filtering can be understood by researchers in all of these fields. In this section we present the basic equations of nonlinear filtering which will be used in the sequel; the relevant concepts of nonlinear systems and Lie algebras will be reviewed in § 2; and these concepts will be applied to nonlinear filtering problems in § 3 through § 6.

It is not the purpose of this paper to give a historical overview of nonlinear filtering; instead, the reader is referred to the books [12], [18], [23], [30], and the papers [11], [34] for up-to-date accounts and derivations of the nonlinear filtering equations. The basic model is the following (all stochastic processes are defined on a probability space (Ω, \mathcal{F}, P) and a finite time interval $[0, T]$, on which there is defined a filtration $\{\mathcal{F}_t, 0 \leq t \leq T\}$).

The signal or state process is assumed to be a homogeneous Markov process with generator L ; thus for ϕ in the domain $D(L)$ of L , $m_t^\phi = \phi(x_t) - \phi(x_0) - \int_0^t L\phi(x_s) ds$ is a martingale, or $\phi(x_t)$ satisfies

$$(3) \quad \phi(x_t) = \phi(x_0) + \int_0^t L\phi(x_s) ds + m_t^\phi$$

(e.g., m^ϕ will be a stochastic integral of Brownian motion if the state equation is corrupted by “Gaussian white noise”). We assume that m^ϕ is a square-integrable martingale for all $\phi \in D(L)$. The observation process y , which is intended to model nonlinear state observations in Gaussian white noise, is assumed to be a scalar valued process satisfying

$$(4) \quad y_t = \int_0^t h(x_s) ds + w_t$$

where w is a standard \mathcal{F}_t -Wiener process. More generally, vector-valued observations can easily be treated [12]. It is assumed throughout this paper that $P\{\int_0^T h_s^2 ds < \infty\} = 1$. Finally, we assume for simplicity that $\{x_t\}$ and $\{w_t\}$ are independent. The problem of nonlinear filtering is then to recursively compute $E[\phi(x_t)|\mathcal{Y}_t]$, where $\mathcal{Y}_t = \sigma\{y_s, 0 \leq s \leq t\}$ is the σ -field generated by $\{y_s, 0 \leq s \leq t\}$.

The solution to this problem can be obtained by the “measure transformation method” (see, e.g., [34], [51]). First a new measure P_0 is defined on (Ω, \mathcal{F}) by

$$\frac{dP_0}{dP} = \exp\left(-\int_0^T h(x_s) dy_s + \frac{1}{2} \int_0^T h^2(x_s) ds\right).$$

Under this new measure, $\{y_t\}$ is a standard Brownian motion process, $\{x_t\}$ and $\{y_t\}$ are independent, and the distributions of $\{x_t\}$ remain invariant. It is easily shown that conditional expectations under P , which we desire to compute, are related to those under P_0 by a version of Bayes’ formula:

$$(5) \quad \pi_t(\phi) := E[\phi(x_t)|\mathcal{Y}_t] = \frac{E_0[\phi(x_t)\Lambda_t|\mathcal{Y}_t]}{E_0[\Lambda_t|\mathcal{Y}_t]} =: \frac{\sigma_t(\phi)}{\sigma_t(1)}$$

where E_0 is the expectation with respect to P_0 and

$$\Lambda_t = E_0\left[\frac{dP}{dP_0}\middle|\mathcal{F}_t\right] = \exp\left(\int_0^t h(x_s) dy_s - \frac{1}{2} \int_0^t h^2(x_s) ds\right).$$

Since σ_t satisfies much simpler equations which are more amenable to analysis, we will concentrate on these throughout the paper. Under some weak additional integrability conditions on ϕ and h , it can be shown that σ_t satisfies the Itô equation

$$(6) \quad \sigma_t(\phi) = \sigma_0(\phi) + \int_0^t \sigma_s(L\phi) ds + \int_0^t \sigma_s(\phi h) dy_s$$

or the corresponding Fisk–Stratonovich equation

$$(7) \quad \sigma_t(\phi) = \sigma_0(\phi) + \int_0^t \sigma_s(\tilde{L}\phi) ds + \int_0^t \sigma_s(\phi h) \circ dy_s$$

where $\tilde{L} = L - \frac{1}{2}h^2$ and \circ denotes the Fisk–Stratonovich integral.

Since $\{\sigma_t(\phi), \phi \in D(L)\}$ determines a measure-valued stochastic process ϕ_t , (6) can be regarded as a recursive (infinite dimensional) stochastic differential equation for an unnormalized conditional measure σ_t of x_t given \mathcal{Y}_t . In addition, $\sigma_t(\phi)$ and $\pi_t(\phi)$ are conditional statistics computed from σ_t in a memoryless fashion (see (1)–(2))—recall that $\pi_t(\phi) = \sigma_t(\phi)/\sigma_t(1)$. In general, however, it is not possible to derive *finite dimensional* recursive filters, even for the conditional (normalized or unnormalized) moments; some special cases in which finite dimensional filters exist will be discussed later in this paper. Notice also that the stochastic differential equation (6) is *linear* in σ_t ; this fact greatly facilitates its analysis and is one of the major reasons why more progress has been made recently in the analysis of this equation than in the analysis of the equation for the (normalized) conditional measure π_t , which can be computed by Ito's rule from (5) and (6).

If we assume that x is a diffusion process and the unnormalized conditional measure has a sufficiently smooth density $\rho(t, x)$ (i.e., $\sigma_t(\phi) = \int \phi(x)\rho(t, x) dx$), then under appropriate hypotheses one can derive from (6) and (7) stochastic partial differential equations for ρ (see, e.g., [4], [23], [29], [30], [44]). In this case, we assume that x satisfies

$$(8) \quad x_t = x_0 + \int_0^t f(x_s) ds + \int_0^t g(x_s) d\tilde{w}_s$$

where $x_t \in \mathbb{R}^n$; $\tilde{w}_t \in \mathbb{R}^m$ is a vector of independent standard Brownian motion processes independent of x_0 ; and $\{\tilde{w}_t\}$ and x_0 are independent of $\{w_t\}$. The generator of x is

$$(9) \quad L = \sum_{i=1}^n f^i(x) \frac{\partial^i}{\partial x^i} + \frac{1}{2} \sum_{i,j=1}^n b^{ij}(x) \frac{\partial^2}{\partial x^i \partial x^j}$$

where $B(x) = [b^{ij}(x)] = g(x)g^T(x)$. Then

$$(10) \quad d\rho(t, x) = L^*\rho(t, x) dt + \rho(t, x)h(x) dy_t,$$

$$(11) \quad d\rho(t, x) = [L^* - \frac{1}{2}h^2(x)]\rho(t, x) dt + \rho(t, x)h(x) \circ dy_t$$

where L^* is the formal adjoint of L . Notice that (10) and (11) are linear in ρ —i.e., they are “bilinear” stochastic partial differential equations. This equation ((10) or (11)) was derived by Duncan [13], Mortensen [39], and Zakai [54], and we will refer to it as the *Duncan–Mortensen–Zakai* (D-M-Z) equation. Notice that

$$(12) \quad \pi_t(\phi) = \int_{-\infty}^{\infty} \phi(x)\rho(t, x) dx / \int_{-\infty}^{\infty} \rho(t, x) dx.$$

Another case of interest to us is that in which x is a finite state Markov process taking values in $S = \{s_i\}_{i=1}^n$. Let $\tilde{p}_t^i = P(x_t = s_i)$, and assume that $\tilde{p}_t = [\tilde{p}_t^1, \dots, \tilde{p}_t^n]^T$ satisfies

$$\frac{d}{dt}\tilde{p}_t = A\tilde{p}_t.$$

(A is the intensity matrix or the transpose of the generator of x ; the columns of A each sum to zero.) Then taking, in (6) and (7), $\phi^i(x)$ as the indicator function $1_{x=s_i}$, $\rho_t^i = \sigma_t(\phi^i)$, $B = \text{diag}(s_1, \dots, s_n)$, and $\rho_t = [\rho_t^1, \dots, \rho_t^n]^T$, yields [11], [34], [53]

$$(13) \quad \rho_t = \rho_0 + \int_0^t A\rho_s ds + \int_0^t B\rho_s dy_s,$$

$$(14) \quad \rho_t = \rho_0 + \int_0^t \left(A - \frac{1}{2} B^2 \right) \rho_s ds + \int_0^t B\rho_s \circ dy_s.$$

Here (13) and (14) are *finite* (n -) *dimensional* bilinear stochastic differential equations for the unnormalized conditional probabilities ρ_t , and

$$(15) \quad \pi_t(\phi) = \sum_{i=1}^n \phi(s_i) \rho_t^i / \sum_{i=1}^n \rho_t^i.$$

In fact, the equations [53] for the normalized conditional probability vector, although more highly nonlinear, are generically $(n-1)$ -dimensional, since the probabilities sum to 1.

The role of nonlinear system theory becomes clear if one notices (as did Brockett [7]) that the D-M-Z equation (10) for the unnormalized conditional density ρ , together with the “output” equation (12) can be viewed as an infinite dimensional realization (with state ρ_t) of an input-output map from “input” functions y to outputs $\pi_t(\phi)$. It is thus natural to investigate conditions under which finite dimensional realizations of this input-output map do or do not exist—i.e., conditions under which finite dimensional filters for the computation of $\pi_t(\phi)$ do or do not exist; these conditions could then be used to classify estimation problems according to their inherent “level of difficulty”.

A *finite dimensional recursive filter* which computes $\pi_t(\phi)$ is a filter of the form

$$(16) \quad \eta_t = \eta_0 + \int_0^t a(\eta_s) ds + \int_0^t b(\eta_s) \circ dy_s,$$

$$(17) \quad \pi_t(\phi) = \gamma(\eta_t)$$

where η evolves on a finite dimensional Euclidean space \mathbb{R}^n (or on an analytic manifold) and a , b , and γ are analytic (for example, (13) and (15) represent a finite dimensional recursive filter for the computation of $\pi_t(\phi)$ for any ϕ). If the estimation problem (4), (8) has a finite dimensional recursive filter for the computation of $\pi_t(\phi)$ then (10), (12) and (16), (17) represent two realizations of the same input-output map; in this case, certain results from nonlinear realization theory suggest that (if (16), (17) is observable) there should be a smooth map from the reachable part of (10) to the reachable part of (16), which generates a map from a certain algebraic object (a Lie algebra) associated with (10) to the corresponding object associated with (16). It is this point of view that will be exploited in the remainder of this paper, in order to shed new light on the nonlinear estimation problem.

2. Lie algebras and nonlinear systems. This section provides an introduction to Lie algebras and nonlinear systems along the lines of [9] and [12].

2.1. Lie algebras.

DEFINITION 2.1. A *Lie algebra over a field k* (\mathbb{R} or \mathbb{C} in this paper) is a triple $(V, +, [\cdot, \cdot])$, where $(V, +)$ is a vector space over k and $[\cdot, \cdot]$ is a bilinear map (called the *Lie bracket*) from $V \times V$ into V such that, for $v_1, v_2, v_3 \in V$,

(i) $[v_1, v_2] = -[v_2, v_1]$ (anticommutativity);

(ii) $[v_1, [v_2, v_3]] + [v_2, [v_3, v_1]] + [v_3, [v_1, v_2]] = 0$ (Jacobi identity).

The Lie algebra will be denoted by V if the bracket is understood.

Example 2.1. Let $C^\infty(M)$ be the vector space of all infinitely differentiable real-valued functions defined on an n -dimensional differentiable manifold M . The vector space of all differential operators (with C^∞ coefficients) $A: C^\infty(M) \rightarrow C^\infty(M)$ can be regarded as a Lie algebra with the Lie bracket defined by $[A, B] = AB - BA$, where AB denotes the ordinary composition of differential operators.

Example 2.2. The vector space $gl(n)$ of all $n \times n$ matrices over a field with $[A, B] = AB - BA$ is a Lie algebra.

Example 2.3. A subclass of all differential operators on $C^\infty(M)$ is the vector space $V(M)$ of all vector fields on M , described locally as first order differential operators $\sum_i f^i(x) \partial/\partial x^i$. With the Lie bracket defined as in Example 2.1, the bracket of two vector fields is again a vector field (a first, not second order, differential operator):

$$(18) \quad \left[\sum_i f^i(x) \frac{\partial}{\partial x^i}, \sum_i g^i(x) \frac{\partial}{\partial x^i} \right] = \sum_i \left(\sum_j f^j(x) \frac{\partial g^i}{\partial x^j}(x) - \sum_j g^j(x) \frac{\partial f^i}{\partial x^j}(x) \right) \frac{\partial}{\partial x^i}.$$

If we locally represent a vector field $\sum_i f^i(x) \partial/\partial x^i$ by the column vector $[f^1(x), \dots, f^n(x)]^T$ of C^∞ functions on M , then the Lie bracket (18) can equivalently be defined by $[f, g] = (\partial g/\partial x)f - (\partial f/\partial x)g$, where $\partial f/\partial x$ and $\partial g/\partial x$ are the Jacobian matrices of f and g , respectively. However, in the remainder of this paper we will define the Lie bracket of vector fields with the *opposite sign*:

$$(19) \quad [f, g] = \frac{\partial f}{\partial x}g - \frac{\partial g}{\partial x}f.$$

$V(M)$ with this bracket is still a Lie algebra (in fact it is isomorphic to $V(M)$ with the previous bracket); this bracket has been defined so that the bracket of two *linear* vector fields $f(x) = Ax$, $g(x) = Bx$ is $[Ax, Bx] = (AB - BA)x$, which agrees with the usual Lie bracket of the matrices A and B defined in Example 2.2.

DEFINITION 2.2. Let L be a Lie algebra. A *Lie subalgebra* V of L is a vector subspace of L that is closed under the Lie bracket of L —i.e., $[u, v] \in V$ for all $u, v \in V$. A Lie subalgebra I of L is an *ideal* of L if for all $u \in I$ and all $v \in L$, $[u, v] \in I$. If I is an ideal, the *quotient algebra* of L by I is the quotient vector space L/I with the induced Lie bracket $[u+I, v+I] = [u, v] + I$.

DEFINITION 2.3. Let L_1 and L_2 be Lie algebras over k . A *Lie algebra homomorphism* $\phi: L_1 \rightarrow L_2$ is a linear map that preserves Lie brackets: $\phi([u, v]) = [\phi(u), \phi(v)]$ for all $u, v \in L_1$. The homomorphism is an *isomorphism* if it is both injective and surjective. The *kernel* of a homomorphism $\phi: L_1 \rightarrow L_2$ is the set $\text{Ker}(\phi) = \{u \in L_1 | \phi(u) = 0\}$. Then $\text{Ker}(\phi)$ is an ideal of L_1 , and $L_1/\text{Ker}(\phi)$ is isomorphic to the *image* of ϕ , $\text{Im}(\phi) = \phi(L_1) = \{u \in L_2 | \phi(v) = u \text{ for some } v \in L_1\}$.

2.2. Nonlinear systems. Consider a control system of the form

$$(20) \quad \dot{x}_t = f(x_t) + \sum_{i=1}^m u_t^i g_i(x_t)$$

where $x_t \in M$, an n -dimensional differentiable manifold, x_0 is given, f and $\{g_i\}_{i=1}^m$ are

vector fields on M , u^1, \dots, u^m are externally applied input or control functions, and $t \in [0, T]$.

DEFINITION 2.4. The *controllability Lie algebra* of (20) is the Lie algebra $\mathcal{L} \triangleq \{f, g_1, \dots, g_m\}_{LA}$ generated by $f, \{g_i\}_{i=1}^m$ —i.e., it is the smallest Lie algebra of vector fields on M containing $f, \{g_i\}_{i=1}^m$. We denote by $\mathcal{L}(x)$ the space of tangent vectors spanned by the vector fields of \mathcal{L} at x .

\mathcal{L} gives some indication as to the structure of the reachable set of (20) (cf. [6], [19]). If $\dim \mathcal{L}(x) = n$ (its maximum value) for all $x \in M$, we say that (20) satisfies the *controllability rank condition*. If (20) is a linear system and $\text{rank}[B: AB: \dots: A^{n-1}B] = n$, then $\dim \mathcal{L}(x) = n$ for all $x \in \mathbb{R}^n$, and the system is in fact controllable.

Example 2.4. Consider the bilinear system

$$(21) \quad \dot{x}_t = Ax_t + \sum_{i=1}^m u_t^i B_i x_t, \quad x_t \in \mathbb{R}^n.$$

Here $f(x) = Ax$, $g_i(x) = B_i x$, and $[f, g_i](x) = (AB_i - B_i A)x$. Each vector field in \mathcal{L} is thus of the form Dx for some $n \times n$ matrix D ; since the set of all $n \times n$ matrices is an n^2 -dimensional vector space, we have $\dim \mathcal{L} \leq n^2$. Also \mathcal{L} is isomorphic to the matrix Lie algebra generated by $A, \{B_i\}_{i=1}^m$.

Example 2.5. $\dot{x}_t = u_t^1(x_t)^2 + u_t^2(x_t)^3$, $x_t \in \mathbb{R}^1$. Here $f = 0$, $g_1(x) = x^2$, $g_2(x) = x^3$; it is easy to see that g_1 and g_2 generate the infinite dimensional Lie algebra of vector fields of the form $x^2 p(x)$, where p is a polynomial. Thus it is possible for \mathcal{L} to be infinite dimensional, even on a one-dimensional manifold.

DEFINITION 2.5. Consider the control system (20), together with the scalar observation

$$(22) \quad y_t = h(x_t)$$

where h is a C^∞ function from M to \mathbb{R} . The system (20), (22) is said to be *observable* if, given any two initial states $x_0, x_1 \in M$, there exists an input function u such that the output from x_0 differs from the output from x_1 .

The following “state space isomorphism” theorem is a special case of the results of [6], [9], [19], [46].

THEOREM 2.1. In addition to (20), (22), consider the system

$$(23) \quad \dot{z}_t = a(z_t) + \sum_{i=1}^m u_t^i b_i(z_t),$$

$$(24) \quad y_t = c(z_t)$$

with $z_t \in M'$. Assume that M, M' are analytic manifolds and that $f, \{g_i\}_{i=1}^m, a, \{b_i\}_{i=1}^m$ are complete analytic vector fields (a vector field f is complete if the solution of $\dot{x}_t = f(x_t)$ with $x_0 = x$ exists for all $-\infty < t < \infty$). Assume also that (20), (22) is observable and satisfies the controllability rank condition. If (20), (22) and (23), (24) realize the same input-output map, then there is an analytic mapping $\phi: M' \rightarrow M$ that preserves trajectories—i.e., $\phi(z_t) = x_t$ for every input u .

A related result, which is useful in its own right, is used in the proof of Theorem 2.1. This theorem proves the existence of “minimal” realizations.

THEOREM 2.2. Assume that M' is an analytic manifold and that $a, \{b_i\}_{i=1}^m$ are complete analytic vector fields. Then there exists a realization (on a manifold M'') of the input-output map generated by (23), (24) which is observable and satisfies the controllability rank condition (and is thus minimal [19]); moreover, there is an analytic mapping $\phi': M' \rightarrow M''$ which preserves trajectories.

Under the conditions of Theorem 2.1, the differential $d\phi$ maps a to f and b_i to g_i , $i=1, \dots, m$; in fact, $d\phi$ extends to a homomorphism of Lie algebras $d\phi: \{a, b_1, \dots, b_m\}_{LA} \rightarrow \{f, g_1, \dots, g_m\}_{LA}$. In general, $d\phi$ is not an isomorphism and may have a nontrivial kernel. The existence of such a homomorphism between the controllability algebras of systems with the same input-output map has motivated the application of Lie algebraic methods to the study of finite dimensional filters, for the reasons given previously. Notice, however, that the observability hypothesis in Theorem 2.1 is in general difficult to check. Also of potential interest in filtering are results which, given a homomorphism between the controllability algebras of two systems, imply the existence of maps between the state spaces which preserve trajectories. One such result is the following (a special case of [25], [45]). First, for any Lie algebra \mathcal{L} of vector fields on a manifold M , we define the isotropy subalgebra \mathcal{L}_x of \mathcal{L} at x by $\mathcal{L}_x = \{f \in \mathcal{L}: f(x) = 0\}$.

THEOREM 2.3. *Let M, M' be analytic manifolds, of dimension n, n' , respectively, and assume that $a, \{b_i\}_{i=1}^m, f, \{g_i\}_{i=1}^m$ are analytic. Assume that the controllability algebras \mathcal{L} and \mathcal{L}' of (20) and (23) satisfy the controllability rank condition at x_0, z_0 , respectively. Then there exists a Lie algebra homomorphism $\alpha: \mathcal{L}' \rightarrow \mathcal{L}$ which maps \mathcal{L}'_{z_0} into \mathcal{L}_{x_0} if and only if there exist neighborhoods U of z_0 and V of x_0 and an analytic map $\phi: U \rightarrow V$ that preserves trajectories. That is, if x_t and z_t are the solutions of (20) and (23) for the same input u and $z_t \in U$ for $|t| < \varepsilon$, then $x_t = \phi(z_t) \in V$ for $|t| < \varepsilon$.*

Global results of this type are also discussed in the preceding references.

2.3. The Wei-Norman equations. Lie algebraic methods can also be used, in certain cases, to obtain explicit representations of solutions of bilinear equations, even if the state space is infinite dimensional (as is the case with the D-M-Z equation). This representation was explicitly computed by Wei and Norman [50] in the following manner. Consider the bilinear system (21); its controllability algebra \mathcal{L} is finite dimensional and isomorphic to the matrix Lie algebra generated by $A, \{B_i\}_{i=1}^m$, which we also denote by \mathcal{L} ; let A_1, \dots, A_d be a basis for \mathcal{L} . Wei and Norman assumed that the solution of (21) can be written in the form

$$(25) \quad x_t = e^{g^1 A_1} e^{g^2 A_2} \dots e^{g^d A_d} x_0$$

where $\{g^i\}_{i=1}^d$ are real-valued functions of t to be determined. The representation (25) is then substituted into (21); the left-hand side of (21) is computed by (omitting explicit time-dependence)

$$\begin{aligned} \frac{d}{dt}(e^{g^1 A_1} \dots e^{g^d A_d}) &= g^1 A_1 e^{g^1 A_1} \dots e^{g^d A_d} + e^{g^1 A_1} g^2 A_2 e^{g^2 A_2} \dots e^{g^d A_d} \\ &\quad + \dots + e^{g^1 A_1} \dots e^{g^{d-1} A_{d-1}} g^d A_d e^{g^d A_d}. \end{aligned}$$

We would like to collect the common factor $e^{g^1 A_1} \dots e^{g^d A_d}$ on the right so as to be able to equate coefficients of the basis elements $\{A_i\}_{i=1}^d$ on both sides of (21); however, this is complicated by the fact that the A_i do not in general commute ($[A_i, A_j] \neq 0$ for $i \neq j$). In order to achieve this, we employ a variation of the Baker-Campbell-Hausdorff formula:

$$(26) \quad e^{tA_i} A_j = \sum_{k=0}^{\infty} \frac{t^k}{k!} \text{ad}_{A_i}^k A_j e^{tA_i}$$

where

$$(27) \quad \text{ad}_A^0 B = B, \quad \text{ad}_A^{k+1} B = [A, \text{ad}_A^k B] = [A, [A, [A \dots [A, B] \dots]] \quad (k+1 \text{ times}).$$

Since \mathcal{L} is finite dimensional, (26) can be rewritten as

$$(28) \quad e^{tA_i} A_j = \sum_{l=1}^d h_l^{ij} A_l e^{tA_i}$$

where the h_l^{ij} are computed from (26). This formula allows us to move the A_i 's past the e^{A_i} 's and collect the common product of exponentials on the right. Equating coefficients of $\{A_i\}_{i=1}^d$ in (21) yields a set of differential equations for $\{g^i\}_{i=1}^d$ (the *Wei-Norman equations*), which have the form

$$(29) \quad \begin{aligned} \dot{g}^1(t) &= f_1(g_t, u_t), \\ &\vdots \\ \dot{g}^d(t) &= f_d(g_t, u_t) \end{aligned}$$

where $\{f_i\}_{i=1}^d$ are nonlinear functions of $g_t = (g_t^1, \dots, g_t^d)$ and $u_t = (u_t^1, \dots, u_t^m)$, and $g_0^1 = \dots = g_0^d = 0$. It can be shown that the Wei-Norman equations can be solved on some interval $|t| \leq \varepsilon$, but in most cases the solution cannot be continued for all time. An important fact is that the functions $\{f_i\}_{i=1}^d$ depend *only* on the structure of the Lie algebra \mathcal{L} generated by $A, \{B_i\}_{i=1}^m$; that is, if $\tilde{A}, \{\tilde{B}_i\}_{i=1}^m$ are matrices of any dimension which generate a Lie algebra isomorphic to \mathcal{L} (and the isomorphism takes $\tilde{A} \mapsto A, \tilde{B}_i \mapsto B_i$), then the Wei-Norman equations for the corresponding bilinear system would be exactly the same. In fact, these results will later be applied to the D-M-Z equation, in which \tilde{A}, \tilde{B}_i are differential operators on an infinite dimensional vector space, but some care is necessary in order to make these results rigorous in that case.

In certain circumstances, the Wei-Norman representation holds globally for all t .

DEFINITION 2.6. A Lie algebra L is *solvable* if the derived series of ideals: $L^{(0)} = L$; $L^{(n+1)} \triangleq [L^{(n)}, L^{(n)}] \triangleq \{[u, v]: u, v \in L^{(n)}\}$, $n \geq 0$, is the trivial ideal $\{0\}$ for some n . L is *nilpotent* if the lower central series of ideals: $L^0 = L$; $L^{n+1} \triangleq [L^n, L^n] \triangleq \{[u, v]: u \in L^n, v \in L^n\}$, $n \geq 0$, is $\{0\}$ for some n . L is *abelian* if $[u, v] = 0$ for all $u, v \in L$. L abelian $\Rightarrow L$ nilpotent $\Rightarrow L$ solvable.

THEOREM 2.4. *If \mathcal{L} is solvable, then there is a basis of \mathcal{L} , and an ordering of this basis, for which the Wei-Norman representation (25) is valid for all t .*

A number of examples of solvable Lie algebras arise naturally in nonlinear filtering, as we shall see in this paper. For a detailed finite dimensional example, see [9].

3. Lie algebras and filtering: Introduction and a first example. The application of the methods of nonlinear systems and Lie algebras to nonlinear filtering problems has led to a number of new results concerning finite dimensional filters and to a deeper understanding of the structure of nonlinear filtering problems in general. Except for problems in which the state is a finite state Markov process, these results rely in part on generalizations of the results of the previous section to an *infinite dimensional* bilinear equation: the D-M-Z equation. Some such generalizations have been proved, but for the most part they remain conjecture; however, they have already provided considerable motivation and guidance in the search for finite dimensional filters, the classification of filtering problems, and the design of useful approximate filters.

Brockett and Clark [10] and Brockett [7]–[9] began the application of nonlinear system theory and Lie algebras to nonlinear estimation problems, and Mitter [37], [38] has emphasized the importance of functional integration and group representations and has shown the connection between certain Lie algebras arising in estimation and those arising in mathematical physics.

The fundamental objects in the approach of Brockett [7]–[9] to nonlinear filtering problems are the equations for the unnormalized conditional distributions—the D-M-Z

equations (10)–(11) and the equations (13)–(14) for filtering a finite state Markov process. For the purpose of using results from nonlinear system theory, we will use the Fisk–Stratonovich equations (11) and (14), since these satisfy the ordinary differential rule. As discussed in § 1, these are bilinear equations, (10)–(11) being infinite dimensional and (13)–(14) being finite dimensional. Also, (11)–(12) and (14)–(15) can be viewed as realizations of the input–output map which takes input functions $\{y_i\}$ into output functions $\{\pi_i(\phi)\}$. However, these realizations may be high (or even infinite) dimensional, and it is of considerable interest from the point of view of implementation to determine when low dimensional realizations of this map—i.e., low dimensional recursive filters of the form (16)–(17)—exist.

Given the results of § 2, it is reasonable to expect that Lie algebraic conditions can be related to the existence of low dimensional filters. Notice that, for filtering finite state Markov processes, the controllability algebra of (14) is isomorphic to the matrix Lie algebra $\mathcal{L} = \{A - \frac{1}{2}B^2, B\}_{LA}$ (cf. Example 2.4), and the input–output map will have a low dimensional minimal realization if the dimension of \mathcal{L} is small. This is due to Theorem 2.2 and the fact that $\dim \mathcal{L} \geq \dim \mathcal{L}(x)$ for all x , and thus $\dim \mathcal{L}$ is an upper bound on the dimension of a minimal realization (i.e., on the dimension of a minimal sufficient statistic). In the filtering of diffusion processes, we do not even know whether finite dimensional filters exist, since the D-M-Z equation is an *infinite* dimensional realization of the input–output map, and we are interested in conditions under which finite dimensional filters do or do not exist. We must also be careful in attempting to apply the results of § 2 to the D-M-Z equation, because they are (as stated there) only for valid *finite* dimensional systems.

This point of view leads naturally to the following questions which can be analyzed by the methods of the previous section:

(I) For a given nonlinear filtering problem, do there exist low (or even finite) dimensional filters which compute some statistics $\pi_i(\phi)$ (or which compute the entire conditional distribution)?

(II) Does there exist a classification of filtering problems which are of “equivalent” complexity in some sense? Are there useful invariants associated with each equivalence class?

(III) Do there exist interesting filtering problems for which the controllability algebra of (14) (or a similar algebra which will be associated with (11)) is low dimensional, thus ensuring the existence of low dimensional minimal filters?

As we indicated above, (I) and (III) are studied most naturally by means of Theorems 2.1 and 2.2 and the resulting homomorphisms of Lie algebras discussed in § 2. We shall see that the same Lie algebras are also useful in classifying estimation problems and thus providing a partial answer to (II). In the remainder of this paper we will provide partial answers to these questions and conclude with a discussion of the utility and limitations of this approach.

As a first example of the answer to (III), we discuss a class of examples of finite state Markov process filtering problems in which the controllability algebra $\mathcal{L} = \{A - \frac{1}{2}B^2, B\}_{LA}$ of (14) is low dimensional; \mathcal{L} will henceforth be called the *estimation (Lie) algebra*. As mentioned in § 1, the computation of the conditional probability vector is generically $(n-1)$ -dimensional; however, in the following development, we will see that, given any $n \geq 2$, there exist n -state Markov processes such that the conditional probability vector can be computed with a 2-dimensional filter. Although not generic, this type of result can obviously represent a significant savings in terms of implementation for certain classes of problems.

In the class of examples to be constructed (due to Brockett and Clark [8]–[10]), \mathcal{L} is isomorphic to the Lie algebra $gl(2)$, the 4-dimensional Lie algebra with basis

$$\left\{ \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\}.$$

We desire an intensity matrix A and a diagonal matrix B such that $\{A - \frac{1}{2}B^2, B\}_{LA} \approx gl(2)$. Notice first that

$$\bar{A} = \begin{bmatrix} -(n-1) & 1 & & & 0 \\ n-1 & -(n-1) & 2 & & \\ & n-2 & -(n-1) & \ddots & \\ & & \ddots & \ddots & n-1 \\ 0 & & & 1 & -(n-1) \end{bmatrix}$$

is obviously an intensity matrix; it is also easily verified that if we define $\bar{B} = \text{diag}(n-1, n-3, \dots, -(n-3), -(n-1))$, then $\{\bar{A} - \alpha I_n, \bar{B}\}_{LA} \approx gl(2)$ for any $\alpha \geq 0$. We next introduce a diagonal change of basis defined by a matrix $R = \text{diag}(r_1, \dots, r_n)$ with positive entries, which takes $\bar{A} \mapsto R\bar{A}R^{-1}$, but $R\bar{B}R^{-1} = \bar{B}$ and $R I_n R^{-1} = I_n$; the change of basis does not change the Lie algebraic structure, so we still have $\{R\bar{A}R^{-1} - \alpha I_n, \bar{B}\}_{LA} \approx gl(2)$ (the transformation $A \mapsto R\bar{A}R^{-1}$ with R diagonal and $r_i > 0$ is the finite dimensional analogue of the gauge transformation to be discussed in § 4). Since we desire A, B such that $\{A - \frac{1}{2}B^2, B\}_{LA} \approx gl(2)$, we define $A = R\bar{A}R^{-1} + \frac{1}{2}B^2 - \alpha I_n$, $B = \bar{B}$, and investigate whether we can find R, α such that A is an intensity matrix. The condition for A to be an intensity matrix is that its columns sum to zero—i.e., $cA = 0$, where $c = (1, \dots, 1)$, or

$$(30) \quad cR(\bar{A} + \frac{1}{2}B^2 - \alpha I_n) = 0.$$

It is a consequence of the Perron–Frobenius theorem on positive matrices that we can find positive $\{r_i\}_{i=1}^n$ and $\alpha \geq 0$ that solve (30).

Example 3.1. Let $A = R\bar{A}R^{-1} + \frac{1}{2}B^2 - \alpha I_n$, $B = \bar{B}$ be defined as above. Then (14) becomes

$$(31) \quad d\rho_t = (R\bar{A}R^{-1} - \alpha I_n)\rho_t dt + B\rho_t \circ dy_t.$$

Consider the 2-dimensional system with controllability algebra $gl(2)$:

$$(32) \quad \begin{bmatrix} d\omega_t \\ d\psi_t \end{bmatrix} = \begin{bmatrix} \frac{1-n-\alpha}{n-1} & 1 \\ 1 & \frac{1-n-\alpha}{n-1} \end{bmatrix} \begin{bmatrix} \omega_t \\ \psi_t \end{bmatrix} dt + \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \omega_t \\ \psi_t \end{bmatrix} \circ dy_t.$$

Then define the n -vector

$$\omega^{[n-1]} = \left[\binom{n-1}{0} \omega^{n-1}, \binom{n-1}{1} \omega^{n-2} \psi, \binom{n-1}{2} \omega^{n-3} \psi^2, \dots, \binom{n-1}{n-1} \psi^{n-1} \right]^T$$

(where $\binom{a}{b}$ is the binomial coefficient); this vector of $(n-1)$ st order homogeneous polynomials in ω and ψ is essentially a normalized symmetric tensor. It is easy to see that $R\omega_t^{[n-1]}$ satisfies (31) if ω_t satisfies (32). Hence if the initial distribution p_0 is of the form $p_0 = \gamma^{[n-1]}$ for some two-vector γ (a binomial distribution), the solution of (31) can be obtained by solving the 2-dimensional equation (32) with initial condition γ , and then setting $\rho_t = R\omega_t^{[n-1]}$. Hence, for this class of examples, all conditional

statistics can be computed in terms of a 2-dimensional sufficient statistic by using the 2-dimensional filter (32), for *any* n ; for example, once ρ_t is obtained we can obtain $\pi_t(\phi)$ for any ϕ from (15).

In addition to providing classes of examples which have low dimensional filters, Brockett and Clark [10] provided some necessary conditions on Lie algebras that can arise as estimation algebras for the estimation of finite state Markov processes. Conversely, Baillieul [1] has characterized the class of finite state Markov process that can give rise to $gl(2)$ (or, more generally, other 4-dimensional) estimation algebras. An analogous class of examples for jump process observations is discussed in [33].

4. Diffusion processes, Weyl algebras, and nonexistence of finite dimensional filters. In the estimation of finite state Markov processes, we are able (after converting to Fisk–Stratonovich form) to directly apply the results from nonlinear systems and Lie algebras, because the unnormalized conditional probability vector satisfies the *finite dimensional* bilinear equation (14). On the other hand, the unnormalized conditional density of a diffusion process satisfies the *infinite dimensional* bilinear stochastic partial differential equation (11). Rigorous analogues of the results of § 2 are difficult to obtain for infinite dimensional systems, for they require much more subtle techniques from analysis (some limited results along these lines will be discussed in this section). However, the same philosophy can be applied and will lead, as we shall see, to a number of new results and insights concerning the structure of finite dimensional filters and the structure of the filtering problem.

By analogy with the finite dimensional case, we can associate with (11) the Lie algebra generated by the operators $L^* - \frac{1}{2}h^2$ and h ; these operators are thought of as operating on C^∞ functions, with h acting by multiplication. This algebra $\mathcal{L} = \{L^* - \frac{1}{2}h^2, h\}_{LA}$, which is in some sense the controllability algebra of (11), will be called the *estimation algebra* associated with (11); its structure will help us to provide partial answers to the questions posed in § 3.

First, the estimation algebra can be useful in recognizing “equivalent” filtering problems; that is, \mathcal{L} is invariant under certain transformations of a filtering problem. First, notice that if we perform a “change of scale” on the unnormalized conditional density function, multiplying it by a nonnegative function ψ taking $\rho \mapsto \tilde{\rho} = \psi\rho$, the D-M-Z equation becomes

$$\begin{aligned} d\tilde{\rho}(t, x) &= \psi(x)(L^* - \tfrac{1}{2}h^2(x))\psi^{-1}(x)\tilde{\rho}(t, x) dt + \tilde{\rho}(t, x)h(x) \circ dy_t \\ (33) \quad &= (\psi(x)L^*\psi^{-1}(x) - \tfrac{1}{2}h^2(x))\tilde{\rho}(t, x) dt + \tilde{\rho}(t, x)h(x) \circ dy_t. \end{aligned}$$

This transformation takes $L^* - \frac{1}{2}h^2 \mapsto \psi(L^* - \frac{1}{2}h^2)\psi^{-1} = \psi L^* \psi^{-1} - \frac{1}{2}h^2$ and $h \mapsto \psi h \psi^{-1} = h$, and the Lie algebras are isomorphic.

THEOREM 4.1. *If $\psi: \mathbb{R}^1 \rightarrow \mathbb{R}^1$ is smooth and positive then the Lie algebras $\mathcal{L} = \{L^* - \frac{1}{2}h^2, h\}_{LA}$ and $\tilde{\mathcal{L}} = \{\psi L^* \psi^{-1} - \frac{1}{2}h^2, h\}_{LA}$ are isomorphic with an isomorphism $\phi: A \mapsto \psi A \psi^{-1}$ for all $A \in \mathcal{L}$.*

Proof. We let $x_t \in \mathbb{R}^1$ for simplicity. Any $A \in \mathcal{L}$ is a sum of differential operators of the form $f(x) \partial^i / \partial x_i = f(x) \partial_i$. To show that this mapping is a Lie algebra homomorphism, we compute, for any smooth function q ,

$$\begin{aligned} [\phi(f_1 \partial_i), \phi(f_2 \partial_j)]q(x) &= [\psi f_1 \partial_i \psi^{-1}, \psi f_2 \partial_j \psi^{-1}]q(x) \\ &= (\psi f_1 \partial_i \psi^{-1} \psi f_2 \partial_j \psi^{-1} - \psi f_2 \partial_j \psi^{-1} \psi f_1 \partial_i \psi^{-1})q(x) \\ &= \psi[f_1 \partial_i, f_2 \partial_j]\psi^{-1}q(x) \\ &= \phi([f_1 \partial_i, f_2 \partial_j]). \end{aligned}$$

Hence ϕ is a homomorphism by Definition 2.3; $\text{Ker}(\phi)$ is clearly the zero operator since ψ is positive, so ϕ is a Lie algebra isomorphism.

The transformation of the previous theorem is the so-called gauge transformation [37]. A related phenomenon occurs when one performs a smooth nonsingular change of variables $z = \alpha(x)$ with inverse $x = \beta(z)$ (here again we assume $x_t \in \mathbb{R}^1$). If, for example, α is strictly increasing, then the unnormalized conditional densities of x and z are related by

$$\rho_x(x) = (\rho_z \circ \alpha)(x) \alpha'(x)$$

or

$$(\rho_z \circ \alpha)(x) = \rho_x(x) / \alpha'(x).$$

Hence, this is a particular example of a gauge transformation, with $\psi(x) = 1/\alpha'(x)$. Thus from (33)

$$(34) \quad d(\rho_z \circ \alpha)(x) = \left(\frac{1}{\alpha'(x)} L^* \alpha'(x) - \frac{1}{2} [h(x)]^2 \right) (\rho_z \circ \alpha)(x) dt + (\rho_z \circ \alpha)(x) h(x) \circ dy_t$$

and the resulting estimation algebras are isomorphic. The formulas become more complicated if the results are expressed in the z -coordinates and if α is not strictly increasing, but the results are the same [8], [12]. In fact, we have the following theorem [8].

THEOREM 4.2. *If the estimation problem (4), (8) is transformed by a smooth nonsingular change of coordinates $z_t = \alpha(x_t)$, so that $\{z_t\}$ has generator L_z , then the mapping*

$$\phi: L^* - \frac{1}{2} h^2 \mapsto L_z^* - \frac{1}{2} (h \circ \beta)^2, \quad \phi: h \mapsto h \circ \beta$$

extends to an isomorphism of the Lie algebras $\{L^ - \frac{1}{2} h^2, h\}_{LA}$ and $\{L_z^* - \frac{1}{2} (h \circ \beta)^2, h \circ \beta\}_{LA}$.*

Since the set of all transformations consisting of successive applications of the two types of transformations described in Theorems 4.1 and 4.2 forms a group under composition, Brockett [8] has called this the *estimation equivalence group*, and he has termed two estimation problems equivalent if their estimation algebras can be transformed into one another by elements of this group. This group is called the (stochastic) invariance group by Hijab [20], who further analyzes its properties. Baras [2] has studied nonlinear filtering via group invariance methods and has introduced a third type of transformation involving the solution of a finite set of ordinary differential equations. The utility of the estimation equivalence group is demonstrated by a filtering problem which is merely a linear filtering problem after a change of coordinates. For example, if $f = 0$, $g = 1$, and $h(x) = x$, the transformed problem for $z_t = \alpha(x_t)$ is

$$(35) \quad \begin{aligned} dz_t &= \frac{1}{2} \alpha''(\beta(z_t)) dt + \alpha'(\beta(z_t)) d\tilde{w}_t, \\ dy_t &= h(\beta(z_t)) dt + dw_t. \end{aligned}$$

The estimation algebra of (35) is a 4-dimensional Lie algebra isomorphic to that of the linear filtering problem for *any* isomorphism α (we will study this Lie algebra later), and the conditional density of z can also be computed with the finite dimensional Kalman-Bucy filter. It may be difficult to recognize (35) as a solvable estimation problem directly, but computation of the estimation algebra gives useful information which points in this direction. Thus, assuming that Theorem 2.3 could be generalized to infinite dimensional systems (i.e., the D-M-Z equations for (35) and the linear filtering problem), the isomorphism of the estimation algebras could be used to conclude that the conditional densities of (35) could be computed from the conditional density

of the linear filtering problem (and hence from the Kalman–Bucy filter). However, we must be very careful here. First, notice that the isotropy subalgebra condition of Theorem 2.3 must be checked. Second, and more importantly, we should point out that there are difficulties in making Theorem 2.3 rigorous in the infinite dimensional case. Therefore, a reasonable approach is to use Lie algebraic techniques to *suggest* a relationship between the conditional densities, but then to use rigorous analytical results from nonlinear filtering to prove this and to show the relationship between the conditional densities.

Since the D-M-Z equation is infinite dimensional, possibly the most important application of the concepts of nonlinear systems and Lie algebras in filtering is to the question of the existence of recursive finite dimensional filters. For the computation of a given statistic $\pi_t(\phi)$, this is equivalent to the question of the existence of a finite dimensional realization of the input–output map $\beta: \{y_t\} \mapsto \{\pi_t(\phi)\}$ generated by (11)–(12). Suppose that, for some given initial density, some statistic $\pi_t(\phi)$ can be calculated with a minimal finite dimensional recursive filter of the form (16)–(17). Since (16)–(17) is a finite dimensional realization of the same input–output map β , Brockett [7] observed that results analogous to Theorem 2.1 (and the resulting Lie algebra homomorphism) should hold. That is, under appropriate hypotheses, the controllability algebra $\mathcal{F} = \{a, b\}_{LA}$ of (16) (with Lie bracket (19)) should be a homomorphic image (quotient) of the estimation algebra \mathcal{L} ; we call this the *homomorphism principle* (recall that \mathcal{F} may be infinite dimensional, even though (16) is a finite dimensional filter—cf. Example 2.5). On the other hand, if there is a homomorphism ϕ of \mathcal{L} into a Lie algebra generated by two complete analytic vector fields on a finite dimensional manifold M , and if ϕ maps the isotropy subalgebra at the initial density into the isotropy subalgebra at a point of M , then this is an indication (possibly via an appropriate infinite dimensional generalization of Theorem 2.3, as mentioned above), that some conditional statistic may be computable by an estimator of the form (16)–(17) (i.e., the statistic is *finite dimensionally computable*). It is not known in what generality such results are valid, especially for cases in which \mathcal{L} is infinite dimensional; some particular results have been proved, and we will return to this and related questions later. However, it is clear (in part, from the examples discussed in this paper) that there is a strong relationship in general between the structure of \mathcal{L} and the existence of finite dimensional filters, and that this point of view gives new insights into the structure of nonlinear filtering problems and guidance in the search for finite dimensional filters.

There is one class of (infinite dimensional) Lie algebras, the Weyl algebras, from which there are *no* nonconstant homomorphisms into Lie algebras of C^∞ or analytic vector fields; hence if a Weyl algebra arises as the estimation algebra of a filtering problem, then (modulo the infinite dimensional proof of the homomorphism principle) *no* nonconstant statistic of the conditional density can be computed with a recursive finite dimensional filter. Although it has long been accepted that the nonlinear filtering problem is generically infinite dimensional, the homomorphism principle and Lie algebraic methods have made possible the first proofs that some filtering problems are truly infinite dimensional (for example, the homomorphism principle has been proved for the cubic sensor problem, Example 4.1—cf. [17], [47], [48]).

The Weyl algebra W_n is the algebra of all polynomial differential operators; i.e., $W_n = \mathbb{R}\langle x_1, \dots, x_n; \partial/\partial x_1, \dots, \partial/\partial x_n \rangle$. A basis for W_n consists of all monomial expressions

$$(36) \quad e_{\alpha, \beta} \triangleq x^\alpha \frac{\partial^\beta}{\partial x^\beta} \triangleq x_1^{\alpha_1} \cdots x_n^{\alpha_n} \frac{\partial^{\beta_1}}{\partial x_1^{\beta_1}} \cdots \frac{\partial^{\beta_n}}{\partial x_n^{\beta_n}}$$

where α, β range over all multi-indices $\alpha = (\alpha_1, \dots, \alpha_n)$, $\beta = (\beta_1, \dots, \beta_n)$, $\alpha, \beta \in \mathbb{N} \cup \{0\}$ (the nonnegative integers). W_n is a Lie algebra under the Lie bracket; as an example, we state the general formula for W_1 :

$$(37) \quad \left[x^i \frac{\partial^j}{\partial x^j}, x^k \frac{\partial^l}{\partial x^l} \right] = \sum_{r=1}^j \binom{j}{r} \binom{k}{r} r! x^{i+k-r} \frac{\partial^{j+l-r}}{\partial x^{j+l-r}} - \sum_{s=1}^l \binom{l}{s} \binom{i}{s} s! x^{i+k-s} \frac{\partial^{j+l-s}}{\partial x^{j+l-s}}$$

where

$$\binom{j}{r} = \frac{j!}{(j-r)! r!}$$

is the binomial coefficient and we have used the convention that $\binom{j}{r} = 0$ if $r < 0$ or $j < r$. As is easily checked, the center of W_n (i.e., the ideal of all elements $Z \in W_n$ such that $[X, Z] = 0$ for all $X \in W_n$) is the one-dimensional space $\mathbb{R} \cdot 1$ with basis $\{1\}$. It is shown in [16] that the Lie algebra $W_n/\mathbb{R} \cdot 1$ is *simple*—i.e., it has no ideals other than $\{0\}$ and $W_n/\mathbb{R} \cdot 1$ itself. Thus we have the algebraic part of a proof that if W_n occurs as the Lie algebra \mathcal{L} for some estimation problem, then either the unnormalized conditional density itself is finite dimensionally computable or no statistic at all is finite dimensionally computable. The next two theorems [16] complete the algebraic part of the argument by showing that in fact neither W_n nor its quotients can be realized by vector fields on a finite dimensional manifold.

Let \hat{V}_m be the Lie algebra of vector fields

$$\hat{V}_m \triangleq \left\{ \sum_{i=1}^m f_i(x_1, \dots, x_m) \frac{\partial}{\partial x_i} \right\}$$

with (formal) power series coefficients $f_i \in \mathbb{R}[[x_1, \dots, x_m]]$, and let $V(M)$ be the Lie algebra of C^∞ -vector fields on a C^∞ -manifold M .

THEOREM 4.3. *Fix $n \neq 0$. Then there are no nonconstant homomorphisms from W_n to \hat{V}_m or from $W_n/\mathbb{R} \cdot 1$ to \hat{V}_m for any m .*

THEOREM 4.4. *Fix $n \neq 0$. Then there are no nonconstant homomorphisms from W_n to $V(M)$ or from $W_n/\mathbb{R} \cdot 1$ to $V(M)$ for any finite dimensional C^∞ -manifold M .*

These results suggest (assuming the appropriate homomorphism principle) that if a system has estimation algebra $\mathcal{L} = W_n$ for some n , then neither the conditional density nor any nonconstant statistic of the conditional density can be computed with a finite dimensional filter of the form (16)–(17) with a and b C^∞ or analytic. This is indeed the case for the cubic sensor problem (Example 4.1 below), as was mentioned before. We will give some examples of such systems, but first we present a general method for showing that $\mathcal{L} = W_n$.

THEOREM 4.5 [16]. *The Lie algebra W_n is generated by the elements $x_i, \partial^2/\partial x_i^2, x_i^2, \partial/\partial x_i, i = 1, \dots, n$; and $x_i x_{i+1}, i = 1, \dots, n-1$.*

Proof. Let \mathcal{L} be the Lie algebra generated by these elements. Since $[x_i^2 \partial/\partial x_i, x_i^k] = kx_i^{k+1}$, \mathcal{L} contains $x_i^k, k \geq 1$. Now, $[\partial^2/\partial x_i^2, x_i] = \partial/\partial x_i$ and $[\partial/\partial x_i, x_i] = 1$. Also,

$$(38) \quad \left[\frac{\partial^2}{\partial x_i^2}, x_i^k \left(\frac{\partial}{\partial x_i} \right)^l \right] = 2kx_i^{k-1} \left(\frac{\partial}{\partial x_i} \right)^{l+1} + k(k-1)x_i^{k-2} \left(\frac{\partial}{\partial x_i} \right)^l, \quad k \geq 2;$$

with $l=0$, (38) implies that $x_i^k \partial/\partial x_i \in \mathcal{L}, k \geq 0$. Then by induction (38) implies that $x_i^k (\partial/\partial x_i)^l \in \mathcal{L}$ for all $k, l \geq 0$. Notice that $[x_i, \partial^2/\partial x_i^2, x_i x_{i+1}] = 2x_i x_{i+1} \partial/\partial x_i$, and commut-

ing this with $x_i^k(\partial/\partial x_i)^l$ gives $x_{i+1} \cdot \mathbb{R}\langle x_i, \partial/\partial x_i \rangle \in \mathcal{L}$. Repeated commutation with $x_{i+1}^2 \partial/\partial x_{i+1}$ and $(\partial/\partial x_{i+1})^2$ yields (as above) $\mathbb{R}\langle x_i, x_{i+1}, \partial/\partial x_i, \partial/\partial x_{i+1} \rangle$. By induction, we have that $\mathcal{L} = W_n$.

Theorem 4.5 provides a relatively systematic method for showing that $\mathcal{L} = W_n$ for a particular estimation problem; one need only show that by taking repeated Lie brackets of $L - \frac{1}{2}h^2$ and h , the generating elements of W_n given in Theorem 4.5 are obtained. Notice that if $n = 1$, the generating elements are $x, \partial^2/\partial x^2$, and $x^2(\partial/\partial x)$. Some interesting examples, which illustrate the computations involved, are the following.

Example 4.1 [16] (the cubic sensor problem). Consider the system

$$dx_t = d\tilde{w}_t, \quad dy_t = x_t^3 dt + dw_t.$$

The Lie algebra \mathcal{L} is generated by the operators

$$e_0 = \frac{1}{2} \frac{\partial^2}{\partial x^2} - \frac{1}{2} x^6, \quad e_1 = x^3.$$

We can compute a sequence of Lie brackets to obtain a sequence of elements $e_i \in \mathcal{L}$, eventually obtaining the desired generators of W_n :

$$[e_0, e_1] = 3x^2 \frac{\partial}{\partial x} + 3x \Rightarrow e_2 = x^2 \frac{\partial}{\partial x} + x,$$

$$ad_{e_2}^k e_1 = 3 \cdot 4 \cdots (k+2) x^{k+3} \Rightarrow x^k \in \mathcal{L}, \quad k \geq 3.$$

Combined with $e_0, x^6 \in \mathcal{L}$ implies that $e_3 = \partial^2/\partial x^2 \in \mathcal{L}$. Continuing,

$$[e_3, e_2] = 4x \frac{\partial^2}{\partial x^2} + 4 \frac{\partial}{\partial x} \Rightarrow e_4 = x \frac{\partial^2}{\partial x^2} + \frac{\partial}{\partial x},$$

$$[e_4, e_2] = 3x^2 \frac{\partial^2}{\partial x^2} + 6x \frac{\partial}{\partial x} + 1 \Rightarrow e_5 = 3x^2 \frac{\partial^2}{\partial x^2} + 6x \frac{\partial}{\partial x} + 1,$$

$$[e_4, e_1] = 6x^3 \frac{\partial}{\partial x} + 9x^2 \Rightarrow e_6 = 2x^3 \frac{\partial}{\partial x} + 3x^2,$$

$$[e_3, e_6] = 12x^2 \frac{\partial^2}{\partial x^2} + 24x \frac{\partial}{\partial x} + 6,$$

which combined with e_5 implies that $e_7 = 1$ and $e_8 = x^2 \partial^2/\partial x^2 + 2x \partial/\partial x$ are in \mathcal{L} . A few more calculations will complete the demonstration:

$$[e_3, e_8] = 4x \frac{\partial^3}{\partial x^3} + 6 \frac{\partial^2}{\partial x^2} \Rightarrow e_9 = x \frac{\partial^3}{\partial x^3},$$

$$[e_1, e_8] = -6x^4 \frac{\partial}{\partial x} - 12x^3 \Rightarrow e_{10} = x^4 \frac{\partial}{\partial x},$$

$$[e_2, e_9] = -5x^2 \frac{\partial^3}{\partial x^3} - 9x \frac{\partial^2}{\partial x^2} \Rightarrow e_{11} = 5x^2 \frac{\partial^3}{\partial x^3} + 9x \frac{\partial^2}{\partial x^2},$$

$$[e_3, e_{10}] = 8x^3 \frac{\partial^2}{\partial x^2} + 12x^2 \frac{\partial}{\partial x} \Rightarrow e_{12} = 2x^3 \frac{\partial^2}{\partial x^2} + 3x^2 \frac{\partial}{\partial x},$$

$$[e_3, e_{12}] = 12x^2 \frac{\partial^3}{\partial x^3} + 24x \frac{\partial^2}{\partial x^2} + 6 \frac{\partial}{\partial x} \Rightarrow e_{13} = 2x^2 \frac{\partial^3}{\partial x^3} + 4x \frac{\partial^2}{\partial x^2} + \frac{\partial}{\partial x}.$$

Now e_{13} , e_{11} , and e_4 are all linear combinations of the elements $x^2 \partial^3/\partial x^3$, $x \partial^2/\partial x^2$, and $\partial/\partial x$, and the coefficient matrix

$$\begin{bmatrix} 0 & 1 & 1 \\ 5 & 9 & 0 \\ 2 & 4 & 1 \end{bmatrix}$$

is nonsingular. It follows that \mathcal{L} contains $e_{14} = \partial/\partial x$, $e_{15} = x \partial^2/\partial x^2$, and $e_{16} = x^2 \partial^3/\partial x^3$. Finally,

$$[e_{14}, e_1] = 3x^2 \Rightarrow e_{17} = x^2,$$

$$[e_{14}, e_{17}] = 2x \Rightarrow e_{18} = x,$$

which combined with e_2 gives $x^2 \partial/\partial x \in \mathcal{L}$; thus by Theorem 4.5, $\mathcal{L} = W_1$.

Analogous computation of selected Lie brackets and the use of Theorem 4.5 yield similar results for the following examples.

Example 4.2 (mixed linear-bilinear type). Consider the system with state equations

$$dx_t = d\tilde{w}_t^1, \quad d\xi_t = x_t dt + x_t d\tilde{w}_t^2,$$

with observations

$$dy_t = \xi_t dt + dw_t.$$

\mathcal{L} is generated by

$$\frac{1}{2} \frac{\partial^2}{\partial x^2} + \frac{1}{2} x^2 \frac{\partial^2}{\partial \xi^2} - x \frac{\partial}{\partial \xi} - \frac{1}{2} \xi^2 \quad \text{and} \quad \xi;$$

it is shown in [16] that $\mathcal{L} = W_2$. The same result is obtained if the $x_t dt$ term is absent in the ξ equation; in that case we have a multiple Wiener integral of Brownian motion observed in Brownian motion noise.

Example 4.3 [16]. Consider the system with state equations

$$dx_t = d\tilde{w}_t, \quad d\xi_t = x_t^2 dt$$

and observations

$$dy_t^1 = x_t dt + dw_t^1, \quad dy_t^2 = \xi_t dt + dw_t^2.$$

\mathcal{L} is generated by

$$\frac{1}{2} \frac{\partial^2}{\partial x^2} - x^2 \frac{\partial}{\partial \xi} - \frac{1}{2} x^2 - \frac{1}{2} \xi^2, \quad x, \quad \text{and} \quad \xi;$$

it is easily shown that $\mathcal{L} = W_2$.

Example 4.4 (the quadratic sensor problem). For the system

$$dx_t = dw_{1t}, \quad dy_t = x_t^2 dt + dw_t,$$

\mathcal{L} is generated by $\frac{1}{2} \partial^2/\partial x^2 - \frac{1}{2} x^4$ and x^2 , and \mathcal{L} is equal to the subalgebra W'_1 of W_1 spanned by all operators $x^i \partial^j/\partial x^j$ with $i-j$ even. Although \mathcal{L} is not the entire Weyl algebra, results analogous to Theorems 4.3 and 4.4 can be proved for W'_1 ; thus, for this problem, the algebraic computations indicate that no nonconstant statistics can be computed with finite dimensional filters.

The rigorous proof that, for problems in which $\mathcal{L} = W_n$, no nontrivial statistics can be computed with finite dimensional filters, depends upon proofs of the

homomorphism principle for these cases. This was done for the cubic sensor in [47] and [48]; proofs for more general problems may result from the method of [48] or from that of [21].

Despite the fact that, for this class of estimation problems, no statistics appear to be *exactly* computable with finite dimensional filters, we shall see in § 6 that the structure of the estimation algebra can also lead to the design of reasonable finite dimensional suboptimal filters for such problems.

5. Examples of Lie algebras and finite dimensional filters. In the previous section, Lie algebraic methods were used to indicate the nonexistence of finite dimensional filters for certain classes of filtering problems. In this section, we present a number of examples in which finite dimensional filters *do* exist, and in which Lie algebraic methods and the homomorphism principle can be used to enhance our understanding of the structure of filtering problems. From another point of view, in each of these examples we verify the homomorphism principle. Notice that \mathcal{L} is finite dimensional in §§ 5.1–5.3, while \mathcal{L} is infinite dimensional in §§ 5.4–5.6; in all cases, however, at least some statistics are finite dimensionally computable—the results are highly dependent upon the particular structure of \mathcal{L} .

5.1. The linear filtering problem. We only consider a simple scalar example (for the general vector case, see [7], [9]):

$$(39) \quad dx_t = d\tilde{w}_t, \quad dy_t = x_t dt + dw_t.$$

The estimation algebra \mathcal{L} is 4-dimensional with basis

$$e_0 = \frac{1}{2} \frac{\partial^2}{\partial x^2} - \frac{1}{2} x^2, x, \frac{\partial}{\partial x}, 1.$$

This algebra is well known in physics, and is called the *oscillator* algebra. Its importance in the estimation problem was first pointed out by Brockett [7] and Mitter [37]. The nonzero commutation relations are

$$[e_0, x] = \frac{\partial}{\partial x}, \quad \left[e_0, \frac{\partial}{\partial x} \right] = x, \quad \left[x, \frac{\partial}{\partial x} \right] = -1.$$

The oscillator algebra is thus the semi-direct sum of $\mathbb{R} \cdot 1$ and the *Heisenberg* algebra spanned by e_0 , x , and $\partial/\partial x$.

If the density of x_0 is Gaussian, then the conditional mean $\hat{x}_t = E[x_t | \mathcal{Y}_t]$ can be computed with the Kalman–Bucy filter [24]:

$$(40) \quad d\hat{x}_t = -P_t \hat{x}_t dt + P_t \cdot dy_t,$$

$$(41) \quad \dot{P}_t = 1 - P_t^2$$

where P_t is the (deterministic) error covariance. Notice that the Kalman–Bucy filter, being linear, has the same form in terms of either the Itô or Fisk–Stratonovich integral. Now, (40)–(41) can be thought of as a 2-dimensional time-invariant system (note that we need a *time-invariant* realization in order to compute the controllability algebra as in § 2). The controllability algebra \mathcal{F} of the Kalman–Bucy filter (viewed as a 2-dimensional control system with input y) is a 3-dimensional Lie algebra with Lie bracket given by (19) and basis

$$f_0 = \begin{bmatrix} -P\hat{x} \\ 1 - P^2 \end{bmatrix}, \quad f_1 = \begin{bmatrix} P \\ 0 \end{bmatrix}, \quad f_2 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}.$$

It can easily be verified that there is a Lie algebra homomorphism (see Definition 2.3) from \mathcal{L} to \mathcal{F} that takes $e_0 \mapsto f_0$, $x \mapsto f_1$, $\partial/\partial x \mapsto f_2$, $1 \mapsto 0$. That the homomorphism has the one-dimensional kernel $\{1\}$ (corresponding to multiplication by a constant) is due to the fact that ρ is an unnormalized density; one point of view is that, since the normalization factor $\sigma_t(1)$ (see (5)) can be computed from ρ by a memoryless operation but must be computed from \hat{x}_t via a *dynamic* equation, this dynamic equation must be appended to the \hat{x} equation for the controllability algebras to be isomorphic. This is indeed the case, for if we append the equation (set $\phi(x) = 1 \cdot x$ in (7))

$$(42) \quad d\sigma_t(1) = -\frac{1}{2}(\hat{x}_t^2 + P_t)\sigma_t(1) dt + \hat{x}_t\sigma_t(1) \circ dy_t$$

to (40)–(41), this set of equations has a controllability algebra isomorphic to \mathcal{L} .

For linear filtering problems with Gaussian initial conditions, the conditional density is also Gaussian and the Kalman–Bucy filter conditional mean estimate provides a sufficient statistic for the conditional density; also the conditional mean together with the normalization factor provides a sufficient statistic for the unnormalized conditional density. However, from another point of view, since \mathcal{L} is finite dimensional and solvable for linear filtering problems, it is natural to attempt to solve the D-M-Z equation via the Wei–Norman representation; this is an alternative method for deriving finite dimensional recursive filters for sufficient statistics, even for *non-Gaussian* initial conditions. The calculation, which can be rigorously justified [40], proceeds as follows for our example (see § 2.3). Assume that the solution of the D-M-Z equation has the form

$$(43) \quad \rho(t, x) = (e^{g_t^1 e_0} e^{g_t^2 x} e^{g_t^3 \partial/\partial x} e^{g_t^4} \rho_0)(x),$$

where $\{e^{te_0}, t \geq 0\}$ is the semigroup defined by solving $\partial u/\partial t = e_0 u_t$ on $L^2(\mathbb{R})$:

$$(44) \quad \begin{aligned} (e^{te_0} \phi)(x) &= \int_{-\infty}^{\infty} G(x, y, t) \phi(y) dy, \\ G(x, y, t) &= (2\pi \sinh t)^{-1/2} \exp \left[-\frac{1}{2}(\coth t)(x^2 + y^2) + (\sinh t)^{-1}xy \right]. \end{aligned}$$

Similarly, $(e^{tx} \phi) = e^{tx} \phi(x)$ and $(e^{t(\partial/\partial x)} \phi)(x) = \phi(x+t)$. The commutation relations and the Baker–Campbell–Hausdorff formula (26) yield

$$(45) \quad \begin{aligned} e^{te_0} x &= (\cosh t) x e^{te_0} + (\sinh t) \frac{\partial}{\partial x} e^{te_0}, \\ e^{te_0} \frac{\partial}{\partial x} &= (\sinh t) x e^{te_0} + (\cosh t) \frac{\partial}{\partial x} e^{te_0}. \end{aligned}$$

Utilizing these identities, the Wei–Norman equations (29) are calculated to be

$$(46) \quad \begin{aligned} g_t^1 &= t, & dg_t^2 &= \cosh t \circ dy_t, \\ dg_t^3 &= -\sinh t \circ dy_t, & dg_t^4 &= -(\sinh t) g_t^2 \circ dy_t. \end{aligned}$$

Finally, by substituting these expressions into (43) and using the explicit form of $\exp(te_0)$ given in (44), we obtain (for the initial density $\rho_0(x) = \delta(x - x_0)$):

$$(47) \quad \rho(t, x) = \int_{-\infty}^{\infty} k(t, z) \exp \left\{ -\frac{1}{2}(\tanh t)^{-1} [x - (\cosh t)^{-1}(z - g_t^3)]^2 \right\} \rho_0(z) dz,$$

$$(48) \quad k(t, x) = \frac{1}{\sqrt{2\pi \sinh t}} \exp \left\{ -\frac{1}{2}(\tanh t)(z - g_t^3)^2 + g_t^2(z - g_t^3) + g_t^4 \right\}.$$

Let $\rho(t, x; z)$ denote the unnormalized conditional transition density starting at $x_0 = z$ (i.e., the integrand of (47) exclusive of $\rho_0(z)$). Then the normalized version of $\rho(t, x; z)$ is obviously Gaussian with covariance $P_t = \tanh t$ and mean $\hat{x}_t = (\cosh t)^{-1}(z - g_t^3)$, and this \hat{x} satisfies the Kalman–Bucy filter equation (40).

Notice that a 2-dimensional minimal sufficient statistic for $\rho(t, x; z)$ (i.e., a minimal realization of the input–output map from $\{y_t\}$ to the transition density $\{\rho(t, x; z)\}$) consists of the conditional mean \hat{x}_t and the normalization factor $\sigma_t(1)$ of (42). An alternative 2-dimensional realization based on the $\{g^i\}$ of the Wei–Norman representation is obtained by noticing that if we define, in (48),

$$(49) \quad \xi_t = g_t^2(z - g_t^3) + g_t^4,$$

then from (46)

$$(50) \quad d\xi_t = (\cosh t)(z - g_t^3) \circ dy_t;$$

hence (g_t^3, ξ_t) obviously comprises a 2-dimensional sufficient statistic for $\rho(t, x; z)$. Finally, we note that an alternative method for deriving sufficient statistics, based upon the recognition that the unnormalized conditional transition density is the exponential of a quadratic form for this problem, is given in [9].

5.2. The Beneš problem [3]. Consider the scalar problem

$$(51) \quad dx_t = f(x_t) dt + d\tilde{w}_t, \quad dy_t = x_t dt + dw_t$$

where f satisfies

$$(52) \quad f'(x) + f^2(x) = ax^2 + bx + c$$

for some $a, b, c \in \mathbb{R}$. It is assumed that the Riccati equation (52) has a global solution on \mathbb{R} (this implies [40] that either $a > 0$ or $a = b = 0, c > 0$). Then an unnormalized conditional transition density of x_t given \mathcal{Y}_t and $x_0 = z$ is given by

$$\begin{aligned} \rho(t, x; 0, z) = \frac{1}{\sqrt{2r_t^{11}}} \exp \left\{ xy_t + \int_z^x [f(u) + ku] du - m_t^T \nu \right. \\ \left. + \frac{1}{2} \nu^T R_t \nu - \frac{(x + (R_t \nu)^1 - m_t^1)^2}{2r_t^{11}} - \left(k + \frac{1}{2}c\right)t \right\} \end{aligned}$$

where $k = (a + 1)^{1/2}$; $R_t = [r_t^{ij}]$ is the 3×3 matrix solution of $R_0 = 0$; $\dot{R}_t = Y_t + R_t A_t^T + A_t R_t$; $m_t \in \mathbb{R}^3$ is the solution of $\dot{m}_t = A_t m_t$; $m_0^T = (z, 0, 0)$; $\nu^T = (0, 1, -1)$;

$$A_t = \begin{bmatrix} -k & 0 & 0 \\ 0 & 0 & 0 \\ ky_t - \frac{1}{2}b & 0 & 0 \end{bmatrix}, \quad Y_t = \begin{bmatrix} 1 & y_t & 0 \\ y_t & y_t^2 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 \\ y_t \\ 0 \end{bmatrix} [1 \quad y_t \quad 0].$$

A finite dimensional recursive filter for the unnormalized conditional transition density consists of the equations for m_t and R_t (along with y_t); thus we have a 10-dimensional sufficient statistic η_t and $\rho(t, x; 0, z) = \gamma(\eta_t; x, z)$ with γ given above. This leads, of course, to a 10-dimensional recursive filter for the computation of the normalized conditional density, since

$$p(t, x) = \int_{-\infty}^{\infty} \gamma(\eta_t; x, z) \rho_0(z) dz \Big/ \int \int_{-\infty}^{\infty} \gamma(\eta_t; x, z) \rho_0(z) dz dx$$

is again only a function of η_t . The same is true for all conditional moments of x_t given \mathcal{Y}_t . In fact, Beneš [3] shows that a 2-dimensional filter is sufficient, and the equations

are similar to those of the Kalman–Bucy filter. Examples of functions f satisfying (52) are (i) $f(x) = (A e^x - B e^{-x}) / (A e^x + B e^{-x})$; for which $a = b = 0$, $c = 1$; and (ii) the linear case $f(x) = \alpha x + \beta$. Thus, this class of problems provides a generalization of the linear filtering problem.

The close relationship between the Beneš problem and the linear problem is further elucidated by an analysis of the estimation algebras. The estimation algebra \mathcal{L}_B of the Beneš problem (51) is generated by $e_0 = \frac{1}{2} \partial^2 / \partial x^2 - (\partial / \partial x) f - \frac{1}{2} x^2$ and x , and has the basis $\{e_0, x, \partial / \partial x - f, 1\}$ with commutation relations

$$[e_0, x] = \frac{\partial}{\partial x} - f, \quad \left[e_0, \frac{\partial}{\partial x} - f \right] = (a+1)x + \frac{b}{2}, \quad \left[x, \frac{\partial}{\partial x} - f \right] = -1.$$

Hence, as in § 5.1, \mathcal{L}_B is 4-dimensional and solvable, and (39) is a special case of the Beneš problem with $f = 0$. In another sense, the Beneš problem is essentially equivalent to a linear problem. This is due to the fact that if we define the gauge transformation [3], [37], [38], [40], $F(x) = \int_0^x f(u) du$, $\psi(x) = \exp(-F(x))$, and define $\tilde{\rho}(t, x) = \psi(x)\rho(t, x)$, then the D-M-Z equation for the Beneš problem is transformed according to (33) into

$$(53) \quad d\tilde{\rho}(t, x) = \left(\frac{1}{2} \frac{\partial^2}{\partial x^2} - \frac{1}{2} [(a+1)x^2 + bx + c] \right) \tilde{\rho}(t, x) + x\tilde{\rho}(t, x) \circ dy_t,$$

which is essentially of the same form as the D-M-Z equation for (39). \mathcal{L}_B is, by Theorem 4.1, isomorphic to the controllability algebra of (53), which has the basis

$$\left\{ \frac{1}{2} \frac{\partial^2}{\partial x^2} - \frac{1}{2} [(a+1)x^2 + bx + c], x, \frac{\partial}{\partial x}, 1 \right\}$$

with the same commutation relations as \mathcal{L} of § 5.1. Hence the unnormalized conditional density of the Beneš problem can be obtained by solving an essentially linear problem and multiplying by $\psi(x)$. Therefore, sufficient statistics for this problem can be obtained by functional integration [3], solution of a linear problem and multiplication by $\psi(x)$, or by solving the D-M-Z equation for this problem directly via a Wei–Norman calculation. Details on reduction of the number of sufficient statistics, verification of the homomorphism principle, and extensions to the multi-dimensional case can be found in [3].

5.3. Some negative results [40], [41]. Although the finite dimensionality of \mathcal{L} (as in the two previous examples) is not necessary for the existence of finite dimensionally computable statistics, this property certainly simplifies the situation; this is because \mathcal{L} can then always be realized by a Lie algebra of vector fields on a finite dimensional manifold, and (if \mathcal{L} is solvable) Wei–Norman techniques can often be justified. However, it is shown by Ocone in [40], [41] that, in the scalar case, the linear and Beneš problems are in some sense the *only* estimation problems in which the Wei–Norman technique can be used to produce finite dimensional filters for the conditional density. More precisely, the following is proved.

THEOREM 5.1. *Consider the estimation problem (4), (8).*

- (a) *Suppose $\dim \mathcal{L} < \infty$. If a function $l \in \mathcal{L}$, then l is a polynomial of degree ≤ 2 .*
 (b) *Let $n = m = 1$ and $g = 1$ in (8). Then $\dim \mathcal{L} < \infty$ only in the following three cases:*

- (i) $h(x) = \alpha x$ and (52) is satisfied;
 (ii) $h(x) = \alpha x^2 + \beta x$, $\alpha \neq 0$, and

$$(54) \quad f'(x) + f^2(x) = -h^2(x) + a(2\alpha x + \beta)^2 + b + c(2\alpha x + \beta)^{-2}$$

or

$$(55) \quad f'(x) + f^2(x) = -h^2(x) + ax^2 + bx + c.$$

(c) If f satisfies (52) with $a < 0$, (54), or (55), then f has a singularity on any unbounded interval. In this case, the D-M-Z equation contains boundary conditions which prevent the Wei–Norman method from producing a finite dimensional filter for the conditional density.

This result has two important consequences. First, even if \mathcal{L} is formally finite dimensional, the Wei–Norman method may not yield a finite dimensional filter, because of singularities in f and boundary conditions associated with the differential operators. Second, the linear and Beneš problems are the only scalar problems in which \mathcal{L} is finite dimensional and the Wei–Norman technique does produce such a filter. On the other hand, it appears that the scalar case is not typical, and that many more examples in which \mathcal{L} is finite dimensional can be found in the case in which x_t is not a scalar.

5.4. A 2-dimensional example [52]. In [52], W. Wong has begun to expose the possibilities inherent in the multi-dimensional case by employing Lie algebraic methods. The general class of (2-dimensional) filtering problems considered is given by

$$(56) \quad dx_t^1 = f(x_t^1, x_t^2) dt + d\tilde{w}_t, \quad dx_t^2 = g(x_t^2) dt, \quad dy_t = x_t^1 dt + dw_t.$$

The class includes as a special case the problem of identification of some linear systems discussed in § 5.5 (but for which \mathcal{L} is in general infinite dimensional). Wong has given two sets of sufficient conditions under which the estimation algebra \mathcal{L} of (56) is finite dimensional; we consider here only the second condition, which is a generalization of (51)–(52): if there exist constants α and β and a C^∞ function γ such that $\frac{1}{2}\partial_1^2 f + f \cdot \partial_1 f + g \cdot \partial_2 f = \alpha + \beta x^1 + \gamma(x^2)$ (where $\partial_i = \partial/\partial x^i$) and $G \equiv \{g \cdot \partial_2, \gamma\}_{LA}$ is finite dimensional, then \mathcal{L} is finite dimensional. Based on this condition, it is shown that the following class of examples has a finite dimensional estimation algebra:

$$\begin{aligned} dx_t^1 &= \left[\eta \left(x_t^1 - \int_0^{x_t^1} \frac{2\gamma(r)+1}{2g(r)} dr \right) + \gamma(x_t^2) \right] dt + d\tilde{w}_t, \\ dx_t^2 &= g(x_t^2) dt, \\ dy_t &= x_t^1 dt + dw_t, \end{aligned}$$

where $\eta(x) = c e^x (1 + c e^x)^{-1}$ and g and γ satisfy certain conditions. In addition, \mathcal{L} is solvable, so the unnormalized conditional density can be found via Wei–Norman methods, which are rigorously justified in [52]. All of this can also be extended to dimensions greater than two.

5.5. Recursive estimation of functionals of Gaussian processes. Since \mathcal{L} is in general infinite dimensional, it is of considerable interest to investigate estimation problems with \mathcal{L} infinite dimensional, but for which finite dimensionally computable statistics exist. This is the case for the class of problems studied in [32], [35], [36], in which it is desired to recursively estimate a polynomial functional of the state of a linear system, given noisy linear observations of the state. The simplest example is one in which the state equations are

$$dx_t = d\tilde{w}_t, \quad d\xi_t = (x_t)^2 dt$$

and the observations are

$$dy_t = x_t dt + dw_t$$

with x_0 Gaussian. The computation of \hat{x}_t is of course straightforward by means of the Kalman–Bucy filter; however, as shown in [35], [36], all conditional moments $E[(\xi_t)^n | \mathcal{Y}_t]$ of ξ_t can also be computed recursively with finite dimensional filters. \mathcal{L} is generated by

$$A_0 = -x^2 \frac{\partial}{\partial \xi} + \frac{1}{2} \frac{\partial^2}{\partial x^2} - \frac{1}{2} x^2 \quad \text{and} \quad B_0 = x;$$

its structure is given in the next theorem.

THEOREM 5.2.

(i) The estimation algebra \mathcal{L} generated by A_0 and B_0 has as basis the elements

$$A_0, B_i = x \frac{\partial^i}{\partial \xi^i}, C_i = \frac{\partial}{\partial x} \frac{\partial^i}{\partial \xi^i}, D_i = \frac{\partial^i}{\partial \xi^i}, \quad i = 0, 1, 2, \dots$$

(ii) The nonzero commutation relations between basis elements are given by $[A_0, B_i] = C_i$, $[A_0, C_i] = B_i + 2B_{i+1}$, $[B_i, C_j] = -D_{i+j}$.

(iii) The center of \mathcal{L} is $\{D_i, i = 0, 1, 2, \dots\}$.

(iv) Every ideal of \mathcal{L} has finite codimension; i.e., for any ideal I , the quotient \mathcal{L}/I is finite dimensional.

(v) Let I_j be the ideal generated by B_j with basis $\{B_i, C_i, D_i; i \geq j\}$. Then $I_0 \supset I_1 \supset \dots$ and $\bigcap_j I_j = \{0\}$, so \mathcal{L} is “pro-finite dimensional” [16].

(vi) \mathcal{L} is solvable.

The Lie algebras of the finite dimensional filters for the conditional moments of x and ξ are analyzed in detail in [32], and the homomorphism principle is verified for this example. In particular, realizations of the quotients \mathcal{L}/I_j are given in terms of a sequence of recursive finite dimensional filters for the conditional moments.

It is interesting to compare this example with Example 4.3, which is the same except for the additional observation $dy_t^2 = \xi_t dt + dw_t^2$; in that case $\mathcal{L} = W_2$ (the Weyl algebra), so that no conditional statistic can be computed *exactly* with a finite dimensional filter. However, it is probable that, due to the additional observation, a *suboptimal* approximate filter (such as the extended Kalman filter) for the conditional mean of ξ_t will result in a lower mean-square error than the *optimal* filter which computes ξ_t in the present example. Thus some care must be taken in interpreting the Lie algebraic structure of a nonlinear estimation problem; this structure has direct implications on the *exact* computation of conditional statistics; its implications on approximate filtering will be discussed later.

As shown by Ocone [42], the Lie algebraic structure for the problem of recursively estimating polynomial functionals of the state of the Beneš system (51) [43] is almost identical to that of the problems discussed in this section. Other examples having properties (iv) and (v) of Theorem 5.2, the so-called pro-finite dimensional Lie algebras, are discussed in [16]. In fact, all known filtering problems which admit finite dimensionally computable statistics have estimation algebras of this type. Another interesting example of such an algebra is that of the next section.

5.6. Linear systems with unknown parameters. The recursive estimation of states and parameters in linear systems with unknown parameters has been treated from a Lie algebraic point of view in [26]–[28]. Problems with and without noise in the state equations, and both constant and jumping parameters are discussed in these references, but we treat here only a simple scalar example which has many of the characteristics of the general problem [27], [28]:

$$(57) \quad dx_t = \theta d\tilde{w}_t,$$

$$(58) \quad dy_t = x_t dt + dw_t$$

where θ is a random variable taking values in a smooth manifold Θ (θ is to be thought of as an additional state variable obtained by augmenting (57) with the state equation $d\theta_t = 0$). Although suboptimal filters have been used, little is known about the existence of finite dimensionally computable statistics for this problem, and Lie algebraic techniques can produce some limited results in this direction.

For this problem, $\mathcal{L} = \{(\theta^2/2) \partial^2/\partial x^2 - x^2/2, x\}_{LA}$ is an infinite dimensional Lie algebra spanned by the set of operators: $(\theta^2/2) \partial^2/\partial x^2 - x^2/2$, $\{\theta^{2n} x\}_{n=0}^\infty$, $\{\theta^{2n} \partial/\partial x\}_{n=1}^\infty$, and $\{\theta^{2n} \cdot 1\}_{n=1}^\infty$. Notice that \mathcal{L} is simply a Lie subalgebra with two generators of the infinite dimensional Lie algebra obtained by tensoring the polynomial ring $\mathbb{R}[\theta^2]$ with the six dimensional Lie algebra of operators $\text{st}(1) = \{\partial^2/\partial x^2, x \partial/\partial x, \partial/\partial x, x^2, x, 1\}$; i.e., $\mathcal{L} \subseteq \mathbb{R}[\theta^2] \otimes \text{st}(1)$. From a slightly different point of view, we know from § 5.1 that if θ is a *known* constant, then we have a linear filtering problem in which $\mathcal{L} \subseteq \text{st}(n)$. In the present problem, however, θ is treated as a variable, each element of \mathcal{L} is a function of θ , and for each θ , each element of \mathcal{L} takes values in $\text{st}(1)$. Indeed, it is shown in [28] that \mathcal{L} is a solvable Lie subalgebra of the "current algebra" $C^\infty(\Theta; \text{st}(1))$ of C^∞ maps from Θ to $\text{st}(1)$. If Θ is a finite set, then clearly $C^\infty(\Theta, \text{st}(1))$ is finite dimensional, and so is \mathcal{L} ; thus finite dimensional recursive filters (consisting essentially of a Kalman filter for each value of θ) which compute \hat{x}_t and $P(\theta = i | \mathcal{Y}_t)$ can be constructed for this problem [22].

For the more interesting problem in which Θ is a general smooth manifold (such as \mathbb{R}^n), the homomorphism principle suggests that the existence of finite dimensional recursive filters is related to the representation of \mathcal{L} by a Lie algebra of vector fields on a finite dimensional manifold. With this motivation, it is shown in [27]–[28] that \mathcal{L} can be realized by such a Lie algebra of vector fields arising from a finite dimensional filter which, for a particular $\theta_0 \in \Theta$, computes $E[x_t | \mathcal{Y}_t, \theta_0]$ and the density of θ given \mathcal{Y}_t , *evaluated at* θ_0 . Although this density can be computed finite dimensionally for each $\theta_0 \in \Theta$, the computation of the entire posterior density function appears not to admit a finite dimensional filter in general (unless Θ is finite). Despite the intrinsic infinite dimensionality of this problem, it does have considerable structure which can be exploited via Lie algebraic methods. For example, the unnormalized joint conditional density $\rho(t, x, \theta)$ of x_t and θ given \mathcal{Y}_t is computed in [27]–[28] via a generalization of the Wei–Norman method; in this generalization, the ordinary differential equations (29) are replaced by first-order partial differential equations (which cannot be computed on-line, of course), and $\rho(t, x, \theta)$ is computed explicitly in terms of the $\{g^i\}$ in a formula similar to (47). By expanding the solutions of the equations for the $\{g^i\}$, it is easily seen that a sufficient statistic for $\rho(t, x, \theta)$ is given by the sequence $\{\int_0^t (\sigma^k/k!) dy_\sigma\}$, thus further elucidating the structure of the problem.

6. Lie algebras and asymptotic expansions. The Lie algebraic results of § 4 on the nonexistence of *exact* finite dimensional filters do not address the issue of nonexact but high-performance suboptimal filters, which is much more important from a practical point of view. This problem is considered from both Lie algebraic and analytical points of view in [5] for linear systems with small nonlinear perturbations. A typical such system has the form

$$\begin{aligned} dx_t &= ax_t dt + d\tilde{w}_t, \\ dy_t^\varepsilon &= [x_t + \varepsilon(x_t)^k] dt + dw_t, \quad k \geq 1, \\ y_0^\varepsilon &= 0 \quad p_0(x) \text{ Gaussian} \end{aligned}$$

where ε is a small positive parameter; as $\varepsilon \downarrow 0$, we recover the linear problem of § 5.1. Let the estimation algebra for this problem, with fixed k, ε , be denoted $\mathcal{L}_{k\varepsilon}$.

As a particular case, consider the “weak cubic sensor” when $k=3$. It can be shown [14] that, just as with the cubic sensor (Example 4.1), the estimation algebra $\mathcal{L}_{3\varepsilon}$ for $\varepsilon \neq 0$ is isomorphic to the Weyl algebra W_1 , and that no nontrivial statistics can be computed exactly with finite dimensional filters. Of course, for $\varepsilon=0$, the Lie algebra \mathcal{L}_{30} is just the 4-dimensional algebra of the linear problem (§ 5.1); in this case, there is a 2-dimensional sufficient statistic which can be evaluated recursively. Thus, as ε passes from zero to $\varepsilon \neq 0$, the filtering problem moves from the simplest to the most difficult class.

To treat the case $\varepsilon \neq 0$ small, it is natural to consider expansions of the conditional density and statistics in powers of ε . From a Lie algebraic point of view, this was considered in [14]. Let $W_1(\varepsilon) = \mathbb{R}\langle x, \varepsilon, d/dx \rangle$ be the Lie algebra of differential operators with coefficients that are polynomials in x and ε . Thus, $W_1(\varepsilon)$ has basis $\{e_{ijl} = \varepsilon^i x^j d^l/dx^l; i, j, l = 0, 1, \dots\}$ (here we regard ε as a “variable”). The estimation algebra $\mathcal{L}_{k\varepsilon}$ may be regarded as a subalgebra of $W_1(\varepsilon)$. Define $\mathcal{L}_{k\varepsilon} \bmod \varepsilon^n$ as the Lie algebra obtained from $\mathcal{L}_{k\varepsilon}$ by setting $\varepsilon^i = 0$ for $i \geq n$. It is shown that $\mathcal{L}_{k\varepsilon} \bmod \varepsilon^n$ is finite dimensional for each k, n (thus giving some reason to believe this procedure may lead to finite dimensional filters); for instance, $\mathcal{L}_{3\varepsilon} \bmod \varepsilon^1$ is \mathcal{L}_0 , and $\mathcal{L}_{3\varepsilon} \bmod \varepsilon^2$ is 14 dimensional.

This development is related to asymptotic expansions in the following way. The unnormalized conditional density $\rho^\varepsilon(t, x)$ of x_t given \mathcal{Y}_t satisfies (10)–(11) with $L^* \rho^\varepsilon = \frac{1}{2} \partial^2 \rho^\varepsilon / \partial x^2 - a \partial(x \rho^\varepsilon) / \partial x$. We assume that ρ^ε has a formal expansion in powers of ε

$$(59) \quad \rho^\varepsilon(t, x) = \rho_0(t, x) + \varepsilon \rho_1(t, x) + \varepsilon^2 \rho_2(t, x) + \dots$$

Substituting this in (10) and equating coefficients of powers of ε gives

$$(60) \quad \begin{aligned} d\rho_0(t, x) &= L^* \rho_0(t, x) dt + x \rho_0(t, x) dy_t^\varepsilon, \\ \rho_0(0, x) &= p_0(x), \end{aligned}$$

$$(61) \quad \begin{aligned} d\rho_l(t, x) &= L^* \rho_l(t, x) dt + x \rho_l(t, x) dy_t^\varepsilon + x^k \rho_{l-1}(t, x) dy_t^\varepsilon, \quad l = 1, 2, \dots, \\ \rho_l(0, x) &= 0. \end{aligned}$$

Writing (60)–(61) in Fisk–Stratonovich form and truncating after $l = n$ gives

$$(62) \quad d \begin{bmatrix} \rho_0 \\ \rho_1 \\ \rho_2 \\ \vdots \\ \rho_n \end{bmatrix} = \begin{bmatrix} L^* - \frac{1}{2} x^2 & 0 & & & \\ -x^{k+1} & L^* - \frac{1}{2} x^2 & 0 & & \\ -\frac{1}{2} x^{2k} & -x^{k+1} & L^* - \frac{1}{2} x^2 & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{2} x^{2k} & -x^{k+1} & L^* - \frac{1}{2} x^2 \end{bmatrix} \begin{bmatrix} \rho_0 \\ \rho_1 \\ \rho_2 \\ \vdots \\ \rho_n \end{bmatrix} dt$$

$$+ \begin{bmatrix} x & & & & \\ x^k & x & & & 0 \\ & & \ddots & \ddots & \\ 0 & & & x^k & x \end{bmatrix} \begin{bmatrix} \rho_0 \\ \rho_1 \\ \vdots \\ \rho_n \end{bmatrix} \circ dy_t^\varepsilon.$$

The controllability Lie algebra of (62) is isomorphic to $\mathcal{L}_{k\varepsilon} \bmod \varepsilon^{n+1}$; thus the Lie algebraic construction discussed above corresponds to the approximation (59) to $n+1$ terms. This is important, because using the additional fact that $\mathcal{L}_{k\varepsilon} \bmod \varepsilon^{n+1}$ is solvable, it is shown in [5] that (62) can be solved by the Wei–Norman method, and the resulting $\{g^i\}$ constitute a finite dimensional recursive filter for the approximate conditional density up to $n+1$ terms. To justify the use of this expansion and the resulting approximate filters, it is proved in [5] that the formal expansion (59)–(61) is in fact a true asymptotic expansion; i.e., it is shown that, in an appropriate norm, the error $\rho^\varepsilon(t, x) - \sum_{i=0}^n \varepsilon^i \rho_i(t, x)$ is of the order of ε^{n+1} for ε small. The corresponding results and filters are also presented for the approximations to the conditional mean, and the mean-square errors of the zeroth and first order filters are compared by simulation in [5] to the extended Kalman filter (EKF); in general, the zeroth order filter performs worse than the EKF, while the first order filter performs better than the EKF. Similar results were obtained by other methods in [49].

7. Conclusions. In this paper we have attempted to demonstrate that methods from nonlinear systems and Lie algebras can provide useful tools for the solution and understanding of nonlinear filtering problems. Equivalent filtering problems can be recognized via their estimation Lie algebras. Some new finite dimensional (or lower dimensional) filters have been constructed via Lie algebraic techniques; these may not all be filtering problems that arise in practice, but these results aid in the understanding of the structure of the filtering problem from a different point of view. In a similar fashion, these methods have provided new insights into known finite dimensional filters, such as the Kalman filter. The Lie algebraic and system theoretic approach has, for the first time, rigorously shown the inherent infinite dimensionality of certain filtering problems. Finally, these methods have aided in the construction of filters which arise in asymptotic expansions of the conditional density.

If one is approaching a particular nonlinear filtering problem, how can these tools be used? First an attempt should be made to compute the Lie algebra; if it is finite dimensional and solvable, finite dimensional filters can usually be computed via the Wei–Norman technique. However, one must be careful to take into account the domain of the differential operators and boundary conditions in computing the Lie algebra if the state is not defined on \mathbb{R}^n (cf. [41]). The unnormalized density obtained by the Wei–Norman technique can be plugged back into the D-M-Z equation to determine if it is a solution. In most cases, the Lie algebra will not be finite dimensional; if it is infinite dimensional, this still does not exclude finite dimensional filters (cf. § 5.5). If one can show that \mathcal{L} is a Weyl algebra, then (modulo a rigorous homomorphism theorem), there are no exact finite dimensional recursive filters for conditional statistics. If \mathcal{L} is infinite dimensional but not a Weyl algebra, the Lie algebraic information is often not of great use, since even if there are homomorphisms from \mathcal{L} to Lie algebras of vector fields on finite dimensional manifolds, it is difficult to determine precisely which conditional statistic the corresponding filter computes. In the infinite dimensional case, it is sometimes possible to gain insight into suboptimal filters through the estimation algebra and its properties (cf. § 6).

There are still a number of open problems which deserve attention. Rigorous versions of the homomorphism theorem (particularly in higher dimensions) are necessary for concluding the nonexistence of finite dimensional filters. In addition, Ocone [42] has proposed formal conditions that a finite dimensionally computable statistic must satisfy; if this were made rigorous, it would become a powerful tool. Along these lines, we note also the Lie algebraic necessary condition presented in [55]

for the existence of a finite dimensional filter which computes all conditional statistics. It also appears that if other examples with \mathcal{L} finite dimensional are to be found, they will occur in higher dimensions (cf. [41] and § 5.4); the search for these should be continued and will certainly lead to new insights. A number of purely algebraic problems concerning the structure of \mathcal{L} are posed in [15], including the structure of subalgebras of the Weyl algebra and the investigation of how \mathcal{L} changes if the observation structure is changed (e.g., when an observation is added, when the output is processed through another system before being observed, etc.). Possibly the most promising and potentially applicable direction is the further study of the uses of the estimation algebra in the design of suboptimal estimators, either along the lines of § 6 or through some type of study of the extended Kalman filter.

REFERENCES

- [1] J. BAILLIEUL, *Estimation problems with low dimensional filters*, Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., Reidel, Dordrecht, The Netherlands, 1981, pp. 559–564.
- [2] J. S. BARAS, *Group invariance methods in nonlinear filtering of diffusion processes*, Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., Reidel, Dordrecht, The Netherlands, 1981, pp. 565–572.
- [3] V. E. BENEŠ, *Exact finite dimensional filters for certain diffusions with nonlinear drift*, Stochastics, 5 (1981), pp. 65–92.
- [4] J.-M. BISMUT AND D. MICHEL, *Diffusions conditionnelles. I. Hypoellipticité partielle*, J. Functional Analysis, 44 (1981), pp. 174–211.
- [5] G. L. BLANKENSHIP, C.-H. LIU AND S. I. MARCUS, *Asymptotic expansions and Lie algebras for some nonlinear filtering problems*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 787–797.
- [6] R. W. BROCKETT, *Nonlinear systems and differential geometry*, Proc. IEEE, 64 (1976), pp. 61–72.
- [7] ———, *Remarks on finite dimensional nonlinear estimation*, Analyse des systèmes, C. Lobry, ed., Bordeaux, 1978; also, Astérisque, 76 (1980), Soc. Math. de France.
- [8] ———, *Classification and equivalence in estimation theory*, Proc. 18th IEEE Conference on Decision and Control, Ft. Lauderdale, FL, 1979, pp. 172–175.
- [9] ———, *Nonlinear systems and nonlinear estimation theory*, Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., Reidel, Dordrecht, The Netherlands, 1981, pp. 441–478.
- [10] R. W. BROCKETT AND J. M. C. CLARK, *The geometry of the conditional density equation*, Analysis and Optimization of Stochastic Systems, O. L. R. Jacobs et al., eds., New York, 1980, pp. 299–309.
- [11] M. H. A. DAVIS AND S. I. MARCUS, *An introduction to nonlinear filtering*, Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., Reidel, Dordrecht, The Netherlands, 1981, pp. 565–572.
- [12] ———, *Nonlinear Estimation*, MIT Press, Cambridge, MA, 1985.
- [13] T. E. DUNCAN, *Probability densities for diffusion processes with applications to nonlinear filtering theory and diffusion theory*, Ph.D. thesis, Stanford Univ., Stanford, CA, 1967.
- [14] M. HAZEWINKEL, *On deformations, approximations, and nonlinear filtering*, Syst. Contr. Lett., 1 (1981), pp. 32–36.
- [15] M. HAZEWINKEL AND S. I. MARCUS, *Some results and speculations on the role of Lie algebras in filtering*, Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., Reidel, Dordrecht, The Netherlands, 1981, pp. 591–604.
- [16] ———, *On Lie algebras and finite dimensional filtering*, Stochastics, 7 (1982), pp. 29–62.
- [17] M. HAZEWINKEL, S. I. MARCUS AND H. J. SUSSMANN, *Nonexistence of finite dimensional filters for conditional statistics of the cubic sensor problem*, Syst. Contr. Lett., 3 (1983), pp. 331–340.
- [18] M. HAZEWINKEL AND J. C. WILLEMS, eds., *Stochastic Systems: The Mathematics of Filtering and Identification and Applications*, Reidel, Dordrecht, The Netherlands, 1981.
- [19] R. HERMANN AND A. J. KRENER, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 728–740.
- [20] O. B. HIJAB, *Minimum energy estimation*, Ph.D. thesis, Univ. of California, Berkeley, 1980.
- [21] ———, *Finite dimensional causal functionals of brownian motion*, Proc. NATO-ASI Nonlinear Stochastic Problems, Reidel, Dordrecht, The Netherlands, 1982.

- [22] ———, *The adaptive LQG problem—Part I*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 171–178.
- [23] G. KALLIANPUR, *Stochastic Filtering Theory*, Springer-Verlag, New York, 1980.
- [24] R. E. KALMAN AND R. S. BUCY, *New results in linear filtering and prediction theory*, Trans. ASME Ser. D. J. Basis Eng., 83 (1961), pp. 95–108.
- [25] A. J. KRENER, *On the equivalence of control systems and the linearization of nonlinear systems*, this Journal, 11 (1973), pp. 670–676.
- [26] P. S. KRISHNAPRASAD AND S. I. MARCUS, *Some nonlinear filtering problems arising in recursive identification*, Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., Reidel, Dordrecht, The Netherlands, 1981, pp. 299–304.
- [27] ———, *System identification and nonlinear filtering: Lie algebras*, Proc. 20th IEEE Conference on Decision and Control, San Diego, CA, 1981, pp. 330–334.
- [28] P. S. KRISHNAPRASAD, S. I. MARCUS AND M. HAZEWINDEL, *Current algebras and the identification problem*, Stochastics, 11 (1983), pp. 65–101.
- [29] H. KUNITA, *Stochastic partial differential equations connected with non-linear filtering*, Nonlinear Filtering and Stochastic Control, S. K. Mitter and A. Moro, eds., Springer-Verlag, New York, 1983, pp. 100–169.
- [30] R. S. LIPTSER AND A. N. SHIRYAEV, *Statistics of Random Processes I*, Springer-Verlag, New York, 1977.
- [31] C.-H. LIU, *Applications of algebraic and approximation methods in nonlinear estimation*, Ph.D. dissertation, Dept. Electrical Engineering, Univ. Texas at Austin, Austin, TX, 1981.
- [32] C.-H. LIU AND S. I. MARCUS, *The Lie algebraic structure of a class of finite dimensional nonlinear filters*, Algebraic and Geometric Methods in Linear Systems Theory, Lectures in Applied Math. 18, C. I. Byrnes and C. F. Martin, eds., American Mathematical Society, Providence, RI, 1980, pp. 277–297.
- [33] S. I. MARCUS, *Low dimensional filters for a class of finite state estimation problems with Poisson observations*, Syst. Contr. Lett., 1 (1982), pp. 237–241.
- [34] ———, *Nonlinear estimation*, to appear in Encyclopedia of Systems and Control, M. Singh, ed., Pergamon, Oxford, 1984.
- [35] S. I. MARCUS, S. K. MITTER AND D. L. OCONE, *Finite dimensional nonlinear estimation for a class of systems in continuous and discrete time*, Analysis and Optimization of Stochastic Systems, O. L. R. Jacobs et al., eds., Academic Press, New York, 1980, pp. 387–406.
- [36] S. I. MARCUS AND A. S. WILLISKY, *Algebraic structure and finite dimensional nonlinear estimation*, SIAM J. Math. Anal., 9 (1978), pp. 312–327.
- [37] S. K. MITTER, *Filtering theory and quantum fields*, Analyse des systèmes, C. Lobry, eds., Bordeaux, 1978; also, Astérisque, 76 (1980).
- [38] ———, *On the analogy between mathematical problems of non-linear filtering and quantum physics*, Recherche di Automatica, 10 (1980), pp. 163–216.
- [39] R. E. MORTENSEN, *Optimal control of continuous-time stochastic systems*, Ph.D. thesis, Univ. California, Berkeley, 1966.
- [40] D. L. OCONE, *Topics in nonlinear filtering theory*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, 1980.
- [41] ———, *Finite dimensional estimation algebras in nonlinear filtering*, Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., Reidel, Dordrecht, The Netherlands, 1981, pp. 629–636.
- [42] ———, *Finite dimensionally computable statistics and estimation algebras in nonlinear filtering*, Proc. 1981 International Symposium on the Mathematical Theory of Networks and Systems, Santa Barbara, CA, 1981.
- [43] D. L. OCONE, J. S. BARAS AND S. I. MARCUS, *Explicit filters for diffusions with certain nonlinear drifts*, Stochastics, 8 (1982), pp. 1–16.
- [44] E. PARDOUX, *Equations du filtrage nonlinéaire, de la prédiction et du lissage*, Stochastics, 6 (1982), pp. 193–232.
- [45] H. SUSSMANN, *An extension of a theorem of Nagano on transitive Lie algebras*, Proc. Am. Math. Soc., 45 (1974), pp. 349–356.
- [46] ———, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Syst. Theory, 10 (1977), pp. 263–284.
- [47] ———, *Rigorous results on the cubic sensor problem*, Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., Reidel, Dordrecht, The Netherlands, 1981, pp. 637–648.
- [48] ———, *Nonexistence of finite dimensional filters for the cubic sensor problem*, preprint, December 1982.

- [49] H. SUSSMANN, *Approximate finite dimensional filters for some nonlinear problems*, Stochastics, 7 (1982), pp. 183–203.
- [50] J. WEI AND E. NORMAN, *On the global representation of the solutions of linear differential equations as a product of exponentials*, Proc. Am. Math. Soc., 15 (1964), pp. 327–334.
- [51] E. WONG, *Stochastic Processes in Information and Dynamical Systems*, McGraw-Hill, New York, 1971.
- [52] W. WONG, *New classes of finite dimensional filters*, Syst. Cont. Lett., 3 (1983), pp. 155–164.
- [53] W. M. WONHAM, *Some applications of stochastic differential equations to optimal nonlinear filtering*, SIAM J. Control, 2 (1965), pp. 347–369.
- [54] M. ZAKAI, *On the optimal filtering of diffusion processes*, Z. Wahrsch. Verw. Geb., 11 (1969), pp. 230–243.
- [55] M. CHALEYAT-MAUREL AND D. MICHEL, *Un théorème de non existence de filtre de dimension finie*, C. R. Acad. Sc. Paris, 296 (1983), Série I, pp. 933–936.

A NEW PROCEDURE FOR STOCHASTIC REALIZATION OF SPECTRAL DENSITY MATRICES*

A. J. VAN DER SCHAFT†‡ AND J. C. WILLEMS†

Abstract. In this paper we consider the problem of obtaining a state space realization of a zero mean gaussian vector process. A new algorithm is presented for the case in which the process is given in terms of its spectral density function. Contrary to the usual procedure followed, which requires a partial fraction expansion, the algorithm presented starts with a (deterministic) realization of the spectral density function itself.

Key words. Spectral densities, stochastic realization, Hamiltonian systems, spectral factorization, Riccati equation

1. Introduction. One of the basic problems in mathematical system theory is the question of obtaining an “internal” state space realization of a system given in “external” form. In the stochastic case the main problem studied in this context may be formulated as follows: *Given a stationary zero mean gaussian vector process y , construct a stationary zero mean Gauss–Markov vector process x and a matrix C such that Cx has the same statistical properties as y .*

The basic formulation of this problem is due to Kalman [6] about fifteen years ago and since that time numerous publications have appeared on it. A good account of the existing theory may be found in the book by Faurre et al. [5]. The existing algorithms assume that the process y is given by its (matrix) autocorrelation function. However, in many applications the process y will be given by its spectral density function and in order to apply the existing algorithms it is then necessary to factor (e.g., by means of a partial fraction expansion) the spectral density into a part which is analytic in $\operatorname{Re} s > 0$ and a part which is analytic in $\operatorname{Re} s < 0$. This is however a nontrivial step in the application of these methods, and more often than not this difficulty is glossed over. Indeed this requires the factorization of a high order polynomial. From a numerical point of view this is a nonlinear problem of the same level of difficulty as solving an algebraic Riccati equation. This difficulty motivates the approach followed in the present paper. We will propose a different solution of the problem of stochastically realizing a spectral density. Our approach is based on the observation that a spectral density matrix $\Phi(s)$ is always *Hamiltonian* ($\Phi(s) = \Phi^T(-s)$) and *passive* ($\Phi(j\omega) + \Phi^T(-j\omega) \geq 0$). Exploiting this structure, one may obtain a special “deterministic” realization of $\Phi(s)$ (viewed itself as a rational function in s) which then yields an algebraic Riccati equation from which a stochastic realization (or a white noise representation) of the original process is readily derived.

2. Realization of autocorrelation functions. In this section we will introduce the problem and review some of the main results previously obtained on it.

Let $y := \{y_t, t \in \mathbb{R}\}$ be a real p -dimensional vector process defined on a probability space $\{\Omega, \mathcal{A}, P\}$. We assume that y is gaussian, zero mean and stationary. In the stochastic realization problem for such processes we look for a real n -dimensional zero mean stationary Gauss–Markov process $x := \{x_t, t \in \mathbb{R}\}$, defined on a probability

* Received by the editors September 12, 1982, and in revised form September 21, 1983.

† Mathematics Institute, P.O. Box 800, 9700 AV Groningen, the Netherlands.

‡ Current address: Department of Applied Mathematics, Twente University of Technology, P.O. Box 217, 7500 AE Enschede, the Netherlands.

space $\{\Omega', \mathcal{A}', P'\}$, and a matrix $C: \mathbb{R}^n \rightarrow \mathbb{R}^p$ such that $Cx \sim y$. Here \sim denotes (*stochastic equivalence*), i.e. the probability distributions of the vectors $(y_{t_1}, y_{t_2}, \dots, y_{t_k})$ and $(Cx_{t_1}, Cx_{t_2}, \dots, Cx_{t_k})$ should be equal for all choices of t_1, t_2, \dots, t_k . Obviously if (x, C) realizes y , then so does (Px, CP^{-1}) for any nonsingular matrix P . We will say that (x, C) defines a *minimal* realization of y if the only matrices $P: \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that (Px, C') also realizes y for some C' are actually the nonsingular matrices.

Note that we do not require that y and x are defined on the same probability space. Often this problem is therefore called the *weak* realization problem in order to distinguish it from the *strong* version in which $\{\Omega', \mathcal{A}', P'\} = \{\Omega, \mathcal{A}, P\}$. Other than in Remark 6 we will in this paper deal exclusively with the weak version.

The marginal probability distributions of y (and thus of all processes which are equivalent to it) are completely specified by its autocorrelation function $R: \mathbb{R} \rightarrow \mathbb{R}^{p \times p}$ defined by $R(t) := \mathbf{E}\{y(t)y^T(0)\}$. The restriction of R to $[0, \infty)$ will be denoted by R^+ . Of course, R^+ specifies R completely since $R(t) = R^T(-t)$. The following basic result from deterministic realization theory is well known:

PROPOSITION 1. *The following conditions are equivalent:*

(i) R^+ is Bohl (i.e., every entry of R^+ is a finite sum of products of a polynomial, an exponential, and a trigonometric function).

(ii) There exists matrices (F, G, H) such that $R^+(t) = He^{Ft}G$, with $F \in \mathbb{R}^{n \times n}$, $G \in \mathbb{R}^{n \times p}$, and $H \in \mathbb{R}^{p \times n}$.

In addition:

(iii) There is a minimal n , n_{\min} , for which the factorization in (ii) is possible. This n_{\min} is called the *McMillan degree* of R^+ , and $n = n_{\min}$ iff (F, G) is controllable and (F, H) is observable. The triple (F, G, H) is then called *minimal*.

(iv) All triples (F, G, H) with $n = n_{\min}$ are obtainable from one by the transformation group

$$(F, G, H) \xrightarrow[\det S \neq 0]{S} (SFS^{-1}, SG, HS^{-1}).$$

Assume that x is an n -dimensional zero mean stationary Gauss–Markov process with $\mathbf{E}\{x(t)x^T(t)\} = Q > 0$ (obviously $Q > 0$ whenever (x, C) is a minimal realization of y). We call such Markov processes *nonsingular*. Then $\mathbf{E}\{x(t)|x(0)\}$ for $t \geq 0$ is of the form $e^{At}x(0)$ for some A . Moreover it is easily seen that Q and A completely specify the marginal distributions of the Markov process x . It is in fact not difficult to see that this way any pair of matrices $\{Q, A\}$ will define a zero mean stationary nonsingular Gauss–Markov process x with $\mathbf{E}\{x(0)x^T(0)\} > 0$ iff $Q = Q^T > 0$ and $AQ + QA^T \leq 0$. We will call Q the *covariance* and A the *infinitesimal generator* of x . Now the marginal distributions of the process Cx are completely specified by (A, Q, C) and it thus makes sense to talk about this triple as defining a realization. Since in our context minimality as defined previously corresponds also to taking the dimension of x as small as possible we will call a realization (x, C) with x n -dimensional and n as small as possible a *minimal* realization.

Two processes x_1 and x_2 will be called *linearly equivalent* if there exists a nonsingular matrix S such that $x_2 \sim Sx_1$. Following this we will call two realizations (A_1, Q_1, C_1) and (A_2, Q_2, C_2) *linearly equivalent* if there exists a nonsingular matrix S such that $(A_2, Q_2, C_2) = (SA_1S^{-1}, SQ_1S^T, C_1S^{-1})$. The following theorem is the basic result in this area. It follows from the basic work by Kalman [6] and Faurre [4] and has later been studied further by Anderson [2], Lindquist and Picci [8], Ruckebush [11] and many others.

THEOREM 2.

(i) *There exists a finite dimensional realization of y iff its autocorrelation function R is Bohl.*

(ii) *All minimal realizations (A, Q, C) can, up to linear equivalence, be obtained from a minimal factorization triple (F, G, H) of R^+ by taking $A = F$, $C = H$, and solving the relations*

$$FG + QF^T \leq 0, \quad QH^T = G$$

for $Q = Q^T$.

The problem thus becomes one of solving this combination of matrix inequalities/equalities. Actually, it may be shown [4], [5], [13] that there exist solutions Q_- , Q_+ such that for every other solution Q , there holds $0 < Q_- \leq Q \leq Q_+ < \infty$. Moreover the solution set is convex and compact. A great deal of additional information on the structure of the solution set of these equations may be found in the above references. Note that choosing $A = F$ and $C = H$ in the above theorem corresponds to fixing the basis in the state space (since (F, H) is observable). Once the basis has been picked, it is only the covariance of x , Q , which remains to be chosen.

The above results are the well-known basic facts of stochastic realization theory. They yield a minimal stochastic realization by the following procedure. This procedure inputs as the

Data. The autocorrelation R^+ of y .

Then it computes as

Step 1. Determine a minimal realization (F, G, H) of R^+ .

Subsequently it proceeds with

Step 2. Solve the linear matrix inequality

$$Q = Q^T, \quad FQ + QF^T \leq 0, \quad QH^T = G.$$

The procedure then returns (F, Q, H) which defines a minimal realization (x, H) of y , where x is a Gauss–Markov process with covariance Q and infinitesimal generator F . For the algorithmic implementation of Step 1 we can use any of the realization theory algorithms of linear systems theory. Also Step 2 has received much attention and may be reduced to solving linear matrix equations and a suitably defined reduced order algebraic Riccati equation.

Remark 1. Some readers may be more familiar with the problem of generating a spectral density function by passing white noise through a linear system (which is called the *shaping filter*). In fact, this problem is solved by a simple extension of the result of Theorem 2. Indeed, let (A, Q, C) be a minimal stochastic realization and let B be such that $AQ + QA^T = -BB^T$. Consider now the system described by the stochastic differential equation

$$dx = Ax \, dt + B \, dw, \quad x(0) = x_0, \\ \tilde{y} = Cx$$

with x_0 zero mean gaussian, and $\mathbf{E}\{x_0 x_0^T\} = Q$, and with $w := \{w_t, t \in [0, \infty)\}$ a normalized Wiener process, independent of x_0 . This defines a white noise driven model which generates a process $\tilde{y} \sim y$.

In the present paper we will develop an algorithm which starts from Φ and proceeds by determining a minimal realization of Φ (considered as a rational function in s). Since in very many applications the spectral density is a more basic design specification than the autocorrelation function, it may be of interest to have this alternative algorithm

available (even though we do not claim any superiority of our algorithm above the previously mentioned one which uses the partial fraction expansion followed by an implementation of Theorem 2).

3. Realization of spectral density matrices. Let $y = \{y_t, t \in \mathbb{R}\}$ be a real gaussian p -dimensional vector process and assume that its autocorrelation function R is integrable. Let $\Phi(s)$ denote the (two-sided) Laplace transform of R . Φ is called the spectral density of y [10], [16]. Obviously Φ is well-defined in a strip containing the imaginary axis. The existence of a finite dimensional realization of y may of course also be deduced from its spectral density Φ :

PROPOSITION 3. *Assume that R is integrable. Then the following conditions are equivalent:*

- (i) R^+ is Bohl.
- (ii) Φ is rational.

We remark that the assumption that R is integrable is in our context equivalent to assuming that the process y is ergodic.

In applications of stochastic realization theory it happens more often than not that one starts with a rational spectral density matrix. Of course, in this case one may proceed by computing a partial fraction expansion of $\Phi(s)$

$$\Phi(s) = Z(s) + Z^T(-s)$$

with $Z(s)$ analytic in $\operatorname{Re} s \geq 0$, realizing $Z(s)$ minimally as $Z(s) = H(sI - F)^{-1}G$, and using the theory of § 2. However, the problem of factoring $\Phi(s)$ into the above form is a highly nontrivial one. It requires factoring a polynomial which may be of high degree, and all together this may very well be the most difficult step in this whole realization procedure.

Our purpose in the present paper is to outline a procedure which proceeds by using a realization of the spectral density $\Phi(s)$ directly (and not of its "causal" part $Z(s)$). It is well known that a rational $(p \times p)$ matrix $\Phi(s)$ is a spectral density matrix of a process y with integrable R if and only if it has the following properties:

- (i) $\Phi(s) = \Phi^T(-s)$,
- (ii) $\Phi(s)$ has no poles on the imaginary axis
- (iii) $\Phi(j\omega) \geq 0 \quad \forall \omega \in \mathbb{R}$,
- (iv) $\lim_{s \rightarrow \infty} \Phi(s) = 0$.

In the sequel we will show how by constructing a special minimal realization of $\Phi(s)$, initially seen as a transfer function, we can solve the stochastic realization problem. The key observation for our procedure is that $\Phi(s)$ is a *Hamiltonian* transfer matrix, i.e. $\Phi(s) = \Phi^T(-s)$ (in the literature this property is also called para-hermitian). It is well known (cf. [3]) that a minimal realization $(\bar{A}, \bar{B}, \bar{C})$ of $\Phi(s)$ will then satisfy

$$\begin{aligned} (1) \quad & \bar{A}^T J + J \bar{A} = 0, \\ (2) \quad & \bar{B}^T J = \bar{C} \end{aligned}$$

for some unique nonsingular matrix J satisfying $J = -J^T$. It follows from the nonsingularity of J that the dimension of the state space is even, say $2n$. Basis free, J is an anti-symmetric bilinear form on \mathbb{R}^{2n} , and is called a *symplectic* form. By Darboux's theorem there exist bases of \mathbb{R}^{2n} in which J has the matrix form $\begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix}$ (cf. [1]).

The second observation which we make is that $\Phi(j\omega) \geq 0$ implies the *passivity* of the system with transfer matrix Φ . (Note that Φ is not asymptotically stable. Passivity here means that for any input/output pair which corresponds to a closed path in state space, the \mathcal{L}_2 inner product of input and output is nonnegative (see (14])). In fact

$\Phi(j\omega) \geq 0$ is equivalent to the existence of a nonsingular symmetric matrix Σ , which is not necessarily positive definite (basis free, Σ is a nondegenerate quadratic form), such that (see [13])

$$(3) \quad \bar{A}^T \Sigma + \Sigma \bar{A} \leq 0,$$

$$(4) \quad \bar{B}^T \Sigma = \bar{C}.$$

We emphasize that the solution Σ of (3) and (4) is in general not unique. The facts that the realizations satisfy the symmetry conditions (1), (2) and the passivity conditions (3), (4) may be combined to yield:

LEMMA 4. *There exists a solution Σ of (3, 4) such that*

$$(5) \quad \Sigma = J \Sigma^{-1} J.$$

Proof. Notice that if Σ satisfies (3) and (4), then also $J \Sigma^{-1} J$ does, because of (1) and (2). Furthermore it is well known that the set of all Σ 's satisfying (3) and (4) is convex and compact. Since the map $\Sigma \rightarrow J \Sigma^{-1} J$, considered as a map on the space of nonsingular symmetric matrices, is continuous it therefore follows from Brouwer's fixed point theorem that there exist Σ 's satisfying (3) and (4) and $\Sigma = J \Sigma^{-1} J$. \square

Remark 2. The proof of Lemma 4 is completely analogous to the existence of reciprocal passive realizations [15].

Now consider for a solution Σ of (3), (4) which satisfies (5), the matrix $J^{-1} \Sigma$ (this is the Hamiltonian matrix corresponding to the energy function $\frac{1}{2} x^T \Sigma x$). Because of (5) and $\Sigma = \Sigma^T$, $J = -J^T$ it follows that

$$(6) \quad (J^{-1} \Sigma)^T J (J^{-1} \Sigma) = -J,$$

$$(7) \quad (J^{-1} \Sigma)^2 = I.$$

In [9] (see also [12]) it is proven that therefore there exist bases of \mathbb{R}^{2n} in which

$$J = \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} \quad \text{and} \quad J^{-1} \Sigma = \begin{pmatrix} I_n & 0 \\ 0 & -I_n \end{pmatrix},$$

or equivalently

$$(8) \quad J = \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} 0 & I_n \\ I_n & 0 \end{pmatrix}.$$

Roughly, this may be seen as follows. From (6) and (7) it follows that the eigenvalues of $J^{-1} \Sigma$ are $+1$ (multiplicity n) and -1 (multiplicity n). Define $\Sigma^+ := \ker(I - J^{-1} \Sigma)$ and $\Sigma^- := \ker(I + J^{-1} \Sigma)$. Take an arbitrary basis q_1, \dots, q_n of Σ^+ . It can be proven that there exists a basis p_1, \dots, p_n of Σ^- such that $q_i^T J p_j = \delta_{ij}$, $i, j = 1, \dots, n$. Furthermore it can be seen that after having fixed q_1, \dots, q_n , the vectors p_i as above are uniquely determined. Then in such a basis $\{q_1, \dots, q_n, p_1, \dots, p_n\}$, J and Σ have the required form (8). (Notice that the transformations which leave J and Σ in the form (8) are exactly the transformations of the form

$$\begin{pmatrix} S & 0 \\ 0 & (S^T)^{-1} \end{pmatrix},$$

with $\det S \neq 0$.)

Proceeding with a basis as explained it follows from (1) and (2) that \bar{A} , \bar{B} and \bar{C} have the form

$$(9) \quad \bar{A} = \begin{pmatrix} F & -P \\ -R & -F^T \end{pmatrix}, \quad \bar{B} = \begin{pmatrix} G \\ H^T \end{pmatrix}, \quad \bar{C} = (H \quad -G^T)$$

with $P = P^T$ and $R = R^T$.

Then (3) is equivalent to $P \geq 0$ and $R \geq 0$, while (4) gives $G = 0$.

We sum this up in

THEOREM 5. *Let $\Phi(s)$ be a matrix of rational functions satisfying*

- (i) $\Phi(s) = \Phi^T(-s)$,
- (ii) Φ has no poles on the imaginary axis,
- (iii) $\Phi(j\omega) \geq 0 \quad \forall \omega \in \mathbb{R}$,
- (iv) $\lim_{s \rightarrow \infty} \Phi(s) = 0$.

Then there exists a minimal realization $(\bar{A}, \bar{B}, \bar{C})$ of $\Phi(s)$ such that

$$(10) \quad \begin{aligned} \bar{A} &= \begin{pmatrix} F & -P \\ -R & -F^T \end{pmatrix}, \quad P = P^T \geq 0, \quad R = R^T \geq 0, \\ \bar{B} &= \begin{pmatrix} 0 \\ H^T \end{pmatrix}, \quad \bar{C} = (H \quad 0). \end{aligned}$$

The next step is to consider the following (n -dimensional) Riccati equation

$$(11) \quad F^T K + KF - KPK + R = 0.$$

We first state

LEMMA 6. *Controllability of (\bar{A}, \bar{B}) implies controllability of (F, P) .*

Proof. Write $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ corresponding to $\begin{pmatrix} F & -P \\ -R & -F^T \end{pmatrix}$, with $x_1 \in X_1$, $x_2 \in X_2$ (with $X_1 \simeq X_2 \simeq \mathbb{R}^n$). Suppose that (F, P) is not controllable. Then there exists a subspace $\mathcal{L} \subset X_1$, and $\mathcal{L} \neq X_1$, such that $\text{Im } P \subset \mathcal{L}$ and \mathcal{L} is F -invariant. Then it can be easily checked that $\mathcal{L} \oplus X_2$ is invariant with respect to $\begin{pmatrix} F & -P \\ -R & -F^T \end{pmatrix}$ and contains $\text{Im} \begin{pmatrix} 0 \\ H^T \end{pmatrix}$. Therefore it follows that

$$\begin{pmatrix} F & -P \\ -R & -F^T \end{pmatrix}, \begin{pmatrix} 0 \\ H^T \end{pmatrix}$$

is not controllable. \square

Remark 3. In general, observability of (\bar{A}, \bar{C}) does *not* imply observability of (F, R) .

Consider now the Riccati equation (11). By Lemma 6 $P = P^T \geq 0$ is such that (F, P) is controllable. Since also $R = R^T \geq 0$, and \bar{A} does not have purely imaginary eigenvalues, this implies [7], [13] that there exists a symmetric nonnegative definite matrix K satisfying (11) and such that $F - PK$ is asymptotically stable. In fact, we have to take the maximal $K = K^T$ satisfying (11) [13].

Applying now the symplectic transformation $\bar{K} = \begin{pmatrix} I & 0 \\ -K & I \end{pmatrix}$ to \bar{A} , \bar{B} and \bar{C} yields

$$(12) \quad \bar{K} \bar{A} \bar{K}^{-1} = \begin{pmatrix} F - PK & -P \\ 0 & -(F - PK)^T \end{pmatrix}, \quad \bar{K} \bar{B} = \begin{pmatrix} 0 \\ H^T \end{pmatrix}, \quad \bar{C} \bar{K}^{-1} = (H \quad 0).$$

Now define $A := F - PK$, $C := H$, and let Q be the symmetric positive definite matrix satisfying

$$(13) \quad AQ + QA^T = -P.$$

(Q is uniquely determined by (13) since $\text{Re } \sigma(A) < 0$ and Q is positive definite since

(A, P) is controllable.) Then (A, Q, C) is a minimal realization of the process with spectral density $\Phi(s)$. We summarize this as

THEOREM 7. *Let $\Phi(s)$ be a rational spectral density matrix. Then $\Phi(s)$ has, as a transfer matrix, a minimal realization $(\bar{A}, \bar{B}, \bar{C})$ of the form*

$$\bar{A} = \begin{pmatrix} F - PK & -P \\ 0 & -(F - PK)^T \end{pmatrix}, \quad \bar{B} = \begin{pmatrix} 0 \\ H^T \end{pmatrix}, \quad \bar{C} = (H \quad 0)$$

with $P = P^T \geq 0$, and $F - PK$ asymptotically stable. Now let $A := F - PK$, $C := H$ and let $Q = Q^T > 0$ be the unique solution of $AQ + QA^T = -P$. Then (A, Q, C) is a minimal realization of the process with spectral density $\Phi(s)$.

Proof. The existence of a minimal realization (A, B, C) as above follows from the previous considerations (Theorem 5 and Lemma 6). Then it is easy to see that

$$(14) \quad \bar{C}(Is - \bar{A})^{-1}\bar{B} = H(Is - A)^{-1}P(Is - A^T)^{-1}H^T.$$

Write $P = BB^T$ and define $W(s) = H(Is - A)^{-1}B$; then it follows that $\Phi(s) = W(s)W^T(-s)$. Furthermore $W(s)$ is analytic in $\operatorname{Re} s \geq 0$ and (A, B, H) is a minimal triple. So we have given a spectral factorization of $\Phi(s)$, which is known to be equivalent to the stochastic realization problem (see for instance [2], [5]). It is well known that in terms of this factorization the covariance Q of the associated Markov process is given as the solution of $AQ + QA^T = -BB^T = -P$. \square

Remark 4. Hence we have constructed a spectral factorization $\Phi(s) = W(s)W^T(-s)$ directly from $\Phi(s)$, instead of first taking a partial fraction expansion $\Phi(s) = Z(s) + Z^T(-s)$.

Remark 5. Write $P = BB^T$; then Theorem 7 gives us immediately a white noise representation of the process

$$dx = Ax \, dt + B \, dw, \quad x(0) = x_0,$$

$$\tilde{y} = Cx$$

with x_0 zero mean gaussian, $\mathbf{E}\{x_0 x_0^T\} = Q$ and $w := \{w_t, t \in [0, \infty)\}$ a normalized Wiener process independent of x_0 (see Remark 1). In this case, since A is asymptotically stable, we can equivalently define x_t by

$$x_t = \int_{-\infty}^t e^{A(t-s)} B \, dw(s).$$

Remark 6. It is well known (cf [4]) that a strong realization corresponds to a P of minimal rank. In our construction we obtain a strong realization by taking $\Sigma = \Sigma^T$ satisfying (3), (4) and (5) such that $\bar{A}^T \Sigma + \Sigma \bar{A}$ restricted to Σ^+ has minimal rank.

Remark 7. Take $Q = Q^T > 0$ as the solution of $AQ + QA^T = -P$. Applying the (symplectic) transformation $\bar{Q} = \begin{pmatrix} I & Q \\ 0 & I \end{pmatrix}$ to

$$\bar{A} = \begin{pmatrix} A & -P \\ 0 & -A^T \end{pmatrix}, \quad \bar{B} = \begin{pmatrix} 0 \\ C^T \end{pmatrix}, \quad \bar{C} = (C \quad 0)$$

yields

$$\bar{Q}\bar{A}\bar{Q}^{-1} = \begin{pmatrix} A & 0 \\ 0 & -A^T \end{pmatrix}, \quad \bar{Q}\bar{B} = \begin{pmatrix} QC^T \\ C^T \end{pmatrix}, \quad \bar{C}\bar{Q}^{-1} = (C \quad -CQ).$$

Then (A, QC^T, C) is a minimal triple with A asymptotically stable. When we define $Z(s) = C(Is - A)^{-1}QC^T$, it follows that we have obtained a partial fraction expansion of $\Phi(s)$ since $\Phi(s) = Z(s) + Z^T(-s)$.

Recapitulating we have obtained the following

ALGORITHM

Data. $\Phi(s)$, the spectral density matrix of the process y .

Step 1. Construct a minimal realization $(\bar{A}, \bar{B}, \bar{C})$ of $\Phi(s)$. Find the unique nonsingular J satisfying $\bar{A}^T J + J \bar{A} = 0$, $\bar{B}^T J = \bar{C}$. Then also $J = -J^T$.

Step 2. Find a nonsingular Σ satisfying $\bar{A}^T \Sigma + \Sigma \bar{A} \leq 0$, $\bar{B}^T \Sigma = \bar{C}$, $\Sigma = J \Sigma^{-1} J$, $\Sigma = \Sigma^T$.

Step 3. Compute $\Sigma^+ = \ker(I - J^{-1} \Sigma)$ and $\Sigma^- = \ker(I + J^{-1} \Sigma)$. Take a basis (q_1, \dots, q_n) for Σ^+ and construct the basis (p_1, \dots, p_n) for Σ^- such that $q_i^T J p_j = \delta_{ij}$, $i, j = 1, \dots, n$.

In this basis we can write

$$\bar{A} = \begin{pmatrix} F & -P \\ -R & -F^T \end{pmatrix}, \quad \bar{B} = \begin{pmatrix} 0 \\ H^T \end{pmatrix}, \quad \bar{C} = (H \quad 0)$$

and we will have $P = P^T \geq 0$ and $R = R^T \geq 0$.

Step 4. Find the maximal symmetric nonnegative definite solution K of

$$F^T K + K F - K P K + R = 0.$$

Step 5. Define $A := F - PK$, $C := H$, and $Q = Q^T > 0$ as the unique solution of $Q A^T + A Q = -P$. Then (A, Q, C) is a minimal realization of y .

Step 6. Let $P = B B^T$. Then

$$dx = A x dt + B dw, \quad x(0) = x_0, \quad \mathbf{E}\{x_0 x_0^T\} = Q,$$

$$\tilde{y} = C x$$

is a white noise representation of y .

Remark 8. With respect to the actual calculation of Σ in Step 2 we can be more specific. Recall that the signature triple of a symmetric matrix consists of the number of positive, negative, respectively zero eigenvalues. We make use of the following

LEMMA 8. Let $(\bar{A}, \bar{B}, \bar{C})$ be a minimal triple (not necessarily satisfying (1) and (2)). Denote the set of symmetric solutions of (3) and (4) by $S(\Sigma)$. Then every element of $S(\Sigma)$ is nonsingular and has the same signature triple. If $(\bar{A}, \bar{B}, \bar{C})$ also satisfies (1) and (2) then every element of $S(\Sigma)$ has n negative eigenvalues and n positive eigenvalues.

Proof. Let $\bar{A}^T \Sigma + \Sigma \bar{A} = -L^T L$. First suppose that (\bar{A}, L) is observable. Now let $x \in \ker \Sigma$. Then $x^T \bar{A}^T \Sigma x + x^T \Sigma \bar{A} x = -x^T L^T L x$ implies $Lx = 0$. Furthermore $\bar{A}^T \Sigma x + \Sigma \bar{A} x = -L^T L x$ implies $\Sigma \bar{A} x = 0$. Therefore $\ker \Sigma$ is \bar{A} -invariant and contained in $\ker L$. By observability of (\bar{A}, L) it follows that $\ker \Sigma = 0$. If (\bar{A}, L) is not observable we proceed as follows. By (3) and (4)

$$(15) \quad (\bar{A} - \bar{B} \bar{C})^T \Sigma + \Sigma (\bar{A} - \bar{B} \bar{C}) = -L^T L - 2 \bar{C}^T \bar{C}.$$

Then: $\{(\bar{A}, \bar{C}) \text{ observable}\} \Leftrightarrow \{(\bar{A}, \bar{C}^T \bar{C}) \text{ observable}\} \Leftrightarrow \{(\bar{A} - \bar{B} \bar{C}, 2 \bar{C}^T \bar{C}) \text{ observable}\} \Rightarrow \{(\bar{A} - \bar{B} \bar{C}, -L^T L - 2 \bar{C}^T \bar{C}) \text{ observable}\}$. Therefore as above we can conclude that Σ satisfying (15) is nonsingular.

Because $S(\Sigma)$ is convex and consists of nonsingular matrices every element of $S(\Sigma)$ has the same signature triple. By Lemma 4 it follows that there exists a $\Sigma \in S(\Sigma)$ satisfying (5). This Σ has n negative and n positive eigenvalues as can be seen from (8). \square

Consider now the following algorithm (see also [15]).

Let Σ_1 be a solution of (3) and (4). Then define

$$(16) \quad \Sigma_{n+1} = \frac{1}{2}(\Sigma_n + J \Sigma_n^{-1} J), \quad n \geq 1.$$

If $\Sigma \in S(\Sigma)$, then also $J\Sigma^{-1}J \in S(\Sigma)$. Therefore since every element of $S(\Sigma)$ is nonsingular and $S(\Sigma)$ is convex, $\Sigma_{n+1} \in S(\Sigma)$ for every $n \geq 1$. From the compactness of $S(\Sigma)$ it follows that $\lim_{n \rightarrow \infty} \Sigma_n \in S(\Sigma)$. Denote $\Sigma_\infty := \lim \Sigma_n$. Then Σ_∞ satisfies $\Sigma_\infty = \frac{1}{2}(\Sigma_\infty + J\Sigma_\infty^{-1}J)$, or equivalently, $\Sigma_\infty = J\Sigma_\infty^{-1}J$. Furthermore note that we can rewrite (16) as

$$(17) \quad \Sigma_{n+1}J = \frac{1}{2}(\Sigma_nJ + (\Sigma_nJ)^{-1}).$$

Since $\lim \frac{1}{2}(\Sigma_nJ + (\Sigma_nJ)^{-1})$ equals $\text{sign}(\Sigma_nJ)$ (see e.g. [15]), Σ_∞ may be written as $(\text{sign}(\Sigma_nJ))J^{-1}$. Concluding, a solution Σ satisfying (3), (4) and (5) may be computed as follows. Take any solution Σ_1 of (3) and (4) (this is a standard problem, see [5]). Compute $\Sigma_\infty := (\text{sign}(\Sigma_1J))J^{-1}$. Then Σ_∞ satisfies (3), (4) and (5).

Remark 9. The covariance Q in Step 5, uniquely determined by $QA^T + AQ = -P$, can also be computed directly from the Riccati equation of Step 4. In fact:

LEMMA 9. *Let*

$$\bar{A} = \begin{pmatrix} F & -P \\ -R & -F^T \end{pmatrix}, \quad \bar{B} = \begin{pmatrix} 0 \\ H^T \end{pmatrix}, \quad C = (H \quad 0)$$

be a minimal realization of a spectral density $\Phi(s)$, with $P \geq 0$, $R \geq 0$ (such a realization exists by Theorem 5). From Theorem 7 it follows that $(\bar{A}, \bar{B}, \bar{C})$ determines a minimal stochastic realization denoted by (A, Q, C) . Denote by K^+ and K^- respectively the maximal and minimal symmetric solutions of the Riccati equation associated to \bar{A} (11):

$$F^TK + KF + KPF - R = 0.$$

Then $Q = (K^+ - K^-)^{-1}$ (under the conditions of Theorem 7, $K^+ - K^-$ is necessarily > 0 , cf. [13]).

Proof. After the basis transformation $\begin{pmatrix} I & 0 \\ -K^+ & I \end{pmatrix}$, \bar{A} is given by

$$\begin{pmatrix} F^+ & -P \\ 0 & -(F^+)^T \end{pmatrix} \quad \text{with } F^+ = F - PK^+.$$

Let $\begin{pmatrix} -Q \\ I \end{pmatrix} x$ be a vector in $\text{Im} \begin{pmatrix} -Q \\ I \end{pmatrix}$. Then, since $F^+Q + QF^+ = -P$ (see (13)):

$$\begin{pmatrix} F^+ & -P \\ 0 & -(F^+)^T \end{pmatrix} \begin{pmatrix} -Q \\ I \end{pmatrix} x = \begin{pmatrix} -F^+Q - P \\ -(F^+)^T \end{pmatrix} x = - \begin{pmatrix} -Q \\ I \end{pmatrix} (F^+)^T x.$$

Therefore, in this new basis, $\text{Im} \begin{pmatrix} -Q \\ I \end{pmatrix}$ equals the positive eigenspace of \bar{A} since F^+ is asymptotically stable. Hence in the original basis the positive eigenspace of \bar{A} is given by

$$\begin{pmatrix} I & 0 \\ K^+ & I \end{pmatrix} \text{Im} \begin{pmatrix} -Q \\ I \end{pmatrix} = \text{Im} \begin{pmatrix} -Q \\ -K^+Q + I \end{pmatrix}.$$

On the other hand we know that the positive eigenspace of \bar{A} is also given by $\text{Im} \begin{pmatrix} I \\ K^- \end{pmatrix}$. Therefore

$$\text{Im} \begin{pmatrix} -Q \\ -K^+Q + I \end{pmatrix} = \text{Im} \begin{pmatrix} I \\ K^- \end{pmatrix}$$

or $(K^+ - K^-)Q = I$. \square

Remark 10. After having fixed a Σ satisfying (3), (4) and (5), the stochastic realization (A, Q, C) obtained in our algorithm is uniquely determined (up to basis transformations). The freedom in the choice of the covariance Q as appearing in Theorem 2 (or alternatively the freedom in the choice of a spectral factorization of

the spectral density, cf. [2] and Remark 7) is therefore equivalent to the freedom in the choice of a Σ satisfying (3), (4) and (5).

Remark 11. Consider a stochastic vector process y with autocorrelation function R which admits a minimal white noise representation $dx = Ax dt + B dw$, $\tilde{y} = Cx$, as explained in Remark 1. Then it is easy to see that the following equivalences hold: $\{y \text{ is ergodic}\} \Leftrightarrow \{\lim_{|t| \rightarrow \infty} R(t) = 0\} \Leftrightarrow \{(A, B) \text{ is controllable}\} \Leftrightarrow \{\operatorname{Re} \sigma(A) < 0\}$.

As we have seen in Theorem 7, if we start from a rational spectral density $\Phi(s)$ we always arrive at a controllable, asymptotically stable white noise representation. Therefore a rational spectral density always corresponds to an ergodic process. On the other hand to give a stochastic realization of an autocorrelation function it is not necessary to assume that the process is ergodic. In some sense this is disadvantage of our approach. However the theory of § 3 may be extended to cover the nonergodic case as well. It can be easily seen that the following is true. Let (F, G, H) be a realization of R , i.e. $R^+(t) = He^{Ft}G$. Define the Hamiltonian system $(\bar{A}, \bar{B}, \bar{C})$ by

$$(18) \quad \bar{A} = \begin{pmatrix} F & 0 \\ 0 & -F^T \end{pmatrix}, \quad \bar{B} = \begin{pmatrix} G \\ H^T \end{pmatrix}, \quad \bar{C} = (H \quad -G^T).$$

Then:

$\{R \text{ is the autocorrelation function of an ergodic process } y\} \Leftrightarrow \{(\bar{A}, \bar{B}, \bar{C}) \text{ is a minimal triple}\}$. If the process is ergodic (or $(\bar{A}, \bar{B}, \bar{C})$ is minimal) then it has spectral density $\Phi(s) = \bar{C}(Is - \bar{A})^{-1}\bar{B}$.

Finally we note that the spectral density of a nonergodic process is in a certain sense irrational. It is in fact of the form $\sum_{i=1}^k \pi(\delta(\omega - \omega_i) + \delta(\omega + \omega_i))$, with $\{j\omega_i\}$ the spectrum of the generator of the nonergodic part (cf. [10]).

4. Conclusions. In this paper we have given a procedure for the stochastic realization of a spectral density matrix which starts by treating the spectral density as a (fictitious) transfer function. As such this transfer function has a symplectic structure and is passive. This structure is then exploited to arrive at a stochastic realization.

In comparison with the usual approach our method avoids having to factor the spectral density additively into its analytic and co-analytic part. However which of the two approaches would algorithmically be the most advantageous remains a matter of study.

REFERENCES

- [1] R. ABRAHAM AND J. E. MARSDEN, *Foundations of Mechanics*, Benjamin/Cummings, New York, 1978.
- [2] B. D. O. ANDERSON, *The inverse problem of stationary covariance generation*, J. Statist. Phys., 1 (1969), pp. 133–147.
- [2] R. W. BROCKETT AND A. RAHIMI, *Lie algebras and linear differential equations*, in Ordinary Differential Equations, L. Weiss, ed., Academic Press, New York, 1972, pp. 379–386.
- [4] P. FAURRE, *Réalisations Markoviennes de processus stationnaires*, Report IRIA, No. 13, 1973.
- [5] P. FAURRE, M. CLERGET AND F. GERMAIN, *Opérateurs rationnels positifs*, Dunod, Paris, 1979.
- [6] R. E. KALMAN, *Linear stochastic filtering theory—Reappraisal and outlook*, Proc. Brooklyn Polytechnic Symposium on System Theory, New York, 1965, pp. 197–205.
- [7] V. KUCERA, *A review of the matrix Riccati equation*, Kybernetika, 9 (1973), pp. 42–61.
- [8] A. LINDQUIST AND G. PICCI, *On the stochastic realization problem*, this Journal, 17 (1979), pp. 365–390.
- [9] K. R. MEYER, *Hamiltonian systems with a discrete symmetry*, J. Differential Equations, 41 (1981), pp. 228–238.
- [10] A. PAPOULIS, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 1965.

- [11] G. RUCKEBUSH, *Représentations markoviennes de processus gaussiens stationnaires*, Thesis, Univ de Paris VI, 1975.
- [12] A. J. VAN DER SCHAFT, *Time-reversible Hamiltonian systems*, Systems & Control Letters, 1 (1982), pp. 295–300.
- [13] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 621–634.
- [14] ———, *Realization of systems with internal passivity and symmetry constraints*, J. Franklin Institute, 301 (1967), pp. 605–621.
- [15] ———, *Dissipative dynamical systems, Part II: Linear systems with quadratic supply rates*, Arch. Rat. Mech. Anal., 45 (1972), pp. 352–393.
- [16] A. M. YAGLOM, *An Introduction to the Theory of Stationary Random Functions*, Dover, New York, 1973.

CONNECTIONS BETWEEN OPTIMAL STOPPING AND SINGULAR STOCHASTIC CONTROL I. MONOTONE FOLLOWER PROBLEMS*

IOANNIS KARATZAS† AND STEVEN E. SHREVE‡

Abstract. The stochastic control problem of tracking a Brownian motion by a nondecreasing process (Monotone Follower) is related to a question of Optimal Stopping. Direct probabilistic arguments are employed to show that the two problems are equivalent, and that both admit optimal solutions.

Key words. Brownian motion, optimal stopping, stochastic control, uniform integrability, weak L^1 -convergence

1. Introduction. This article establishes an equivalence between a problem of stochastic control and a problem of optimal stopping for Brownian motion. The control problem has the state process

$$X_t = x + W_t - \xi_t, \quad 0 \leq t \leq \tau,$$

where $W = \{W_t; t \geq 0\}$ is a standard Brownian motion and $\xi = \{\xi_t; t \geq 0\}$ is an adapted, nondecreasing and left-continuous process with $\xi_0 = 0$. It is desired to choose ξ so as to minimize the expected cost

$$E \left[\int_0^\tau h(t, X_t) dt + \int_{[0, \tau)} f(t) d\xi_t + g(X_\tau) \right]$$

where $h(t, \cdot)$ and $g(\cdot)$ are convex functions. We call this problem the *Monotone Follower Stochastic Control Problem* and denote its value function by $V(\tau, x)$. Special cases of this problem have been studied in [4], [13].

The optimal processes in these works can be characterized by two regions in (t, x) space: an open region of inaction and its complement, the region of action. If the initial time-state pair is in the latter region, the *optimal control process* causes the state to jump immediately to the closest point on the boundary demarcating the two regions; it acts thereafter only when the state is on the boundary, and pushes only enough to prevent a crossing of this boundary into the interior of the region of action. Thus, it acts like the *local time* of the (optimally controlled) state process at the boundary. Control problems whose optimal processes exhibit this behaviour have been called “singular” and have been studied, first by Bather and Chernoff [3], and recently by Beneš, Shepp and Witsenhausen [4], Karatzas [13], [14], Shreve, Lehoczky and Gaver [23], Harrison and Taylor [12], Harrison and Taksar [11].

Using the data of the Monotone Follower Problem, we pose a question of *Optimal Stopping*: to find a stopping time $\sigma \leq \tau$ which minimizes the risk

$$E \left[\int_0^\sigma h_x(t, x + W_t) dt + f(\sigma) 1_{\{\sigma < \tau\}} + g'(x + W_\tau) 1_{\{\sigma = \tau\}} \right].$$

We denote by $u(\tau, x)$ the optimal risk for this problem.

* Received by the editors March 17, 1983, and in revised form September 1, 1983.

† Department of Mathematical Statistics, Columbia University, New York, New York 10027. The research of this author was supported in part by the National Science Foundation under grant NSF MCS-81-03435-A01.

‡ Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213. The research of this author was supported by the U.S. Air Force under grant AFOSR 82-0259.

Questions of optimal stopping arise in many sequential decision problems and have been studied extensively (a very partial list of references includes [1], [2], [5], [8]–[10], [18], [21], [22], [24]). The optimal stopping rule can usually be described by two regions in the (t, x) plane: an (open) optimal continuation set and its complement, the optimal stopping set. Determination of these regions leads to very interesting analytical (variational inequality and free boundary) problems.

Bather and Chernoff [3] were the first to notice the connection between Singular Stochastic Control and Optimal Stopping. These authors posed a specific control problem, introduced a related stopping problem, and argued on heuristic grounds that the optimal risk of the latter ought to be the gradient of the value function of the former: $u = V_x$. Furthermore, the optimal continuation region in the stopping problem ought to be the region of inaction in the control problem. Additional special cases of the relationship between singular control and optimal stopping were developed (rigorously) by Karatzas [14], albeit again in a mostly analytical way.

In the present article we show by purely probabilistic arguments that, under continuity and growth conditions on h , f and g , the Monotone Follower and Optimal Stopping problems are equivalent in the sense described above. Again by probabilistic means, we establish existence of an optimal process in the control problem. The above-mentioned equivalence then allows us to deduce the existence of an optimal stopping time. General results concerning the existence of optimal stopping rules have been established heretofore mainly by analytical methods (Van Moerbeke [24], Friedman [8], [9], Grigelionis and Shiryaev [10]).

One benefit of the equivalence between stochastic control and optimal stopping problems is that bounds on the continuation region of the stopping problem, obtained by posing and “solving” more favourable and less favourable problems, translate into bounds on the region of inaction in the control problem. This comparison idea goes back to Bather [1] and was carried out in the realm of singular stochastic control by Bather and Chernoff [3], [5] and Karatzas [14].

A second benefit of the equivalence between control and stopping problems is that it sheds light on the heuristic “principle of smooth fit” advanced in [4], which suggests that $V_{xx}(\tau, x)$ should be continuous across the boundary that demarcates the regions of action and inaction. Although no formal justification of this principle seems to have been offered, it has been observed to hold in numerous examples and to play a fundamental role in determining the value function $V(\tau, x)$. Our results suggest that this principle is a manifestation of the better understood fact that the optimal risk $u(\tau, x)$ for any “reasonable” optimal stopping problem has a continuous gradient; see Grigelionis and Shiryaev [10], Bather [2], Van Moerbeke [24], Friedman [8], [9] and Shiryaev [22] for proofs of this result under various conditions.

A third advantage concerns the fact that control policies are more easily topologized than stopping times, and thus more amenable to the continuity and compactness arguments frequently used in existence proofs. We exploit this fact in our proof of existence of an optimal control process, which leads immediately to the existence of an optimal stopping time.

In the sequel [15] to this paper we establish by similar methods the equivalence between Reflected Follower Problems and Optimal Stopping for a Brownian motion with absorption at the origin, a much more delicate matter than the one treated here. The original pair of problems studied by Bather and Chernoff [3] was of this type.

2. Summary. Section 3 establishes the relation $u = V_x$ under the assumption that the control problem admits an optimal process (Theorem 3.4). A recurrent, constructive

element in our approach, which appears in different guises both here and in the sequel [15], is the principle of tracking, at some distance, an optimal or nearly optimal path up to a certain stopping time, and then jumping on it. This technique enables us to make very effective comparisons of expected costs at nearby points in the control problem, and is thus essential in computing $V_x(\tau, x)$.

We broach the question of existence of optimal processes for the Monotone Follower Problem in § 4, under stronger conditions on the cost functions (Theorem 4.1). These conditions enable us to establish the uniform integrability of a minimizing sequence of processes (Proposition 4.2), and to construct an optimal process by a weak compactness argument.

Instrumental in this approach is a sufficient condition for uniform integrability, as well as a Fatou lemma for weak L^1 -convergence. Both are of a certain independent interest, and appear in § 5 (Appendix). A second Appendix (§ 6) deals with the nature of the optimal control process and the optimal stopping rule, under even stronger conditions on the cost functions; it is independent of the rest of the work, and is included for the sake of completeness.

3. The Monotone Follower Problem. Let us consider a probability space $(\Omega, \mathcal{F}, P; \mathcal{F}_t)$ where the family of σ -fields $\{\mathcal{F}_t; 0 \leq t \leq \tau\}$ is increasing, with $\mathcal{F} = \mathcal{F}_\tau$, and satisfies the “usual conditions”: right-continuity and completion by P -negligible sets. We suppose that this space is rich enough to accommodate a Brownian motion process $W = \{W_t; 0 \leq t \leq \tau\}$, and we consider the class \mathcal{A} of *admissible control processes*, consisting of all $\{\mathcal{F}_t\}$ -adapted, nondecreasing and left-continuous processes $\xi = \{\xi_t; 0 \leq t \leq \tau\}$ with $\xi_0 = 0$, a.s. P . Corresponding to such a process ξ , and to an initial position $x \in \mathbb{R}$, the state process is

$$(3.1) \quad X_t = x + W_t - \xi_t, \quad 0 \leq t \leq \tau.$$

We formulate now an optimal control problem. The ingredients are:

- (3.2) (i) a real-valued, continuous function $f(t)$ on $[0, \tau]$, representing the *running cost of controlling effort per unit time*;
- (3.2) (ii) a real-valued, continuous and continuously differentiable function $g(x)$ on \mathbb{R} such that $g'(x)$ is nondecreasing, representing a *terminal cost on the state*;
- (3.2) (iii) a function $h(t, x): [0, \tau] \times \mathbb{R} \rightarrow \mathbb{R}$ which is continuous on its domain, with gradient $h_x(t, x)$ continuous on $[0, \tau] \times \mathbb{R}$ and nondecreasing in the space variable x , representing a *running cost per unit time on the state*.

It will be assumed that the functions $g'(x)$ and $h_x(t, x)$ satisfy a polynomial growth condition in the space variable: for some $m \geq 1$, $K > 0$,

$$(3.2) \quad (iv) \quad |h_x(t, x)| + |g'(x)| \leq K(1 + |x|^m) \quad \text{on } [0, \tau] \times \mathbb{R}.$$

The *Monotone Follower Stochastic Control Problem* is to choose a control process $\xi \in \mathcal{A}$ in such a way as to minimize the expected total cost

$$(3.3) \quad V(\tau, x) = \inf_{\xi} E \left[\int_0^\tau h(t, X_t) dt + \int_{[0, \tau)} f(t) d\xi_t + g(X_\tau) \right].$$

Similarly, we can formulate a *Problem of Optimal Stopping* for the Brownian motion W , where we impose a terminal cost $g'(x)$ and a continuation cost $h_x(t, x)$ per unit time on the state, as well as a cost $f(t)$ for premature termination. The problem

is then to select an $\{\mathcal{F}_t\}$ -stopping time $\sigma: \{\sigma \leq t\} \in \mathcal{F}_t, \forall 0 \leq t \leq \tau$ and $P(0 \leq \sigma \leq \tau) = 1$, so as to minimize the risk

$$(3.4) \quad u(\tau, x) = \inf_{0 \leq \sigma \leq \tau} E \left[\int_0^\sigma h_x(t, x + W_t) dt + f(\sigma) 1_{\{\sigma < \tau\}} + g'(x + W_\tau) 1_{\{\sigma = \tau\}} \right].$$

It will be shown that these two problems are intimately connected. The basic result of this section (Theorem 3.4 below) asserts that, if there exists an optimal process for the Control Problem, then the relationship $u = V_x$ holds, and the stopping time σ^* given by (3.18) is optimal for the Stopping Problem. Conditions which guarantee the existence of an optimal process are imposed in § 4.

The equivalence $u = V_x$ has been presaged by the works of Bather and Chernoff [3] and Karatzas [14]; it was used there in the opposite direction, to construct the optimal process for the Bounded Variation Follower from the solution of an Optimal Stopping Problem with Absorption. The equivalence between these two problems is taken up in the sequel [15] to this paper.

Throughout this work, we shall use the following notation for the four derivatives of the function $V(\tau, \cdot)$ at x :

$$\Delta^\pm V(\tau, x) \triangleq \lim_{\delta \rightarrow 0^\pm} \frac{V(\tau, x + \delta) - V(\tau, x)}{\delta},$$

$$\Delta_\pm V(\tau, x) \triangleq \lim_{\delta \rightarrow 0^\pm} \frac{V(\tau, x + \delta) - V(\tau, x)}{\delta}.$$

PROPOSITION 3.1. $\Delta^+ V(\tau, x) \leq u(\tau, x)$.

Proof. For each $\varepsilon > 0$, there exists a control process $\xi^\varepsilon = \{\xi_t^\varepsilon; 0 \leq t \leq \tau\}$ in \mathcal{A} , such that, with $X_t^\varepsilon = x + W_t - \xi_t^\varepsilon, 0 \leq t \leq \tau$,

$$(3.5) \quad V(\tau, x) + \varepsilon \geq E \left[\int_0^\tau h(t, X_t^\varepsilon) dt + \int_{[0, \tau)} f(t) d\xi_t^\varepsilon + g(X_\tau^\varepsilon) \right].$$

Let us fix a stopping time σ with $P(0 \leq \sigma \leq \tau) = 1$; for each $\delta > 0$, we construct the process

$$Y_{t,\delta}^\varepsilon = \begin{cases} X_t^\varepsilon + \delta, & 0 \leq t \leq \sigma, \\ X_t^\varepsilon, & \sigma < t \leq \tau, \end{cases}$$

according to the principle of “tracking the X^ε path at a distance $\delta > 0$ apart, and then jumping on it at time σ ”. The new process can be expressed in the form (3.1) as: $Y_{t,\delta}^\varepsilon = (x + \delta) + W_t - \eta_{t,\delta}^\varepsilon, 0 \leq t \leq \tau$, with

$$\eta_{t,\delta}^\varepsilon = \begin{cases} \xi_t^\varepsilon, & 0 \leq t \leq \sigma, \\ \delta + \xi_t^\varepsilon, & \sigma < t \leq \tau, \end{cases}$$

a control process in \mathcal{A} : left-continuous, nondecreasing, null at zero. The performance of this process is certainly suboptimal for the control problem at $(\tau, x + \delta)$, so

$$(3.6) \quad V(\tau, x + \delta) \leq E \left[\int_0^\sigma h(t, X_t^\varepsilon + \delta) dt + \int_\sigma^\tau h(t, X_t^\varepsilon) dt + \int_{[0, \tau)} f(t) d\xi_t^\varepsilon + \delta f(\sigma) 1_{\{\sigma < \tau\}} + g(X_\tau^\varepsilon) 1_{\{\sigma < \tau\}} + g(X_\tau^\varepsilon + \delta) 1_{\{\sigma = \tau\}} \right].$$

Combining (3.5) and (3.6), and using the fact that $X_t^\varepsilon + \delta \leq x + \delta + W_t$, for all $t \geq 0$,

along with the properties of h and g , we get

$$(3.7) \quad \frac{V(\tau, x + \delta) - V(\tau, x)}{\delta} \leq E \left[\int_0^\sigma h_x(t, x + \delta + W_t) dt + f(\sigma) 1_{\{\sigma < \tau\}} + g'(x + \delta + W_\tau) 1_{\{\sigma = \tau\}} \right] + \frac{\varepsilon}{\delta}.$$

One can now pass successively to the limit in (3.7), as $\varepsilon \downarrow 0$ and $\delta \downarrow 0$. The polynomial growth of $h_x(t, \cdot)$ and $g'(\cdot)$, together with dominated convergence, guarantee that

$$(3.8) \quad \Delta^+ V(\tau, x) \leq E \left[\int_0^\sigma h_x(t, x + W_t) dt + f(\sigma) 1_{\{\sigma < \tau\}} + g'(x + W_\tau) 1_{\{\sigma = \tau\}} \right],$$

for any stopping time σ , $P(0 \leq \sigma \leq \tau) = 1$. The result follows by taking the infimum of the right-hand side in (3.8) over such stopping times. \square

PROPOSITION 3.2. *Suppose there exists an optimal process $\xi^* = \xi^*(\tau, x)$ for the control problem at (τ, x) . Then:*

$$\Delta_- V(\tau, x) \geq u(\tau, x).$$

Proof. Let the optimal state process be: $X_t = x + W_t - \xi_t$; $0 \leq t \leq \tau$, where $\xi = \xi^*$ is the optimal control process whose existence is being assumed (we omit the stars);

$$(3.9) \quad V(\tau, x) = E \left[\int_0^\tau h(t, X_t) dt + \int_{[0, \tau)} f(t) d\xi_t + g(X_\tau) \right].$$

For $\delta > 0$, we consider the random variables

$$\sigma = \inf \{0 \leq t \leq \tau; \xi_t > 0\},$$

$$\sigma^\delta = \inf \{0 \leq t \leq \tau; \xi_t \geq \delta\} = \inf \{0 \leq t \leq \tau; x - \delta + W_t \geq X_t\},$$

with the convention $\inf \emptyset = \tau$. We observe that $\{\sigma^\delta \leq t\} = \{\xi(t+) \geq \delta\} \in \mathcal{F}_{t+} = \mathcal{F}_t$, so the σ^δ 's are indeed stopping times for the family of fields $\{\mathcal{F}_t; 0 \leq t \leq \tau\}$. On the other hand, it is clear that:

$$(3.10) \quad \begin{aligned} \sigma^\delta &= \sigma && \text{for all } 0 < \delta \leq \xi(\sigma+), && \text{on the event } \{\xi(\sigma+) > 0\}, \\ \sigma^\delta &\downarrow \sigma && \text{as } \delta \downarrow 0, && \text{on the event } \{\xi(\sigma+) = 0\}. \end{aligned}$$

Consequently: $\sigma^\delta \downarrow \sigma$ as $\delta \downarrow 0$, a.s. P . Now we construct the new state process

$$(3.11) \quad \begin{aligned} Y_t^\delta &= x - \delta + W_t, && 0 \leq t \leq \sigma^\delta, \\ &= X_t, && \sigma^\delta < t \leq \tau, \end{aligned}$$

according to the principle: "follow the Brownian path starting at $x - \delta$ until an amount δ of controlling effort has already been spent on the neighbouring optimal trajectory for (τ, x) , and then switch to the X path." Putting Y^δ in the standard form (3.1):

$$Y_t^\delta = x - \delta + W_t - \eta_t^\delta, \quad 0 \leq t \leq \tau$$

with

$$\begin{aligned} \eta_t^\delta &= 0, && 0 \leq t \leq \sigma^\delta, \\ &= \xi_t - \delta, && \sigma^\delta < t \leq \tau, \end{aligned}$$

we observe that the switching of paths might involve a leftward jump of size $\xi(\sigma^\delta+) - \delta$, which is a nonnegative number.

Clearly, the performance of the new process is suboptimal for $(\tau, x - \delta)$, so

$$(3.12) \quad V(\tau, x - \delta) \leq E \left[\int_0^{\sigma^\delta} h(t, x - \delta + W_t) dt + \int_{\sigma^\delta}^\tau h(t, X_t) dt \right. \\ \left. + \int_{[0, \tau)} f(t) d\eta_t^\delta + g(X_\tau) 1_{\{\sigma^\delta < \tau\}} + g(x - \delta + W_\tau) 1_{\{\sigma^\delta = \tau\}} \right].$$

Combining (3.9) with (3.12), and using the fact that

$$\int_{[0, \tau)} f(t) d\xi_t - \int_{[0, \tau)} f(t) d\eta_t^\delta = \int_{[0, \sigma^\delta)} f(t) d\xi_t + f(\sigma^\delta) [\delta - \xi(\sigma^\delta)] 1_{\{\sigma^\delta < \tau\}},$$

we obtain

$$(3.13) \quad V(\tau, x) - V(\tau, x - \delta) \geq E \left[\int_0^{\sigma^\delta} \{h(t, X_t) - h(t, x - \delta + W_t)\} dt \right. \\ \left. + \delta f(\sigma^\delta) 1_{\{\sigma^\delta < \tau\}} + \left\{ \min_{0 \leq t \leq \sigma^\delta} f(t) - f(\sigma^\delta) \right\} \xi(\sigma^\delta) 1_{\{\sigma^\delta < \tau\}} \right. \\ \left. + \{g(X_\tau) - g(x - \delta + W_\tau)\} 1_{\{\sigma^\delta = \tau\}} \right].$$

We proceed to obtain lower bounds for the two quantities in braces on the right-hand side of (3.13). On $\{0 \leq t \leq \sigma^\delta\}$, one has $0 \leq \xi_t \leq \delta$. Coupled with the fact that $h_x(t, \cdot)$, $g'(\cdot)$ are increasing, this yields

$$h(t, x + W_t - \xi_t) - h(t, x - \delta + W_t) \geq (\delta - \xi_t) h_x(t, x - \delta + W_t) \quad \text{on } \{0 \leq t \leq \sigma^\delta\}$$

and

$$g(x + W_\tau - \xi_\tau) - g(x - \delta + W_\tau) \geq (\delta - \xi_\tau) g'(x - \delta + W_\tau) \quad \text{on } \{\sigma^\delta = \tau\}.$$

Substituting these estimates in (3.13) and rearranging terms, we obtain

$$(3.14) \quad \frac{V(\tau, x) - V(\tau, x - \delta)}{\delta} \\ \geq E \left[\int_0^\sigma h_x(t, x + W_t) dt + f(\sigma) 1_{\{\sigma < \tau\}} + g'(x + W_\tau) 1_{\{\sigma = \tau\}} \right] + \sum_{j=1}^6 EI_j(\delta),$$

where

$$I_1(\delta) \equiv \int_0^\sigma \{h_x(t, x - \delta + W_t) - h_x(t, x + W_t)\} dt,$$

$$I_2(\delta) \equiv \int_\sigma^{\sigma^\delta} h_x(t, x - \delta + W_t) \left(1 - \frac{\xi_t}{\delta}\right) dt,$$

$$I_3(\delta) \equiv [g'(x - \delta + W_\tau) - g'(x + W_\tau)] 1_{\{\sigma = \tau\}},$$

$$I_4(\delta) \equiv g'(x - \delta + W_\tau) \left[1 - \frac{\xi(\sigma^\delta)}{\delta}\right] 1_{\{\sigma < \sigma^\delta = \tau\}},$$

$$I_5(\delta) \equiv f(\sigma^\delta) 1_{\{\sigma^\delta < \tau\}} - f(\sigma) 1_{\{\sigma < \tau\}},$$

and

$$I_6(\delta) \equiv \left\{ \min_{\sigma \leq t \leq \sigma^\delta} f(t) - f(\sigma^\delta) \right\} \frac{\xi(\sigma^\delta)}{\delta} 1_{\{\sigma^\delta < \tau\}}.$$

The next step is to check that $E I_j(\delta) \rightarrow 0$, as $\delta \downarrow 0$, for each $1 \leq j \leq 6$. Verification is straightforward; it rests on the continuity and growth properties of the functions $h_x(t, x)$, $g'(x)$ and $f(t)$, on the dominated convergence theorem, as well as on the fact

$$\{\sigma < \sigma^\delta = \tau\} \downarrow \emptyset$$

as $\delta \downarrow 0$, which is a direct consequence of (3.10). The details are omitted.

Finally, a passage to the limit as $\delta \downarrow 0$ in (3.14) yields the desired result. \square

We shall need the following property of the value function $V(\tau, x)$.

LEMMA 3.3. *The function $V(\tau, \cdot): \mathbb{R} \rightarrow \mathbb{R}$ is convex.*

Proof. Let us introduce the notation

$$(3.15) \quad J(\tau, x; \xi) = E \left[\int_0^\tau h(t, x + W_t - \xi_t) dt + \int_{[0, \tau)} f(t) d\xi_t + g(x + W_\tau - \xi_\tau) \right].$$

For any $x_1, x_2 \in \mathbb{R}$, any processes $\xi_1, \xi_2 \in \mathcal{A}$ and $0 \leq \lambda \leq 1$, we have by virtue of convexity of the cost functions $h(t, \cdot)$, $g(\cdot)$:

$$V(\tau, x) \leq J(\tau, x; \xi) \leq \lambda J(\tau, x_1; \xi_1) + (1 - \lambda) J(\tau, x_2; \xi_2),$$

where $x = \lambda x_1 + (1 - \lambda)x_2$, $\xi = \lambda \xi_1 + (1 - \lambda)\xi_2 \in \mathcal{A}$. Taking the infimum of the right-hand side successively over $\xi_1 \in \mathcal{A}$, $\xi_2 \in \mathcal{A}$, we obtain

$$(3.16) \quad V(\tau, x) \leq \lambda V(\tau, x_1) + (1 - \lambda) V(\tau, x_2). \quad \square$$

THEOREM 3.4. *Under conditions (3.2) on the cost functions, let us suppose that there exists an optimal process $\xi^* = \xi^*(\tau, x)$ for the Monotone Follower Control problem at (τ, x) . Then the gradient of the value function $V(\tau, x)$ exists at (τ, x) , and*

$$(3.17) \quad V_x(\tau, x) = u(\tau, x).$$

Besides, the stopping time

$$(3.18) \quad \begin{aligned} \sigma^* &= \inf \{0 \leq t \leq \tau; \xi_t^* > 0\} \\ &= \tau \quad \text{if } \{\cdot \cdot \cdot\} = \emptyset \end{aligned}$$

is then optimal for the stopping problem.

Proof. The results of Propositions 3.1 and 3.2 imply a fortiori that

$$(3.19) \quad \begin{aligned} \Delta_+ V(\tau, x) &\leq \Delta^+ V(\tau, x) \leq u(\tau, x) \\ &\leq E \left[\int_0^{\sigma^*} h_x(t, x + W_t) dt + f(\sigma^*) 1_{\{\sigma^* < \tau\}} + g'(x + W_\tau) 1_{\{\sigma^* = \tau\}} \right] \\ &\leq \Delta_- V(\tau, x) \leq \Delta^- V(\tau, x). \end{aligned}$$

Now for any $x \in \mathbb{R}$ and positive numbers δ_1, δ_2 we have from (3.16) with $x_1 = x - \delta_1$, $x_2 = x + \delta_2$ and $\lambda = \delta_2 / (\delta_1 + \delta_2)$:

$$\frac{V(\tau, x) - V(\tau, x - \delta_1)}{\delta_1} \leq \frac{V(\tau, x + \delta_2) - V(\tau, x)}{\delta_2},$$

whence

$$(3.20) \quad \Delta^- V(\tau, x) \leq \Delta_+ V(\tau, x).$$

The two inequalities (3.19), (3.20) establish the existence of $V_x(\tau, x)$, the fact that it is equal to the optimal risk $u(\tau, x)$ for the stopping problem, and the optimality of the time σ^* .

COROLLARY 3.5. *If there is no optimal stopping time for the stopping problem, then there is no optimal process in \mathcal{A} for the control problem.*

It is well known in the theory of Optimal Stopping (cf. Grigelionis and Shiryaev [10], Bather [2], Van Moerbeke [24], Friedman [8], [9]) that the gradient $u_x(\tau, x)$ of the optimal risk is continuous on $\mathbb{R}^+ \times \mathbb{R}$, under appropriate conditions on the cost functions $h_x(t, x)$, $f(t)$ and $g'(x)$ (see Remark at the end of this section). This is called the “principle of smooth fit” for the Optimal Stopping Problem; in view of relation (3.17), it implies a “principle of smooth fit” for the Monotone Follower Stochastic Control Problem, namely continuity of the second derivative $V_{xx}(\tau, x)$ on $\mathbb{R}^+ \times \mathbb{R}$. Such a principle was advanced on heuristic grounds by Beneš, Shepp and Witsenhausen in [4], and was employed in solving explicitly the Monotone Follower Problem with $h(t, x) = x^2$, $f(t) = g(x) \equiv 0$.

These authors showed that the optimal process for their problem is of the form

$$\xi_t^* = \max [0, \max_{0 \leq u \leq t} \{x + W_u - \delta(\tau - u)^{1/2}\}], \quad 0 < t \leq \tau$$

where δ is a positive constant characterized in terms of an integral equation. Theorem 3.4 above shows that the stopping time

$$\sigma^* = \inf \{0 \leq t \leq \tau; \xi_t^* > 0\} = \inf \{0 \leq t \leq \tau; x + W_t \geq \delta(\tau - t)^{1/2}\}$$

is optimal for the problem of minimizing the risk

$$E \int_0^\sigma (x + W_t) dt$$

over stopping times σ such that $P(0 \leq \sigma \leq \tau) = 1$. This question was posed by Shepp in [21] and was answered by Miroshnichenko in [18]. Beneš et al. [4] noted the similarity of their solution to that provided by Miroshnichenko. Theorem 3.4 and the existence of an optimal control process for the monotone follower problem of [4] show that these two problems are actually *equivalent*.

Remark. Sufficient conditions for the existence of an optimal stopping time and for the continuity of the gradient $u_x(\tau, x)$ of the optimal risk are derived by Friedman in [9]. In addition to (3.2), these would entail, in our context, that:

(3.21) the derivatives $h_{tx}(t, x)$ and $g'''(x)$ exist and satisfy a polynomial growth condition on $[0, \tau] \times \mathbb{R}$ and \mathbb{R} , respectively;

(3.22) $f(t)$ is nondecreasing on $[0, \tau]$, with $\sup_{x \in \mathbb{R}} g'(x) \leq f(0)$; and

(3.23) $\sup_{x \in \mathbb{R}} |g'(x)| \leq \text{const.} < \infty$.

Condition (3.22) is unnecessarily strong for existence; it should be compared with condition (4.2) of the next section.

4. Existence of optimal processes. The main result of § 3, Theorem 3.4, was based on the assumption that an optimal admissible process exists for the Monotone Follower Problem. One cannot expect this to be the case, however, in the general setting of § 3; for example, the control problem with $h(t, x) = 0$, $f(t) = 1$ and $g(x) = x^2$ admits no optimal process. Indeed, one can check rather easily (see also the discussion on condition (4.2) below) that the stopping problem with $h_x(t, x) = 0$, $f(t) = 1$ and $g'(x) = 2x$ has no optimal solution; this example was suggested by S. P. Lalley. By Corollary 3.5, no optimal process exists for the control problem.

In this section we shall deal with the question of existence of optimal processes under the following additional assumptions:

$$(4.1) \quad 0 < c \leq f(t) \leq C \quad \forall t \in [0, \tau] \quad \text{for some constants } c, C;$$

$$(4.2) \quad \sup_{x \in \mathbb{R}} g'(x) \leq f(\tau);$$

$$(4.3) \quad h(t, x) \geq 0, g(x) \geq 0 \quad \forall (t, x) \in [0, \tau] \times \mathbb{R}.$$

Condition (4.2) was violated in the above counter-example. To understand its function, let us assume that (4.1) and (4.3) hold, but (4.2) fails; then, both sets $S_0 = \{x \in \mathbb{R}; g'(x) < c\}$ and $S_1 = \{x \in \mathbb{R}; g'(x) > f(\tau)\}$ are nonempty. Indeed, if $S_0 = \emptyset$, then $g(0) - g(a) \geq -ca$ for $a < 0$, which means that for sufficiently negative a , $g(a)$ would be negative, contradicting (4.3). In the stopping problem it would then be profitable to terminate in S_0 (i.e., to take $\sigma = \tau$ if $x + W_\tau \in S_0$) and to stop before time τ rather than terminate in S_1 . However, stopping strictly prior to τ foregoes a chance of terminating in S_0 . Thus there can exist admissible stopping rules approaching optimality which mandate stopping at time $\tau - 1/n$ if $x + W_{\tau-1/n} \in S_1$, but their “limit” has $\sigma = \tau$ if $x + W_\tau \in S_1$, which is suboptimal. Under such circumstances, an optimal stopping rule can fail to exist. In the control problem, condition (4.2) ensures that the cost of a leftward jump near the final time is at least as big as the reduction in terminal cost expected as a consequence of the jump.

Condition (4.1) amounts to imposing a nontrivial penalty for the use of controlling effort. It is instrumental in establishing our uniform integrability result (Proposition 4.2 below); the rest of the argument does not depend on it.

In this section, we shall take as our sample space $\Omega = C[0, \tau]$ the set of continuous functions ω on $[0, \tau]$ with $\omega(0) = 0$, $W_t(\omega) \triangleq \omega(t)$ for all $0 \leq t \leq \tau$, and $P \equiv$ Wiener measure on $C[0, \tau]$. For each $0 \leq t \leq \tau$, we define $\mathcal{F}_t^0 = \sigma\{\omega(s); 0 \leq s \leq t\}$; \mathcal{F}_τ^0 is the smallest σ -field which makes all projections measurable, and coincides with the Borel σ -field generated by the sup-norm topology (Parthasarathy [19, p. 212, Thm. 2.1]). We shall denote by \mathcal{F}_t the completion $\bar{\mathcal{F}}_t^0$ of \mathcal{F}_t^0 by events of P -measure zero, for all $0 \leq t \leq \tau$, and put $\mathcal{F} \equiv \mathcal{F}_\tau$. It is known (cf. Liptser and Shiryaev [16, Thm. 4.3, p. 87]) that the family $\{\mathcal{F}_t; t \geq 0\}$ is right-continuous, and so the “usual conditions” are satisfied.

Our probability space is now $(\Omega, \mathcal{F}, P; \mathcal{F}_t)$ as above. We recall also the definition of the class \mathcal{A} of admissible control processes: a process $\xi = \{\xi_t(\omega); 0 \leq t \leq \tau, \omega \in \Omega\}$ is in \mathcal{A} , if $\xi_0 = 0$ a.s. P and

- (i) $\xi_t(\cdot)$ is an \mathcal{F}_t -measurable random variable for every $0 \leq t \leq \tau$ (i.e., ξ is $\{\mathcal{F}_t\}$ -adapted),
- (ii) for P -a.e. ω , $\xi_t(\omega)$ is nondecreasing and left-continuous.

It follows (cf. Dellacherie [6, Chap. IV]) that a process $\xi \in \mathcal{A}$ is progressively measurable, hence *jointly measurable*:

$$(4.4) \quad \xi \text{ is Borel}_{[0, \tau]} \otimes \mathcal{F}_\tau\text{-measurable.}$$

In this section we shall prove the following result.

THEOREM 4.1. *Under conditions (3.2) and (4.1)–(4.3) on the cost functions, there exists an optimal process for the Monotone Follower Control Problem.*

The proof of the theorem will be accomplished in a series of propositions and lemmata. In order to set the stage, let us consider a fixed pair $(\tau, x) \in \mathbb{R}^+ \times \mathbb{R}$ and select a minimizing sequence $\{\xi^n\}_{n=1}^\infty \subseteq \mathcal{A}$:

$$(4.5) \quad \varepsilon_n \triangleq J(\tau, x; \xi^n) - V(\tau, x) \downarrow 0 \quad \text{as } n \rightarrow \infty.$$

The definitions (3.15), (4.5) and condition (4.1) imply that

$$(4.6) \quad \sup_{n \geq 1} E \xi_\tau^n \triangleq M \leq \frac{V(\tau, x) + \varepsilon_1}{c} < \infty.$$

One can actually establish a stronger result as follows.

PROPOSITION 4.2. *The sequence $\{\xi_\tau^n\}_{n=1}^\infty$ is uniformly integrable.*

We introduce first the required notation. For each integer $n \geq 1$ and $\lambda > 0$, we define the stopping time

$$(4.7) \quad \begin{aligned} T_n(\lambda) &= \inf \{t \geq 0; \xi_t^n \geq \lambda\} \\ &= +\infty \quad \text{if } \{\cdot \cdot \cdot\} = \emptyset, \end{aligned}$$

as well as the process $\eta^n(\lambda) \in \mathcal{A}$

$$(4.8) \quad \begin{aligned} \eta_t^n(\lambda) &= \xi_t^n, \quad 0 \leq t \leq T_n(\lambda) \wedge \tau, \\ &= \xi_{T_n(\lambda)}^n, \quad T_n(\lambda) < t \leq \tau \quad \text{on } \{T_n(\lambda) < \tau\}, \end{aligned}$$

with associated expected cost

$$(4.9) \quad J(\tau, x; \eta^n(\lambda)) = V(\tau, x) + \varepsilon_n(\lambda),$$

where $\varepsilon_n(\lambda)$ is a nonnegative number.

LEMMA 4.3. *For every $n \geq 1$, $\lim_{\lambda \rightarrow \infty} \varepsilon_n(\lambda) = \varepsilon_n$.*

Proof. From (4.5), (4.8) and (4.9) we obtain

$$(4.10) \quad \varepsilon_n(\lambda) - \varepsilon_n = I_1(n, \lambda) - I_2(n, \lambda) + I_3(n, \lambda),$$

where

$$(4.11) \quad \begin{aligned} I_1(n, \lambda) &\triangleq E \int_{T_n(\lambda) \wedge \tau}^{\tau} [h(t, x + W_t - \xi_{T_n(\lambda)}^n) - h(t, x + W_t - \xi_t^n)] dt \\ &= \int \int_{[0, \tau] \times \Omega} [h(t, x + W_t - \xi_{T_n(\lambda)}^n) - h(t, x + W_t - \xi_t^n)] 1_{(T_n(\lambda) \wedge \tau, \tau)}(t) dP dt, \end{aligned}$$

$$(4.12) \quad I_2(n, \lambda) \triangleq E \int_{[T_n(\lambda) \wedge \tau, \tau)} f(t) d\xi_t^n = E \left[1_{\{\xi_\tau^n > \lambda\}} \int_{[T_n(\lambda), \tau)} f(t) d\xi_t^n \right],$$

$$(4.13) \quad I_3(n, \lambda) \triangleq E([g(x + W_\tau - \xi_{T_n(\lambda)}^n) - g(x + W_\tau - \xi_\tau^n)] 1_{\{T_n(\lambda) < \tau\}}).$$

As $\lambda \uparrow \infty$, we have $T_n(\lambda) \rightarrow \infty$, P -a.s., and the integrand in (4.11) converges to zero ($\text{meas} \times P$)-a.e., where “meas” means “Lebesgue measure”. The convexity and non-negativity of h imply

$$(4.14) \quad h(t, x + W_t - \xi_{T_n(\lambda)}^n) \leq h(t, x + W_t - \xi_t^n) + h(t, x + W_t), \quad t \geq T_n(\lambda),$$

and therefore the integrand in (4.11) is bounded in modulus by: $2h(t, x + W_t - \xi_t^n) + h(t, x + W_t)$, a $(\text{meas} \times P)$ -integrable function, since by (3.2)(iv)

$$E \int_0^\tau h(t, x + W_t) dt < \infty \quad \text{and} \quad E \int_0^\tau h(t, x + W_t - \xi_t^n) dt \leq V(\tau, x) + \varepsilon_1.$$

By dominated convergence, $\lim_{\lambda \rightarrow \infty} I_1(n, \lambda) = 0$, and a similar argument yields

$\lim_{\lambda \rightarrow \infty} I_3(n, \lambda) = 0$. As for $I_2(n, \lambda)$, we have the bounds

$$0 \leq \int_{[T_n(\lambda) \wedge \tau, \tau]} f(t) d\xi_t^n \leq \int_{[0, \tau]} f(t) d\xi_t^n \quad \text{a.s. } P,$$

$$E \int_{[0, \tau]} f(t) d\xi_t^n \leq V(\tau, x) + \varepsilon_1 \quad \text{and} \quad 0 \leq \int_{[T_n(\lambda) \wedge \tau, \tau]} f(t) d\xi_t^n \leq C[\xi_\tau^n - \xi_{T_n(\lambda) \wedge \tau}^n].$$

The latter converges to zero, P -a.s. as $\lambda \rightarrow \infty$, and again dominated convergence completes the proof that $\lim_{\lambda \rightarrow \infty} I_2(n, \lambda) = 0$. \square

Proof of Proposition 4.2. The quantity $I_2(n, \lambda)$ admits the lower bound

$$c \int_{\{\xi_\tau^n > \lambda\}} (\xi_\tau^n - \lambda) dP.$$

On the other hand, the Markov inequality gives, in conjunction with (4.1),

$$\lambda P(\xi_\tau^n \geq \lambda) \leq E\xi_\tau^n \leq \frac{1}{c} E \int_{[0, \tau]} f(t) d\xi_t^n \leq \frac{V(\tau, x) + \varepsilon_1}{c} \quad \forall n \geq 1, \quad \lambda > 0,$$

and we obtain (with the help of (4.14) and of the Cauchy inequality) the upper bound

$$I_1(n, \lambda) \leq E \left[1_{\{\xi_\tau^n > \lambda\}} \int_0^\tau h(t, x + W_t) dt \right]$$

$$\leq \left[\frac{V(\tau, x) + \varepsilon_1}{\lambda c} \right]^{1/2} \left(\tau \cdot E \int_0^\tau h^2(t, x + W_t) dt \right)^{1/2}.$$

A similar argument applied to $I_3(n, \lambda)$ gives

$$I_3(n, \lambda) \leq \left[\frac{V(\tau, x) + \varepsilon_1}{\lambda c} \right]^{1/2} (Eg^2(x + W_\tau))^{1/2}.$$

The identity (4.10) implies then

$$(4.10)' \quad c \int_{\{\xi_\tau^n > \lambda\}} (\xi_\tau^n - \lambda) dP \leq \varepsilon_n - \varepsilon_n(\lambda) + \frac{Q(\tau, x)}{\lambda^{1/2}} \quad \forall n \geq 1, \quad \lambda > 0,$$

where

$$Q(\tau, x) \triangleq \left[\frac{V(\tau, x) + \varepsilon_1}{c} \right]^{1/2} \left[(Eg^2(x + W_\tau))^{1/2} + \left(\tau E \int_0^\tau h^2(t, x + W_t) dt \right)^{1/2} \right].$$

Now we choose $\delta > 0$; since $\varepsilon_n \downarrow 0$, we can select an integer $N > 1$ such that for all $n \geq N$, we have $\varepsilon_n < c\delta/2$. By virtue of Lemma 4.3, we can also choose $\Lambda > 0$ in such a way that for all $\lambda \geq \Lambda$ we have

$$\varepsilon_n - \varepsilon_n(\lambda) \leq \frac{c}{2} \delta, \quad n = 1, 2, \dots, N-1$$

and

$$Q(\tau, x) \lambda^{-1/2} \leq \frac{c}{2} \delta.$$

It follows that, for all $\lambda \geq \Lambda$, $\sup_{n \geq 1} (\varepsilon_n - \varepsilon_n(\lambda)) \leq c\delta/2$ and

$$\sup_{n \geq 1} \int_{\{\xi_\tau^n > \lambda\}} (\xi_\tau^n - \lambda) dP < \delta.$$

Uniform integrability of the sequence of random variables $\{\xi_\tau^n\}_{n=1}^\infty$ is now a consequence of Lemma 5.1 in the Appendix. \square

Proposition 4.2 implies that the minimizing sequence of processes $\{\xi_t^n(\omega)\}_{n=1}^\infty \subseteq \mathcal{A}$ is (meas $\times P$)-uniformly integrable:

$$\sup_{n \geq 1} \iint_{\{(t, \omega); \xi_t^n(\omega) > \lambda\}} \xi_t^n(\omega) dt dP \leq \tau \cdot \sup_{n \geq 1} \int_{\{\omega; \xi_\tau^n(\omega) > \lambda\}} \xi_\tau^n(\omega) dP \xrightarrow{\lambda \rightarrow \infty} 0.$$

By the Dunford–Pettis compactness criterion (Dunford and Schwartz [7, p. 294], Meyer [17, p. 20]) there exists a jointly measurable and integrable process $\xi = \{\xi_t(\omega); 0 \leq t \leq \tau, \omega \in \Omega\}$ such that $\{\xi_t^n\}_{n=1}^\infty$ converges weakly to ξ , possibly along a (relabelled) subsequence, in the sense that

$$(4.15) \quad \int_0^\tau \int_\Omega \xi_t^n(\omega) \eta(t, \omega) dP dt \xrightarrow{n \rightarrow \infty} \int_0^\tau \int_\Omega \xi_t(\omega) \eta(t, \omega) dP dt$$

holds for any bounded, jointly measurable $\eta = \{\eta(t, \omega); 0 \leq t \leq \tau, \omega \in \Omega\}$.

DEFINITION 4.4. Let $\xi, \tilde{\xi}$ be two jointly measurable processes. The process $\tilde{\xi}$ is called a *modification* of ξ (and vice-versa), if

$$(4.16) \quad P[\omega \in \Omega] \text{ meas } \{0 \leq t \leq \tau; \xi_t(\omega) \neq \tilde{\xi}_t(\omega)\} = 0 = 1.$$

LEMMA 4.5. The process ξ in (4.15) admits a modification $\tilde{\xi}$ which is nondecreasing.

Proof. Consider the jointly measurable set

$$N = \left\{ (t, \omega) \in [0, \tau] \times \Omega; \lim_{k \rightarrow \infty} k \int_t^{t+1/k} \xi_s(\omega) ds \neq \xi_t(\omega) \right\}$$

and its sections $N(\omega) = \{0 \leq t \leq \tau; (t, \omega) \in N\}$. By the fundamental theorem of calculus, meas $N(\omega) = 0$ for P -a.e. $\omega \in \Omega$, so N has product measure zero. Let us choose two rational numbers $t_1 < t_2$ in $[0, \tau]$ and an integer $k \geq 1$, and define the event

$$A_{t_1, t_2, k} \triangleq \left\{ \omega \in \Omega; \int_{t_1}^{t_1+1/k} \xi_t(\omega) dt > \int_{t_2}^{t_2+1/k} \xi_t(\omega) dt \right\}.$$

Relation (4.15) with

$$\eta_i(t, \omega) = 1_{[t, t_1+1/k] \times A_{t_1, t_2, k}}(t, \omega), \quad i = 1, 2,$$

yields

$$\int_{A_{t_1, t_2, k}} \left(\int_{t_i}^{t_i+1/k} \xi_t^n(\omega) dt \right) dP \xrightarrow{n \rightarrow \infty} \int_{A_{t_1, t_2, k}} \left(\int_{t_i}^{t_i+1/k} \xi_t(\omega) dt \right) dP$$

for $i = 1, 2$. Subtracting the resulting two relations memberwise, and recalling that each $\xi_t^n(\omega)$ is P -a.s. nondecreasing, we obtain

$$\int_{A_{t_1, t_2, k}} \left(\int_{t_2}^{t_2+1/k} \xi_t(\omega) dt - \int_{t_1}^{t_1+1/k} \xi_t(\omega) dt \right) dP \geq 0,$$

whence $P(A_{t_1, t_2, k}) = 0$. Now we define the event

$$A \triangleq \bigcup_{\substack{0 \leq t_1 < t_2 \leq \tau \\ t_1, t_2 \text{ rational} \\ k \geq 1}} A_{t_1, t_2, k}$$

and the process

$$(4.17) \quad \begin{aligned} \tilde{\xi}_t(\omega) &= 0, & \omega \in A, \quad 0 \leq t \leq \tau, \\ &= \xi_t(\omega), & \omega \notin A, \quad t \in [0, \tau] \setminus N(\omega), \\ &= \lim_{s \uparrow t, s \notin N(\omega)} \xi_s(\omega), & \omega \notin A, \quad t \in N(\omega). \end{aligned}$$

For any $\omega \notin A$, any rational numbers $t_1 < t_2$ in $[0, \tau]$ and any integer $k \geq 1$, we have

$$(4.18) \quad k \int_{t_2}^{t_2+1/k} \xi_t(\omega) dt \geq k \int_{t_1}^{t_1+1/k} \xi_t(\omega) dt.$$

By continuity in the time argument of $\int_t^{t+1/k} \xi_s(\omega) ds$ for all $k \geq 1$, $\omega \in \Omega$, (4.18) can be extended to hold for all t_1, t_2 in $[0, \tau]$ such that $t_1 < t_2$, not necessarily rational. In particular, if we require that $t_1, t_2 \notin N(\omega)$, we obtain from (4.18) by a passage to the limit as $k \rightarrow \infty$: $\xi_{t_1}(\omega) \leq \xi_{t_2}(\omega)$. It follows that the path $\tilde{\xi}_t(\omega)$ is nondecreasing, for any $\omega \notin A$. It is easily checked from (4.17) that $\tilde{\xi}$ is a modification of ξ , i.e. that (4.16) holds, since obviously $P(A) = 0$. \square

Let us observe that the process ξ has been modified only on $\{[0, \tau] \times A\} \cup \{([0, \tau] \times A^c) \cap N\}$, a jointly measurable set of product measure zero, and that (4.15) holds if ξ is replaced by its modification $\tilde{\xi}$.

LEMMA 4.6. *The process $\tilde{\xi}$ admits the modification $\xi^* \in \mathcal{A}$.*

Proof. For every $\omega \in \Omega$, we define

$$(4.19) \quad \begin{aligned} \xi_t^*(\omega) &\triangleq \lim_{s \uparrow t} \tilde{\xi}_s(\omega), & 0 < t \leq \tau, \\ &0, & t = 0. \end{aligned}$$

It is easily verified that ξ^* is a left-continuous, nondecreasing process with $\xi_0^* = 0$. To prove that $\xi^* \in \mathcal{A}$, it thus remains to show that ξ^* is $\{\mathcal{F}_t\}$ -adapted.

Let us fix $t \in (0, \tau]$, and denote by $\xi^n|_{[0,t]}$, $\tilde{\xi}|_{[0,t]}$ the restrictions of ξ^n , $\tilde{\xi}$, respectively, to $[0, t] \times \Omega$. Because each ξ^n ; $n \geq 1$ is $\{\mathcal{F}_t\}$ -progressively measurable, $\xi^n|_{[0,t]}$ is $\text{Borel}_{[0,t]} \otimes \mathcal{F}_t$ -measurable, but $\tilde{\xi}|_{[0,t]}$ is only known to be $\text{Borel}_{[0,t]} \otimes \mathcal{F}_\tau$ -measurable; cf. (4.4).

On the other hand, it is not hard to see from (4.15)—with ξ replaced by $\tilde{\xi}$ —that the sequence $\{\xi^n|_{[0,t]}\}_{n=1}^\infty$ is weakly convergent to $\tilde{\xi}|_{[0,t]}$. Lemma 3.4, p. 72 in Liptser and Shiryaev [16] shows then that

$$\tilde{\xi}|_{[0,t]} \text{ is } \overline{\text{Borel}_{[0,t]} \otimes \mathcal{F}_t}\text{-measurable,}$$

and it follows from Royden [20, Chap. 12, § 4] that, for a.e. $s \in [0, t]$, $\tilde{\xi}_s$ is \mathcal{F}_t -measurable. Since we can let s approach t from below in such a way that $\tilde{\xi}_s$ is \mathcal{F}_t -measurable for every s , we see that $\xi_t^* = \lim_{s \uparrow t} \tilde{\xi}_s$ is \mathcal{F}_t -measurable. It follows that ξ^* is jointly measurable (cf. (4.4)). According to Lemma 5.4 in the Appendix, it is a modification of $\tilde{\xi}$. \square

So far we have chosen a minimizing sequence $\{\xi^n\}_{n=1}^\infty \subseteq \mathcal{A}$ and shown that there is a subsequence $\{\xi^{n_k}\}_{k=1}^\infty$ and a $\xi^* \in \mathcal{A}$ such that ξ^{n_k} converges weakly to ξ^* . For convenience of notation, we have relabelled the weakly convergent subsequence $\{\xi^n\}_{n=1}^\infty$, so the weak convergence condition becomes

$$(4.15)' \quad \int_0^\tau \int_\Omega \xi_t^n(\omega) \eta(t, \omega) dP dt \xrightarrow{n \uparrow \infty} \int_0^\tau \int_\Omega \xi_t^*(\omega) \eta(t, \omega) dP dt$$

for any bounded, jointly measurable $\eta = \{\eta(t, \omega); 0 \leq t \leq \tau, \omega \in \Omega\}$. The process ξ^* in (4.19) is our candidate optimal process for the control problem.

Note that the weak convergence in (4.15)' is for the processes regarded as L^1 functions on $[0, \tau] \times \Omega$. We do not know that for fixed t , the sequence of random variables $\{\xi_t^n\}_{n=1}^\infty$ converges weakly to the random variable ξ_t^* . The next lemma addresses this issue.

LEMMA 4.7. *For almost every $t \in [0, \tau]$, the sequence of random variables $\{\xi_t^n\}_{n=1}^\infty$ is uniformly integrable and converges weakly to the random variable ξ_t^* .*

Proof. Since $0 \leq \xi_t^n \leq \xi_\tau^n$ holds P -a.s. for every $t \in [0, \tau]$, Proposition 4.2 implies $\{\xi_t^n\}_{n=1}^\infty$ is uniformly integrable for every t .

To prove $\{\xi_t^n\}_{n=1}^\infty$ converges weakly to ξ_t^* , it suffices to show

$$(4.20) \quad \lim_{n \rightarrow \infty} E(\xi_t^n 1_A) = E(\xi_t^* 1_A) \quad \forall A \in \mathcal{F}_\tau$$

Toward that end, choose $A \in \mathcal{F}_\tau$ and define $\varphi_n(t) = E(\xi_t^n 1_A)$, $\varphi(t) = E(\xi_t^* 1_A)$. Since ξ_t^* is left-continuous, so is φ , and the set of continuity points of φ on $[0, \tau]$ has full measure. Let t be a continuity point of φ , and $h > 0$. Taking $\eta(s, \omega) = 1_{[t, t+h] \times A}(s, \omega)$ in (4.15)' and using the fact that φ_n and φ are nondecreasing, we obtain

$$\overline{\lim}_{n \rightarrow \infty} \varphi_n(t) \leq \frac{1}{h} \lim_{n \rightarrow \infty} \int_t^{t+h} \varphi_n(s) ds = \frac{1}{h} \int_t^{t+h} \varphi(s) ds \leq \varphi(t+h).$$

Letting $h \downarrow 0$, we see that $\overline{\lim}_{n \rightarrow \infty} \varphi_n(t) \leq \varphi(t)$.

Taking $\eta(s, \omega) = 1_{[t-h, t] \times A}(s, \omega)$, we obtain by a similar argument $\underline{\lim}_{n \rightarrow \infty} \varphi_n(t) \geq \varphi(t)$, and (4.20) is established. \square

The following Proposition concludes the proof of Theorem 4.1.

PROPOSITION 4.8. *The process ξ^* of Lemma 4.5 is optimal for the Monotone Follower Control Problem.*

Proof. Let G be the set of points t in $[0, \tau]$ for which $\{\xi_t^n\}_{n=1}^\infty$ is uniformly integrable and converges weakly to ξ_t^* . Note that 0 is in G , but τ may not be. According to Lemma 4.7, G has full measure. For $t \in G$, Lemma 5.2 in the Appendix implies

$$Eh(t, x + W_t - \xi_t^*) \leq \underline{\lim}_{n \rightarrow \infty} Eh(t, x + W_t - \xi_t^n),$$

and from Fubini and Fatou we obtain

$$(4.21) \quad E \int_0^\tau h(t, x + W_t - \xi_t^*) dt \leq \underline{\lim}_{n \rightarrow \infty} E \int_0^\tau h(t, x + W_t - \xi_t^n) dt.$$

On the other hand, given any $\varepsilon > 0$, one can find a step function $\varphi(t) = \sum_{i=0}^{m-1} c_i 1_{[a_i, a_{i+1})}(t)$ such that:

$$\begin{aligned} 0 &= a_0 < a_1 < \cdots < a_m = \tau, \\ a_i &\in G \quad \forall i = 0, 1, \dots, m-1, \\ c_i &\geq 0 \quad \forall i = 0, 1, \dots, m-1, \\ |f(t) - \varphi(t)| &\leq \varepsilon \quad \forall t \in [0, \tau]. \end{aligned}$$

From

$$\int_{[0, \tau)} f(t) d\zeta_t = \int_{[0, \tau)} [f(t) - \varphi(t)] d\zeta_t + \int_{[0, \tau)} \varphi(t) d\zeta_t,$$

with $\zeta \equiv \xi^n$ and $\zeta \equiv \xi^*$, and with the interpretation of integrals as in Remark 5.3 of the Appendix, we have

$$\begin{aligned} E \int_{[0, \tau)} f(t) d\xi_t^n - E \int_{[0, \tau)} f(t) d\xi_t^* \\ \geq -\varepsilon (\sup_{n \geq 1} E\xi_\tau^n + E\xi_\tau^*) + \sum_{i=0}^{m-1} c_i [E\xi_{a_{i+1}}^n - E\xi_{a_{i+1}}^*] - \sum_{i=0}^{m-1} c_i [E\xi_{a_i}^n - E\xi_{a_i}^*]. \end{aligned}$$

Letting $n \rightarrow \infty$, we obtain

$$(4.22) \quad \begin{aligned} & \overline{\lim}_{n \rightarrow \infty} E \int_{[0, \tau)} f(t) d\xi_t^n - E \int_{[0, \tau)} f(t) d\xi_t^* \\ & \geq -\varepsilon (\sup_{n \geq 1} E\xi_\tau^n + E\xi_\tau^*) + c_{m-1} [\overline{\lim}_{n \rightarrow \infty} E\xi_\tau^n - E\xi_\tau^*]. \end{aligned}$$

Recalling (4.6), we can let $t_j \uparrow \tau$, with $t_j \in G$ for every $j \geq 1$, and obtain

$$E\xi_{t_j}^* = \lim_{n \rightarrow \infty} E\xi_{t_j}^n \leq \overline{\lim}_{n \rightarrow \infty} E\xi_\tau^n \leq M,$$

so

$$(4.23) \quad \sup_{n \geq 1} E\xi_\tau^n + E\xi_\tau^* \leq 2M < \infty.$$

As $\varepsilon \downarrow 0$: $c_{m-1} \rightarrow f(\tau)$ and so, using (4.23), (4.22) becomes

$$(4.24) \quad \overline{\lim}_{n \rightarrow \infty} E \int_{[0, \tau)} f(t) d\xi_t^n \geq E \int_{[0, \tau)} f(t) d\xi_t^* + f(\tau) [\overline{\lim}_{n \rightarrow \infty} E\xi_\tau^n - E\xi_\tau^*].$$

Combining this with (4.21) and using the inequality

$$(4.25) \quad \overline{\lim} (x_n + y_n) \geq \overline{\lim} x_n + \underline{\lim} y_n,$$

we obtain

$$(4.26) \quad \begin{aligned} & \overline{\lim}_{n \rightarrow \infty} E \left[\int_0^\tau h(t, x + W_t - \xi_t^n) dt + \int_{[0, \tau)} f(t) d\xi_t^n \right] \\ & \geq f(\tau) [\overline{\lim}_{n \rightarrow \infty} E\xi_\tau^n - E\xi_\tau^*] + E \left[\int_0^\tau h(t, x + W_t - \xi_t^*) dt + \int_{[0, \tau)} f(t) d\xi_t^* \right]. \end{aligned}$$

Let us consider, as before, a strictly increasing sequence $\{t_j\}_{j=1}^\infty \subseteq G$, such that $\lim_{j \rightarrow \infty} t_j = \tau$. For each $j \geq 1$, $n \geq 1$ we have, by virtue of condition (4.2):

$$g(x + W_\tau - \xi_{t_j}^n) - g(x + W_\tau - \xi_\tau^n) \leq f(\tau)(\xi_\tau^n - \xi_{t_j}^n), \quad P\text{-a.s.},$$

whence

$$(4.27) \quad \overline{\lim}_{n \rightarrow \infty} [Eg(x + W_\tau - \xi_{t_j}^n) - Eg(x + W_\tau - \xi_\tau^n)] \leq f(\tau) \overline{\lim}_{n \rightarrow \infty} (E\xi_\tau^n - E\xi_{t_j}^n) \quad \forall j \geq 1.$$

From (4.25) in the form: $\overline{\lim} (y_n - x_n) \geq \underline{\lim} y_n - \underline{\lim} x_n$, the left-hand side of (4.27) is seen to dominate

$$Eg(x + W_\tau - \xi_{t_j}^*) - \underline{\lim}_{n \rightarrow \infty} Eg(x + W_\tau - \xi_\tau^n)$$

by virtue of Lemma 5.2, since $g(\cdot)$ is nonnegative and convex. On the other hand, from (4.25) in the form: $\overline{\lim} (z_n - x_n) \leq \overline{\lim} z_n - \underline{\lim} x_n$, the right-hand side of (4.27) is bounded above by

$$f(\tau) [\underline{\lim}_{n \rightarrow \infty} E\xi_\tau^n - E\xi_{t_j}^*].$$

We have shown, therefore, that

$$(4.28) \quad Eg(x + W_\tau - \xi_{t_j}^*) - \underline{\lim}_{n \rightarrow \infty} Eg(x + W_\tau - \xi_\tau^n) \leq f(\tau) [\overline{\lim}_{n \rightarrow \infty} E\xi_\tau^n - E\xi_{t_j}^*]$$

holds for any $j \geq 1$. Passing to the limit as $j \rightarrow \infty$ in (4.28) we obtain the crucial inequality

$$(4.29) \quad E g(x + W_\tau - \xi_\tau^*) - \lim_{n \rightarrow \infty} E g(x + W_\tau - \xi_\tau^n) \leq f(\tau) [\overline{\lim}_{n \rightarrow \infty} E \xi_\tau^n - E \xi_\tau^*].$$

In conjunction with (4.25) and (4.26), inequality (4.29) yields

$$\overline{\lim}_{n \rightarrow \infty} J(\tau, x; \xi^n) \geq J(\tau, x; \xi^*),$$

and because $\{\xi^n\}_{n=1}^\infty$ is a minimizing sequence for (τ, x) , ξ^* is optimal. \square

As a corollary of Theorems 3.4 and 4.1 we obtain the following existence result for the Optimal Stopping problem.

THEOREM 4.9. *Under conditions (3.2) and (4.1)–(4.3) on the cost functions, there exists a stopping time σ^* which is optimal for the Stopping Problem (3.4).*

It is noteworthy that this result has been established by *purely probabilistic* arguments, using the control problem as an intermediary.

5. Appendix: On uniform integrability and weak L^1 -convergence. This section contains an (apparently novel) sufficient condition for uniform integrability and a Fatou lemma for weak L^1 -convergence; both results are of some independent interest. We conclude with a remark on Riemann–Stieltjes integration and a real variable Lemma.

LEMMA 5.1. *On uniform integrability.*

Let $\{X_n\}_{n=1}^\infty$ be a sequence of nonnegative random variables on the probability space (Ω, \mathcal{F}, P) , satisfying

$$(5.1) \quad \sup_{n \geq 1} \int_{\{X_n > \lambda\}} (X_n - \lambda) dP \rightarrow 0 \quad \text{as } \lambda \rightarrow \infty.$$

The sequence $\{X_n\}_{n=1}^\infty$ is then uniformly integrable.

Proof. First, we show that condition (5.1) implies

$$(5.2) \quad \lambda \sup_{n \geq 1} P(X_n > \lambda) \rightarrow 0 \quad \text{as } \lambda \rightarrow \infty.$$

Suppose not; then there exists a number $\varepsilon > 0$ and two sequences $\{\lambda_k\}_{k=1}^\infty$ and $\{n_k\}_{k=1}^\infty$ of positive integers, increasing strictly to infinity, such that

$$\lambda_k P(X_{n_k} > \lambda_k) \geq \varepsilon > 0 \quad \forall k \geq 1.$$

For a fixed integer $k^* \geq 1$, and any $k \geq k^*$, we have

$$\sup_{n \geq 1} \int_{\{X_n > \lambda_{k^*}\}} (X_n - \lambda_{k^*}) dP \geq \int_{\{X_{n_k} > \lambda_{k^*}\}} (X_{n_k} - \lambda_{k^*}) dP \geq \varepsilon \left(1 - \frac{\lambda_{k^*}}{\lambda_k}\right).$$

Letting $k \rightarrow \infty$ on the right-hand side we obtain

$$\sup_{n \geq 1} \int_{\{X_n > \lambda_{k^*}\}} (X_n - \lambda_{k^*}) dP \geq \varepsilon \quad \forall k^* \geq 1,$$

a contradiction to (5.1).

With (5.2) established, uniform integrability is a consequence of (5.1) and of the obvious inequality

$$\sup_{n \geq 1} \int_{\{X_n > \lambda\}} X_n dP \leq \sup_{n \geq 1} \int_{\{X_n > \lambda\}} (X_n - \lambda) dP + \lambda \sup_{n \geq 1} P(X_n > \lambda). \quad \square$$

The Dunford–Pettis compactness criterion (see Dunford and Schwartz [7, p. 294] or Meyer [17, p. 20]) asserts that every uniformly integrable sequence of random

variables $\{X_n\}_{n=1}^\infty$ contains a subsequence $\{X_{n_k}\}_{k=1}^\infty$ which converges to an integrable random variable X weakly in L^1 , i.e., $\lim_{k \rightarrow \infty} E(X_{n_k}Y) = E(XY)$, for any bounded random variable Y .

LEMMA 5.2. *A Fatou lemma for weak L^1 -convergence.*

Let $\{X_n\}_{n=1}^\infty$ be a uniformly integrable sequence of random variables on (Ω, \mathcal{F}, P) and X be an integrable random variable, such that $\{X_n\}_{n=1}^\infty$ converges to X weakly in L^1 . We have then

$$(5.3) \quad Eh(X) \leq \varliminf_{n \rightarrow \infty} Eh(X_n)$$

for any nonnegative, convex function h on \mathbb{R} .

Proof. We start with the special case

$$h(x) = 0 \vee \max_{1 \leq k \leq m} (\alpha_k x + \beta_k),$$

where $\{\alpha_k\}_{k=1}^m, \{\beta_k\}_{k=1}^m$ are finite sets of real numbers and m is an integer. The sequence

$$\{Z_n \triangleq h(X_n); n \geq 1\}$$

is obviously uniformly integrable. Let $\{Z_{n_j}\}_{j=1}^\infty$ be a subsequence such that

$$\lim_{j \rightarrow \infty} EZ_{n_j} = \varliminf_{n \rightarrow \infty} Eh(X_n).$$

By the Dunford–Pettis compactness criterion, there exists an integrable random variable Z and a further subsequence $\{Z_{n_{j_i}}\}_{i=1}^\infty$ which converges to Z weakly in L^1 . It follows that

$$\lim_{j \rightarrow \infty} EZ_{n_j} = EZ,$$

and it remains to show

$$(5.4) \quad Eh(X) \leq EZ.$$

For notational simplicity, the subsequence which converges weakly to Z will be denoted by $\{Z_n\}_{n=1}^\infty$. Using weak convergence of $\{X_n\}_{n=1}^\infty$ to X , we have for any fixed integer k , $1 \leq k \leq m$:

$$\begin{aligned} E(\alpha_k X + \beta_k) 1_{\{Z < \alpha_k X + \beta_k\}} &= \lim_{n \rightarrow \infty} E(\alpha_k X_n + \beta_k) 1_{\{Z < \alpha_k X + \beta_k\}} \\ &\leq \varliminf_{n \rightarrow \infty} EZ_n 1_{\{Z < \alpha_k X + \beta_k\}} \\ &= EZ 1_{\{Z < \alpha_k X + \beta_k\}} \\ &\leq E(\alpha_k X + \beta_k) 1_{\{Z < \alpha_k X + \beta_k\}}, \end{aligned}$$

and so the last inequality must be an equality. This implies $\alpha_k X + \beta_k \leq Z$, a.s. P , for any $1 \leq k \leq m$. Besides, $\lim_{n \rightarrow \infty} EZ_n 1_{\{Z < 0\}} = EZ 1_{\{Z < 0\}}$ yields $Z \geq 0$ and therefore $Z \geq h(X)$, a.s. P ; relation (5.4) follows.

A general nonnegative, convex function h can be written as the supremum of countably many linear functions:

$$h(x) = \sup_{m \geq 1} h_m(x) \quad \text{where } h_m(x) = 0 \vee \max_{1 \leq k \leq m} (\alpha_k x + \beta_k).$$

By what has already been shown,

$$Eh_m(X) \leq \varliminf_{n \rightarrow \infty} Eh_m(X_n) \leq \lim_{n \rightarrow \infty} Eh(X_n)$$

holds for every $m \geq 1$. Now let $m \rightarrow \infty$ and use monotone convergence to obtain (5.3) in the general case. \square

Remark 5.3. Integration with respect to nondecreasing functions.

Let $\xi(t)$ be a nondecreasing function on $[0, \tau]$, with $\xi(0) = 0$, and $k(t)$ be bounded, continuous on $[0, \tau]$. The Riemann–Stieltjes integral $\int_{[0, \tau]} k(t) d\xi(t)$ is defined by

$$(5.5) \quad \lim_{\|\alpha\| \downarrow 0} \left[\sum_{i=0}^{m-2} k(b_i) \{\xi(\alpha_{i+1}) - \xi(\alpha_i)\} + k(b_{m-1}) \{\xi(\alpha_m) - \xi(\alpha_{m-1})\} \right],$$

where α is a partition $0 = \alpha_0 < \alpha_1 < \dots < \alpha_m = \tau$ of the interval $[0, \tau]$ with mesh $\|\alpha\| = \max_{0 \leq i \leq m-1} |\alpha_{i+1} - \alpha_i|$ and $\alpha_i \leq b_i \leq \alpha_{i+1}$; $i = 0, 1, \dots, m-1$.

If $k(\cdot)$ is right continuous with finitely many points of discontinuity, the above limit still exists and can be taken as the definition of $\int_{[0, \tau]} k(t) d\xi(t)$, provided the partitions α are chosen to include the discontinuities of $k(\cdot)$.

If $k(\cdot)$ is continuous and $\xi_1(t) = \xi_2(t)$ for a.e. $t \in [0, \tau]$, then

$$(5.6) \quad \int_{[0, \tau]} k(t) d\xi_1(t) = \int_{[0, \tau]} k(t) d\xi_2(t).$$

LEMMA 5.4. Let $\varphi: [0, \tau] \rightarrow \mathbb{R}$ be bounded and nondecreasing, and extend it by

$$\varphi(t) = \varphi(0), \quad t \leq 0, \quad \varphi(t) = \varphi(\tau), \quad t \geq \tau.$$

Then the function

$$\varphi_L(t) \triangleq \lim_{s \uparrow t} \varphi(s)$$

is left-continuous, the function

$$\varphi_R(t) \triangleq \lim_{s \downarrow t} \varphi(s)$$

is right-continuous, and

$$(5.7) \quad \varphi_L(t) \leq \varphi(t) \leq \varphi_R(t) \quad \forall t \in [0, \tau].$$

Besides, φ_L and φ_R have the same set of continuity points, and equality holds in (5.7) on this set. In particular,

$$\varphi_L(t) = \varphi(t) = \varphi_R(t) \quad \text{a.e. on } [0, \tau].$$

The proof is left to the diligent reader.

6. Appendix: On the nature of the optimal control process. Theorem 4.1 guarantees the existence of an optimal process $\xi^* \in \mathcal{A}$ for the control problem but provides no information about its nature. We shall address this question by pursuing the equivalence between the problems of control and stopping in the opposite direction. Namely, strong conditions are imposed in this section on the data h, f and g to ensure that the optimal stopping rule is expressible in terms of a monotone moving boundary $s(t)$; $0 \leq t \leq \tau$:

$$(6.1) \quad \begin{aligned} \sigma^* &= \inf \{0 \leq t \leq \tau; x + W_t \geq s(t)\} \\ &= \tau \quad \text{if } \{\cdot \cdot \cdot\} = \emptyset. \end{aligned}$$

Suitable arguments are then employed to show that the process $\tilde{\xi}^* \in \mathcal{A}$ given by

$$(6.2) \quad \xi_t^* = \max [0, \max_{0 \leq u \leq t} \{x + W_u - s(u)\}], \quad t > 0$$

is optimal for the control problem. This is the approach taken in [3], [14]; it relies heavily on the powerful analytical tools which have been developed for the stopping problem, and is presented here for completeness. It shows, in particular, that almost all sample functions of the optimal process ξ^* are *singular* with respect to Lebesgue measure and flat off $\{0 \leq t \leq \tau; X_t^* \geq s(t)\}$, where

$$(6.3) \quad X_t^* = x + W_t - \xi_t^*, \quad 0 \leq t \leq \tau,$$

is the optimal state process: a Brownian motion reflected leftwards along the moving boundary for $t > 0$ and with a possible jump at $t = 0$, so that: $X_{0+}^* = \min\{x, s(0)\}$. Thus, ξ^* can be viewed as the *local time* of X^* at the moving boundary, for $t > 0$.

Let us suppose with Van Moerbeke [24] that, in addition to (3.2), the functions h , f and g satisfy the following conditions:

(6.4) there exists a function $H(t, x)$ which is continuous on $[0, \tau] \times \mathbb{R}$, with partial derivatives of the form $(\partial^{s+r}/\partial t^s \partial x^r)H(t, x)$ which are continuous on $[0, \tau] \times \mathbb{R}$ whenever $2s + r \leq 5$. The gradient $H_x(t, x) = (\partial/\partial x)H(t, x)$ obeys a polynomial growth condition in the space variable x , and we have

$$H_t(t, x) + \frac{1}{2}H_{xx}(t, x) = h_x(t, x) \quad \text{on } [0, \tau] \times \mathbb{R};$$

(6.5) $f(t)$ is twice continuously differentiable on $[0, \tau]$;

$$(6.6) \quad h_{tx}(t, x) + f''(t) \geq 0 \quad \text{on } [0, \tau] \times \mathbb{R};$$

(6.7) there exists a number $b > 0$ such that

$$h_x(t, x) + f'(t) > 0 \quad \text{on } [0, \tau] \times (b, \infty);$$

(6.8) the function $g(x)$ is four times continuously differentiable on $(-\infty, b)$, and satisfies:

- (i) $g''(b-) = 0$;
- (ii) $g'(x) < f(\tau)$, $x < b$;
- (iii) $-\frac{1}{2}g'''(b-) = h_x(\tau, b) + f'(\tau-) > 0$;
- (iv) $\frac{1}{2}g'''(x) + h_x(t, x) + f'(\tau-) \leq 0$, $x < b$.

For any $t \in [0, \tau]$ and any stopping time σ such that $P(0 \leq \sigma \leq \tau - t) = 1$, we have by an application of Itô's rule:

$$E \int_0^\sigma h_x(t + \theta, x + W_\theta) d\theta = EH(t + \sigma, x + W_\sigma) - H(t, x).$$

We define the optimal risk for a stopping problem as in (3.4), but now on the interval $[t, \tau]$, by

$$(6.9) \quad \begin{aligned} \nu(t, x; \tau) \triangleq & \inf_{0 \leq \sigma \leq \tau - t} E[H(t + \sigma, x + W_\sigma) + f(t + \sigma)1_{\{\sigma < \tau - t\}} \\ & + g'(x + W_\sigma)1_{\{\sigma = \tau - t\}}] - H(t, x), \end{aligned}$$

for $0 \leq t \leq \tau$, $x \in \mathbb{R}$. Obviously, $\nu(0, x; \tau) \equiv u(\tau, x)$. Under conditions (3.2) and (6.4)–(6.8), Theorem 3 and its corollary in Van Moerbeke [24] guarantee the existence of a decreasing, continuously differentiable function $s(t)$, $0 \leq t \leq \tau$ with $s(\tau) = b$, such that the stopping time:

$$\begin{aligned} \sigma_t^* &= \inf \{0 \leq \theta \leq \tau - t; x + W_\theta \geq s(\theta)\} \\ &= \tau - t \quad \text{if } \{\cdot \cdot \cdot\} = \emptyset, \end{aligned}$$

is optimal for the stopping problem in (6.9). In addition, the moving boundary $s(t)$ and the optimal risk $\nu(t, x) = \nu(t, x; \tau)$ solve the following Free Boundary Problem:

$$(6.10) \quad \frac{1}{2}\nu_{xx}(t, x) + \nu_t(t, x) + h_x(t, x) = 0, \quad 0 \leq t < \tau, \quad x < s(t),$$

$$(6.11) \quad \geq 0, \quad 0 \leq t < \tau, \quad x > s(t),$$

$$(6.12) \quad \nu(\tau, x) = g'(x), \quad x \in \mathbb{R},$$

$$(6.13) \quad \nu(t, x) < f(t), \quad 0 \leq t \leq \tau, \quad x < s(t),$$

$$(6.14) \quad \nu(t, x) = f(t), \quad 0 \leq t \leq \tau, \quad x \geq s(t),$$

$$(6.15) \quad \nu_x(t, s(t)) = 0, \quad 0 \leq t \leq \tau.$$

It is then not hard to see that the function $M(t, x) \equiv M(t, x; \tau)$ defined on $[0, \tau] \times \mathbb{R}$ by

$$(6.16) \quad M(t, x; \tau) \triangleq g(b) + \int_t^\tau \{h(u, s(u)) - f(u)s'(u)\} du - \int_x^{s(t)} \nu(t, y) dy,$$

is continuous along with its gradient $M_x(t, x)$ on $[0, \tau] \times \mathbb{R}$, has derivatives $M_t(t, x)$, $M_{xx}(t, x)$ which are continuous on $[0, \tau] \times \mathbb{R}$, and satisfies the relations:

$$(6.17) \quad \frac{1}{2}M_{xx}(t, x) + M_t(t, x) + h(t, x) = 0, \quad 0 \leq t < \tau, \quad x \leq s(t),$$

$$(6.18) \quad > 0, \quad 0 \leq t < \tau, \quad x > s(t),$$

$$(6.19) \quad M(\tau, x) = g(x), \quad x \in \mathbb{R},$$

$$(6.20) \quad M_x(t, x) < f(t), \quad 0 \leq t \leq \tau, \quad x < s(t),$$

$$(6.21) \quad M_x(t, x) = f(t), \quad 0 \leq t \leq \tau, \quad x \geq s(t),$$

$$(6.22) \quad M_{xx}(t, x) \geq 0, \quad 0 \leq t \leq \tau, \quad x \in \mathbb{R}.$$

With $M(t, x)$ defined as in (6.16), relations (6.17)–(6.21) can be verified directly, in conjunction with conditions (6.10)–(6.15) and the monotonicity of $s(t)$. In order to check (6.22) we observe (Van Moerbeke [24, Lemma 5]) that the function $m(t, x) \triangleq M_{xx}(t, x) = \nu_{xx}(t, x)$ satisfies

(a) the differential equation:

$$\frac{1}{2}m_{xx}(t, x) + m_t(t, x) + h_{xx}(t, x) \equiv 0, \quad 0 \leq t < \tau, \quad x < s(t);$$

(b) the terminal condition: $m(\tau, x) = g''(x) \geq 0; x \leq b$;

(c) the lateral condition: $m(t, s(t)) = 0; 0 \leq t \leq \tau$.

Condition (6.22) follows then from the stochastic representation

$$m(t, x) = E \left[\int_0^{\sigma_t^*} h_{xx}(t + \theta, x + W_\theta) d\theta + g''(x + W_{\tau-t}) 1_{\{\sigma_t^* = \tau-t\}} \right],$$

$$0 \leq t \leq \tau, \quad x < s(t),$$

because the functions $h(t, \cdot)$, $g(\cdot)$ are convex.

We can now show the relevance of the function $M(t, x)$ to our control problem. With $\xi \in \mathcal{A}$ an arbitrary, admissible control process, apply the Doléans Dade–Meyer change of variable formula (see, for instance, [14, § 9]) to the function $M(t, X_t)$ of

the semimartingale $X_t = x + W_t - \xi_t$; $0 \leq t \leq \tau$ to obtain, P -almost surely:

$$\begin{aligned}
 g(X_\tau) - M(0, x) &= \int_0^\tau \{ \tfrac{1}{2} M_{xx}(t, X_t) + M_t(t, X_t) \} dt \\
 (6.23) \quad &+ \int_0^\tau M_x(t, X_t) dW_t - \int_{[0, \tau)} M_x(t, X_t) d\xi_t \\
 &+ \sum_{0 \leq t < \tau} \{ M(t, X_{t+}) - M(t, X_t) - (X_{t+} - X_t) M_x(t, X_t) \}.
 \end{aligned}$$

From (6.17)–(6.22) we have the almost sure bounds

$$\begin{aligned}
 (6.24) \quad &\int_0^\tau \{ \tfrac{1}{2} M_{xx}(t, X_t) + M_t(t, X_t) \} dt \geq - \int_0^\tau h(t, X_t) dt, \\
 &\int_{[0, \tau)} f(t) d\xi_t \geq \int_{[0, \tau)} M_x(t, X_t) d\xi_t,
 \end{aligned}$$

whereas the last (summation) term in (6.23) is nonnegative, by convexity of $M(t, \cdot)$ (relation (6.22)). Therefore, by taking expectations on both sides of (6.23), we obtain

$$(6.25) \quad M(0, x; \tau) \leq J(\tau, x; \xi) \quad \forall \xi \in \mathcal{A}.$$

For the special choice $\xi = \xi^*$ and $X = X^*$ as in (6.2), (6.3), the inequalities (6.24) hold as identities, and (6.23) becomes:

$$\begin{aligned}
 (6.23)^* \quad &M(0, x) = \int_0^\tau h(t, X_t^*) dt + \int_{[0, \tau)} f(t) d\xi_t^* + g(X_\tau^*) \\
 &+ \int_0^\tau M_x(t, X_t^*) dW_t - [M(0, X_{0+}^*) - M(0, x) - (X_{0+}^* - x) M_x(0, x)],
 \end{aligned}$$

P -almost surely. It is not hard to see that the last term (in brackets) is zero, and thus by taking expectations on both sides of (6.23)*:

$$(6.25)^* \quad M(0, x; \tau) = J(\tau, x; \xi^*).$$

It follows from the last relation and (6.25) that $M(0, x; \tau) = V(\tau, x)$ and that the process ξ^* in (6.2) is optimal for the control problem. We have established the following result.

PROPOSITION 6.1. *Suppose that the cost functions h, f and g satisfy conditions (3.2) and (6.4)–(6.8). Then there exists a decreasing, continuously differentiable function $s(t)$; $0 \leq t \leq \tau$ with $s(\tau) = b$, such that the process ξ^* given by (6.2) is optimal for the Monotone Follower Control Problem, and the stopping time σ^* given by (6.1) (or, equivalently, by (3.18) in terms of ξ^*) is optimal for the Stopping Problem.*

REFERENCES

- [1] J. A. BATHER, *Bayes sequential procedures for deciding the sign of a normal mean*, Proc. Cambridge Philos. Soc., 58 (1962), pp. 599–620.
- [2] ———, *Optimal stopping problems for Brownian motion*, Adv. Appl. Prob., 2 (1970), pp. 259–286.
- [3] J. A. BATHER AND H. CHERNOFF, *Sequential decisions in the control of a spaceship*, Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, 3 (1966), pp. 181–207.
- [4] V. E. BENEŠ, L. A. SHEPP AND H. S. WITSENHAUSEN, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 134–160.
- [5] H. CHERNOFF, *Optimal stochastic control*, Sankhyā, Ser. A, 30 (1968), pp. 221–252.

- [6] C. DELLACHERIE, *Capacités et Processus Stochastiques*, Springer-Verlag, Berlin, 1972.
- [7] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators. Part I: General Theory*, Wiley-Interscience, New York, 1963.
- [8] A. FRIEDMAN, *Stochastic games and variational inequalities*, Arch. Rational Mech. Anal., 51 (1973), pp. 321–346.
- [9] ———, *Regularity theorems for variational inequalities in unbounded domains and applications to stopping time problems*, Arch. Rational Mech. Anal., 52 (1973), pp. 134–160.
- [10] B. GRIGELIONIS AND A. N. SHIRYAEV, *On Stefan's problem and optimal stopping rules for Markov processes*, Theory Probability Appl., 11 (1966), pp. 541–558.
- [11] J. M. HARRISON AND M. I. TAKSAR, *Instantaneous control of a Brownian motion*, Technical Report, Dept. Operations Research, Stanford Univ., Stanford, CA., 1981.
- [12] J. M. HARRISON AND A. J. TAYLOR, *Optimal control of a Brownian storage system*, Stoch. Proc. Appl., 6 (1978), pp. 179–194.
- [13] I. KARATZAS, *The monotone follower problem in stochastic decision theory*, Appl. Math. Optim., 7 (1981), pp. 175–189.
- [14] ———, *A class of singular stochastic control problems*, Adv. Appl. Probability, 15 (1983), pp. 225–254.
- [15] I. KARATZAS AND S. E. SHREVE, *Connections between optimal stopping and singular stochastic control II: Reflected follower problems*, this Journal, 23 (1985), to appear.
- [16] R. S. LIPTSER AND A. N. SHIRYAEV, *Statistics of Random Processes. Part I: General Theory*, Springer-Verlag, Berlin, 1977.
- [17] P. A. MEYER, *Probability and Potentials*, Blaisdell, Waltham, MA., 1966.
- [18] T. P. MIROSHNICHENKO, *Optimal stopping of the integral of a Wiener process*, Theory Probability Appl., 20 (1975), pp. 387–391.
- [19] K. R. PARTHASARATHY, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.
- [20] H. L. ROYDEN, *Real Analysis*, Macmillan, New York, 1963.
- [21] L. A. SHEPP, *Explicit solutions to some problems of optimal stopping*, Ann. Math. Stat., 40 (1969), pp. 993–1010.
- [22] A. N. SHIRYAEV, *Optimal Stopping Rules*, Springer-Verlag, Berlin, 1978.
- [23] S. E. SHREVE, J. P. LEHOCZKY AND D. P. GAVER, *Optimal consumption for general diffusions with absorbing and reflecting barriers*, this Journal, 22 (1984), pp. 55–75.
- [24] P. VAN MOERBEKE, *On optimal stopping and free boundary problems*, Arch. Rational Mech. Anal., 60 (1976), pp. 101–148.

APPROXIMATION OF ITÔ INTEGRALS ARISING IN STOCHASTIC TIME-DELAYED SYSTEMS*

ARUNABHA BAGCHI†

Abstract. Likelihood functional for stochastic linear time-delayed systems involve Itô integrals with respect to the observed data. Since the Wiener process appearing in the standard observation process model for such systems is not realizable and the physically observed process is smooth, one needs to study approximation of such integrals by means of a smooth process; e.g., a band-limited process with no frequency components outside a finite, although large, band. This approximation is studied in the present paper.

Key words. Itô integral, band limited process, white noise, time-delayed systems

1. Introduction. Approximation of a stochastic integral in the “time domain” and its limiting behavior has been extensively studied in the recent book of Ikeda and Watanabe [1]. From the physical point of view, however, “band-limited” (or “frequency domain”) approximation is often more appropriate. One such approximation has been studied by Balakrishnan [2]. The basic limiting result in that paper motivated Mackevičius in [3] to define a symmetric stochastic integral and study its approximations. One practical motivation for introducing band-limited approximation arises in the identification problem for a stochastic linear dynamical system with unknown parameters. The likelihood functional for such problems involves Itô integrals with respect to observations and this Itô integral can be shown to be the limit of an appropriate band-limited approximation with an additional “correction” term. In processing real data, one has to use the approximated integral together with the “correction” term in place of the Itô integral appearing in the likelihood functional. A similar situation arises if one studies the parameter identification problem for a stochastic linear time-delayed system, except that the “correction” term does not follow from the corresponding result in systems without time-delays. The filtered state is now given by a stochastic partial differential equation and obtaining the limiting behavior of band-limited approximation becomes complicated. The correct formula for the limit of band limited approximation in this situation is derived in the present paper.

2. Mathematical preliminaries. Let us consider the following linear stochastic time-delayed system

$$(2.1) \quad x(t; \omega) = \sum_{i=0}^k \int_0^t A_i x(\sigma - h_i; \omega) d\sigma + \int_0^t B dW(\sigma; \omega),$$

$$x(t; \omega) = 0 \quad \text{for } t \leq 0,$$

$$(2.2) \quad Y(t; \omega) = \sum_{i=0}^k \int_0^t C_i x(\sigma - h_i; \omega) d\sigma + \int_0^t D dW(\sigma; \omega),$$

where $x(t; \omega)$ and $Y(t; \omega)$ are n - and m -dimensional “state” and “observation” respectively; $W(t; \omega)$ is a p -dimensional Wiener process; $0 = h_0 < h_1 < \dots < h_k = b$ are the time-delays and $A_i, B, C_i, D, i = 0, \dots, k$ are appropriate dimensional matrices. Assume that $BD^* = 0$ (state and observation noises independent) and $DD^* > 0$ is known which, without loss of generality, may be taken to be the identity matrix. $A_i,$

* Received by the editors March 22, 1983, and in revised form October 5, 1983.

† Department of Applied Mathematics, Twente University of Technology, P.O. Box 217, 7500 AE Enschede, the Netherlands.

B , C_i , $i=0, \dots, k$, have some unknown components. To estimate these unknown parameters by the method of maximum likelihood, based on the data $Y(t; \omega)$, $0 \leq t \leq T$, one has to first evaluate the likelihood functional for the problem. The likelihood functional is the Radon–Nikodym derivative of the measure induced by the process $Y(\cdot; \omega)$ on the space of \mathbb{R}^m -valued continuous functions on $[0, T]$ with respect to Wiener measure thereon, evaluated at the actual sample trajectory of the observation. Let $\beta(t)$ denote the smallest σ -algebra generated by $Y(\sigma; \omega)$, $0 \leq \sigma \leq t$, completed with respect to the sets of measure zero and define

$$(2.3) \quad \hat{x}(t, \theta | \tau) = E[x(t - \theta) | \beta(\tau)].$$

We denote $\hat{x}(t, \theta | t)$ by $\hat{x}(t, \theta)$. With this notation, the likelihood functional for the problem is given by

$$(2.4) \quad L_T(Y(\cdot; \omega)) = \exp \left(-\frac{1}{2} \left(\int_0^T \left\| \sum_{i=0}^k C_i \hat{x}(t, h_i) \right\|^2 dt - 2 \int_0^T \left[\sum_{i=0}^k C_i \hat{x}(t, h_i), dY(t; \omega) \right] \right) \right).$$

The smoothed estimates $\hat{x}(t, \theta)$ for the present model have been obtained in [4] and for a somewhat more general hereditary system model in [5]. They are given by

$$(2.5a) \quad d_t \hat{x}(t, \theta) + \frac{\partial x(t, \theta)}{\partial \theta} dt = K(t, \theta, t) dZ_0(t; \omega),$$

$$(2.5b) \quad d_t \hat{x}(t, 0) - \sum_{i=0}^k A_i \hat{x}(t, h_i) dt = K(t, 0, t) dZ_0(t; \omega),$$

$$\hat{x}(0, \theta) = 0, \quad \theta \geq 0.$$

$$(2.6) \quad K(t, \theta, \tau) = \sum_{i=0}^k P(\tau, \tau - (t - \theta), h_i) C_i^*,$$

$$(2.7) \quad Z_0(t; \omega) = Y(t; \omega) - \sum_{i=0}^k \int_0^t C_i \hat{x}(\sigma, h_i; \omega) d\sigma,$$

and $P(t, \theta_1, \theta_2)$ satisfies

$$(2.8a) \quad \frac{\partial P(t, \theta_1, \theta_2)}{\partial t} + \frac{\partial P(t, \theta_1, \theta_2)}{\partial \theta_1} + \frac{\partial P(t, \theta_1, \theta_2)}{\partial \theta_2}$$

$$= - \sum_{i,j=0}^k P(t, \theta_1, h_i) C_i^* C_j P(t, h_j, \theta_2), \quad \theta_1 \geq 0, \theta_2 \geq 0$$

in the domain $[0, T] \times (0, b] \times (0, b]$;

$$(2.8b) \quad \frac{\partial P(t, \theta_1, 0)}{\partial t} + \frac{\partial P(t, \theta_1, 0)}{\partial \theta_1} - \sum_{i=0}^k P(t, \theta_1, h_i) A_i^*$$

$$= - \sum_{i,j=0}^k P(t, \theta_1, h_i) C_i^* C_j P(t, h_j, 0), \quad \theta_1 \geq 0,$$

$$(2.8c) \quad \frac{\partial P(t, 0, \theta_2)}{\partial t} + \frac{\partial P(t, 0, \theta_2)}{\partial \theta_2} - \sum_{i=0}^k A_i P(t, h_i, \theta_2)$$

$$= - \sum_{i,j=0}^k P(t, 0, h_i) C_i^* C_j P(t, h_j, \theta_2), \quad \theta_2 \geq 0$$

in the domains $[0, T] \times (0, b]$;

$$\begin{aligned}
 \frac{dP(t, 0, 0)}{dt} &= \sum_{i=0}^k A_i P(t, h_i, 0) + \sum_{i=0}^k P(t, 0, h_i) A_i^* + BB^* \\
 &\quad - \sum_{i,j=0}^k P(t, 0, h_i) C_i^* C_j P(t, h_j, 0), \\
 P(0, \theta_1, \theta_2) &= 0 \quad \text{for } (\theta_1, \theta_2) \in [0, b] \times [0, b].
 \end{aligned}
 \tag{2.8d}$$

Two other related results proved in [4] will be useful in the sequel. We put them together in the following lemma.

LEMMA 1.

$$(a) \quad \sum_{i=0}^k C_i \hat{x}(t, h_i; \omega) = \int_0^t h(t, s) dY(s; \omega), \quad 0 \leq t \leq T$$

where $\int_0^T \int_0^t \|h(t, s)\|^2 ds dt < \infty$.

$$(b) \quad \hat{x}(t, \theta) = \int_0^t K(t, \theta, \tau) dZ_0(\tau; \omega)$$

where $K(t, \theta, \tau)$ is given by [4, eq. (4.4)].

Remark. The results mentioned above remain unchanged if A_i , B , C_i , D , $i = 0, \dots, k$ are taken as time-varying matrices. Nonzero initial condition $x(t; \omega)$, $-h_k \leq t \leq 0$, introduces only minor modification into the smoothing equations.

Going back to the likelihood functional, (2.4), we see that it involves Itô integrals with respect to the observation $Y(t; \omega)$, $0 \leq t \leq T$. Unfortunately, what is observed in practice is not $Y(\cdot; \omega)$ given by (2.2) but a band-limited version thereof. Of course, the band-width must be very large to justify the use of (2.2) in theory and obtain the smoothing equations (2.5). However, the Itô integral appearing in (2.4) is *not* the limit of the integral corresponding to the band-limited version of the observation process as the band-width increases without bound. One needs an additional correction term and this will be explicitly determined in this paper.

To obtain this correction term, we need to use the central result in [2] which we state without proof in the following theorem.

THEOREM 1 ([2, Thm. 2.1]). *Let H be the real separable Hilbert space of $n \times 1$ square integrable matrices $L_2^n \equiv L_2([0, T]; \mathbb{R}^n)$. Let $L(t, s)$ be a $n \times n$ matrix-valued function, Lebesgue measurable in t and s , such that*

$$\int_0^T \int_0^t \|L(t, s)\|^2 ds dt < \infty.$$

Define the linear operator \mathcal{L} mapping H into itself by

$$\mathcal{L}f = g, \quad g(t) = \int_0^t L(t, s)f(s) ds, \quad 0 \leq t \leq T.$$

Suppose that $(\mathcal{L} + \mathcal{L}^)$ is nuclear (or, trace class). See [6] for details. Then*

$$\begin{aligned}
 \lim_{m \rightarrow \infty} \int_0^T \left[\int_0^t L(t, s) w^m(s; \omega) ds, w^m(t; \omega) \right] dt \\
 = \int_0^T \left[\int_0^t L(t, s) dW(s; \omega), dW(t; \omega) \right] + \frac{1}{2} \text{Tr}(\mathcal{L} + \mathcal{L}^*),
 \end{aligned}
 \tag{2.9}$$

where

$$\begin{aligned} w^m(t; \omega) &= \int_{-\infty}^{\infty} M(t-s) dW(s; \omega), \\ M(s) &= I_n \int_{-m}^m e^{i2\pi fs} df = I_n \frac{(\sin 2\pi ms)}{\pi s}, \\ I_n &= n \times n \text{ identity matrix,} \end{aligned}$$

and the limit is taken in the L_2 -sense.

3. Main result. Our previous discussion shows that in evaluating the likelihood functional, (2.4), using *real data*, one is confronted with studying the limiting behavior of

$$(3.1) \quad \int_0^T \left[\sum_{i=0}^k C_i \hat{x}^m(t, h_i; \omega), y^m(t; \omega) \right] dt$$

as $m \rightarrow \infty$, where $y^m(t; \omega) = \int_{-\infty}^{\infty} M(t-s) dY(s; \omega)$ with $M(\cdot)$ as defined in Theorem 1 and $\hat{x}^m(t, \theta)$ satisfies

$$\begin{aligned} (3.2) \quad & \frac{\partial \hat{x}^m(t, \theta)}{\partial t} + \frac{\partial \hat{x}^m(t, \theta)}{\partial \theta} = K(t, \theta, t) z_0^m(t, \omega), \\ & \frac{d \hat{x}^m(t, 0)}{dt} - \sum_{i=0}^k A_i \hat{x}^m(t, h_i) = K(t, 0, t) z_0^m(t; \omega), \\ & \hat{x}^m(0, \theta) = 0, \quad \theta \geq 0, \\ & z_0^m(t; \omega) = y^m(t; \omega) - \sum_{i=0}^k C_i \hat{x}^m(t, h_i; \omega), \end{aligned}$$

and $K(t, \theta, \tau)$ is given by (2.6).

THEOREM 2 (main result). *With the same notation as above,*

$$\begin{aligned} (3.3) \quad & \lim_{m \rightarrow \infty} \int_0^T \left[\sum_{i=0}^k C_i \hat{x}^m(t, h_i; \omega), y^m(t; \omega) \right] dt \\ &= \int_0^T \left[\sum_{i=0}^k C_i \hat{x}(t, h_i; \omega), dY(t; \omega) \right] + \int_0^T \text{Tr} \left(\sum_{i,j=0}^k C_i P(t, h_i, h_j) C_j^* \right) dt. \end{aligned}$$

The proof is based on the following lemmas.

LEMMA 2. *Let $K(t, \tau) = \sum_{i=0}^k C_i K(t, h_i, \tau)$ and \mathcal{H}, \mathcal{K} be operators mapping L_2^n into itself, defined by*

$$\begin{aligned} \mathcal{H}f &= g; \quad g(t) = \int_0^t h(t, s) f(s) ds, \text{ the kernel } h(t, s) \text{ being as given in Lemma 1a,} \\ \mathcal{K}f &= g; \quad g(t) = \int_0^t K(t, s) f(s) ds. \end{aligned}$$

Then $(\mathcal{H} + \mathcal{K}^)$ is nuclear if and only if $(\mathcal{H} + \mathcal{K}^*)$ is nuclear and they have the same trace.*

Proof. From Lemma 1a and (2.7), we get

$$\begin{aligned} \sum_{i=0}^k C_i \hat{x}(t, h_i; \omega) &= \int_0^t h(t, s) dY(s; \omega) \\ &= \int_0^t h(t, s) \left[dZ_0(s; \omega) - \sum_{i=0}^k C_i \hat{x}(s, h_i; \omega) ds \right] \end{aligned}$$

so that

$$(3.4) \quad \sum_{i=0}^k \left[C_i \hat{x}(t, h_i; \omega) + \int_0^t h(t, s) C_i \hat{x}(s, h_i; \omega) ds \right] = \int_0^t h(t, s) dZ_0(s; \omega).$$

Using Lemma 1b, one gets

$$\begin{aligned} C_i \hat{x}(t, h_i; \omega) + \int_0^t h(t, s) C_i \hat{x}(s, h_i; \omega) ds \\ = \int_0^t C_i K(t, h_i, \tau) dZ_0(\tau; \omega) + \int_0^t h(t, s) \int_0^s C_i K(s, h_i, \tau) dZ_0(\tau; \omega) ds \\ = \int_0^t \left[C_i K(t, h_i, \tau) + \int_\tau^t h(t, s) C_i K(s, h_i, \tau) ds \right] dZ_0(\tau; \omega), \end{aligned}$$

and (3.4) together with $K(t, \tau) = \sum_{i=0}^k C_i K(t, h_i, \tau)$ yields

$$\int_0^t \left[K(t, \tau) + \int_\tau^t h(t, s) K(s, \tau) ds \right] dZ_0(\tau; \omega) = \int_0^t h(t, \tau) dZ_0(\tau; \omega).$$

This implies that

$$h(t, \tau) = K(t, \tau) + \int_\tau^t h(t, s) K(s, \tau) ds$$

or, in operator form.

$$(3.5) \quad (I + \mathcal{H})^{-1} = I - \mathcal{H} \quad (\text{note that } (I + \mathcal{H}) \text{ always has a bounded inverse}).$$

Suppose that $(\mathcal{H} + \mathcal{H}^*)$ is nuclear. Then from the theorem in [7, p. 232],

$$(3.6) \quad \log \det [(I - \mathcal{H})(I - \mathcal{H}^*)] = -\text{Tr} (\mathcal{H} + \mathcal{H}^*),$$

and using (3.5),

$$(3.7) \quad \log \det [(I + \mathcal{H}^*)(I + \mathcal{H})] = -\log \det [(I - \mathcal{H})(I - \mathcal{H}^*)] < \infty,$$

implying that $(\mathcal{H} + \mathcal{H}^*)$ is also nuclear. Similarly, $(\mathcal{H} + \mathcal{H}^*)$ nuclear implies that $(\mathcal{H} + \mathcal{H}^*)$ is also nuclear. Finally, it readily follows from (3.6) and (3.7) that

$$\text{Tr} (\mathcal{H} + \mathcal{H}^*) = \text{Tr} (\mathcal{H} + \mathcal{H}^*).$$

LEMMA 3. Let \mathcal{L} be the operator as defined in Theorem 1. If $L(t, s)$ has the form

$$L(t, s) = \int_s^t M(t, \sigma) d\sigma$$

such that

$$\int_0^T \int_0^t \|M(t, \sigma)\|^2 d\sigma dt < \infty,$$

then \mathcal{L} is nuclear.

Proof. See the theorem in [7, p. 228].

LEMMA 4. Let \mathcal{L} be the operator as defined in Theorem 1. If $L(t, s)$ has the form

$$L(t, s) = \int_s^t M(\sigma, s) d\sigma$$

such that

$$\int_0^T \int_0^t \|M(t, s)\|^2 ds dt < \infty,$$

then \mathcal{L} is nuclear.

Proof. Define operators \mathcal{P} and \mathcal{S} as follows:

$$\begin{aligned}\mathcal{P}f &= g, & g(t) &= \int_0^t M(t, s)f(s) ds, \\ \mathcal{S}f &= g, & g(t) &= \int_0^t f(s) ds.\end{aligned}$$

Then it is easy to verify that $\mathcal{L} = \mathcal{S}\mathcal{P}$. \mathcal{S} and \mathcal{P} being both Hilbert–Schmidt, the result follows.

LEMMA 5. Let \mathcal{P}_{ij} be the operator mapping L_2^n into itself, defined by

$$\mathcal{P}_{ij}f = g, \quad g(t) = \int_0^t K_{ij}(t, \tau)f(\tau) d\tau,$$

where $K_{ij}(t, \tau) = E[(x(t - h_i) - E(x(t - h_i)|\beta(\tau)))(x(\tau - h_j) - E(x(\tau - h_j)|\beta(\tau)))^*]$. Then $(\mathcal{P}_{ij} + \mathcal{P}_{ij}^*)$ is nuclear and

$$(3.8) \quad \text{Tr}(\mathcal{P}_{ij} + \mathcal{P}_{ij}^*) = \int_0^T \text{Tr} P(t, h_i, h_j) dt.$$

Proof. As shown in [4, p. 204], we may write, for $\tau \leq t$,

$$\begin{aligned}K_{ij}(t, \tau) &= E \left[\left(\int_0^\tau \Phi(t, \sigma) B dW(\sigma) - E \left(\int_0^\tau \Phi(t, \sigma) B dW(\sigma) | \beta(\tau) \right) \right) \right. \\ &\quad \left. \cdot (x(\tau - h_j) - E(x(\tau - h_j) | \beta(\tau)))^* \right]\end{aligned}$$

where the transition matrix $\Phi(t, \tau)$ satisfies

$$\begin{aligned}\frac{d}{dt}\Phi(t, \tau) &= \sum_{i=0}^k A_i \Phi(t - h_i, \tau) \quad \text{for } t \geq \tau, \\ \Phi(\tau, \tau) &= I, \\ \Phi(t, \tau) &= 0 \quad \text{for } t < \tau.\end{aligned}$$

Using the result that for any random vector x ,

$$E(x|\beta(\tau)) = \int_0^\tau \left[\frac{\partial}{\partial \sigma} E x W(\sigma)^* \right] dW(\sigma)$$

(see [8], for example), the formula for covariance of Wiener integrals and the fact that $\Phi(t, \tau)$ is differentiable in t for $t \geq \tau$, it is easy to verify that $K_{ij}(t, \tau)$ and $\partial K_{ij}(t, \tau)/\partial t$ are continuous in $0 \leq \tau \leq t \leq T$. Therefore $g(t)$ is differentiable w.r.t. t and

$$\dot{g}(t) = K_{ij}(t, t) + \int_0^t \frac{\partial K_{ij}(t, \tau)}{\partial t} f(\tau) d\tau.$$

Note that $K_{ij}(t, t) = P(t, h_i, h_j)$ and thus,

$$\begin{aligned} g(t) &= \int_0^t P(s, h_i, h_j) f(s) ds + \int_0^t \int_0^s \frac{\partial K_{ij}(s, \tau)}{\partial s} f(\tau) d\tau ds \\ &= \int_0^t P(s, h_i, h_j) f(s) ds + \int_0^t \left(\int_\tau^t \frac{\partial K_{ij}(s, \tau)}{\partial s} ds \right) f(\tau) d\tau \\ &= P(t, h_i, h_j) \int_0^t f(s) ds - \int_0^t \left(\int_s^t \frac{\partial P(\tau, h_i, h_j)}{\partial \tau} d\tau \right) f(s) ds \\ &\quad + \int_0^t \left(\int_s^t \frac{\partial K_{ij}(\tau, s)}{\partial \tau} d\tau \right) f(s) ds. \end{aligned}$$

Thus

$$\mathcal{P}_{ij} = \mathcal{P}_{ij}^1 + \mathcal{P}_{ij}^2, \quad \text{where}$$

$$\mathcal{P}_{ij}^1 f = g, \quad g(t) = P(t, h_i, h_j) \int_0^t f(s) ds,$$

and

$$\begin{aligned} \mathcal{P}_{ij}^2 f = g, \quad g(t) &= - \int_0^t \left(\int_s^t \frac{\partial P(\tau, h_i, h_j)}{\partial \tau} d\tau \right) f(s) ds \\ &\quad + \int_0^t \left(\int_s^t \frac{\partial K_{ij}(\tau, s)}{\partial \tau} d\tau \right) f(s) ds. \end{aligned}$$

From Lemmas 3 and 4, \mathcal{P}_{ij}^2 is nuclear and being a Volterra operator, has zero trace. Thus, $(\mathcal{P}_{ij}^2 + \mathcal{P}_{ij}^{2*})$ also has zero trace. It is easy to see that $(\mathcal{P}_{ij}^1 + \mathcal{P}_{ij}^{1*})$ is trace class and

$$\text{Tr}(\mathcal{P}_{ij}^1 + \mathcal{P}_{ij}^{1*}) = \int_0^T \text{Tr} P(t, h_i, h_j) dt.$$

It follows that $(\mathcal{P}_{ij} + \mathcal{P}_{ij}^*)$ is trace class and

$$\text{Tr}(\mathcal{P}_{ij} + \mathcal{P}_{ij}^*) = \int_0^T \text{Tr} P(t, h_i, h_j) dt.$$

Proof of the theorem. Lemma 5 readily implies that with \mathcal{H} as defined in Lemma 2, $(\mathcal{H} + \mathcal{H}^*)$ is nuclear and

$$\text{Tr}(\mathcal{H} + \mathcal{H}^*) = \int_0^T \text{Tr} \left(\sum_{i,j=0}^k C_i P(t, h_i, h_j) C_j^* \right) dt.$$

It follows from Lemma 2 that $(\mathcal{H} + \mathcal{H}^*)$ is also nuclear and has the same trace as above. The theorem then follows by proceeding analogously as the proof of [2, Thm. 2.2].

Example. Let us illustrate, by means of a simple scalar example, how we can use the result of Theorem 2 for the problem of estimating parameters of a time-delayed system. Consider the following one-dimensional system

$$(3.9) \quad \begin{aligned} x(t; \omega) &= \int_0^t a x(\sigma; \omega) d\sigma + b W_1(t; \omega), \\ x(t; \omega) &= 0, \quad t \leq 0, \end{aligned}$$

$$(3.10) \quad Y(t; \omega) = \int_0^t [x(\sigma; \omega) + x(\sigma - 1; \omega)] d\sigma + W_2(t; \omega),$$

where $\{W_1(t; \omega)\}$ and $\{W_2(t; \omega)\}$ are two scalar independent Wiener processes. Writing

$$W(t; \omega) = \begin{bmatrix} W_1(t; \omega) \\ W_2(t; \omega) \end{bmatrix}, \quad B = \begin{bmatrix} b & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 1 \end{bmatrix},$$

we may express (3.9)–(3.10) exactly in the form (2.1)–(2.2). Suppose that a and b are unknown system parameters which have to be estimated on the basis of the observed data $Y(t; \omega)$, $0 \leq t \leq T$. We first determine $P(t, \theta_1, \theta_2)$ for $(t, \theta_1, \theta_2) \in [0, T] \times (0, 1] \times (0, 1]$ which is scalar and independent of the observed data. This can be done by solving the set of equations (2.8a)–(2.8d) with $k = 1$, $h_1 = 1$, $A_0 = a$, $A_1 = 0$, $C_0 = 1$, $C_1 = 1$ and $B = \begin{bmatrix} b & 0 \end{bmatrix}$. We write the solution as $P(t, \theta_1, \theta_2; a, b)$ to denote explicit dependence on the parameters a and b .

Now the observation equation (3.10) is only a mathematical idealization. In practice, one never observes $Y(t; \omega)$, $0 \leq t \leq T$, but a band-limited version of $\dot{Y}(t; \omega)$, $0 \leq t \leq T$; that is, one really observes $y^m(t; \omega) = \int_{-\infty}^{\infty} M(t-s) dY(s; \omega)$ with some $M(\cdot)$ as defined in Theorem 1, albeit with m large enough to justify, in theory, the use of (3.10). In practical identification problems, one has to use this real observation to calculate the filter and the likelihood functional. What Theorem 2 implies is that the expression (approximation of the true likelihood functional) that one has to maximize to determine the unknown parameters a and b for this problem is given by

$$\begin{aligned} L_T^m(a, b) = \exp \bigg(& -\frac{1}{2} \left(\int_0^T [\hat{x}^m(t, 0; a, b) + \hat{x}^m(t, 1; a, b)]^2 dt \right. \\ & - 2 \int_0^T [\hat{x}^m(t, 0; a, b) + \hat{x}^m(t, 1; a, b)] y^m(t) dt \\ & + 2 \int_0^T [P(t, 0, 0; a, b) + P(t, 0, 1; a, b) \\ & \quad \left. + P(t, 1, 0; a, b) + P(t, 1, 1; a, b)] dt \right) \bigg), \end{aligned}$$

where $\hat{x}^m(t, \theta; a, b)$, $\theta \in [0, 1]$ is determined from

$$\begin{aligned} \frac{\partial x^m(t, \theta; a, b)}{\partial t} + \frac{\partial \hat{x}(t, \theta; a, b)}{\partial \theta} &= K(t, \theta, t; a, b) z_0^m(t; a, b), \\ \frac{d\hat{x}^m(t, 0; a, b)}{dt} - a\hat{x}^m(t, 0; a, b) &= K(t, 0, t; a, b) z_0^m(t; a, b), \\ \hat{x}^m(0, \theta) &= 0 \quad \text{for } \theta \in [0, 1], \\ z_0^m(t; a, b) &= y^m(t) - [\hat{x}^m(t, 0; a, b) + \hat{x}^m(t, 1; a, b)], \\ K(t, \theta, t) &= P(t, \theta, 0) + P(t, \theta, 1). \end{aligned}$$

In fact, what Theorem 2 really asserts is that

$$\lim_{m \rightarrow \infty} L_T^m(a, b) = L_T(a, b)$$

where $L_T(a, b)$ is the idealized likelihood functional given by (2.4). We must also note that, in practice, $y^m(t; \omega)$, $0 \leq t \leq T$, only means the real observation available to us which is used to calculate all the terms in the approximate likelihood functional.

4. Relation with “white noise model”. If, instead of the standard model for linear time delayed systems described by equations (2.1) and (2.2), one uses the white noise model of Balakrishnan (see [6] for details), stochastic integrals in the smoothing equations and likelihood functional disappear, while the correction term appears in the expression for the likelihood functional. This is not surprising, since physically the white noise model is indeed the true limiting form (idealization) of band-limited approximate noise models. We outline briefly the steps involved in evaluating the smoothing equations and the likelihood functional in the white noise model.

The state and observation equations are given by

$$(4.1) \quad \begin{aligned} \dot{x}(t; n) &= \sum_{i=0}^k A_i x(t - h_i; n) + Bn(t), \\ x(t; n) &= 0, \quad -h_k \leq t \leq 0, \end{aligned}$$

$$(4.2) \quad y(t; n) = \sum_{i=0}^k C_i x(t - h_i; n) + Dn(t),$$

where $n(\cdot)$ is “white noise” on L_2^p as defined in [6]. Assume, as before, that $BD^* = 0$ and $DD^* = I$. $\mathcal{L}_C(t)$ is the bounded linear transformation from \mathcal{L}_2^p into $\mathcal{L}_{2t}^m \equiv L_2([0, t]; \mathbb{R}^m)$:

$$\mathcal{L}_C(t)n = y, \quad y(s; n) = \sum_{i=0}^k C_i \int_0^s \Phi(s - h_i, \sigma) Bn(\sigma) d\sigma, \quad 0 \leq s \leq t,$$

where the transition matrix $\Phi(t, \tau)$ satisfies

$$\begin{aligned} \frac{d}{dt} \Phi(t, \tau) &= \sum_{i=0}^k A_i \Phi(t - h_i, \tau), \quad t \geq \tau, \\ \Phi(\tau, \tau) &= I, \\ \Phi(t, \tau) &= 0 \quad \text{for } t < \tau. \end{aligned}$$

Let \mathcal{D} be defined by

$$(\mathcal{D}n)(t) = D(t)n(t).$$

Then $\eta(t; n) \in L_{2t}^m$, and the weak random variable (see [6]) $y(s; n)$, $0 \leq s \leq t$, has the nonsingular covariance operator $\mathcal{L}_C(t)\mathcal{L}_C(t)^* + I$. Let $\mathcal{H}(t) = I - (I + \mathcal{L}_C(t)\mathcal{L}_C(t)^*)^{-1}$ which can be shown to be an integral operator with kernel $\bar{K}(t, s, \sigma)$. Let $\mathcal{M}(t)$ be the bounded linear operator from L_2^p into \mathbb{R}^n :

$$\mathcal{M}(t)n = x, \quad x = \int_0^t \Phi(t, s) Bn(s) ds,$$

and let

$$\mathcal{R}_x(t, s) = E[x(t; n)x(s; n)^*].$$

Then one can show that

$$\begin{aligned} \hat{x}(t, \theta | \tau) &= E[x(t - \theta; n) | \eta(\tau; n)] \\ &= \mathcal{M}(t - \theta)(\mathcal{L}_C(\tau) + \mathcal{D})^*(I + \mathcal{L}_C(\tau)\mathcal{L}_C(\tau)^*)^{-1} \eta(\tau; n) \\ &= \mathcal{M}(t - \theta)\mathcal{L}_C(\tau)^*(I - \mathcal{H}(\tau))\eta(\tau; n). \end{aligned}$$

Some calculations yield

$$(4.3) \quad \sum_{i=0}^k C_i \mathcal{M}(t-h_i) \mathcal{L}_C(t)^*(I-\mathcal{H}(t))f = \int_0^t \bar{K}(t, t, s) f(s) ds.$$

Define

$$z_0(t; n) = y(t; n) - \sum_{i=0}^k C_i \hat{x}(t, h_i; n).$$

One can show that

$$E \int_0^T ([z_0(t; n), f(t)]) dt \int_0^T ([z_0(t; n), g(t)]) dt = \int_0^T [f(t), g(t)] dt$$

so that $z_0(\cdot; n)$ is a white noise process in L_2^m and furthermore, one gets from (4.3) that

$$z_0(t; n) = y(t; n) - \int_0^t \bar{K}(t, t, s) y(s; n) ds.$$

Thus one can write

$$y(\cdot; n) = (I - J) z_0(\cdot; n),$$

and therefore one has the integral representation

$$(4.4) \quad \hat{x}(t, \theta) = \int_0^t K(t, \theta, \tau) z_0(\tau; n) d\tau.$$

If $P(s, \theta_1, \theta_2) = E[(x(s - \theta_1) - \hat{x}(s, \theta_1))(x(s - \theta_2) - \hat{x}(s, \theta_2))^*]$, $P(s, \theta_1, \theta_2)$ satisfies the same set of equations as (2.8). While using the representation (4.4) and whiteness of $z_0(\cdot; n)$, one gets

$$K(t, \theta, \tau) = \sum_{i=0}^k P(\tau, \tau - (t - \theta), h_i) C_i^*,$$

the same expression as (2.6). Direct differentiation and (2.8) yield

$$\begin{aligned} \frac{\partial \hat{x}(t, \theta)}{\partial t} + \frac{\partial \hat{x}(t, \theta)}{\partial \theta} &= K(t, \theta, t) z_0(t; n), \\ \frac{d\hat{x}(t, 0)}{dt} - \sum_{i=0}^k A_i \hat{x}(t, h_i) &= K(t, 0, t) z_0(t; n), \end{aligned}$$

the precise limiting form of (3.1) and (3.2). To evaluate the likelihood functional, note that

$$\begin{aligned} y(t; n) &= \sum_{i=0}^k C_i \hat{x}(t, h_i; n) + z_0(t; n) \\ &= \sum_{i=0}^k C_i \int_0^t K(t, h_i, \tau) dz_0(\tau; n) + z_0(t; n). \end{aligned}$$

Thus, with \mathcal{H} as defined in Lemma 2,

$$y(\cdot; n) = (I + \mathcal{H}) z_0(\cdot; n)$$

and

$$z_0(\cdot; n) = (I - \mathcal{H}) y(\cdot; n),$$

where from the defining relation for $z_0(t; n)$, it is obvious that

$$(\mathcal{H}y(\cdot; n))(t) = \sum_{i=0}^k C_i \hat{x}(t, h_i; n).$$

The Radon–Nikodym derivative of the weak distribution induced by $y(t; n)$, $0 \leq t \leq T$, on L_2^m with respect to normal distribution thereon is given by ([6, pp. 300–302])

$$L(\omega) = \exp \left(-\frac{1}{2} (\| (I - \mathcal{H})\omega \|^2 - \|\omega\|^2 + \text{Tr}(\mathcal{H} + \mathcal{H}^*)) \right), \quad \omega \in L_2^m,$$

so that the likelihood functional is

$$\begin{aligned} L_T(y(\cdot; n)) &= \exp \left(-\frac{1}{2} (\|\mathcal{H}y(\cdot; n)\|^2 - 2[\mathcal{H}y(\cdot; n), y(\cdot; n)] + \text{Tr}(\mathcal{H} + \mathcal{H}^*)) \right) \\ &= \exp \left(-\frac{1}{2} \left(\int_0^T \left\| \sum_{i=0}^k C_i \hat{x}(t, h_i) \right\|^2 - 2 \left[\sum_{i=0}^k C_i \hat{x}(t, h_i), y(t; n) \right] \right) dt \right. \\ &\quad \left. + \int_0^T \text{Tr} \left(\sum_{i,j=0}^k C_i P(t, h_i, h_j) C_j^* \right) dt \right) \end{aligned}$$

so that the “correction” term appears naturally in the likelihood functional in the white noise model.

5. Conclusion. We have studied band-limited approximation of Itô integrals arising in the expression for likelihood functional in stochastic linear time-delayed systems. It is shown that the “correction” term that one has to take into account while working with the approximate model appears naturally in the alternate formulation of the problem via the white noise model.

REFERENCES

- [1] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland/Kondanska, Amsterdam/Tokyo, 1981.
- [2] A. V. BALAKRISHNAN, *On the approximation of Itô integrals using band-limited processes*, this Journal, 12 (1974), pp. 237–251.
- [3] V. MACKEVICUS, *Symmetric stochastic integrals and their approximations*, Stochastics, 8, (1982), pp. 121–138.
- [4] A. BAGCHI, *A martingale approach to state estimation in delay-differential systems*, J. Math. Anal. Appl., 56 (1976), pp. 195–210.
- [5] J. Y. OUVARD, *Linear filtering in Hilbert spaces II: an application to the smoothing theory for hereditary systems with observation delays*, this Journal, 16 (1978), pp. 938–952.
- [6] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.
- [7] ———, *Stochastic Differential Systems*, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, 1973.
- [8] T. KAILATH, *A note on least squares estimation by the innovations methods*, this Journal, 10 (1972), pp. 477–486.

H*-CONTROLLABILITY AND OBSERVABILITY OF LINEAR PERIODIC SYSTEMS

SERGIO BITTANTI[†], PATRIZIO COLANERI[‡] AND GUIDO GUARDABASSI[†]

Abstract. With reference to linear periodic systems, the classical notion of controllability introduced by Kalman (*K*-controllability) is shown to be equivalent to the characterization of controllability (*H*-controllability) first proposed by Hewer as a natural extension to the periodic case of an eigenvalue-eigenvector controllability condition independently introduced by several authors for time-invariant systems only. The proof of such an equivalence also leads to the conclusion that if a *T*-periodic system *S* is *K*-controllable and *k* is the degree of the minimal polynomial of the monodromy matrix associated with *S*, then, at any time *t*, *S* is controllable over $(t, t + kT)$. Since *k* is lower than or equal to the system order, a well-known result due to Brunovsky is slightly strengthened. By duality, the corresponding observability results follow.

Key words. continuous-time linear systems, periodic systems, controllability and observability, modal controllability conditions, periodic Riccati equation

1. Introduction. This paper deals with the controllability of linear periodic systems. Since the corresponding observability results are readily obtained by duality, no specific attention will further be given to them in the sequel.

For the linear time-invariant case, an eigenvalue-eigenvector characterization of controllability, equivalent to the classical one due to Kalman, was introduced by many authors, such as Johnson [1], Popov [2], Belevitch [3], Hautus [4]. A first extension to the periodic case of this notion, henceforth referred to as *H*-controllability, was made by Hewer [5, § 2], who also developed some arguments to prove that *H*-controllability is equivalent to Kalman controllability (*K*-controllability) in the periodic case as well. Unfortunately, Hewer's arguments were not technically sound since his Theorem 2.15, on which they were essentially based, has subsequently been proved to be false [6]. A slightly different concept of *H*-controllability for periodic systems, totally equivalent to the one proposed by Hewer, has been adopted by Kano and Nishimura [7] to study the existence of periodic solutions to periodic Riccati differential equations.

The main purpose of the present paper is to prove that, for linear periodic systems, *H*-controllability as defined by Hewer, Kano and Nishimura is in fact equivalent to standard *K*-controllability. The paper is organized as follows. In § 2 some preliminary definitions and results are briefly recalled. The main result is proposed in § 3, while some concluding remarks are collected in the last section.

2. Preliminaries. Consider the linear periodic system

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

where $A: R \rightarrow R^{n \times n}$, $B: R \rightarrow R^{n \times m}$ and $u: R \rightarrow R^m$ are piecewise continuous functions. Furthermore, *A* and *B* are by assumption periodic functions of period *T*. Denoting by $\Phi(t, \tau)$ the system transition matrix, it is apparent that

$$(2) \quad \Phi(t + T, \tau + T) = \Phi(t, \tau).$$

* Received by the editors January 27, 1983, and in revised form September 15, 1983. This research was partially supported by CNR, Centro di Teoria dei Sistemi, and by M.P.I.

[†] Dipartimento di Elettronica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy.

[‡] Centro di Teoria dei Sistemi, CNR, c/o Dipartimento di Elettronica, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy.

The matrix $\Phi(T, 0)$, which plays a major role in the machinery of linear periodic systems, is usually named *monodromy matrix*. For easy reference, the definition of K -controllability is first recalled.

DEFINITION 1. System (1) is K -controllable over a time interval (t, τ) if, for any $\bar{x} \in \mathbb{R}^n$, there exists an input $u_{[t, \tau]}(\cdot)$ such as to carry the state $x(t) = \bar{x}$ into the origin at time τ .

System (1) is K -controllable if, for any t , there exists $\tau > t$ such that (1) is K -controllable over (t, τ) . \square

Contrary to what is stated in [5, § 2] and in [8, Prop. 2.26], if system (1) is K -controllable, it may not be controllable over $(0, T)$. A counterexample is given in [6]. Instead, it is true that, if system (1) is K -controllable, then it is controllable over $(0, nT)$, where n is the system order. This fact is an obvious corollary of the following result, proved by Brunovsky in [9].

THEOREM 1. System (1) is K -controllable if and only if the matrix

$$(3) \quad W_i = \int_0^{iT} \Phi(0, t) B(t) B(t)' \Phi(0, t)' dt$$

is positive definite for $i = n$.

As for H -controllability, the following characterization is given in [7].

DEFINITION 2. System (1) is H -controllable if, for each eigenvalue λ of the monodromy matrix, $\Phi(T, 0)' \eta = \lambda \eta$ and $B(t)' (\Phi^{-1}(t, 0))' \eta = 0$ for a.e. $t \in [0, T]$ imply $\eta = 0$. \square

This definition is equivalent to the one considered in [5]. In fact, consider the Floquet representation [10] of the transition matrix

$$\Phi(t, 0) = F(t) e^{Jt}.$$

where F is periodic of period T and J is a constant matrix. Then, system (1) is apparently H -controllable if and only if, for each eigenvalue ν of J , $J' \eta = \nu \eta$ and $B(t)' (F^{-1}(t))' \eta = 0$ for a.e. $t \in [0, T]$ imply $\eta = 0$.

In the case of time invariant systems, $A(t) = A$, $B(t) = B$, it can be taken $F(t) = I$ and $J = A$. This leads to the conclusion that an invariant system is H -controllable if and only if for each eigenvector of A' , $B' \eta = 0$ implies $\eta = 0$. In other words, when specialized to the time-invariant case, Definition 2 is consistent with the characterization of controllability given in [1]–[4].

3. H -controllability of linear periodic systems. The main result of this paper is that, for periodic systems, H and K -controllability are equivalent (see Theorem 2 below). To prove this, three lemmas are first introduced, which make the proofs of the subsequent Theorems 2 and 3 straightforward.

LEMMA 1. The matrix (3) can be given the following expression:

$$(4) \quad W_{i+1} = W_i + \Phi_i W_1 \Phi_i',$$

where

$$(5) \quad \Phi_i := \Phi(0, iT).$$

Proof. From (3) it is apparent that

$$(6) \quad W_{i+1} = W_i + \int_{iT}^{(i+1)T} \Phi(0, t) B(t) B(t)' \Phi(0, t)' dt.$$

Let $\tilde{t} = t - iT$. Then, in view of (2) and (5),

$$(7) \quad \Phi(0, t) = \Phi(0, iT)\Phi(iT, t) = \Phi_i\Phi(0, \tilde{t}).$$

Since $B(t) = B(\tilde{t})$, from (3), (6) and (7), (4) follows.

Remark 1. In view of (2) and (5), $\Phi_i = \Phi_1^i$. Hence (4) can be equivalently written as follows

$$(8) \quad W_{i+1} = W_i + \Phi_1^i W_1 \Phi_1^{i'}.$$

LEMMA 2. *If system (1) is K-controllable, then it is H-controllable.*

Proof. By contradiction, let $\eta \neq 0$ be an eigenvector of $\Phi(T, 0)'$ associated with eigenvalue λ such that

$$(9) \quad B(t)'\Phi(0, t)'\eta = 0 \quad \text{for a.e. } t \in [0, T].$$

We then prove that matrix W_i defined by (3) is singular for any positive integer i . Precisely, by induction, we show that $\eta^* W_i \eta = 0$, $\forall i$, where $*$ denotes conjugate transpose.

Since

$$\eta^* W_1 \eta = \int_0^T \|B(t)'\Phi(0, t)'\eta\|^2 dt,$$

from (9) the conclusion that

$$(10) \quad \eta^* W_1 \eta = 0$$

follows.

Now assume that $\eta^* W_i \eta = 0$. Then (8) entails that

$$\eta^* W_{i+1} \eta = \eta^* \Phi_1^i W_1 \Phi_1^{i'} \eta.$$

As η is an eigenvector of the transpose of the monodromy matrix, the inverse of which is matrix Φ_1 ,

$$\Phi_1^{i'} \eta = \lambda^{-i} \eta.$$

Therefore

$$\eta^* W_{i+1} \eta = |\lambda|^{-2i} \eta^* W_1 \eta = 0.$$

By induction, the singularity of W_i is proved for each integer i , and this contradicts the assumed K -controllability. \square

The converse of the above lemma is stated as follows.

LEMMA 3. *If system (1) is H-controllable, then it is K-controllable over $(0, kT)$ where k is the degree of the minimal polynomial of the monodromy matrix.*

Proof. By contradiction, assume that the matrix W_k is singular. Then, since the null space of W_k , $i < k$, is a subspace of the null space of W_i , there exists a $\xi \neq 0$ such that

$$W_i \xi = 0, \quad i = 1, 2, \dots, k.$$

As Φ_1 is nonsingular, from (8) it then follows that

$$(11) \quad W_1 \Phi_1^{i'} \xi = 0, \quad i = 0, 1, \dots, k-1.$$

Now, let $\alpha(z)$ be a polynomial of minimal degree such that

$$(12) \quad \alpha(\Phi_1') \xi = 0.$$

Obviously, such a polynomial has degree not greater than the one of the minimal polynomial of Φ_1 . Notice that, since $\Phi(T, 0) = \Phi_1^{-1}$, such a degree is equal to k . Let λ be a zero of $\alpha(z)$ and let

$$(13) \quad \alpha(z) = \beta(z)(z - \lambda).$$

Define also

$$(14) \quad \eta = \beta(\Phi_1')\xi.$$

The degree of polynomial $\beta(z)$ is lower than the one of $\alpha(z)$. Had $\eta = 0$ held true, $\alpha(z)$ would not have been a polynomial of minimal degree satisfying (12). Thus $\eta \neq 0$.

Equations (12)–(14) entail that η is an eigenvector of Φ_1' . Moreover, (11) and (14) lead to the conclusion that

$$(15) \quad W_1\eta = 0.$$

In turn, (15) is equivalent to

$$B(t)'\Phi(0, t)'\eta = 0 \quad \text{for a.e. } t \in [0, T].$$

This contradicts the assumption of H -controllability.

THEOREM 2. *System (1) is H -controllable if and only if it is K -controllable.*

THEOREM 3. *If system (1) is K -controllable, it is K -controllable over $(0, kT)$, where k is the degree of the minimal polynomial of the monodromy matrix.*

4. Concluding remarks. In this paper the equivalence between K -controllability and H -controllability for periodic systems has been established (Theorem 2) and a well-known result of Brunovsky has been slightly strengthened (Theorem 3).

The results of § 3 and, in particular, Theorem 2 combined with previous results by Kano and Nishimura [7, Thm. 6] provide in fact an indirect proof of an important result first implicitly stated but incorrectly proved by Hewer [5, Thm. 1.3]. The result is as follows. Consider the T -periodic matrix Riccati differential equation

$$\dot{P}(t) = -A(t)'P(t) - P(t)A(t) + P(t)B(t)R^{-1}(t)B(t)'P(t) - C(t)'C(t)$$

where $R(t)$ is positive definite for all t . If (A, B) is K -controllable and (A, C) is K -observable, then i) there exists a unique positive definite T -periodic solution \bar{P} of the Riccati equation above, and ii) the system

$$\dot{x}(t) = [A(t) - B(t)R^{-1}(t)B(t)'\bar{P}(t)]x(t)$$

is asymptotically stable.

Acknowledgment. The authors gratefully acknowledge Professors A. Locatelli and N. Schiavoni for stimulating discussions.

REFERENCES

- [1] C. D. JOHNSON, *Invariant hyperplanes for linear dynamical systems*, IEEE Trans. on Automat. Control, AC-11 (1966), pp. 113–116.
- [2] V. M. POPOV, *Hyperstability of Control Systems*, Springer, Berlin, 1973 (translation of Rumanian ed., 1966).
- [3] V. BELEVITCH, *Classical Network Theory*, Holden-Day, San Francisco, 1968.
- [4] M. L. J. HAUTUS, *Controllability and observability conditions of linear autonomous systems*, Indag. Math., 72 (1969), pp. 443–448.
- [5] G. A. HEWER, *Periodicity, detectability and the matrix Riccati equation*, this Journal, 13 (1975), pp. 1235–1251.

- [6] S. BITTANTI, G. GUARDABASSI, C. MAFFEZZONI AND L. SILVERMAN, *Periodic systems: controllability and the matrix Riccati equation*, this Journal, 16 (1978), pp. 37–40.
- [7] H. KANO AND T. NISHIMURA, *Periodic solutions of matrix Riccati equations with detectability and stabilizability*, Int. J. Control, 29 (1979), pp. 471–487.
- [8] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1979.
- [9] P. BRUNOVSKY, *Controllability and linear closed-loop controls in linear periodic systems*, J. Differential Equations, 6 (1969), pp. 296–313.
- [10] A. HALANAY, *Differential Equations*, Academic Press, New York, 1966.

EXISTENCE OF VALUE IN GENERALIZED PURSUIT-EVASION GAMES*

L. S. ZAREMBA†

Abstract. The results presented here relate both to a pursuit-evasion game and to a game of fixed time duration. The dynamics of both games is governed by a system of differential inclusions, and the state variables x, y have to satisfy some general type constraints. Player I can apply any lower 17-strategy [15, p. 400] and player II can apply any strategy in the Varaiya-Lin sense. The main result is the existence of a value (in both the games) and an optimal strategy for player II. The key assumption is the condition that from any sequence of player II's trajectories emanating from a convergent sequence of points one can choose a subsequence convergent on a fixed segment $[t_0, T]$ to some player II's trajectory. The condition holds when the set $F_2(t, y)$ is convex for any $(t, y) \in R^{1+k}$. This paper is an outgrowth of work conducted by the author [14]–[17]; it extends the results obtained in [14], [16].

Key words. pursuit-evasion game, value of a game, strategy, nonanticipating operator, lower strategy

1. Introduction. The paper is concerned with both a generalized pursuit-evasion game (G_1) and a game of fixed time duration (G_2) in which the dynamics of both players is modeled by a system of differential inclusions

$$(1.1) \quad \dot{x}(t) \in F_1(t, x(t)), \quad x(t_*) = x_*, \quad x \in R^n,$$

$$(1.2) \quad \dot{y}(t) \in F_2(t, y(t)), \quad y(t_*) = y_*, \quad y \in R^k.$$

Similarly as in [14]–[17], we investigate the case where the state variables x, y are restricted in the sense that

$$(1.3) \quad x(t) \in N_1(t), \quad y(t) \in N_2(t),$$

where $N_1 = \{(N_1(t), t): t \in R\}$ and $N_2 = \{(N_2(t), t): t \in R\}$ are closed subsets in suitable Euclidean spaces. Game G_1 ends when, for some $T_M[x(\cdot), y(\cdot)]$, the triplet $(t, x(t), y(t))$ hits a fixed target set $M \subset R^{1+n+k}$. Player II strives to minimize the amount

$$(1.4) \quad P = P[t_*, x(\cdot), y(\cdot)] = \int_{t_*}^{T_M} h(t, x(t), y(t)) dt, \quad h \geq 0,$$

of “energy” consumed in forcing $(t, x(t), y(t))$ into the terminal set M , while player I is purposeful in maximizing (1.4). In game G_2 is concerned, player II minimizes

$$(1.5) \quad P = P[t_*, x(\cdot), y(\cdot)] = g(T, x(T), y(T)) + \int_{t_*}^T h(t, x(t), y(t)) dt,$$

contrary to player I who makes the payoff function (1.5) as large as possible (T is a fixed number greater than t_*).

The problem of the existence of a value in generalized pursuit-evasion differential games, as well as in more general differential games of survival, has received rather scant attention in the literature. Some existence theorems for differential games of survival without restricted phase coordinates have been obtained in [1], [2], [5], [6], [11] and [7, pp. 37–38].

Differential games with restricted phase coordinates were also investigated in the literature. Games of fixed time duration were examined in [5], [6], [10], pursuit-evasion games in [3], [4] (see also [12], where the dynamics was governed by a general system

* Received by the editors February 18, 1983, and in revised form November 10, 1983.

† Department of Mathematics, Agricultural and Pedagogical University, Siedlce, 08-110, Poland.

satisfying some axioms) and generalized pursuit–evasion differential games in [5, pp. 243–244].

The present paper is an outgrowth of work conducted by the author [14]–[17]. It extends the results obtained in [14], [16], where the payoffs are of the form

$$P[t_*, x(\cdot), y(\cdot)] = \int_{t_*}^{T_M} dt, \quad P[t_*, x(\cdot), y(\cdot)] = \int_{t_*}^T h(t, x(t), y(t)) dt,$$

respectively, and the dynamics is governed by a system of ordinary differential equations.

In § 2, we make assumptions on games G_1 , G_2 . The key assumption (condition (2.7)) states that from every sequence of player II's trajectories emanating from a convergent sequence of points one can choose a subsequence convergent on $[t_0, T]$ to some player II's trajectory. This assumption is fulfilled unless the set $F_2(t, y)$ is convex for any $(t, y) \in [t_*, T] \times R^k$. The sets of admissible strategies are defined as in the author's previous works [14]–[17]. The existence of a value and an optimal strategy for player II having complete information is proved for both game G_1 (Theorem 5.1) and game G_2 (Theorem 5.2).

2. Assumptions and preliminary conclusions. Throughout this paper, t_* , x_* , y_* and T are specified values with $t_* < T$; (t_*, x_*, y_*) is the initial point. We assume that all points (t_0, x_0, y_0) occurring in the paper are admissible, that is, $x_0 \in N_1(t_0)$, $y_0 \in N_2(t_0)$. Below, we make the following assumptions on games G_1 , G_2 .

(2.1) The sets N_1 , N_2 , M are closed in R^{n+1} , R^{k+1} , R^{n+k+1} , respectively.

It is plain that the following two assumptions refer to game G_1 whereas the third one concerns game G_2 .

(2.2) The target set M contains the hyperplane $P = \{(t, x, y): t = T\}$.

(2.3) The function h appearing in (1.4) is nonnegative and continuous with respect to all its arguments.

(2.4) The functions g and h occurring in (1.5) are continuous with respect to all their arguments.

Both F_1 and F_2 are required to satisfy the following condition; it is stated in terms of F_1 only.

- (2.5) (i) For each $(t, x) \in [t_*, T] \times R^n$, the set $F_1(t, x)$ is compact;
 (ii) for each $x \in R^n$, $F_1(t, x)$ is measurable in t ;
 (iii) $h(F_1(t, x_1), F_1(t, x_2)) \leq K \|x_1 - x_2\|$, $t_* \leq t \leq T$, where h is the Hausdorff metric;
 (iv) there exist $x^* \in R^n$ and a positive number M such that, for all $t \in [t_*, T]$, $h(F_1(t, x^*), \{0\}) \leq M$.

It is known [8, p. 163] that the system (1.1), (1.2) has a solution defined on $[t_*, T]$. For each bounded set D in R^n and t_0 satisfying $t_* \leq t_0 \leq T$, denote by $X(t_0, D)$ the space of all admissible (satisfying (1.3)) trajectories of player I defined on $[t_0, T]$ and emanating from a set D . We consider $X(t_0, D)$ as a subset of $C^n[t_0, T]$, the Banach space of all continuous mappings from $[t_0, T]$ into R^n . In case $D = \{x_0\}$, we simply write $X(t_0, x_0)$; dual notations apply to player II with $X(t_0, D)$ replaced by $Y(t_0, E)$.

COROLLARY 2.1. *For each bounded set $D(E)$ in $R^n(R^k)$, the space $X(t_0, D)$ (resp. $Y(t_0, E)$) is bounded and equicontinuous subset of $C^n[t_0, T]$. Therefore, the closures of $X(t_0, D)$, $Y(t_0, E)$ are nonempty compact subsets in suitable Banach spaces.*

Proof. For each $(t, x) \in [t_*, T] \times R^n$, we have (below, $|F| = h(F, \{0\})$)

$$(2.6) \quad |F_1(t, x)| \leq |F_1(t, x^*)| + K \|x - x^*\| + M_1(1 + \|x\|), \quad x \in R^n,$$

with $M_1 = \max(K, M + K\|x^*\|)$. Let $x(\cdot)$ be a trajectory of player I emanating from a set D . Taking into account that

$$\|x(t)\| \leq \|x(t_0)\| + M_1(t - t_0) + \int_{t_0}^t \|x(s)\| ds$$

we infer from the Gronwall–Bellman inequality that $X(t_0, D)$ is bounded if D is bounded in R^n . Finally, $\|x(t) - x(t')\| \leq M_2(t - t')$, where $M_2 = M_1(1 + M_3)$, $M_3 = \sup \{\|x(t)\| : t_0 \leq t' \leq t \leq T, x(\cdot) \in X(t_0, D)\}$. \square

COROLLARY 2.2. *The following inequalities hold: (i) $\langle x, y \rangle \leq L(1 + \|x\|^2)$, $y \in F_1(t, x)$; (ii) $\langle x, y \rangle \leq L(1 + \|y\|^2)$, $x \in F_2(t, y)$, where $\langle x, y \rangle$ stands for the scalar product of vectors x, y .*

Proof. We recall from (2.6) that $|F_1(t, x)| \leq M_1(1 + \|x\|)$; hence $|F_1(t, x)| \leq M_1(\|x\| - 1) + 2M_1$, i.e., $|F_1(t, x)| \leq 2M_1\|x\|$ if $\|x\| \geq 1$ and $|F_1(t, x)| \leq 2M_1$ otherwise. Therefore, for each $y \in F_1(t, x)$, we have $\langle x, y \rangle \leq 2M_1\|x\|^2$ if $\|x\| \geq 1$ and $\langle x, y \rangle \leq 2M_1$ if $\|x\| \leq 1$. Finally, $\langle x, y \rangle \leq 2M_1(1 + \|x\|^2)$, $y \in F_1(t, x)$, which completes the proof. \square

The following condition is the key assumption in our approach (cf. Comment 5.1).

(2.7) Every sequence of trajectories $y_n(\cdot) \in Y(t_0, y_n)$ for which $y_n(t_0)$ converge to some point y_0 contains a subsequence convergent to a certain $y(\cdot) \in Y(t_0, y_0)$.

This condition is fulfilled if the set $F_2(t, y)$ is convex for any $(t, y) \in [t_*, T] \times R^k$. In fact, it is known that, for each compact set E in R^k , the space $Y(t_0, E)$ is nonempty [9, Thm. 2] and compact [9, Thm. 3] in $C^k[t_0, T]$ whenever conditions (i)–(iv), are met: (i) For each $(t, y) \in [t_*, T] \times R^k$, the set $F_2(t, y)$ is compact and convex; (ii) $F_2(t, y)$ is measurable in t and upper semicontinuous in y ; (iii) $F_2(t, y)$ carries bounded sets into bounded sets (this condition follows from (2.6)); (iv) $\langle y, z \rangle \leq C(1 + \|y\|^2)$, $C > 0$, for any $z \in F_2(t, y)$.

3. Strategies. Following [12, p. 157], [13, p. 143] an operator

$$\beta = \beta(t_0, x_0, y_0): X(t_0, x_0) \rightarrow Y(t_0, y_0)$$

is said to be a strategy for player II if β is a nonanticipating operator, that is, $x_1(t) = x_2(t)$, $t_0 \leq t \leq t_1$, implies that

$$\beta[x_1(\cdot)](t) = \beta[x_2(\cdot)](t), \quad t_0 \leq t \leq t_1.$$

An operator $\alpha = \alpha(t_0, x_0, y_0): Y(t_0, y_0) \rightarrow X(t_0, x_0)$ is called a player I's strategy (cf. [15]–[17]) if it is a lower π -strategy for some partition $\pi = (t_i)$ of the segment $[t_0, T]$, which means that from the equality

$$y_1(t) = y_2(t), \quad t_0 \leq t \leq \bar{t} < T,$$

where $t_i \leq \bar{t} < t_{i+1} \leq T$, it follows

$$\alpha[y_1(\cdot)](t) = \alpha[y_2(\cdot)](t) \quad \text{in } t_0 \leq t \leq t_{i+1}.$$

The set of such defined strategies (π varies over all partitions of $[t_0, T]$) are designated by $B(t_0, x_0, y_0)$, $A(t_0, x_0, y_0)$, respectively.

Remark 3.1. The concepts of a strategy defined above coincide with those given in [16], [17]. However, there is a little difference between them and those given in [15] since in [15] by $X(t_0, x_0)$ (resp. $Y(t_0, y_0)$) is meant the space of all player I's (II's) trajectories defined on $[t_0, \infty)$.

It is clear that from each pair of strategies, a unique outcome $[x(\cdot), y(\cdot)]$ results satisfying [15, p. 400]

$$(3.1) \quad \alpha[y(\cdot)] = x(\cdot), \quad \beta[x(\cdot)] = y(\cdot).$$

The notions of lower and upper values as well as value for games G_1, G_2 are defined in a standard way [15, p. 400], namely

$$(3.2) \quad \underline{V} = \underline{V}(t_0, x_0, y_0) = \sup \{P^-(\alpha) : \alpha \in A(t_0, x_0, y_0)\},$$

$$(3.3) \quad \bar{V} = \bar{V}(t_0, x_0, y_0) = \inf \{P^+(\beta) : \beta \in B(t_0, x_0, y_0)\},$$

where

$$(3.4) \quad P^-(\alpha) = \inf \{P[t_0, \alpha[y(\cdot)], y(\cdot)] : y(\cdot) \in Y(t_0, y_0)\},$$

$$(3.5) \quad P^+(\beta) = \sup \{P[t_0, x(\cdot), \beta[x(\cdot)]] : x(\cdot) \in X(t_0, x_0)\}.$$

We say that game $G_1(G_2)$ has a value V if $V = \underline{V} = \bar{V}$. Similarly as in [15]–[17], by a global lower strategy $\alpha_E(t_0, x_0, y_0)$,

$$(3.6) \quad \alpha_E : Y(t_0, E) = \bigcup_{y \in E} Y(t_0, y) \rightarrow X(t_0, x_0), \quad y_0 \in E,$$

is understood a family of lower π -strategies $\alpha_y \in A(t_0, x_0, y)$, $y \in E$, indexed by elements of some open ball $E \subset R^k$ which as a rule is not the same (although it may be) for different global lower strategies. The set of all global lower π -strategies (π varies over all partitions of $[t_0, T]$) is designated by $A_1(t_0, x_0, y_0)$.

4. Auxiliary considerations. The remaining part of the paper is built up as follows. We start by defining a set valued function $W_N(\cdot)$, $N \subset R^1$, and next prove some stability properties of $W_N(\cdot)$. The main results of this paper easily follow from these properties.

If it is not otherwise stated, the consideration below concern games G_1, G_2 simultaneously. Below, and elsewhere we designate by $P = P[t_0, x(\cdot), y(\cdot)]$ the payoff given by (1.4) or (1.5). The Euclidean distance of a point z from a set K is denoted by $\rho[z, K]$. For a given (admissible) state (t_0, x_0, y_0) and a global lower strategy $\alpha_E(t_0, x_0, y_0)$, by

$$Z(t_0, x_0, y_0, \alpha_E)$$

is understood the set of all (finite) limits p of the form

$$(4.1) \quad p = \lim_{n \rightarrow \infty} P[t_0, \alpha[y_n(\cdot)], y_n(\cdot)], \quad y_n(\cdot) \in Y(t_0, E), \quad \lim_{n \rightarrow \infty} y_n(t_0) = y_0.$$

The following proposition is a consequence of (2.2)–(2.4) and Corollary 2.1.

PROPOSITION 4.1. *Let a strategy $\alpha_E \in A_1(t_0, x_0, y_0)$ and a closed subset N of the real line be given. If $p \notin N$ for $p \in Z(t_0, x_0, y_0, \alpha_E)$ then, for some $\varepsilon > 0$, we have $\rho(P[t_0, \alpha[y(\cdot)], y(\cdot)], N) \geq \varepsilon$, $y(\cdot) \in Y(t_0, E')$ where $E' \subset E$ is a ball in R^k containing y_0 .*

Put

$$(x_0, y_0, r_0) \in W_N(t_0), \quad t_* \leq t_0 \leq T$$

if and only if $x_0 \in N_1(t_0)$, $y_0 \in N_2(t_0)$ and to each global lower strategy $\alpha_E \in A_1(t_0, x_0, y_0)$ there exists a point $p \in Z(t_0, x_0, y_0, \alpha_E)$ satisfying $r_0 + p \in N$. The proof of Lemma 4.1 is modeled after the proof of [16, Lemma 5.1].

LEMMA 4.1. *To each $(x_0, y_0, r_0) \in W_N(t_0)$, each trajectory $x_0(\cdot) \in X(t_0, x_0)$, and $\delta > 0$, there corresponds $y_0(\cdot) \in Y(t_0, y_0)$ such that either $T_M[x_0(\cdot), y_0(\cdot)] \leq t_0 + \delta$ or $(x_0(t_0 + \sigma), y_0(t_0 + \delta), r\delta) \in W_N(t_0 + \delta)$, where*

$$(4.2) \quad r_\delta = r_0 + \int_{t_0}^{t_0 + \delta} h(t, x_0(t), y_0(t)) dt.$$

Proof. Assuming the contrary, put

$$(4.3) \quad K = \{(y, r): y = y(t_0 + \delta), r = r(t_0 + \delta), y(\cdot) \in Y(t_0, y_0)\},$$

where

$$r(t_0 + \delta) = r_0 + \int_{t_0}^{t_0 + \delta} h(t, x_0(t), y(t)) dt.$$

It is plain that $T_M[x_0(\cdot), y(\cdot)] > t_0 + \delta$, $y(\cdot) \in Y(t_0, y_0)$ and, by the definition of K ,

$$(x_0(t_0 + \delta), z) \notin W_N(t_0 + \delta) \quad \text{for } z = (y(t_0 + \delta), r(t_0 + \delta)) \in K.$$

By the definition of $W_N(\cdot)$ and Proposition 4.1, there exists a global lower strategy $\alpha_z \in A_1(t_0 + \delta, x_0(t_0 + \delta), y(t_0 + \delta))$, an open ball E_z containing $y(t_0 + \delta)$ and an open set $R_z \ni r$, such that the inequality

$$\rho(r + P[t_0 + \delta, \alpha_z[y(\cdot)], y(\cdot)], N) \geq \varepsilon_z > 0$$

holds for $y(\cdot) \in Y(t_0 + \delta, E_z)$ and $r \in R_z$. By virtue of (2.3), (2.4) and (2.7), the set K is compact in R^{n+1} . It is well known from elementary topology that K is contained in a finite number, say m , of open sets $E_{z_i} \times R_{z_i} = E_i \times R_i$, $1 \leq i \leq m$; here \times denotes the Cartesian product. Besides, there exist m global strategies $\alpha_{z_i} = \alpha_i \in A_1(t_0 + \delta, x_0(t_0 + \delta), y_i(t_0 + \delta))$, $z_i = (y_i(t_0 + \delta), r_i) \in E_i \times R_i$, such that

$$(4.4) \quad \rho(r + P[t_0 + \delta, \alpha_i[y(\cdot)], y(\cdot)], N) \geq \varepsilon_i$$

for all $y(\cdot) \in Y(t_0 + \delta, E_i)$ and $r \in R_i$, $1 \leq i \leq m$. By invoking (2.7) we infer that, for some ball $E_0 \ni y_0$,

$$(4.5) \quad (y(t_0 + \delta), r(t_0 + \delta)) \in \bigcup_{i=1}^m E_i \times R_i, \quad y(\cdot) \in Y(t_0, E_0).$$

Now, it is easy to define a global strategy $\alpha_0 \in A_1(t_0, x_0, y_0)$ with the property

$$(4.6) \quad \rho(r_0 + P[t_0, \alpha_0[y(\cdot)], y(\cdot)], N) \geq \min_{1 \leq i \leq m} \varepsilon_i > 0$$

for all $y(\cdot) \in Y(t_0, E_0)$ [16, pp. 588–589]. In this way we have completed the proof of the lemma because the inequality (4.6) is contradictory to the assumption $(x_0, y_0, r_0) \in W_N(t_0)$. \square

The concept of a global lower strategy is used in the proof of the proposition below.

PROPOSITION 4.2. *If $(x, y_n, r_n) \in W_N(\bar{t})$, $n = 1, 2, \dots$, with $t_* \leq \bar{t}$, and the sequence (y_n, r_n) converges to (y, r) , then $(x, y, r) \in W_N(\bar{t})$, too.*

For the proof, it suffices to assume the contrary, apply Proposition 4.1 and observe that if $\alpha \in A_1(\bar{t}, x, y)$, then $\alpha \in A_1(\bar{t}, x, y_n)$ whenever n is a sufficiently large number.

LEMMA 4.2. *Let $(x_0, y_0, r_0) \in W_N(t_0)$ and a motion $\bar{x}(\cdot) \in X(t_0, x_0)$ be given. Then, for some trajectory $\bar{y}(\cdot) \in Y(t_0, y_0)$ and*

$$(4.7) \quad \bar{r}(t) = r_0 + \int_{t_0}^t h(s, \bar{x}(s), \bar{y}(s)) ds, \quad t \geq t_0,$$

we have $(\bar{x}(t), \bar{y}(t), \bar{r}(t)) \in W_N(t)$ for $t_0 \leq t \leq T_M[\bar{x}(\cdot), \bar{y}(\cdot)]$.

Proof. Similarly as in [15, p. 402], we apply Lemma 4.1 k time with $\delta = 2^{-n}$ and make use of (2.7) with $y_n = y_0$, $n = 1, 2, \dots$, arriving at (by virtue of Proposition 4.2)

$$(\bar{x}(t_n^k), \bar{y}(t_n^k), \bar{r}(t_n^k)) \in W_N(t_n^k)$$

whenever $t_0 \leq t_n^k = t_0 + k2^{-n} \leq T_M[\bar{x}(\cdot), \bar{y}(\cdot)]$. Assume that the assertion of the lemma is false, i.e., for a certain $T_1 \in [t_0, T_M]$, we have $(\bar{x}(T_1), \bar{y}(T_1), \bar{r}(T_1)) \notin W_N(T_1)$. By virtue of Proposition 4.1, there exists a global strategy $\delta_E \in A_1(T_1, \bar{x}(T_1), \bar{y}(T_1))$ with the property that

$$(4.8) \quad \rho(\bar{r}(T_1) + P[T_1, \alpha_E[y(\cdot)], y(\cdot)], N) \geq \varepsilon_1 > 0, \quad y(\cdot) \in Y(T_1, E)$$

with E being a ball containing $\bar{y}(T_1)$.

Let now $T_1 < T_M[\bar{x}(\cdot), \bar{y}(\cdot)]$. It follows from (2.1) that the distance between a point $(t, \bar{x}(t), \bar{y}(t))$, $t_0 \leq t \leq T_1$, and the terminal set M is greater than some number $\varepsilon > 0$. Based on (2.7) and the proof of Corollary 2.1, one can select such number $t' \in [t_0, T_1]$ and an open ball E' containing $\bar{y}(t')$ that $y'(T_1) \in E$ whenever $y'(\cdot) \in Y(t', E')$; also $T_M[\bar{x}(\cdot), y(\cdot)] > T_1$, $y(\cdot) \in Y(t', E')$. Let $y'_\delta(\cdot)$ designates the portion of $y'(\cdot)$ on the ray $[t' + \delta = T_1, \infty)$; putting

$$\alpha_{E'}[y'(\cdot)](s) = \begin{cases} \bar{x}(s), & t' \leq s \leq T_1, \\ \alpha_E[y'_\delta(\cdot)](s), & T_1 \leq s \leq T \end{cases}$$

we get a global strategy $\alpha_{E'} \in A_1(t', \bar{x}(t'), \bar{y}(t'))$ guaranteeing, by virtue of (2.3), (2.4), (4.8) and Corollary 2.1,

$$(4.9) \quad \rho(\bar{r}(t') + P[t', \alpha_{E'}[y'(\cdot)], y'(\cdot)], N) \geq \frac{1}{2} \varepsilon_1, \quad y'(\cdot) \in Y(t', E'),$$

which is impossible because t' may be equal to one of the numbers $t_n^k = t_0 + k \cdot 2^{-n}$.

The argument presented above is not valid in the case $T_M[\bar{x}(\cdot), \bar{y}(\cdot)] = T_1$ since then it may happen (for game G_1 only) that $T_M[\bar{x}(\cdot), y'(\cdot)] < T_1 = T_M[\bar{x}(\cdot), \bar{y}(\cdot)]$ for some trajectory $y'(\cdot) \in Y(t', E')$. In the first part of the proof, we have shown that $(\bar{x}(t), \bar{y}(t), \bar{r}(t)) \in W_N(t)$, $t_0 \leq t < T_M[\bar{x}(\cdot), \bar{y}(\cdot)] = T_M$. Therefore, to each t , $t_0 \leq t < T_M$, and an arbitrary global strategy $\alpha' \in A_1(t, \bar{x}(t), \bar{y}(t))$, there corresponds $p' \in Z(t, \bar{x}(t), \bar{y}(t), \alpha')$ such that

$$(4.10) \quad \bar{r}(t) + p' \in N,$$

where $p' = \lim P[t, \alpha'[y_n(\cdot)], y_n(\cdot)]$, $\lim y_n(t) = y(t)$, $n \rightarrow \infty$. Taking into account (4.10) and (4.8) with $T_1 = T_M$ and setting

$$\alpha'_D[y(\cdot)](s) = \begin{cases} \bar{x}(s), & t \leq s \leq T_M, \\ \alpha_E[y_\delta(\cdot)](s), & T_M \leq s \leq T, \end{cases}$$

where D is a ball containing $\bar{y}(t)$ (t is temporarily fixed) such that $y(T_M) \in E$ for all $y(\cdot) \in Y(t, D)$ and $y_\delta(\cdot)$ means the portion of $y(\cdot)$ on $[t + \delta = T_M, \infty)$, we conclude (taking a subsequence of $y_n(\cdot)$ if necessary) that $\lim T_M[\alpha'_D[y_n(\cdot)], y_n(\cdot)] \leq T_M[\bar{x}(\cdot), \bar{y}(\cdot)] = T_M$ unless the difference $T_M - t$ is sufficiently small. Hence $\bar{r}(T_M) \in N$ since N is closed and the difference $\bar{r}(t) + P[t, \alpha'_D[y_n(\cdot)], y_n(\cdot)] - \bar{r}(T_M)$ is a sufficiently small number. On the other hand, it follows from (4.8) that $\rho(\bar{r}(T_M), N) \geq \varepsilon$ because $P[T_M, \alpha_E[y(\cdot)], y(\cdot)] = 0$ for $y(\cdot) \in Y(T_M, \bar{y}(T_M))$. This contradiction completes the proof. \square

5. Basic results. The arguments employed in the proof of the lemma below are borrowed from [14, Thm. 5.1] and [16, Thm. 6.1].

LEMMA 5.1. *If $(x_0, y_0, r_0) \in W_N(t_0)$, with $N = (-\infty, z]$, then there exists a strategy $\beta_0 \in B(t_0, x_0, y_0)$ ensuring*

$$(5.1) \quad r_0 + P(t_0, x(\cdot), \beta_0[x(\cdot)](t)) \leq z, \quad x(\cdot) \in X(t_0, x_0).$$

Proof. First, we are going to show that, for some strategy $\beta_0 \in B(t_0, x_0, y_0)$,

$$(5.2) \quad (x(t), \beta_0[x(\cdot)](t), r(t)) \in W_N(t), \quad x(\cdot) \in X(t_0, x_0), \quad t_0 \leq t \leq T_M(x(\cdot), \beta_0[x(\cdot)])$$

where

$$r(t) = r_0 + \int_{t_0}^t h(s, x(s), \beta_0[x(\cdot)](s)) ds.$$

To this end, denote by H the family of such pairs (γ, C) for which $C \subset X(t_0, x_0)$, $\gamma: C \rightarrow Y(t_0, y_0)$ and

- (i) $(x(t), \gamma[x(\cdot)](t), r(t)) \in W_N(t)$, $t_0 \leq t \leq T_M(x(\cdot), \beta_0[x(\cdot)](t))$, $x(\cdot) \in C$,
- (ii) γ is a nonanticipating operator.

The set H (nonempty by Lemma 4.2) is partially ordered by the relation α : $(\gamma_1, C_1) \alpha (\gamma_2, C_2)$ if and only if $C_1 \subset C_2$ and $\gamma_2[x(\cdot)] = \gamma_1[x(\cdot)]$, $x(\cdot) \in C_1$. Denote by (γ_0, C_0) a maximal element in H (existing by the Kuratowski-Zorn lemma). To show (5.2), it suffices to prove that $C_0 = X(t_0, x_0)$. Assuming that there exists $x_0(\cdot) \in X(t_0, x_0) \setminus C_0$, set

$$T_1 = \sup \{ \bar{t} \geq t_0 : \exists x(\cdot) \in C_0, x(t) = x_0(t), t_0 \leq t \leq \bar{t} \}.$$

Similarly as in [16, p. 592], we arrive at

$$(5.3) \quad (x_0(t), y_0(t), r_0(t)) \in W_N(t), \quad t_0 \leq t < T_1$$

with $y_0(\cdot) \in Y(t_0, y_0)$ and

$$r_0(t) = r_0 + \int_{t_0}^t h(s, x_0(s), y_0(s)) ds, \quad t \geq t_0.$$

Besides [14, p. 143], $T_1 < T_M[x_0(\cdot), y_0(\cdot)]$. To show

$$(5.4) \quad (x_0(T_1), y_0(T_1), r_0(T_1)) \in W_N(T_1),$$

assume the contrary and repeat the same arguments as those given in the first part of the proof of Lemma 4.2 obtaining (4.9) for those t' for which the difference $T_1 - t'$ is a sufficiently small number. It means that $(x_0(t'), y_0(t'), r_0(t')) \notin W_N(t')$, which is impossible in view of (5.3). Applying Lemma 4.2 with (x_0, y_0, r_0) replaced by $(x_0(T_1), y_0(T_1), r_0(T_1))$ and $\bar{x}(\cdot) \in X(t_0, x_0)$ replaced by $x_0(\cdot) \in X(T_1, x_0(T_1))$ one can easily show [14, p. 143] that (γ_0, C_0) is not a maximal element in H .

In this way we have proved (5.2) with $\beta_0 = \gamma_0$. Let $x(\cdot)$ be a fixed trajectory belonging to $X(t_0, x_0)$. In the case of game G_1 , we apply (5.2) with $t = T_M(x(\cdot), \beta_0[x(\cdot)](t)) = T_M$ obtaining $r(T_M) + p \leq z$ for some $p \geq 0$ since $h \geq 0$. In the case of game G_2 , we also apply (5.2) with $t = T_M = T$, arriving at

$$r(T) + g(T, x(T), \beta_0[x(\cdot)](T)) = r_0 + P(t_0, x(\cdot), \beta_0[x(\cdot)](t)) \leq z$$

since for an arbitrary global strategy $\alpha \in A_1(T, x(T), \beta_0[x(\cdot)](T))$ we have

$$Z(T, x(T), \beta_0[x(\cdot)](T), \alpha) = g(T, x(T), \beta_0[x(\cdot)](T)). \quad \square$$

THEOREM 5.1. *For the game (1.1)–(1.4) let assumptions (2.1)–(2.3) and (2.5), (2.7) be met. The game (G_1) then has a value and there exists an optimal strategy for the pursuer (having complete information).*

THEOREM 5.2. *For the game (1.1)–(1.3), (1.5) let assumptions (2.1), (2.4), (2.5), (2.7) be satisfied. The game (G_2) then has a value and there exists an optimal strategy for player II (having complete information).*

Proof of theorems. Observe that $\underline{V} = \underline{V}(t_*, x_*, y_*) \leq \bar{V}(t_*, x_*, y_*) = \bar{V}$ since from every pair of strategies α, β it uniquely results an outcome (cf. (3.1)). Note that

$$(5.5) \quad (x_*, y_*, 0) \in W_N(t_*) \quad \text{if } N = (-\infty, \underline{V}];$$

otherwise, we would get a contradiction with the definition of the number \underline{V} . By invoking (5.5) and Lemma 4.2, we obtain an optimal player II's strategy and the desired inequality $\bar{V} \leq \underline{V}$. \square

Comment 5.1. In our approach we need the powerful assumption (2.7). We feel that assumption (2.7) or its slightly weaker version is necessary to prove the existence of an optimal strategy for player II. If we are interested in proving the existence of a value only, condition (2.7) seems to be a little artificial.

REFERENCES

- [1] R. J. ELLIOTT AND N. J. KALTON, *The existence of value in differential games of pursuit and evasion*, J. Differential Equations, 12 (1972), pp. 504–523.
- [2] ———, *Cauchy problems for certain Isaac–Bellman equations and games of survival*, Trans. Amer. Math. Soc., 198 (1974), pp. 45–72.
- [3] J. FLYNN, *Lion and man: the boundary constraints*, this Journal, 11 (1973), pp. 397–411.
- [4] ———, *Lion and man: the general case*, this Journal, 12 (1974), pp. 581–597.
- [5] A. FRIEDMAN, *Differential Games*, John Wiley, New York, 1971.
- [6] ———, *Remarks on differential games of survival*, J. Differential Equations, 14 (1973), pp. 121–128.
- [7] ———, *Differential Games*, CBMS Regional Conference Series in Applied Mathematics 18, American Mathematical Society, Providence, RI, 1974.
- [8] C. J. HIMMELBERG AND F. S. VAN VLECK, *Lipschitzian generalized differential equations*, Rend. Sem. Mat. Univ. Padova, 48 (1973), pp. 159–169.
- [9] N. KIKUCHI, *Control problems of contingent equations*, Publ. Res. Inst. Math. Sc., Ser. A3 (1967), pp. 85–99.
- [10] R. C. SCALZO, *Differential games with restricted phase coordinates*, this Journal, 12 (1974), pp. 426–435.
- [11] R. J. STERN, *Differential games of survival with space-like terminal set*, this Journal, 12 (1974), pp. 167–177.
- [12] P. P. VARAIYA, *On the existence of solutions to a differential games*, this Journal, 5 (1967), pp. 153–162.
- [13] P. P. VARAIYA AND J. LIN, *Existence of saddle points in differential games*, this Journal, 7 (1969), pp. 142–157.
- [14] L. S. ZAREMBA, *On the existence of value in the Varaiya–Lin sense in differential games of pursuit and evasion*, J. Optim. Theory Appl., 29 (1979), pp. 135–145.
- [15] ———, *Existence of value in a differential game*, Systems Control Lett. 1 (1982), pp. 399–403.
- [16] ———, *Existence of value in differential games with fixed time duration*, J. Optim. Theory Appl., 38 (1982), pp. 581–598.
- [17] ———, *Existence of value in differential games with terminal cost function*, J. Optim. Theory Appl., 39 (1983), pp. 89–104.

AN $O(N^3)$ ALGORITHM FOR OPTIMAL REPLACEMENT PROBLEMS*

SHMUEL GAL†

Abstract. This paper considers the general replacement problem of an item which can be in any one of N operating states. At the beginning of each period, one can either keep the item for (at least) one more period or sell it and get a new item. The change of state during each period is assumed to be Markovian. In contrast to most of the previous works, we assume that the state of the item is determined by several measurable parameters and not only by its age. In this case the optimal replacement rule is much more complicated than the usual simple rule of replacing the item at a fixed critical age.

The above mentioned replacement problem can be formulated in terms of dynamic programming. In general, however, the usual methods for obtaining the optimal policy, such as the policy iteration method, could require many iterations. In our paper we present an algorithm which requires a number of iterations not exceeding the number of states in which a replacement decision has to be taken. We show that the overall number of arithmetic operations needed is $O(N^3)$ and is actually less than twice the effort needed for solving a set of N linear equations with N variables. Using this fact it is demonstrated that the algorithm is optimal from the complexity point of view.

This algorithm is also applicable for optimal stopping problems.

Key words. optimal replacement, Markov decision process, policy iterations, computational complexity

1. Introduction. Consider the following decision problem: At the beginning of each time period $t = 1, 2, \dots$, a certain item is inspected. The state of this item depends on some measurable parameters. It is assumed that the number of these states is finite. Thus, each state can be associated with an integer i , $i = 0, 1, \dots, N$. At the beginning of each period the decision maker has two options: The first option is to keep the item for (at least) one more period. In this case the change of state during this period is Markovian, i.e., for each i there is a probability P_{ij} , $j = 0, 1, \dots, N$ of transition from state i to state j . The expected gain from the item during this period is denoted by q_i .

The other option (in case that $i \neq 0$) is to sell the item at price s_i and immediately get a new item. (It is convenient to deduct from the selling price the cost of the new item; thus s_i may be negative.) It can either be assumed that the new item is a standard one, or that the new item is chosen randomly from a known population of new items. The state i of a new item is chosen as $i = 0$ and the net gain from this item during the first period is denoted by q_0 . The probabilities of transitions of a new item are P_{0j} , $j = 0, 1, \dots, N$.

If the item is at any state other than $i = 0$ (which can happen in case the item broke down during the previous period), one has to decide whether to replace it or not. The objective of the decision maker is to maximize the expected total profit for an infinite horizon.

In §§ 2–4 a discount factor β ($0 < \beta < 1$) will be assumed while in § 5 the case of no discounting will be considered.

Finding an optimal decision policy for the problem described above was discussed in several papers (see e.g. Derman [1] and Luss [3]). Most of these papers, however, assume that the item progresses stochastically through increasingly bad levels of deterioration. Under this assumption there is a natural ranking of the states and this special structure usually leads to a simple optimal policy (i.e., replace the item at a certain critical age). In this work no such structure is assumed. The situation to be considered is the case when the state of the item is determined by several crucial parameters and not only by its age. Such a situation, which motivated the present

* Received by the editors February 11, 1983, and in revised form September 15, 1983.

† IBM Israel Scientific Center, Meyer Advanced Technology Center, Technion City, Haifa 32000, Israel.

work, occurs, for example, in a dairy cow replacement problem. Here the state of each cow is determined not only its age but also by some other important traits such as its expected milk yield, weight, etc. In this problem there is no simple ranking of the states, and the optimal policy may have a complicated structure.

The above optimal replacement problem can be formulated in terms of dynamic programming. In general, however, the usual methods for obtaining the optimal solutions, such as the policy iteration method, could use many iterations. (In principle, the number of iterations needed can grow exponentially with the number of states.) In this work we present an algorithm which requires a number of iterations not exceeding the number of states in which a replacement decision has to be made. Furthermore, all these iterations except the first require a relatively small number of computations so that the overall number of arithmetic operations needed is $O(N^3)$. Actually, the total computational effort needed by the algorithm is less than twice the effort needed for solving a set of linear equations with N variables. Thus, it is expected that this algorithm should be efficient.

2. The basic idea. The model described in the introduction falls into the class of classical Markovian decision processes. For these problems a stationary policy is optimal. Thus, the problem of finding the optimal replacement policy is equivalent to identifying an optimal set of "replacement states" $R^* \subset \{1, 2, \dots, N\}$ such that the item is replaced if $i \in R^*$, and is kept for at least one more period if $i \notin R^*$. (This definition implies that the "new item" state $i = 0$ does not belong to R^* .) The expected discounted profit vector (the "value vector")

$$\mathbf{u} = \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_N \end{pmatrix}$$

for this policy satisfies the following set of linear equations:

$$(1) \quad u_i = \begin{cases} q_i + \beta \sum_{j=0}^N P_{ij} u_j & \text{for } i \notin R^*, \\ u_0 + s_i & \text{for } i \in R^* \end{cases}$$

where β is the discount factor ($0 < \beta < 1$).

The optimality condition that the value vector determined by (1) also satisfies for the above policy is (see, e.g. Howard [2])

$$(2) \quad \begin{aligned} u_i &\geq u_0 + s_i && \text{for } i \notin R^*, \\ u_0 + s_i &\geq q_i + \beta \sum_{j=0}^N P_{ij} u_j && \text{for } i \in R^*. \end{aligned}$$

Our algorithm for finding the optimal set of replacement states is based on the following principle: Start with the policy $R = \emptyset$, i.e., the item is kept for all $i \neq 0$. The value vector \mathbf{v} of this policy satisfies

$$v_i = q_i + \beta \sum_{j=0}^N P_{ij} v_j, \quad i = 0, 1, \dots, N.$$

Then, rank $v_i - s_i$ $i = 1, \dots, N$ by their magnitude:

$$v_{i_1} - s_{i_1} \leq v_{i_2} - s_{i_2} \leq \dots \leq v_{i_N} - s_{i_N}.$$

If $v_{i_1} - s_{i_1} \geq v_0$ then condition (2) is satisfied and thus $R^* = \emptyset$. Otherwise, there exist some states i_1, i_2, \dots, i_J with $v_{i_1} - s_{i_1} \leq v_{i_2} - s_{i_2} \leq \dots \leq v_{i_J} - s_{i_J} < v_0$.

It does not always follow that the optimal decision for all these states is to replace the item (see the example in § 4). It is true, however, that in state i_1 the optimal decision is to replace the item. Thus an element of R^* has been identified. The algorithm then continues making this type of iteration until $v_i \geq v_0 + s_i$ for all states which do not belong to R . Thus, the number of iterations of this algorithm is (at most) equal to the number of states in R^* . A rigorous presentation of this algorithm and a proof that it leads to the optimal policy is given in the next section.

3. Presentation of the algorithm. The algorithm is based on the following theorem which is used in order to identify states belonging to the optimal replacement set R^* .

THEOREM 1. *Let R be a set of integers (possibly empty) satisfying $R \subset R^*$. Determine the value vector \mathbf{v} corresponding to R by*

$$(3) \quad v_i = \begin{cases} q_i + \beta \sum_{j=0}^N P_{ij} v_j & \text{for } i \notin R, \\ v_0 + s_i & \text{for } i \in R. \end{cases}$$

Denote

$$(4) \quad v_I - s_I = \min_{i \notin R} (v_i - s_i).$$

If $v_I < v_0 + s_I$, then $I \in R^*$.

Using Theorem 1, the following iterative scheme for finding the optimal replacement set, R^* , is now described: Start using Theorem 1 with $R = \emptyset$ and obtain a state I_1 (or several such states if the minimum in (4) is not unique) where I_1 is the index I at the left-hand side of (4). Then use Theorem 1 with $R = \{I_1\}$ and obtain another state I_2 ; then use Theorem 1 with $R = \{I_1, I_2\}$, etc. At a certain stage, say when $R = \{I_1, I_2, \dots, I_k\}$, $v_i \geq v_0 + s_i$ for all $i \notin R$. At this state $R = R^*$ (and $\mathbf{u} = \mathbf{v}$) because of the following argument: It is already known that the optimal decision for states I_1, I_2, \dots, I_k is to replace the item. Thus, one can consider only the subfamily of replacement policies in which there is only the replacement option for states I_1, I_2, \dots, I_k . Within this subfamily there exists no improvement to the policy which replaces the item if and only if the state belongs to the present replacement set R . Thus, the present policy has to be optimal.

The proof of Theorem 1 is based on the following lemma:

LEMMA 1. *Let R' be any subset of $\{1, 2, \dots, N\}$ and consider the following two sets of equations*

$$(5) \quad v_i = \begin{cases} q_i + \beta \sum_{j=0}^N P_{ij} v_j & \text{for } i \notin R', \\ v_0 + s_i - b_i & \text{for } i \in R' \end{cases}$$

where

$$(6) \quad b_i \leq c \quad \text{and} \quad c > 0$$

and

$$(7) \quad u_i = \begin{cases} q_i + \beta \sum_{j=0}^N P_{ij} u_j & \text{for } i \notin R', \\ u_0 + s_i & \text{for } i \in R'. \end{cases}$$

Define

$$(8) \quad w_i = u_i - v_i$$

and assume that

$$(9) \quad w_i \geq 0 \quad \text{for all } i = 0, 1, \dots, N.$$

Then

$$(10) \quad \begin{aligned} w_i &\leq w_0 + c \quad \text{for all } i = 1, \dots, N, \\ w_i &< w_0 + c \quad \text{for all } i \notin R'. \end{aligned}$$

Proof. Subtract (5) from (7) and obtain

$$(11) \quad w_i = \beta \sum_{j=0}^N P_{ij} w_j \quad \text{for } i \notin R',$$

and

$$(12) \quad w_i = w_0 + b_i \quad \text{for } i \in R'.$$

It immediately follows from (6) and (12) that (10) holds for $i \in R'$. We now show that (10) holds for $i \notin R'$. Denote

$$w_L = \max_{i \notin R'} w_i.$$

Now, if (10) were not true, then one would have $w_L \geq w_0 + c$ but this would contradict (11) because the equation

$$w_L = \beta \sum_{j=0}^N P_{Lj} w_j$$

would imply that $w_L (w_L > 0)$ is equal to $\beta (0 < \beta < 1)$ multiplied by a convex combination of nonnegative numbers not exceeding w_L , which is clearly impossible. Thus $w_L < w_0 + c$. Q.E.D.

The proof of Theorem 1 now follows:

Proof of Theorem 1. Denote

$$(13) \quad c = v_0 + s_I - v_I > 0$$

(see (4)). Note that for any i , and in particular for $i \in R^*$

$$(14) \quad v_0 + s_i - v_i \leq v_0 + s_I - v_I = c.$$

Let R^* be the optimal replacement set. Since $R \subset R^*$ then it follows from (3) and (14) that the value vector \mathbf{v} which corresponds to the replacement set R satisfies

$$(15) \quad v_i = \begin{cases} q_i + \beta \sum_{j=0}^N P_{ij} v_j, & i \notin R^*, \\ v_0 + s_i - b_i, & i \in R^*, \end{cases}$$

where $b_i \leq c$ and $c > 0$.

Denote the value vector which corresponds to the optimal replacement set R^* by \mathbf{u} . Obviously, $\mathbf{u} \geq \mathbf{v}$ and \mathbf{u} satisfies

$$(16) \quad u_i = \begin{cases} q_i + \beta \sum_{j=0}^N P_{ij} u_j, & i \notin R^*, \\ u_0 + s_i, & i \in R^*. \end{cases}$$

Using Lemma 1 it follows that $u_i - v_i < u_0 - v_0 + c$ for all $i \notin R^*$. Now, if $I \notin R^*$ then, by (13)

$$u_I < u_0 + v_I - v_0 + c = u_0 + s_I,$$

which would contradict the optimality condition (2). Thus, $I \in R^*$. Q.E.D.

The algorithm can be concisely described by the flow chart in Fig. 1.

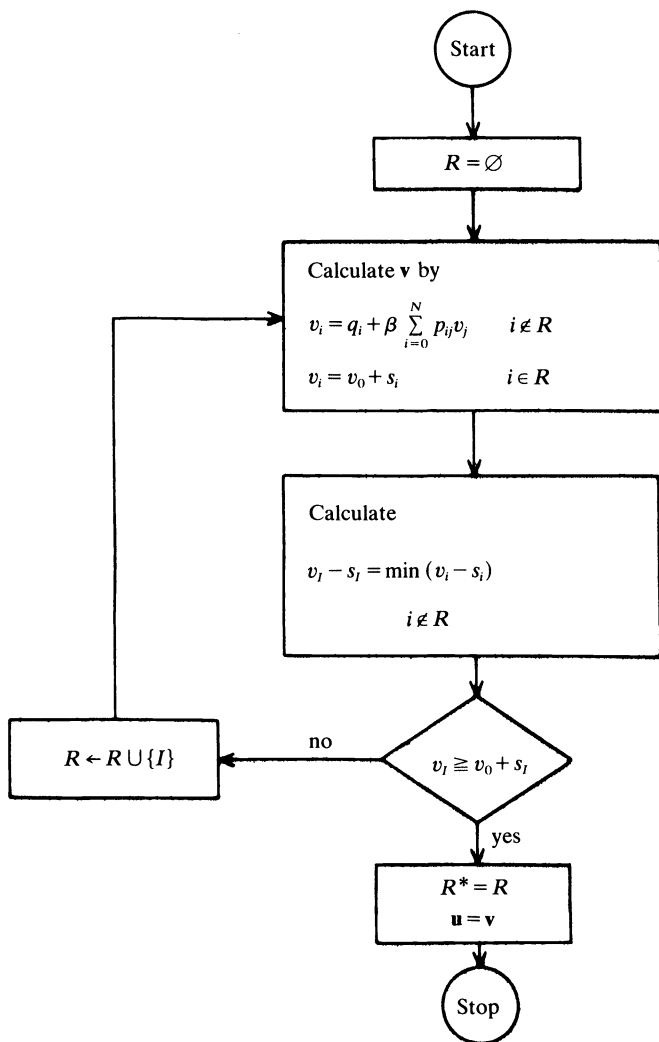


FIG. 1

It should be noted that in each iteration of the algorithm, except for the first one, the set of linear equations to be solved is different from the previous one only by one equation. Thus, using the calculations of the previous iteration, e.g., the inverse of the matrix of coefficients of v_i , makes it possible to obtain the new solution by only $O(N^2)$ arithmetic operations, rather than $O(N^3)$ operations practically needed for solving a set of N linear equations. (Such an idea for updating the inverse of a matrix which is changed by one column is used in the simplex method of linear programming, see e.g., Simmonard [4, p. 78].) Thus, the overall number of arithmetic operations required by

our algorithm is $O(N^3)$. Since any algorithm has to determine the value of at least one policy, which involves solving a set of N linear equations, it follows that our algorithm is the best one achievable from the complexity point of view.

4. Examples and remarks. The application of the algorithm described in the previous chapter to a simple example is now presented. In this example $N = 3$,

$$\mathbf{q} = \begin{pmatrix} 40 \\ 10 \\ 10 \\ 55 \end{pmatrix}, \quad P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & .5 & .5 \\ 0 & .9 & .1 & 0 \\ .1 & 0 & 0 & .9 \end{pmatrix},$$

$\beta = .9$ and $s_i = 0$, $i = 1, 2, 3$.

Starting with $R = \emptyset$ the corresponding value vector is

$$\mathbf{v} = \begin{pmatrix} 379.5 \\ 377.2 \\ 346.7 \\ 469.2 \end{pmatrix}$$

and $I = 2$. Since $v_2 < v_0$ the algorithm proceeds with $R = 2$. The corresponding value vector is now

$$\begin{pmatrix} 412.3 \\ 413.7 \\ 412.3 \\ 484.8 \end{pmatrix}.$$

Since $I = 1$ and $v_1 > v_0$ it follows that $R^* = \{2\}$ so that the optimal policy is to replace the item only at state 2.

Note that starting with the policy $R = \emptyset$, the usual policy iteration algorithm would next choose $R = \{1, 2\}$ and would arrive at the optimal replacement set $R^* = \{2\}$ only at the next iteration.

Remark 1. It should be noted that, in general, a small value of $v_I - v_0 - s_I$ does not imply that there exists a policy with $I \notin R$ which is "close" to the optimum. In order to demonstrate this fact consider the following simple example

$$\mathbf{q} = \begin{pmatrix} 1 + \varepsilon \\ 1 \end{pmatrix}, \quad P = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad s_1 = 0.$$

Starting with $R = \emptyset$ yields

$$v_1 = \frac{1}{1 - \beta}, \quad v_0 = \varepsilon + \frac{1}{1 - \beta}, \quad \text{thus } v_1 = v_0 - \varepsilon.$$

On the other hand $R = \{1\}$ yields a value vector with $v_1 = (1 + \varepsilon)/(1 - \beta)$ and $v_0 = (1 + \varepsilon)/(1 - \beta)$. Thus, the gain over the initial policy is $\varepsilon/(1 - \beta)$ which can be large if β is close to 1.

It follows that, in general, one has to continue the iterations until $v_I - v_0 - s_I \leq 0$.

Remark 2. The optimal stopping time problem is a special case of the replacement problem considered in this paper. In this problem $q_0 = 0$ and $P_{00} = 1$. For all the other states $i = 1, \dots, N$ the decision maker can either stop the process and obtain a profit of s_i , or continue for at least one more time period. All the results that we obtained

remain valid with v_0 replaced by 0 and the same algorithm (in a simpler version) described in Fig. 1 leads to the optimal stopping rule.

In the case of optimal stopping with no discounting: assuming that the value vector \mathbf{u} is finite then it satisfies (1) with $\beta = 1$ and the optimality condition is (2) with $\beta = 1$. Thus, the same algorithm can be used. (Only a slight modification in the proof of Theorem 1 is needed.)

Remark 3. Our algorithm is a modification of Howard's policy iteration method. It is easy to show (using, for example, a proof similar to the one presented by Howard [2]) that our algorithm improves the value vector at each iteration.

Remark 4. Theorem 1 can be improved as follows: Let $v_j - s_j < v_I - s_I + (u_j - v_j)(1 - \beta)/\beta$. Then $j \in R^*$.

Proof. It can actually be obtained from Lemma 1 that $w_L \leq \beta(w_0 + c)$. Thus if $j \notin R^*$ then $u_j - v_j \leq \beta(u_0 - v_0 + c) = \beta(u_0 + s_I - v_I)$, hence $v_j - s_j < (u_j - v_j)(1 - \beta)/\beta + v_I - s_I \leq (1 - \beta)(u_0 + s_I - v_I) + v_I - s_I = u_0 - \beta(u_0 + s_I - v_I) \leq u_0 + v_j - u_j \leq v_j - s_j$. Q.E.D.

5. The replacement problem with no discounting. It is now shown that a simple modification to the algorithm yields the optimal replacement policy in the case that the goal is to maximize the long-run average gain per unit time, denoted by g . For convenience of presentation it is assumed that the process is completely ergodic. In this case the benefit of a policy which corresponds to the replacement set R^* is determined by the following set of equations:

$$(1') \quad \begin{aligned} g^* + u_i &= q_i + \sum_{j=0}^N P_{ij} u_j \quad \text{for } i \notin R^*, \\ u_i &= u_0 + s_i \quad \text{for } i \in R^*. \end{aligned}$$

Here u_i $i = 0, 1, \dots, N$ denote the relative value of state i . Equations (1') can be solved for g^*, u_1, \dots, u_N by setting a fixed value for u_0 (e.g., $u_0 = 0$). The optimality condition of the above policy is

$$(2') \quad \begin{aligned} u_i &\geq u_0 + s_i \quad \text{for } i \notin R^*, \\ g^* + u_0 + s_i &\geq q_i + \sum_{j=0}^N P_{ij} u_j \quad \text{for } i \in R^*. \end{aligned}$$

The modified algorithm is based on the following theorem.

THEOREM 1'. Let $R \subset R^*$. Determine g, v_1, v_2, \dots, v_N by

$$(3') \quad \begin{aligned} g + v_i &= q_i + \sum_{j=0}^N P_{ij} v_j \quad \text{for } i \notin R, \\ v_i &= v_0 + s_i \quad \text{for } i \in R, \end{aligned}$$

setting $v_0 = 0$.

Denote

$$(4') \quad v_I - s_I = \min_{i \notin R} (v_i - s_i).$$

If $v_I < v_0 + s_I$ then I belongs to the optimal replacement set R^* .

Using Theorem 1' yields an iterative algorithm of the same type described in § 3.

The proof of Theorem 1' is based on the following lemma:

LEMMA 1'. Let R' be any subset of $\{1, 2, \dots, N\}$ and consider the following two sets of equations

$$(5') \quad \begin{aligned} g + v_i &= q_i + \sum_{j=0}^N P_{ij}v_j \quad \text{for } i \notin R', \\ v_i &= v_0 + s_i - b_i \quad \text{for } i \in R' \end{aligned}$$

where

$$(6') \quad b_i \leq c \quad \text{and} \quad c > 0$$

and

$$(7') \quad \begin{aligned} g^* + u_i &= q_i + \sum_{j=0}^N P_{ij}u_j \quad \text{for } i \notin R', \\ u_i &= u_0 + s_i \quad \text{for } i \in R'. \end{aligned}$$

Define

$$(8') \quad w_i = u_i - v_i.$$

Assume that

$$(9') \quad g^* - g > 0;$$

then

$$(10') \quad \begin{aligned} w_i &\leq w_0 + c \quad \text{for all } i = 1, \dots, N, \\ w_i &< w_0 + c \quad \text{for } i \notin R'. \end{aligned}$$

Proof. Subtract (5') from (7') and obtain

$$(11') \quad g^* - g + w_i = \sum_{j=0}^N P_{ij}w_j \quad \text{for } i \notin R',$$

$$(12') \quad w_i = w_0 + b_i \quad \text{for } i \in R'.$$

It immediately follows that (10') holds for $i \in R'$. Denote

$$w_L = \max_{i \notin R'} w_i.$$

Then (11') implies that

$$w_L = \sum_{j=0}^N P_{Lj}w_j - (g^* - g).$$

Now, $w_L \geq w_0 + c$ would lead to a contradiction because $w_i \leq w_0 + c$ for $i \in R'$, $w_i \leq w_L$ for $i \notin R'$, and $g^* - g > 0$. Thus, $w_L < w_0 + c$ Q.E.D.

Proof of Theorem 1'. Denote

$$(13') \quad c = v_0 + s_I - v_I > 0.$$

For any $i \in R^*$

$$(14') \quad v_0 + s_i - v_i \leq v_0 + s_I - v_I = c.$$

Since $R \subset R^*$ it follows from (3') and (14') that

$$(15') \quad \begin{aligned} g + v_i &= q_i + \sum_{j=0}^N P_{ij}v_j \quad \text{for } i \notin R^*, \\ v_i &= v_0 + s_i - b_i \quad \text{for } i \in R^* \end{aligned}$$

where

$$b_i \leq c \quad \text{and} \quad c > 0.$$

Denote the average gain per step of the optimal policy by g^* and the relative value vector by \mathbf{u} . Then

$$(16') \quad \begin{aligned} g^* + u_i &= q_i + \sum_{j=0}^N P_{ij} u_j \quad \text{for } i \notin R^*, \\ u_i &= u_0 + s_i \quad \text{for } i \in R^*. \end{aligned}$$

Since $g^* > g$ one can use Lemma 1 and obtain: $u_i - v_i < u_0 - v_0 + c$ for all $i \notin R^*$. Now, if $I \notin R^*$ then by (13') $u_I < u_0 + s_i$ which would contradict the optimality condition (2'). Thus $I \in R^*$. Q.E.D.

Acknowledgment. The author is grateful to Dr. Dieter Kadelka for the improved version of Theorem 1 as appearing in § 4 Remark 4.

REFERENCES

- [1] C. Derman, *Optimal replacement rules when changes of states are Markovian*, in Mathematical Optimization Techniques, R. Bellman, ed., University of California Press, Berkeley, 1963, pp. 201–210.
- [2] R. A. Howard, *Dynamic Programming and Markov Processes*, Technology Press, Wiley, New York, 1960.
- [3] H. Luss, *Maintenance policies when deterioration can be observed by inspection*, Oper. Res., 24 (1976), pp. 359–366.
- [4] M. Simonard, *Linear Programming*, Prentice-Hall, Englewood Cliffs, NJ, 1966.

HYPERCAUSAL LINEAR OPERATORS*

ALAN HOPENWASSER†

Abstract. Several different classes of hypercausal operators are useful in linear system theory. It is the purpose of this note to provide, for each pair of classes of hypercausal operators, necessary and sufficient conditions on the Hilbert resolution space to guarantee equality of the classes. In addition, the effect of similarity transforms on each class is discussed.

Key words. causal operator, hypercausal operator, nest algebra

In linear system theory the concept of physical realizability, or causality, of an operator corresponds to the mathematical concept of a nest algebra. The reader is referred to [4] and to the bibliography cited therein for a detailed account of the rationale behind the identification of the causal operators as the operators in a nest algebra. Three separate hypercausality concepts are discussed in [4], each expressing in some fashion the notion that the present output of a system does not depend upon the present input. The strongest, strict causality, coincides with the Jacobson radical of the nest algebra. The other two are, in order of strength, strong causality (introduced in [3]) and strong strict causality. In between these two lies Larson's ideal \mathcal{R}_N^∞ . We shall define all four of these concepts below, using a single coherent scheme, and give necessary and sufficient conditions on the nest for each pair of concepts to coincide.

Throughout this paper \mathcal{H} will denote a separable Hilbert space. A *nest* (or *resolution of the identity*) is a subset of the set of orthogonal projections on \mathcal{H} which contains 0 and I , is totally ordered under the usual ordering for projections, and is closed in the strong operator topology. The pair $(\mathcal{H}, \mathcal{N})$ is called a *Hilbert resolution space* and the causal operators are, by definition, just the operators in nest algebra, $\text{Alg } \mathcal{N} = \{T \in \mathcal{B}(\mathcal{H}) \mid TP = PTP, \text{ for all } P \in \mathcal{N}\}$.

A projection E in $\mathcal{B}(\mathcal{N})$ is called an *interval from \mathcal{N}* if E can be written as $E = P - Q$, where $P, Q \in \mathcal{N}$ and $Q < P$. If E is an interval, then the projections P and Q are uniquely determined. They are called the *upper* and *lower endpoints* of E . There is a natural partial order \ll on the set of intervals from \mathcal{N} : we say that $E \ll F$ if the upper endpoint of E is a subprojection of (or equal to) the lower endpoint of F . A partition $\mathcal{P} = \{E_i\}_{i \in \mathcal{J}}$ is a family of pairwise orthogonal intervals from \mathcal{N} such that $\sum_{i \in \mathcal{J}} E_i = I$. (The sum converges in the strong operator topology over the net of finite subsets of the index set \mathcal{J} .) Since the Hilbert space is separable, the index set \mathcal{J} is always finite or countably infinite. If E and F are two orthogonal intervals from \mathcal{N} , then either $E \ll F$ or $F \ll E$; consequently, each partition \mathcal{P} is totally ordered by \ll . It is easy to construct an example of a partition with any given countable order type. If (\mathcal{P}, \ll) is order isomorphic to a subset of the integers, with the usual ordering, then we say that \mathcal{P} is an *integer-ordered* partition. If $\mathcal{P}' = \{F_j\}_{j \in \mathcal{J}}$ and $\mathcal{P} = \{E_i\}_{i \in \mathcal{J}}$ are partitions, we say that \mathcal{P}' is a refinement of \mathcal{P} and write $\mathcal{P} < \mathcal{P}'$ if each F_j is a subprojection of some E_i . This gives a partial order on the family of all partitions. Each of the three families, the set of all partitions of \mathcal{N} , the set of integer-ordered

* Received by the editors February 8, 1983, and in revised form March 1, 1983. This research was partially supported by grants from the National Science Foundation and the United States Educational Foundation (Norway).

† Department of Mathematics, University of Alabama, University, Alabama 35486.

partitions, and the set of finite partitions becomes a directed set under ordering by refinement. Each of these directed sets will serve as the index set for convergent nets used in the definition of distinct notions of hypercausality.

For finite partitions, the more customary definition of partition can be obtained by replacing the intervals in the partition by the endpoints of the intervals. For integer-ordered partitions, the endpoints of the intervals form a generalized partition, as defined in [4, Chap. 2, § C]. In each case the two approaches are equivalent; it is more convenient for us to define partitions in terms of intervals so that we can accommodate arbitrary partitions without change of notation.

If $\mathcal{P} = \{E_i\}_{i \in \mathcal{I}}$ is a partition of \mathcal{N} and $A \in \text{Alg } \mathcal{N}$ is a causal operator, let $A_{\mathcal{P}} = \sum_{i \in \mathcal{I}} E_i A E_i$. (When infinite, the sum converges in the strong operator topology over the net of finite partial sums.) For each causal operator A we thereby obtain three distinct nets of operators, depending on whether we take as the index set the finite partitions, the integer-ordered partitions, or the arbitrary partitions. A class of hypercausal operators is obtained by considering all causal operators A such that $A_{\mathcal{P}} \rightarrow 0$ with respect to one of these index sets with convergence in one of the five natural topologies on $\text{Alg } \mathcal{N}$. Fortunately (at least from the point of view of reducing the tedium), the a priori possibility that there are 15 distinct notions of hypercausality does not, in fact, occur. Indeed, there are at most five (and at least four) separate notions.

In what follows, \lim will denote convergence with respect to the norm topology and (fin)-, (int)-, or (arb)- preceding the word \lim will indicate whether the index set is the directed set of finite partitions, integer-ordered partitions or arbitrary partitions. Convergence in the strong operator topology will be denoted by $s\text{-}\lim$ and in the weak operator topology (which we shall have little cause to discuss) by $w\text{-}\lim$. The remaining two topologies, the ultrastrong and the ultraweak, yield nothing new: indeed, the strong and ultrastrong (respectively, weak and ultraweak) topologies agree on bounded sets and each of the nets $A_{\mathcal{P}}$ is bounded. The following proposition further limits the number of hypercausality concepts:

PROPOSITION 1. *Let $A \in \text{Alg } \mathcal{N}$. Then the following are equivalent:*

- (i) $(\text{fin})\text{-}s\text{-}\lim_{\mathcal{P}} A_{\mathcal{P}} = 0$.
- (ii) $(\text{int})\text{-}s\text{-}\lim_{\mathcal{P}} A_{\mathcal{P}} = 0$.
- (iii) $(\text{arb})\text{-}s\text{-}\lim_{\mathcal{P}} A_{\mathcal{P}} = 0$.

Remark. The basic facts in Proposition 1 are alluded to in [2, § 3]. Since no proof is given there, a full proof will be given here.

Proof. Let \mathcal{D} denote either the directed set of integer-ordered partitions or the directed set of arbitrary partitions. We shall show (i) \Rightarrow (ii) and (i) \Rightarrow (iii) simultaneously; the two arguments are identical and each implication is obtained by giving \mathcal{D} the appropriate interpretation. Assume that (i) holds and let $\varepsilon > 0$ and $x \in \mathcal{H}$ be given. We must prove that there exists a partition \mathcal{P} in \mathcal{D} such that if \mathcal{Q} is a partition in \mathcal{D} which refines \mathcal{P} then $\|A_{\mathcal{Q}}x\|^2 < \varepsilon$.

Let $\delta = \varepsilon(1 + \|A\|^2)^{-1}$. Let $\mathcal{P} = \{E_1, \dots, E_n\}$ be a finite partition such that for any finite refinement \mathcal{Q}' of \mathcal{P} , we have $\|A_{\mathcal{Q}'}x\|^2 < \delta$. Now suppose that $\mathcal{Q} = \{F_j\}_{j \in \mathcal{J}}$ is a partition in \mathcal{D} which refines \mathcal{P} . We will show that $\|A_{\mathcal{Q}}x\|^2 < \varepsilon$. Since $\sum_{j \in \mathcal{J}} F_j = I$, there is a finite subset $\mathcal{J}_0 \subseteq \mathcal{J}$ such that $\|(\sum_{j \notin \mathcal{J}_0} F_j)x\|^2 < \delta$. For each $j \in \mathcal{J}_0$, F_j is a subprojection of some E_i in \mathcal{P} , hence there exists a finite partition $\mathcal{G} = \{G_1, \dots, G_n\}$ such that $\{F_j | j \in \mathcal{J}_0\} \subseteq \mathcal{G}$ and \mathcal{G} refines \mathcal{P} . Therefore,

$$\sum_{j \in \mathcal{J}_0} \|F_j A F_j x\|^2 \leq \sum_{i=1}^n \|G_i A G_i x\|^2 = \|A_{\mathcal{G}}x\|^2 < \delta.$$

Consequently, we have

$$\begin{aligned}\|A_{\mathcal{Q}}x\|^2 &= \sum_{j \in \mathcal{J}} \|F_j A F_j x\|^2 = \sum_{j \in \mathcal{J}_0} \|F_j A F_j x\|^2 + \sum_{j \notin \mathcal{J}_0} \|F_j A F_j x\|^2 \\ &< \delta + \sum_{j \notin \mathcal{J}_0} \|A\|^2 \|F_j x\|^2 = \delta + \|A\|^2 \sum_{j \notin \mathcal{J}_0} \|F_j x\|^2 \\ &< \delta + \|A\|^2 \delta = \varepsilon.\end{aligned}$$

This completes the proof that (i) \Rightarrow (ii) and (i) \Rightarrow (iii).

The converses, (ii) \Rightarrow (i) and (iii) \Rightarrow (i) will also be proven simultaneously. So assume either (ii) or (iii), i.e. assume $s\text{-}\lim_{\mathcal{P}} A_{\mathcal{P}} = 0$, where $\mathcal{P} \in \mathcal{D}$. We must prove that $(\text{fin})\text{-}s\text{-}\lim_{\mathcal{P}} A_{\mathcal{P}} = 0$. Let $\varepsilon > 0$ and $x \in \mathcal{H}$ be given. We must find a finite partition \mathcal{P} such that $\|A_{\mathcal{P}}x\|^2 < \varepsilon$ for any finite partition \mathcal{P}' which refines \mathcal{P} .

Let $\mathcal{Q} = \{E_i\}_{i \in \mathcal{I}}$ be a partition in \mathcal{D} such that for any refinement \mathcal{Q}' in \mathcal{D} , $\|A_{\mathcal{Q}'}x\|^2 < \delta$. There exists a finite subset $\mathcal{J}_0 \subseteq \mathcal{I}$ such that

$$\left\| \left(\sum_{i \notin \mathcal{J}_0} E_i \right) x \right\|^2 = \sum_{i \notin \mathcal{J}_0} \|E_i x\|^2 < \delta.$$

Let \mathcal{P} be the finite partition obtained by arranging in order the right and left endpoints of the intervals E_i , $i \in \mathcal{J}_0$ and taking successive differences. \mathcal{P} is, in fact, the smallest finite partition such that $\{E_i | i \in \mathcal{J}_0\} \subseteq \mathcal{P}$. Let $\mathcal{P}' = \{G_1, \dots, G_n\}$ be any finite partition which refines \mathcal{P} . We shall prove that $\|A_{\mathcal{P}'}x\|^2 < \varepsilon$.

Now, every projection G_j is either a subprojection of some E_i with $i \in \mathcal{J}_0$ or is orthogonal to each E_i with $i \in \mathcal{J}_0$. Let

$$\mathcal{J}_0 = \{j | G_j \leq E_i, \text{ for some } i \in \mathcal{J}_0\}$$

and

$$\mathcal{J}_1 = \{j | G_j E_i = 0, \text{ for all } i \in \mathcal{J}_0\} = \left\{ j \mid G_j \leq \sum_{i \notin \mathcal{J}_0} E_i \right\}.$$

Then $\mathcal{J}_0 \cap \mathcal{J}_1 = \emptyset$ and $\mathcal{J}_0 \cup \mathcal{J}_1 = \{1, 2, \dots, n\}$. Let \mathcal{Q}' be a partition in \mathcal{D} which is a common refinement of \mathcal{Q} and \mathcal{P}' and has the property that $\{G_j | j \in \mathcal{J}_0\} \subseteq \mathcal{Q}'$. (Such refinements exist since every G_j with $j \in \mathcal{J}_0$ is a subprojection of some E_i in \mathcal{P} .) Since \mathcal{Q}' refines \mathcal{Q} , we have $\|A_{\mathcal{Q}'}x\| < \delta$. But $\{G_j | j \in \mathcal{J}_0\} \subseteq \mathcal{Q}'$; so we obtain $\sum_{j \in \mathcal{J}_0} \|G_j A G_j x\|^2 < \delta$. Therefore,

$$\begin{aligned}\|A_{\mathcal{P}'}x\|^2 &= \sum_{j=1}^n \|G_j A G_j x\|^2 \\ &= \sum_{j \in \mathcal{J}_0} \|G_j A G_j x\|^2 + \sum_{j \in \mathcal{J}_1} \|G_j A G_j x\|^2 \\ &< \delta + \|A\|^2 \sum_{j \in \mathcal{J}_1} \|G_j x\|^2 \\ &= \delta + \|A\|^2 \left\| \left(\sum_{j \in \mathcal{J}_1} G_j \right) x \right\|^2 \\ &= \delta + \|A\|^2 \left\| \left(\sum_{j \notin \mathcal{J}_0} E_i \right) x \right\|^2 \\ &< \delta + \|A\|^2 \delta = \varepsilon.\end{aligned}$$

This ends the proof of the proposition.

Remark. Essentially the same argument as the one given above, with a few fairly routine modifications, proves the equivalence of the three conditions $(\text{fin})\text{-}w\text{-}\lim_{\mathcal{P}} A_{\mathcal{P}} = 0$, $(\text{int})\text{-}w\text{-}\lim_{\mathcal{P}} A_{\mathcal{P}} = 0$, and $(\text{arb})\text{-}w\text{-}\lim_{\mathcal{P}} A_{\mathcal{P}} = 0$, for $A \in \text{Alg } \mathcal{N}$. It is also immediate that, for $A \in \text{Alg } \mathcal{N}$, $s\text{-}\lim_{\mathcal{P}} A_{\mathcal{P}} = 0$ implies $w\text{-}\lim_{\mathcal{P}} A_{\mathcal{P}} = 0$. It is not known if, and perhaps not likely that, the converse holds. The three possible limits in the uniform topology are, in general, distinct. They are however easier to work with than the strong or weak topology limits. The reason is that if \mathcal{P} is any partition which refines \mathcal{Q} , then $\|A_{\mathcal{P}}\| \leq \|A_{\mathcal{Q}}\|$. Thus, for each of the three nets, to prove $\lim_{\mathcal{P}} A_{\mathcal{P}} = 0$, it is sufficient to find, for $\varepsilon > 0$ given, an appropriate partition \mathcal{P} such that $\|A_{\mathcal{P}}\| < \varepsilon$.

DEFINITION. Let \mathcal{N} be a nest. Define the following families of causal operators:

$$\mathcal{R}_{\mathcal{N}} = \{A \in \text{Alg } \mathcal{N} \mid (\text{fin})\text{-}\lim_{\mathcal{P}} A_{\mathcal{P}} = 0\},$$

$$\mathcal{R}_{\mathcal{N}}^{\text{int}} = \{A \in \text{Alg } \mathcal{N} \mid (\text{int})\text{-}\lim_{\mathcal{P}} A_{\mathcal{P}} = 0\},$$

$$\mathcal{R}_{\mathcal{N}}^{\infty} = \{A \in \text{Alg } \mathcal{N} \mid (\text{arb})\text{-}\lim_{\mathcal{P}} A_{\mathcal{P}} = 0\},$$

$$\mathcal{S}_{\mathcal{N}} = \{A \in \text{Alg } \mathcal{N} \mid (\text{fin})\text{-}s\text{-}\lim_{\mathcal{P}} A_{\mathcal{P}} = 0\}.$$

In view of the remark above, $\mathcal{R}_{\mathcal{N}}$ consists of exactly those causal operators which satisfy Ringrose's criterion for membership in the radical of $\text{Alg } \mathcal{N}$ [6], thus $\mathcal{R}_{\mathcal{N}}$ is precisely the Jacobson radical of $\text{Alg } \mathcal{N}$. $\mathcal{R}_{\mathcal{N}}^{\text{int}}$ is exactly the set of *strongly strictly causal* operators, as defined in [4, Chap. 2, § B]. As shown in [4], $\mathcal{R}_{\mathcal{N}}^{\text{int}}$ is a uniformly closed two-sided ideal in $\text{Alg } \mathcal{N}$. $\mathcal{R}_{\mathcal{N}}^{\infty}$, another uniformly closed two-sided ideal, was introduced by Larson in [5] and plays an important role in the study of similarities of nest algebras. $\mathcal{S}_{\mathcal{N}}$ is the uniformly closed left ideal of *strongly causal* operators, as defined in [3] or [4]. With the aid of Proposition 1, the relation between the strongly strictly causal operators and the strongly causal operators now becomes clear: $\mathcal{R}_{\mathcal{N}}^{\text{int}} \subseteq \mathcal{S}_{\mathcal{N}}$. Indeed, in view of the remark above, the following relations are evident:

$$\mathcal{R}_{\mathcal{N}} \subseteq \mathcal{R}_{\mathcal{N}}^{\text{int}} \subseteq \mathcal{R}_{\mathcal{N}}^{\infty} \subseteq \mathcal{S}_{\mathcal{N}}.$$

These relations, the further fact that each one may be proper, and information about several related ideals can all be found in [2].

Propositions 2, 3 and 5 below will provide appropriate necessary and sufficient conditions on the nest \mathcal{N} to ensure that each containment is, in fact, an equality.

Each of the four ideals above can be viewed as the operators which have, in an appropriate sense, zero diagonal part. The *diagonal* of a nest algebra is the subalgebra $\text{Alg } \mathcal{N} \cap (\text{Alg } \mathcal{N})^*$; the operators in the diagonal are the *memoryless* operators. If $A \in \text{Alg } \mathcal{N}$ and the net $A_{\mathcal{P}}$ is convergent in any of the senses above, then the limit, D , commutes with each projection in \mathcal{N} . Thus the limit, when it exists, is in the diagonal and may be thought of as the diagonal (or memoryless) part of A . In this case, of course, $A - D$ belongs to the ideal which corresponds to the sense in which the net converges.

It is instructive to look, in particular, at the behavior “at atoms”. An *atom* is an interval $E = P - Q$ from \mathcal{N} where Q is the immediate predecessor of P in the order of the nest. Suppose that E is an atom and that $A \in \mathcal{S}_{\mathcal{N}}$. Let x be any vector in E . If \mathcal{P} is any partition which contains E , then $A_{\mathcal{P}}x = EAEx$. Since every partition has a refinement which contains E , we see that $\|EAEx\| < \varepsilon$, for every $\varepsilon > 0$. Thus $EAE = 0$. We may view EAE as the part of the diagonal of A corresponding to the atom E . In particular, suppose that \mathcal{N} is *purely atomic*, i.e. that $I = \sum_{i \in \mathcal{I}} E_i$, where $\mathcal{P} = \{E_i\}_{i \in \mathcal{I}}$ is

the set of atoms from \mathcal{N} . In the net of arbitrary partitions of \mathcal{N} , \mathcal{P} is the terminal element. Therefore, $(\text{arb})\text{-lim}_{\mathcal{P}} A_{\mathcal{P}}$ and $(\text{arb})\text{-s-lim}_{\mathcal{P}} A_{\mathcal{P}}$ always exist and both are equal to $\sum_{i \in \mathcal{P}} E_i A E_i$. Thus, for a totally atomic nest \mathcal{N} , $\mathcal{R}_{\mathcal{N}}^{\infty} = \mathcal{S}_{\mathcal{N}}$ and each consists of the causal operators with diagonal part zero, i.e. $A \in \mathcal{R}_{\mathcal{N}}^{\infty} = \mathcal{S}_{\mathcal{N}}$ if, and only if, $EAE = 0$ for every atom E from \mathcal{N} . We now proceed to the propositions which classify when the various ideals of hypercausal operators are equal.

PROPOSITION 2. *Let \mathcal{N} be a nest. The following are equivalent:*

- (i) *0 has an immediate successor and I has an immediate predecessor.*
- (ii) $\mathcal{R}_{\mathcal{N}} = \mathcal{R}_{\mathcal{N}}^{\text{int}}$.

Proof. Assume (i) holds. It is then clear that any partition \mathcal{P} of \mathcal{N} must have a first and a last element with respect to the order \ll . If \mathcal{P} is integer-ordered, then \mathcal{P} is necessarily finite. Thus the directed set of finite partitions coincides with the directed set of integer ordered partitions and so $\mathcal{R}_{\mathcal{N}} = \mathcal{R}_{\mathcal{N}}^{\text{int}}$.

Now assume that (i) does not hold. Suppose, for example, that I has no immediate predecessor (the argument is essentially the same if 0 has no immediate successor). Then there is an increasing sequence $0 < P_1 < P_2 < \dots$ of projections in \mathcal{N} which converges strongly to I . Let $E_1 = P_1$ and $E_i = P_i - P_{i-1}$, for $i \geq 2$. Then $\mathcal{P} = \{E_i\}_{i \in \mathbb{N}}$ is an integer-ordered partition. Let x_i be a unit vector in E_i , for each i , and let $A = \sum_{i=2}^{\infty} x_i \otimes x_{i-1}$. (The rank one operator $x_i \otimes x_{i-1}$ is defined by $(x_i \otimes x_{i-1})y = \langle y, x_i \rangle x_{i-1}$. It is easy to check that the infinite sum converges in the strong operator topology.) Since any integer-ordered partition possesses an integer-ordered refinement which is also a refinement of \mathcal{P} and since $A_{\mathcal{P}} = 0$, it is clear that $A \in \mathcal{R}_{\mathcal{N}}^{\text{int}}$. On the other hand, if \mathcal{Q} is a finite partition and if E is the last interval in \mathcal{Q} (namely, the interval which has I as its upper endpoint), then $EE_i = E_i$ for all i greater than some integer i_0 . Therefore $\|EAE\| = 1$ and so $\|A_{\mathcal{Q}}\| = 1$. Thus $A \notin \mathcal{R}_{\mathcal{N}}$ and so $\mathcal{R}_{\mathcal{N}} \neq \mathcal{R}_{\mathcal{N}}^{\text{int}}$.

PROPOSITION 3. *Let \mathcal{N} be a nest. The following are equivalent:*

- (i) *Each element of \mathcal{N} excepting 0 and I has an immediate predecessor and an immediate successor.*
- (ii) $\mathcal{R}_{\mathcal{N}}^{\text{int}} = \mathcal{R}_{\mathcal{N}}^{\infty}$.

Proof. Condition (i) is equivalent to the statement that \mathcal{N} is order isomorphic to a subset of the extended integers, $\{-\infty\} \cup \mathbb{Z} \cup \{\infty\}$. When this holds \mathcal{N} is totally atomic and the set of atoms, \mathcal{P} , forms an integer-ordered partition. Furthermore, \mathcal{P} is the terminal element in the directed net of arbitrary partitions; in particular, every partition is integer-ordered. Therefore $\mathcal{R}_{\mathcal{N}}^{\text{int}} = \mathcal{R}_{\mathcal{N}}^{\infty}$ whenever (i) holds.

The proof that (ii) implies (i) is, in spirit, similar to the proof of the preceding proposition. Suppose that $P \neq 0$, I is an element of \mathcal{N} with no immediate successor. (An analogous argument works if P has no immediate predecessor.) Then there is a sequence P_n of projections in \mathcal{N} such that $P_1 = I$, $P_n > P_{n+1}$, for all n , and $\lim_{n \rightarrow \infty} P_n = P$. Let $E_n = P_n - P_{n+1}$, for all n . Let x_n be a unit vector in E_n and let $A = \sum_{n=1}^{\infty} x_n \otimes x_{n+1}$. Then $A \in \text{Alg } \mathcal{N}$ and $\|A\| = 1$. Also note that if Q is any projection in \mathcal{N} which is greater than P , then $\|(Q - P)A(Q - P)\| = 1$.

The set of intervals $\mathcal{P} = \{E_n | n = 1, 2, \dots\} \cup \{P\}$ is a partition of \mathcal{N} and it is easy to check that $A_{\mathcal{P}} = 0$. Thus $A \in \mathcal{R}_{\mathcal{N}}^{\infty}$. On the other hand, if $Q = \{F_n\}$ is an integer-ordered partition, then there is an integer k such that $F_k = Q_k - R_k$ with $Q_k, R_k \in \mathcal{N}$ and $R_k \leq P < Q_k$. Therefore $\|F_k A F_k\| = 1$, hence $\|A_{\mathcal{Q}}\| = 1$. Since \mathcal{Q} is an arbitrary integer-ordered partition, we see that $A \notin \mathcal{R}_{\mathcal{N}}^{\text{int}}$. Thus (ii) \Rightarrow (i).

Remark. The known fact that $\mathcal{R}_{\mathcal{N}} = \mathcal{R}_{\mathcal{N}}^{\infty}$ if, and only if \mathcal{N} is a finite nest also follows from Propositions 2 and 3.

Proposition 5 will characterize the nests for which $\mathcal{R}_{\mathcal{N}}^{\infty} = \mathcal{S}_{\mathcal{N}}$. The most essential ingredient is contained in the lemma below. This lemma was proven by Erdos in [2].

For the sake of self-containment, we will provide a (variant) proof. A *continuous nest* is a nest which has no atoms. Every continuous nest is order isomorphic to the interval $[0, 1]$. (Indeed, if x is a separating vector for the abelian von Neumann algebra generated by \mathcal{N} , then the mapping $P \rightarrow \langle Px, x \rangle$ is an order isomorphism of \mathcal{N} onto $[0, 1]$.) Thus when \mathcal{N} is continuous, we may use $[0, 1]$ as index set for the elements of \mathcal{N} .

LEMMA 4 (Erdos). *If \mathcal{N} is a continuous nest, then $\mathcal{R}_{\mathcal{N}}^{\infty}$ is a proper subset of $\mathcal{S}_{\mathcal{N}}$.*

Proof. Let $\mathcal{N} = \{P_r | r \in [0, 1]\}$ be a continuous nest. Enumerate the rational numbers in $(0, 1)$, i.e. write $Q \cap (0, 1)$ as a sequence $\{r_k | k = 1, 2, 3, \dots\}$. We will choose by induction two sequences, $(t_n)_{n=1,2,\dots}$ and $(\varepsilon_n)_{n=1,2,\dots}$ with the following properties:

- (i) $0 < \varepsilon_n < 1/2^{n+1}$, for all n .
- (ii) The intervals $[t_n - \varepsilon_n, t_n + \varepsilon_n]$, $n = 1, 2, 3, \dots$ are pairwise disjoint subintervals of $[0, 1]$.
- (iii) Each $t_n \in Q \cap (0, 1)$. If $t_n = r_h$ and $j < h$, then

$$r_j \in \bigcup_{i < n} [t_i - \varepsilon_i, t_i + \varepsilon_i].$$

Indeed, let $t_1 = r_1$ and let $\varepsilon_1 < 1/4$ be sufficiently small that $[t_1 - \varepsilon_1, t_1 + \varepsilon_1] \subseteq [0, 1]$. Suppose t_1, \dots, t_{n-1} and $\varepsilon_1, \dots, \varepsilon_{n-1}$ have been chosen satisfying (1)–(3). Let h be the smallest integer such that $r_h \notin \bigcup_{i=1}^{n-1} [t_i - \varepsilon_i, t_i + \varepsilon_i]$ and let $t_n = r_h$. Since the complement of $\bigcup_{i=1}^{n-1} [t_i - \varepsilon_i, t_i + \varepsilon_i]$ in $[0, 1]$ is open and $t_n \neq 0, 1$, there is a number ε_n such that $0 < \varepsilon_n < 1/2^{n+1}$ and $[t_n - \varepsilon_n, t_n + \varepsilon_n]$ is disjoint from $\bigcup_{i=1}^{n-1} [t_i - \varepsilon_i, t_i + \varepsilon_i]$.

For each pair of numbers $r, s \in [0, 1]$ with $r < s$, let $E[r, s] = P_s - P_r$. For each $n = 1, 2, \dots$, let x_n be a unit vector in $E[t_n, t_n + \varepsilon_n]$ and let y_n be a unit vector in $E[t_n - \varepsilon_n, t_n]$. Let $A = \sum_{n=1}^{\infty} x_n \otimes y_n$. It is easy to check that the sum converges in the strong operator topology, that $A \in \text{Alg } \mathcal{N}$ and that $\|A\| = 1$. We shall finish the proof by showing that $A \in \mathcal{S}_{\mathcal{N}}$ and $A \notin \mathcal{R}_{\mathcal{N}}^{\infty}$.

To prove that $A \in \mathcal{S}_{\mathcal{N}}$, let $w \in \mathcal{H}$ and $\varepsilon > 0$ be given. Since the projections $E[t_n - \varepsilon_n, t_n + \varepsilon_n]$ are pairwise orthogonal, there is an integer m such that

$$\sum_{n > m} \|E[t_n - \varepsilon_n, t_n + \varepsilon_n]w\|^2 < \varepsilon.$$

Let $P = \sum_{n > m} E[t_n - \varepsilon_n, t_n + \varepsilon_n]$ and $Q = \sum_{n \leq m} E[t_n - \varepsilon_n, t_n + \varepsilon_n]$. Then P and Q are disjoint projections and $A = A(P + Q)$. Now let \mathcal{P}' be a partition which contains the intervals $E[t_n - \varepsilon_n, t_n]$ and $E[t_n, t_n + \varepsilon_n]$, $n = 1, 2, \dots, m$ among its elements. Let $\mathcal{P} = \{F_i\}_{i \in \mathcal{J}}$ be any refinement of \mathcal{P}' . We need merely show that $\|A_{\mathcal{P}}w\|^2 < \varepsilon$.

If $F_i \leq E[t_n - \varepsilon_n, t_n]$ or $F_i \leq E[t_n, t_n + \varepsilon_n]$, then $F_i A F_i = 0$. Let $\mathcal{J} = \{i \in \mathcal{J} | F_i \leq I - Q\}$. Then we have:

$$\begin{aligned} \|A_{\mathcal{P}}w\|^2 &= \sum_{i \in \mathcal{J}} \|F_i A F_i w\|^2 = \sum_{i \in \mathcal{J}} \|F_i A (P + Q) (I - Q) F_i w\|^2 \\ &= \sum_{i \in \mathcal{J}} \|F_i A F_i P w\|^2 \leq \sum_{i \in \mathcal{J}} \|F_i P w\|^2 \\ &\leq \|P w\|^2 < \varepsilon. \end{aligned}$$

To show that $A \notin \mathcal{R}_{\mathcal{N}}^{\infty}$, we shall prove that $\|A_{\mathcal{P}}\| = 1$ for any partition \mathcal{P} . So let $\mathcal{P} = \{F_i\}$ be a partition. Each element F_k in \mathcal{P} is of the form $E[l_k, h_k]$, for uniquely determined elements l_k, h_k in $[0, 1]$. By the choice of the ε_n , the set $\bigcup_{n=1}^{\infty} [t_n - \varepsilon_n, t_n + \varepsilon_n]$ has Lebesgue measure strictly smaller than 1. On the other hand, since $\sum_{k=1}^{\infty} F_k = I$, the set $\bigcup_{k=1}^{\infty} (l_k, h_k)$ has measure 1. Therefore, there exists a number $q \in [0, 1]$ and an index k such that q belongs to the open interval (l_k, h_k) , but does not belong to $\bigcup_{n=1}^{\infty} [t_n - \varepsilon_n, t_n + \varepsilon_n]$. Since $q \neq h_k$, there is a $\delta > 0$ so that $(q, q + \delta) \subseteq (l_k, h_k)$. Let r

be a rational number in the interval $(q, q + \delta)$. Since $Q \cap (0, 1) \subseteq \bigcup_{n=1}^{\infty} [t_n - \varepsilon_n, t_n + \varepsilon_n]$, r lies in some interval $[t_n - \varepsilon_n, t_n + \varepsilon_n]$. But $q \notin [t_n - \varepsilon_n, t_n + \varepsilon_n]$, so we must have $q < t_n - \varepsilon_n < r < q + \delta$. We would like to have $t_n + \varepsilon_n < q + \delta$, but this may not be true. The situation is easily rectified by repeating the procedure once again: let s be a rational number in the interval $(q, t_n - \varepsilon_n)$ and let m be such that $s \in [t_m - \varepsilon_m, t_m + \varepsilon_m]$. This time we obtain

$$q < t_m - \varepsilon_m < s < t_m + \varepsilon_m < t_n - \varepsilon_n < q + \delta.$$

In particular, $E[t_m - \varepsilon_m, t_m + \varepsilon_m] \subseteq E[q, q + \delta] \subseteq F_k$. Therefore, $F_k A F_k x_m = y_m$; in particular, $\|F_k A F_k\| = 1$. Thus $\|A_{\mathcal{P}}\| = 1$ and $A \notin \mathcal{R}_{\mathcal{N}}^{\infty}$.

PROPOSITION 5. *Let \mathcal{N} be a nest. The following are equivalent:*

- (i) \mathcal{N} is totally atomic.
- (ii) $\mathcal{R}_{\mathcal{N}}^{\infty} = \mathcal{S}_{\mathcal{N}}$.

Proof. The easy implication (i) \Rightarrow (ii) has already been given in the paragraph immediately preceding Proposition 2. So suppose that \mathcal{N} is not totally atomic; we must show $\mathcal{R}_{\mathcal{N}}^{\infty} \neq \mathcal{S}_{\mathcal{N}}$.

Let $\{E_i\}_{i \in \mathcal{I}}$ be the (possibly empty) set of atoms from \mathcal{N} . Let $E = I - \sum_{i \in \mathcal{I}} E_i$. By hypothesis, $E \neq 0$. Let \mathcal{H} be the range of the projection E . Define a nest \mathcal{N}_E on the Hilbert space \mathcal{H} by $\mathcal{N}_E = \{PE|_{\mathcal{H}} | P \in \mathcal{N}\}$. Observe that \mathcal{N}_E is a continuous nest. Each operator A in $\mathcal{B}(\mathcal{H})$ has a unique bounded linear extension to \mathcal{H} which vanishes on the orthogonal complement, \mathcal{H}^{\perp} , of \mathcal{H} . We denote this extension by \tilde{A} . Note that $A \in \text{Alg } \mathcal{N}_E$ if, and only if, $\tilde{A} \in \text{Alg } \mathcal{N}$.

Since \mathcal{N}_E is a continuous nest, $\mathcal{R}_{\mathcal{N}_E}^{\infty}$ is a proper subset of $\mathcal{S}_{\mathcal{N}_E}$. Fix an element A of $\mathcal{S}_{\mathcal{N}_E}$ which is not in $\mathcal{R}_{\mathcal{N}_E}^{\infty}$. We shall show that $\tilde{A} \in \mathcal{S}_{\mathcal{N}}$ and $\tilde{A} \notin \mathcal{R}_{\mathcal{N}}^{\infty}$.

To prove that $\tilde{A} \in \mathcal{S}_{\mathcal{N}}$, let $x \in \mathcal{H}$ and $\varepsilon > 0$ be given. If Q is a projection in \mathcal{N}_E , then one can, by adding appropriate atoms of \mathcal{N} to Q , obtain a projection P in \mathcal{N} so that $Q = PE|_{\mathcal{H}}$. If $Q_0 = 0 < Q_1 < \cdots < Q_n = I_{\mathcal{H}}$ is a finite subnest of \mathcal{N}_E , then we can obtain projections $P_0 = 0 < P_1 < \cdots < P_n = I_{\mathcal{H}}$ in \mathcal{N} so that $Q_i = P_i E|_{\mathcal{H}}$, $i = 1, \dots, n$. Since $\tilde{A}x = EAEx$, we have

$$\sum_{i=1}^n \|(P_i - P_{i-1})\tilde{A}(P_i - P_{i-1})x\|^2 = \sum_{i=1}^n \|(Q_i - Q_{i-1})A(Q_i - Q_{i-1})Ex\|^2.$$

From these remarks it is clear that $\tilde{A} \in \mathcal{S}_{\mathcal{N}}$.

Finally, we need to show that $\tilde{A} \notin \mathcal{R}_{\mathcal{N}}^{\infty}$. Assume the contrary; i.e. assume that $\tilde{A} \in \mathcal{R}_{\mathcal{N}}^{\infty}$. Let $\varepsilon > 0$ be given. Then there is a partition $\mathcal{P} = \{F_i\}_{i \in \mathcal{I}}$ of \mathcal{N} such that $\|F_i \tilde{A} F_i\| < \varepsilon$, for all i . The set $\mathcal{P}_E = \{F_i E | i \in \mathcal{I} \text{ and } F_i E \neq 0\}$ is a partition of \mathcal{N}_E and $\|F_i E A F_i E\| = \|\tilde{F}_i A \tilde{F}_i\| < \varepsilon$, for all i . Thus

$$\tilde{A} \in \mathcal{R}_{\mathcal{N}_E}^{\infty},$$

contrary to hypothesis. This completes the proof of the proposition.

Remark. From Propositions 3 and 5 and the first sentence of the proof of Proposition 3, we see that the strongly strictly causal operators and the strongly causal operators on a nest \mathcal{N} coincide if, and only if, \mathcal{N} is order isomorphic to a subset of the extended integers $\{-\infty\} \cup \mathbb{Z} \cup \{\infty\}$.

We conclude this note with a discussion of the effect of similarities on each of the classes of hypercausal operators considered above. The significance of similarities for system theory is indicated by the fact in Larson's theorem [5] that any two continuous nests are similar implies that there exist positive definite hermitian operators which do not admit spectral factorization. (See [4] for a discussion of factorization problems.)

If \mathcal{N} is a nest and T is an invertible operator in $\mathcal{B}(\mathcal{H})$, then, for each $P \in \mathcal{N}$, TPT^{-1} is an idempotent (not necessarily self-adjoint). Let $\phi_T(P)$ be the orthogonal projection on the range of TPT^{-1} . Thus, T maps the range space of P onto the range space of $\phi_T(P)$. Let $T\mathcal{N}$ denote the nest $\{\phi_T(P) | P \in \mathcal{N}\}$. We say that two nests \mathcal{M} and \mathcal{N} are similar if $\mathcal{M} = T\mathcal{N}$ for some invertible $T \in \mathcal{B}(\mathcal{H})$. The map $\phi_T: \mathcal{N} \rightarrow \mathcal{M}$ induced by T is an order isomorphism of \mathcal{N} onto \mathcal{M} . If ϕ is any order isomorphism of \mathcal{N} onto \mathcal{M} , then ϕ has a natural extension to a map from the set of intervals from \mathcal{N} to the set of intervals from \mathcal{M} : define $\phi(P - Q)$ to be $\phi(P) - \phi(Q)$. (We denote the extension by the same symbol.) In particular, atoms from \mathcal{N} correspond to atoms from \mathcal{M} . If corresponding atoms have the same dimension, we say that ϕ *preserves dimension*. It is evident that each order isomorphism of the form ϕ_T preserves dimension. Recently, Davidson [1] has proven the converse to this: if ϕ is an order isomorphism of \mathcal{N} onto \mathcal{M} which preserves dimension, then there is an invertible operator T such that $\mathcal{M} = T\mathcal{N}$ and $\phi = \phi_T$.

Fix a nest \mathcal{N} and an invertible operator T and let $\mathcal{M} = T\mathcal{N}$. Then the two nest algebras $\text{Alg } \mathcal{N}$ and $\text{Alg } \mathcal{M}$ are similar: $\text{Alg } \mathcal{N} = T^{-1}(\text{Alg } \mathcal{M})T$. Furthermore, $\mathcal{R}_{\mathcal{N}} = T^{-1}\mathcal{R}_{\mathcal{M}}T$ and $\mathcal{R}_{\mathcal{N}}^{\text{int}} = T^{-1}\mathcal{R}_{\mathcal{M}}^{\text{int}}T$, i.e. the strictly causal operators and the strongly strictly causal operators are preserved by similarities. The first of these two facts is completely trivial—it follows immediately from the definition of the radical as the intersection of the kernels of all the irreducible representations of the algebra. It follows equally rapidly from the characterization of the radical as the largest ideal consisting entirely of quasi-nilpotent elements. Yet a third proof is available: both similarity results stated above follow from a lemma of Larson [5] which asserts that if E is any interval from \mathcal{N} and $K = \|T\|\|T^{-1}\|$, then for any $A \in \text{Alg } \mathcal{N}$, $\|EAE\| \leq K\|\phi_T(E)TAT^{-1}\phi_T(E)\|$ and $\|\phi_T(E)TAT^{-1}\phi_T(E)\| \leq K\|EAE\|$. To obtain the two similarity results one need merely observe that if $\mathcal{P} = \{E_i\}_{i \in \mathcal{I}}$ is a finite or integer-ordered partition of \mathcal{N} , then $\{\phi_T(E_i)\}_{i \in \mathcal{I}}$ is a finite or integer-ordered partition of \mathcal{M} .

If \mathcal{P} is an arbitrary partition, then it is not necessarily the case that $\{\phi_T(E_i)\}$ is a partition. As a consequence, $\mathcal{R}_{\mathcal{N}}^{\infty}$ need not be preserved by similarities. A detailed discussion of this may be found in [5].

Finally, we turn to $\mathcal{S}_{\mathcal{N}}$. In light of Larson's results on $\mathcal{R}_{\mathcal{N}}^{\infty}$, it is not surprising that we find that $\mathcal{S}_{\mathcal{N}}$ need not be preserved by similarities.

Example. We use the following standard construction to produce a pair of similar nests. If μ is a finite Borel measure on $[0, 1]$, let $\mathcal{H}_{\mu} = L^2([0, 1], \mu)$. For each $t \in [0, 1]$, we let P_t^{μ} (respectively, P_{t-}^{μ}) denote the multiplication operator by the characteristic function of $[0, t]$ (respectively, $[0, t)$). Let \mathcal{N}^{μ} denote nest consisting of all the projections P_t^{μ} and P_{t-}^{μ} .

Let ν be a purely atomic measure on $[0, 1]$ with support equal to $Q \cap (0, 1)$. So, in the nest \mathcal{N}^{ν} , we find that $P_t^{\nu} \neq P_{t-}^{\nu}$ if, and only if $t \in Q \cap (0, 1)$. The nest is totally atomic and the atoms are the intervals $E_t^{\nu} = P_t^{\nu} - P_{t-}^{\nu}$, $t \in Q \cap (0, 1)$. Let m be Lebesgue measure on $[0, 1]$ and let $\lambda = m + \nu$. In the nest \mathcal{N}^{λ} the atoms are once again the intervals $E_t^{\lambda} = P_t^{\lambda} - P_{t-}^{\lambda}$, $t \in Q \cap (0, 1)$, but this time the nest is not totally atomic. Indeed, $\mathcal{H}_{\lambda} = \mathcal{H}_m \oplus \mathcal{H}_{\nu}$ and the sum of the atoms from \mathcal{N}^{λ} is the projection on \mathcal{H}_{ν} , not the identity on the whole Hilbert space \mathcal{H}_{λ} . The map $\phi: \mathcal{N}^{\nu} \rightarrow \mathcal{N}^{\lambda}$ given by $\phi(P_t^{\nu}) = P_t^{\lambda}$ and $\phi(P_{t-}^{\nu}) = P_{t-}^{\lambda}$, for all t , is an order isomorphism which preserves dimension (all atoms are one-dimensional). By Davidson's theorem [1], $\phi = \phi_T$ for some invertible operator T . So $\mathcal{N}^{\lambda} = T\mathcal{N}^{\nu}$ and $\text{Alg } \mathcal{N}^{\nu} = T^{-1}(\text{Alg } \mathcal{N}^{\lambda})T$. We shall show that $\mathcal{S}_{\mathcal{N}^{\nu}} \neq T^{-1}\mathcal{S}_{\mathcal{N}^{\lambda}}T$.

Let A be a nonzero operator in $(\text{Alg } \mathcal{N}^{\lambda}) \cap (\text{Alg } \mathcal{N}^{\lambda})^*$ with the property that $EAE = 0$ for every atom from \mathcal{N}^{λ} . (A is simply a multiplication operator by a function

$f \in L^\infty([0, 1], \lambda)$ with the property that $f(r) = 0$, for all $r \in Q \cap (0, 1)$.) Since A is memoryless, it commutes with each projection in \mathcal{N}^λ ; therefore $A_\varphi = A$ for any partition \mathcal{P} . Thus $A \notin \mathcal{S}_{\mathcal{N}}\lambda$.

Let $B = T^{-1}AT$. Then $B \in \text{Alg } \mathcal{N}^\nu$. By Larson's lemma [5], $\|FBF\| \leq \|T\| \|T^{-1}\| \|\phi_T(F)TBT^{-1}\phi_T(F)\| = \|T\| \|T^{-1}\| \|\phi_T(F)A\phi_T(F)\| = 0$, for every atom F from \mathcal{N}^ν . Since \mathcal{N}^ν is totally atomic $B \in \mathcal{S}_{\mathcal{N}^\nu}$. But $B = T^{-1}AT \notin T^{-1}\mathcal{S}_{\mathcal{N}^\lambda}T$, so $\mathcal{S}_{\mathcal{N}^\nu} \neq T^{-1}\mathcal{S}_{\mathcal{N}^\lambda}T$ as desired.

REFERENCES

- [1] K. R. DAVIDSON, *Similarity and compact perturbations of nest algebras*, J. Reine Angew. Math. to appear.
- [2] J. A. ERDOS, *Ideals of causal operators*, Preprint.
- [3] A. FEINTUCH, *Strictly and strongly strictly causal linear operators*, SIAM J. Math. Anal., 10 (1979), pp. 603–613.
- [4] A. FEINTUCH AND R. SAEKS, *System Theory: A Hilbert Space Approach*, Academic Press, New York, 1982.
- [5] D. R. LARSON, *Nest algebras and similarity transformations*, Ann. Math., to appear.
- [6] J. R. RINGROSE, *On some algebras of operators*, Proc. London Math. Soc., (3) 15 (1965), pp. 61–83.

STOCHASTIC PRODUCTION PLANNING WITH PRODUCTION CONSTRAINTS*

A. BENSOUSSAN†, S. P. SETHI‡, R. VICKSON§ AND N. DERZKO||

Abstract. This paper considers an infinite horizon stochastic production planning problem with the constraint that production rate must be nonnegative. It is shown that an optimal feedback solution exists for the problem. Moreover, this solution is characterized and is then compared with the solution of the unconstrained problem. Also obtained, by using a policy iteration procedure, are computational solutions to the related problems with upper bounds on the production rate.

Key words. production-inventory model, aggregate planning, stochastic optimal control, control constraints, policy iteration

1. Introduction. Thompson and Sethi [12] consider a production-inventory model which determines production rates over time to minimize an integral representing a discounted quadratic loss function. The loss function is defined in terms of the deviations of production and inventory levels from their rated or factory-optimal values. The model is solved both with and without nonnegative production constraints. A stochastic extension of this paper involving a white noise process [1] is analyzed by Sethi and Thompson [9], [10]. Closed-form solutions for optimal feedback production policy for both finite and infinite horizon versions of the model without production constraints are obtained. In particular, the model must allow negative production rates or disposals.

In this paper, we consider the stochastic production planning problem with the constraint that production rates must be nonnegative. Only the infinite horizon problem is treated. It is shown that an optimal feedback solution exists for the problem. This solution is characterized. Also obtained, by using a policy iteration procedure, are computational solutions to the related problems with upper bounds on the production rate.

2. The model. Consider a factory producing a homogeneous good and having an inventory warehouse. Define the following quantities:

$y(t)$ = inventory level at time t (state variable)
 $p(t)$ = production rate at time t (control variable); $p \geq 0$
 ξ = the constant demand rate at time t ; $\xi \geq 0$
 y_1 = factory-optimal inventory level
 p_1 = factory-optimal production rate
 y_0 = initial inventory level
 h = inventory holding cost coefficient
 c = production cost coefficient
 α = the constant discount rate; $\alpha > 0$
 $w(t)$ = the standard Wiener process; see [1]
 σ = the constant diffusion coefficient; see [1]

* Received by the editors June 3, 1983. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada grants A4617 and A9304.

† INRIA, Rocquencourt, France.

‡ Faculty of Management Studies, University of Toronto, Toronto, Ontario, Canada M5S 1V4.

§ Department of Management Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

|| Department of Mathematics, University of Toronto, Toronto, Ontario, Canada M5S 1V4.

We now state the conditions of the model. The first condition is the stockflow equation stated as an Itô stochastic differential equation (see [1], [4]):

$$(1) \quad dy = (p - \xi) dt + \sigma dw, \quad y(0) = y_0.$$

We note that process $dw(t)$ can be formally expressed as $z(t) dt$, where $z(t)$ is considered to be the white noise process [1]. It can be interpreted as “sales returns,” “inventory spoilage,” etc. which are random in nature. The second is the objective function:

$$(2) \quad \min_{p \geq 0} E \left\{ \int_0^\infty e^{-\alpha t} [c(p - p_1)^2 + h(y - y_1)^2] dt \right\}.$$

3. The Hamilton–Jacobi–Bellman equation. The solution of the above model will be carried out via the development of the Hamilton–Jacobi–Bellman equation satisfied by a certain “value function.” To simplify the notation, we assume that

$$(3) \quad y_1 = p_1 = 0 \quad \text{and} \quad h = c = 1.$$

This assumption results in no loss of generality as the following analysis can be extended in a parallel manner for the case without (3). With (3), we restate the stochastic production planning problem:

$$(4) \quad \min_{p \geq 0} E \left[\int_0^\infty (p^2 + y^2) e^{-\alpha t} dt \right],$$

subject to the Itô equation

$$(5) \quad dy = (p - \xi) dt + \sigma dw, \quad y(0) = y_0.$$

Let $u = u(x)$ denote the expected current-valued value of the control problem (4), (5) with initial value x in (5) so that $u(y_0)$ represents the value of the objective function in (4) subject to the state-equation (5). Then, it can be shown that $u(x)$ is \mathcal{C}^2 and satisfies the following Hamilton–Jacobi–Bellman equation; see [2], [4]:

$$(6) \quad -\frac{1}{2}\sigma^2 u'' + \xi u' + \alpha u = x^2 + \inf_{p \geq 0} (u'p + p^2).$$

This equation can be simplified by noting that the infimum of the infimand is attained at

$$(7) \quad p = \max \left[0, -\frac{u'}{2} \right] = -\frac{u'^-}{2}$$

so that equation (6) can be written as

$$(8) \quad -\frac{1}{2}\sigma^2 u'' + \xi u' + \frac{1}{4}(u'^-)^2 + \alpha u = x^2.$$

Equation (8) is known as the Hamilton–Jacobi–Bellman equation. It is this equation that we are interested in solving. In the next section, we prove that there exists a unique solution $u \geq 0$ in \mathcal{C}^2 with quadratic growth.

4. Existence and uniqueness. In order to solve (8), we define a function ϕ satisfying the linear equation

$$(9) \quad -\frac{1}{2}\sigma^2 \phi'' + \xi \phi' + \alpha \phi = x^2$$

and a function u_M satisfying

$$(10) \quad -\frac{1}{2}\sigma^2 u_M'' + \xi u_M' + \alpha u_M = x^2 + \inf_{0 \leq p \leq M} (u_M'p + p^2)$$

for $M \in [0, \infty)$. From Ladyzhenskaya and Uraltseva [8] (see also, for instance Bensoussan [2]), we know that there exists unique ϕ and u_M in \mathcal{C}^2 with quadratic growth. We remark that ϕ and u_M are the value functions of (4), (5) with $p = 0$ and $p \in [0, M]$, respectively. Note also that $u_0 = \phi$ and it is easily shown that

$$(11) \quad \phi(x) = (1/\alpha)x^2 - (2\xi/\alpha^2)x + (1/\alpha^2)[2\xi^2/\alpha + \sigma^2].$$

LEMMA 1. $0 \leq u_M \leq \phi$.

Proof. Define $\tilde{u}_M = u_M - \phi$. Clearly \tilde{u}_M is \mathcal{C}^2 with quadratic growth. Also,

$$-\frac{1}{2}\sigma^2\tilde{u}_M'' + \xi\tilde{u}_M' + \alpha\tilde{u}_M = \inf_{0 \leq p \leq M} (u_M'p + p^2) \leq 0.$$

This implies that $\tilde{u}_M \leq 0$, which completes the proof.

LEMMA 2. Let $M > N \in [0, \infty)$. Then $u_M \leq u_N$. Furthermore, $u_M \downarrow u$ with $0 \leq u \leq \phi$.

Proof. By definition,

$$(12) \quad -\frac{1}{2}\sigma^2 u_N'' + \xi u_N' + \alpha u_N = x^2 + \inf_{0 \leq p \leq N} (u_N'p + p^2).$$

Let $p_N \in [0, N]$ denote the value of p for which the infimum in (12) is attained. Clearly, $0 \leq p_N \leq N < M$. Define $\tilde{u} = u_M - u_N$. From (10) and (12), we have

$$\begin{aligned} -\frac{1}{2}\sigma^2\tilde{u}'' + \xi\tilde{u}' + \alpha\tilde{u} &= \inf_{0 \leq p \leq M} (u_M'p + p^2) - \inf_{0 \leq p \leq N} (u_N'p + p^2) \\ &\leq u_M'p_N + p_N^2 - u_N'p_N - p_N^2 \\ &= \tilde{u}'p_N. \end{aligned}$$

Thus,

$$-\frac{1}{2}\sigma^2\tilde{u}'' + (\xi - p_N)\tilde{u}' + \alpha\tilde{u} \leq 0,$$

which implies that $\tilde{u} \leq 0$. By the monotone convergence theorem, u_M has a limit $u \geq 0$. For reference, we write

$$(13) \quad u_M \downarrow u \quad \text{and} \quad 0 \leq u \leq \phi.$$

This completes the proof.

We now have a candidate u for a solution of (8). The rest of the section is devoted to proving that u of (13) satisfies (8).

Since u_M is the solution of (10), we know from Bensoussan [2] that it is the value function for the following optimal control problem.

$$\begin{aligned} (14) \quad u_M(x) &= \inf_{0 \leq p(\cdot) \leq M} E \int_0^\infty e^{-\alpha t} (p^2 + y_x^2) dt \\ &\equiv \inf_{0 \leq p(\cdot) \leq M} J_{M,x}[p(\cdot)] \end{aligned}$$

where y_x is defined as the solution of

$$(15) \quad dy = (p - \xi) dt + \sigma dw, \quad y(0) = x.$$

Furthermore, the optimal feedback policy p_M for problem (14), (15) is given by

$$(16) \quad p_M = \begin{cases} 0 & \text{if } u_M' \geq 0, \\ -u_M'/2 & \text{if } -2M < u_M' < 0, \\ M & \text{if } u_M' \leq -2M. \end{cases}$$

To obtain a uniform bound for u'_M in M , we examine the following difference for a given $p(\cdot)$.

$$\begin{aligned}
 J_{M,x}[p(\cdot)] - J_{M,x'}[p(\cdot)] &= E \int_0^\infty e^{-\alpha t} [y_x^2(t) - y_{x'}^2(t)] dt \\
 &= E \int_0^\infty e^{-\alpha t} [y_x(t) - y_{x'}(t)][y_x(t) + y_{x'}(t)] dt \\
 (17) \quad &= (x - x') E \int_0^\infty e^{-\alpha t} \left[x + x' + 2 \int_0^t (p - \xi) ds \right] dt \\
 &= \frac{(x - x')^2}{\alpha} + 2(x - x') E \int_0^\infty e^{-\alpha t} \left[\int_0^t (p - \xi) ds \right] dt \\
 &\leq 2(x - x') \left[-\xi \int_0^\infty t e^{-\alpha t} dt + E \int_0^\infty e^{-\alpha t} \left[\int_0^t p ds \right] dt \right].
 \end{aligned}$$

To be able to perform the integration in (17) by parts, we prove the following lemma.

LEMMA 3. For a given x , $p(\cdot)$ can be restricted to satisfy

$$(18) \quad E \int_0^\infty e^{-\alpha t} p^2 dt \leq C(x^2 + 1)$$

for some given C . Furthermore,

$$(19) \quad E e^{-\alpha t} \int_0^t p ds \rightarrow 0 \quad \text{as } t \rightarrow \infty.$$

Proof. Since $p=0$ is feasible, it is clear from (14) and (15) that any optimal trajectory $p(\cdot)$ will satisfy

$$E \int_0^\infty (p^2 + y_x^2) dt \leq E \int_0^\infty e^{-\alpha t} \left[x - \xi t + \int_0^t \sigma dw \right]^2 dt.$$

We can now establish (18) as follows:

$$\begin{aligned}
 E \int_0^\infty e^{-\alpha t} p^2 dt &\leq E \int_0^\infty e^{-\alpha t} \left[x - \xi t + \int_0^t \sigma dw \right]^2 dt \\
 &= E \int_0^\infty e^{-\alpha t} \left[(x - \xi t)^2 + \left(\int_0^t \sigma dw \right)^2 \right] dt \\
 &= \int_0^\infty e^{-\alpha t} (x - \xi t)^2 + \int_0^\infty e^{-\alpha t} \left(\int_0^t \sigma^2 ds \right) dt \\
 &\leq C(x^2 + 1).
 \end{aligned}$$

Using the Cauchy-Schwarz inequality since $p \geq 0$, and then the Jensen's inequality since the square root function is a concave function, we have

$$\begin{aligned}
 E e^{-\alpha t} \int_0^t p ds &\leq E e^{-\alpha t} \left[\int_0^t p^2 ds \right]^{1/2} \sqrt{t} \\
 &= e^{-\alpha t/2} \sqrt{t} E \left[e^{-\alpha t} \int_0^t p^2 ds \right]^{1/2} \\
 &\leq e^{-\alpha t/2} \sqrt{t} \left[E e^{-\alpha t} \int_0^t p^2 ds \right]^{1/2} \\
 &\leq e^{-\alpha t/2} \sqrt{t} \left[E \int_0^t e^{-\alpha s} p^2 ds \right]^{1/2}.
 \end{aligned}$$

From (18), the right-hand side goes to zero as $t \rightarrow \infty$ and, therefore, we have established (19).

LEMMA 4. *There exists a constant C independent of M , such that*

$$(20) \quad |u'_M(x)| \leq C(|x| + 1),$$

$$(21) \quad |u''_M(x)| \leq C(|x|^2 + 1).$$

Proof. In view of (19), we can integrate by parts to obtain

$$E \int_0^\infty e^{-\alpha t} \left(\int_0^t p \, ds \right) dt = E \int_0^\infty \frac{1}{\alpha} e^{-\alpha t} p \, dt = \frac{1}{\alpha} E \int_0^\infty e^{-\alpha t/2} e^{-\alpha t/2} p \, dt.$$

By using the Schwarz inequality and (18), we have

$$(22) \quad E \int_0^\infty e^{-\alpha t} \left(\int_0^t p \, ds \right) dt \leq \frac{1}{\alpha} \left[\int_0^\infty e^{-\alpha t} dt \right]^{1/2} \left[\int_0^\infty e^{-\alpha t} p^2 dt \right]^{1/2} \\ \leq C\sqrt{x^2 + 1} \leq C(|x| + 1) \leq C(|x| + |x'| + 1).$$

Using (22) in (17), we have

$$J_{M,x}[p(\cdot)] - J_{M,x'}[p(\cdot)] \leq C|x - x'|(|x| + |x'| + 1).$$

Thus,

$$|u_M(x) - u_M(x')| \leq C|x - x'|(|x| + |x'| + 1),$$

which implies (20).

To establish (21), we observe from (16) that

$$u'_M p_M + p_M^2 \geq -\frac{u'_M}{4}.$$

From the above and from (20), we have

$$0 \geq \inf_{0 \leq p \leq M} (u'_M p + p^2) \geq -\frac{u'_M}{4} \geq -C(|x| + 1),$$

which is independent of M . Thus, from (10),

$$-\frac{1}{2}\sigma^2 u''_M + \xi u'_M + \alpha u_M = x^2 - h_M(x),$$

where

$$0 \leq h_M(x) \leq C(|x| + 1).$$

This, with Lemma 1 and (20), implies (21).¹

From (13), (20), and (21), it is easy to establish the following result.

LEMMA 5. *For all x , $u'_M(x)$ converges to $u'(x)$ pointwise as $M \rightarrow \infty$.*

Proof. Indeed, use the formula

$$\frac{u'_M(x)}{1 + |x|^2} = \int_{-\infty}^x \frac{u''_M(z) \, dz}{(1 + |z|^2)^2} - 4 \int_{-\infty}^x \frac{u'_M(z) \, dz}{(1 + |z|^2)^3} \\ \rightarrow \int_{-\infty}^x \frac{u''(z) \, dz}{(1 + |z|^2)^2} - 4 \int_{-\infty}^x \frac{u'(z) \, dz}{(1 + |z|^2)^3} \quad \text{as } M \rightarrow \infty \\ = \frac{u'(x)}{1 + |x|^2}.$$

¹ This estimate is not very sharp but it is sufficient for the remainder.

In the following lemma, we show that the optimal feedback control (16) for problem (14), (15) converges to the optimal feedback control (7) of the problem under consideration.

LEMMA 6. For all x , $\inf_{0 \leq p \leq M} (u'_M p + p^2) \rightarrow -\frac{1}{4}(u'^-)^2$ pointwise as $M \rightarrow \infty$.

Proof. Fix x and select

$$M > \frac{1}{2}C(|x| + 1).$$

From (20), we have

$$2M > |u'_M(x)|.$$

Let p_M denote the unique value for which

$$\inf_{0 \leq p \leq M} (u'_M p + p^2) = u'_M p_M + p_M^2.$$

Then from (16), it is easily seen that

$$u'_M p_M + p_M^2 = -\frac{1}{4}(u'^-)^2$$

which converges to $-\frac{1}{4}(u'^-)^2$ in view of Lemma 5. This completes the proof.

We can now state the main result of this section.

THEOREM 1. The function u obtained in (13) is the value function of the optimal control problem (4, 5). Furthermore, u is the unique positive solution of the Hamilton-Jacobi equation (8) with quadratic growth.

Proof. In Lemmas 1–6, we have obtained a function u as in (13) and have shown that it is a \mathcal{C}^2 -function with quadratic growth (i.e., $0 \leq u \leq \phi$) and is a solution of (8).

From Bensoussan [2], it follows that u is the value function of the problem (4), (5). Moreover, u is the unique \mathcal{C}^2 -solution of (8) with quadratic growth since any such solution will coincide with the value function of the stochastic optimal control problem.

In the next section, we characterize the solution u .

5. Characterization of the solution. Although we are not able to obtain a closed-form for u , we can characterize its behavior.

THEOREM 2. The value function $u(x)$ is strictly convex.

Proof. First, we prove that $J_x[p(\cdot)]$ is strictly convex in x and $p(\cdot)$. Let $\theta \in (0, 1)$. Let x_1 and x_2 and $p_1(\cdot)$ and $p_2(\cdot)$ be arbitrarily chosen. Define

$$\hat{x} = \theta x_1 + (1 - \theta)x_2 \quad \text{and} \quad \hat{p} = \theta p_1 + (1 - \theta)p_2,$$

and the solutions of (5) with different initial conditions:

$$\begin{aligned} y_{\hat{x}} &= \hat{x} + \int_0^t (\hat{p} - \xi) dt + \int_0^t \sigma dw, \\ y_{x_1} &= x_1 + \int_0^t (p_1 - \xi) dt + \int_0^t \sigma dw, \\ y_{x_2} &= x_2 + \int_0^t (p_2 - \xi) dt + \int_0^t \sigma dw. \end{aligned}$$

Then, we have

$$y_{\hat{x}} = \theta y_{x_1} + (1 - \theta)y_{x_2},$$

and

$$\begin{aligned} y_{\bar{x}}^2 &= \theta^2 y_{x_1}^2 + (1-\theta)^2 y_{x_2}^2 + 2\theta(1-\theta)y_{x_1}y_{x_2} \\ &< [\theta^2 + \theta(1-\theta)]y_{x_1}^2 + [(1-\theta)^2 + \theta(1-\theta)]y_{x_2}^2 \\ &= \theta y_{x_1}^2 + (1-\theta)y_{x_2}^2. \end{aligned}$$

Similarly,

$$\hat{p}^2 < \theta p_1^2 + (1-\theta)p_2^2.$$

Now, we have

$$\begin{aligned} J_{\bar{x}}[\hat{p}(\cdot)] &= E \int_0^\infty e^{-\alpha t} [y_{\bar{x}}^2 + \hat{p}^2] dt \\ &\leq E \int_0^\infty e^{-\alpha t} [\theta y_{x_1}^2 + (1-\theta)y_{x_2}^2 + \theta p_1^2 + (1-\theta)p_2^2] dt \\ &= \theta J_{x_1}[p_1(\cdot)] + (1-\theta)J_{x_2}[p_2(\cdot)]. \end{aligned}$$

This proves the convexity of $J_x[p(\cdot)]$.

Now, we let $p_1^*(\cdot)$ and $p_2^*(\cdot)$ denote the optimal controls for initial conditions x_1 and x_2 , respectively. Therefore,

$$\begin{aligned} u[\theta x_1 + (1-\theta)x_2] &= \inf_{p \geq 0} J_{\theta x_1 + (1-\theta)x_2}(p(\cdot)) \\ &\leq J_{\theta x_1 + (1-\theta)x_2}[\theta p_1^*(\cdot) + (1-\theta)p_2^*(\cdot)] \\ &< \theta J_{x_1}(p_1^*(\cdot)) + (1-\theta)J_{x_2}(p_2^*(\cdot)) \\ &= \theta u(x_1) + (1-\theta)u(x_2). \end{aligned}$$

This proves the strict convexity of $u(x)$.

COROLLARY 1. \exists a unique \bar{x} such that $u'(\bar{x}) = 0$ and that u satisfies

$$\begin{aligned} -\frac{1}{2}\sigma^2 u'' + \xi u' + \frac{1}{4}u'^2 + \alpha u &= x^2, & x \leq \bar{x}, \\ -\frac{1}{2}\sigma^2 u'' + \xi u' + \alpha u &= x^2, & x \geq \bar{x}. \end{aligned}$$

Also, the optimal production

$$p^* = \begin{cases} -u'/2 & \text{if } x \leq \bar{x}, \\ 0 & \text{if } x \geq \bar{x}. \end{cases}$$

Proof. The proof follows from the strict convexity of $u(x)$. The point \bar{x} is the minimum point of $u(x)$ and, therefore, it satisfies $u'(\bar{x}) = 0$.

We now turn to obtaining the asymptotic behavior of $u(x)$. For this, we first define

$$\phi_M(x) = (1/\alpha)x^2 + [2(M-\xi)/\alpha^2]x + (1/\alpha)[2(M-\xi)^2/\alpha^2 + \sigma^2/\alpha + M^2],$$

which is the value function for the production planning problem with $p \equiv M$. We note that $\phi_M(x)$ is the unique \mathcal{C}^2 solution with quadratic growth of

$$-\frac{1}{2}\sigma^2 \phi_M'' + (\xi - M)\phi_M' + \alpha \phi_M = x^2 + M^2.$$

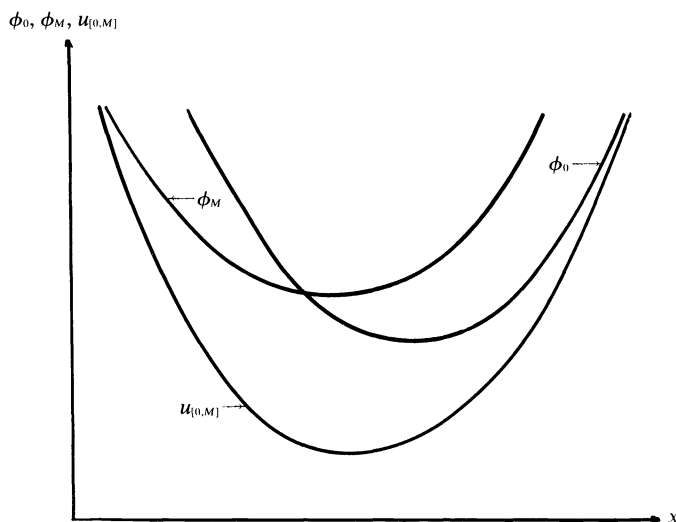
Moreover, $\phi_0 = \phi$ defined in (11).

From the quadratic growth property of u_M , which we now write as $u_{[0,M]}$, ϕ_0 , and ϕ_M , we can infer that (see Fig. 1)

$$u_{[0,M]} \sim \phi_0 \quad \text{as } x \rightarrow +\infty,$$

and

$$u_{[0,M]} \sim \phi_M \quad \text{as } x \rightarrow -\infty.$$

FIG. 1. Asymptotic behavior of $u_M \equiv u_{[0,M]}$.

The intuitive justification of this result is that for large inventory levels, it is clearly optimal to produce nothing, i.e., set $p=0$. Likewise, for large stockout levels, it is optimal to produce at the highest rate possible, i.e., set $p=M$.

To obtain the asymptotic behavior of $u(x)$, we recall that $u(x)$ is the limit of $u_{[0,M]}$ as $M \rightarrow \infty$; we rewrite $u(x)$ as $u_{[0,\infty)}$. It can be easily shown (see Appendix) that

$$u_{[0,\infty)} \sim \phi_0 \quad \text{as } x \rightarrow \infty.$$

In the other direction, however,

$$\phi_\infty = \lim_{M \rightarrow \infty} \sup \phi_M = \infty,$$

and is, therefore, of no help in concluding the behavior of $u_{[0,\infty)}$ as $x \rightarrow -\infty$.

At this point, it seems reasonable to examine the possibility of $u_{[0,\infty)}$ approaching u^* , also written as $u_{(-\infty,\infty)}$, as $x \rightarrow -\infty$, where

$$u^*(x) = mx^2 + nx + q$$

with

$$m = \frac{\sqrt{\alpha^2 + 4} - \alpha}{2} > 0, \quad n = -2m^2\xi \leq 0, \quad q = \frac{\sigma^2 m - m^2 \xi^2 (m^2 - 2)}{\alpha},$$

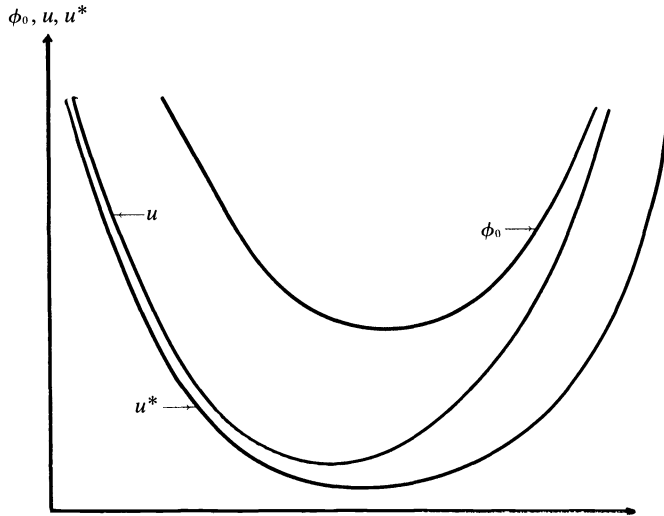
is the solution of the unconstrained problem obtained in Sethi and Thompson [9]. The result of this analysis, which appears in the Appendix is the following theorem; see Fig. 2.

THEOREM 3. *The asymptotic behavior of the value function $u(x)$ is characterized as follows:*

$$u(x) = u_{[0,\infty)}(x) \sim \phi_0(x) \equiv \phi(x) \quad \text{as } x \rightarrow \infty,$$

$$u(x) = u_{[0,\infty)}(x) \sim u_{(-\infty,\infty)}(x) \equiv u^*(x) \quad \text{as } x \rightarrow -\infty.$$

Before concluding this section, we would like to remark that the analysis of the problem considered in this paper started out with the statement of Theorem 3 as a conjecture. Intuitively, when there is a very large amount of shortage, it should be

FIG. 2. Asymptotic behavior of u .

optimal to produce at a very high rate in both the constrained as well as the unconstrained problem. We had therefore thought that a solution of (8) satisfying the asymptotic boundary conditions given above in Theorem 3 should supply us with the value function $u(x)$.

In the next section, we present our computational results.

6. Numerical results for bounded control. By an argument similar to the proof of Theorem 2 and Corollary 1, it follows that for given $M > 0$, u_M is strictly convex and the optimal feedback control $p_M(x)$ is (from (16)):

$$(23) \quad p_M(x) = \begin{cases} 0 & \text{if } x \geq x_0, \\ -u'_M(x)/2 & \text{if } x_M < x < x_0, \\ M & \text{if } x \leq x_M, \end{cases}$$

where x_0 and x_M are the unique roots of $u'_M(x_0) = 0$ and $u'_M(x_M) = -2M$, respectively. (The M -dependence of the switching point x_0 is suppressed for notational simplicity.) We employ a stable finite-difference technique to approximate the derivatives of u_M and we use policy iteration to determine approximate policy switching points x_0 and x_M , as well as the form of the control $p_M = -u'_M/2$ between the switching points. In the following we denote an *admissible* control as p and the *optimal* control as p_M . Furthermore, we denote the optimal value function for the unconstrained problem with $p(\cdot) \in (-\infty, \infty)$ as u^* and the corresponding optimal feedback control as $p^*(x)$. Sethi and Thompson [9] compute u^* and p^* explicitly. Arguments similar to Lemma 1 imply $0 < u^*(x) \leq u_M(x)$ for all x and all finite $M > 0$.

For $x > x_0$ the general solution of (10) having quadratic growth is

$$(24) \quad u_M(x) = a_0 x^2 / 2 + b_0 x + c_0 + A_0 e^{-r_0 x},$$

where

$$(25) \quad a_0 = 2/\alpha, \quad b_0 = -2\xi/\alpha^2, \quad c_0 = \sigma^2/\alpha^2 + 2\xi^2/\alpha^3,$$

and

$$(26) \quad r_0 = [(\xi^2 + 2\sigma^2\alpha)^{1/2} - \xi]/\sigma^2.$$

Since $u'_M(x_0+) = 0$ the constant A_0 is determined as

$$(27) \quad A_0 = (a_0 x_0 + b_0) e^{r_0 x_0 / r_0}.$$

Similarly, for $x < x_M$ we have

$$(28) \quad u_M(x) = a_1 x^2 / 2 + b_1 x + c_1 + A_1 e^{r_1 x},$$

where

$$(29) \quad a_1 = 2/\alpha, \quad b_1 = 2(M - \xi)/\alpha^2, \quad c_1 = [\sigma^2 \alpha + 2(M - \xi)^2 + M^2 \alpha^2]/\alpha^3,$$

$$(30) \quad r_1 = [((M - \xi)^2 + 2\sigma^2 \alpha)^{1/2} - (M - \xi)]/\sigma^2,$$

and

$$(31) \quad A_1 = (2M - a_1 x_M - b_1) e^{-r_1 x_M / r_1}.$$

In the finite-difference approximation we follow Kushner [7] and Smith [11], replacing $u'_M(x)$ and $u''_M(x)$ by

$$(32) \quad \begin{aligned} u'_M &\doteq [u_M(x+h) - u_M(x-h)]/2h, \\ u''_M(x) &\doteq [u_M(x+h) + u_M(x-h) - 2u_M(x)]/h^2, \end{aligned}$$

for step-size $h > 0$. Letting $V(x)$ denote the finite-difference approximation to $u_M(x)$ on a countable set of grid points G spaced h apart, (10) and (32) yield

$$(33) \quad V(x) = \min_{p \in [0, M]} [r(x, p) + \beta q^+(p) V(x+h) + \beta q^-(p) V(x-h)], \quad x \in G,$$

where

$$(34) \quad r(x, p) = h^2(x^2 + p^2)/(\sigma^2 + \alpha h^2), \quad \beta = \sigma^2/(\sigma^2 + \alpha h^2),$$

and

$$(35) \quad q^\pm(p) = (1/2)[1 \pm (p - \xi)h/\sigma^2].$$

Note that if $h \leq \min[\sigma^2/(M - \xi), \sigma^2/\xi]$ then $q^\pm(p) \geq 0$ for all $p \in [0, M]$, and (33) is the dynamic programming functional equation for a discrete “time,” discounted cost Markovian decision problem having countable state space G and uncountable action space $[0, M]$: the cost in state x under action p is $r(x, p)$, the discount rate is $\beta < 1$ and the transition probabilities are $q^\pm(p)$. The condition that (33) be a Markov decision problem is the same as the condition that the numerical method be stable. If h exceeds the upper bound given above, the solution of (33) may involve highly oscillatory behavior that renders it totally useless as an approximation to u_M ; see, e.g., Smith [11]. This implies that the problem will be increasingly difficult to solve numerically as M becomes very large because a very small grid spacing $h \leq \sigma^2/(M - \xi)$ will be needed. In the limit $M \rightarrow \infty$ the finite difference scheme cannot be used at all. This is corroborated by our early experiences in attempting to solve the unbounded case directly, by solving (8) backward from $x = \bar{x}$ (for various \bar{x}) using available codes for numerical solution of ordinary differential equations. Even a highly developed code based on the work of Gear [6]—involving automatic step-size adjustment, internal convergence tests, stiffness tests, A-stability checks and automatic truncation error reduction—was unable to deal with (8). Equation (8) seems to be unstable with respect to arbitrarily small perturbations and might not be solvable *directly* by any numerical method on a digital computer having finite word length. This suspicion was our original motivation for dealing with the case of bounded control. A fortuitous dividend of this

decision is our relatively concrete proof of existence and uniqueness of an optimal feedback policy for the unbounded control case as given in § 4. Although the bounded case is probably more relevant than the unbounded case in modelling real production systems, the unbounded case presents a theoretically challenging problem of numerical analysis.

In practice we must solve (10) on a truncated, compact state space $[L, U]$, where $L < x_M$ and $U > x_0$. Equation (10) then describes an optimally controlled process in the presence of reflecting barriers. We choose boundary conditions at L and U so as to ensure that the optimal control laws and the optimal value functions in the bounded and unbounded state space problems agree on (L, U) . As shown in Kushner [7], the solution of (33) for $x \in [L, U] \cap G$ converges as $h \rightarrow 0$ to a (weak) solution of (10) on $[L, U]$. Furthermore, the corresponding sequence of discrete time-controlled Markov processes converges weakly to the optimal continuous time process on $[L, U]$, the discrete time control laws converge weakly to the optimal continuous time control law, and $\lim_{h \rightarrow 0} V$ is the expected discounted cost of this process.

We solve (33) on the finite set $F \subset G$, where $F = \{L, L+h, \dots, U-h, U\}$. Given a control law $p(\cdot)$ such that $p(L) = M$ and $p(U) = 0$, we use (24) and (28) to evaluate $V(U+h)$ and $V(L-h)$ in (33) as $V(U+h) = 2V(U) - V(x-h) + f_U$ and $V(L-h) = 2V(L) - V(L+h) + f_L$, where f_U and f_L are computable explicitly from (24), (27), (28) and (31). These boundary conditions uniquely determine the solution of (33) on F , which we obtain by a recursive method that is numerically stable against roundoff errors for any candidate control law $p(\cdot)$ on F .

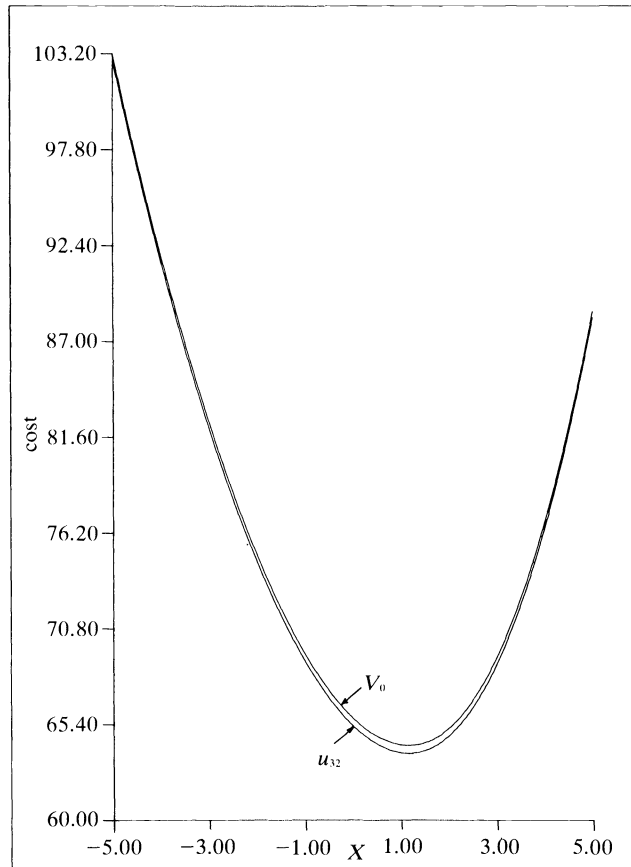
We solve (33) by policy iteration, starting with the truncated unconstrained optimal control law $p_M^0(x) = \max[0, \min[p^*(x), M]]$. In iteration $n = 0, 1, \dots$ we determine the value function V_n on F corresponding to control law p_M^n as outlined above, then obtain a new control law as $p_M^{n+1}(x) = \max[0, \min[M, (V_n(x-h) - V_n(x+h))/4h]]$, which minimizes the right-hand side in (33) with $V = V_n$.

It can be shown that the mapping $T: V_n \rightarrow V_{n+1}$ is a contraction mapping (see, e.g., Denardo [3] or Van Nunen [13] for a proof), so $\lim_{n \rightarrow \infty} V_n = V$, where V solves (33). In this case the action space is not a finite set, so the policy iteration procedure does not converge in a finite number of steps. We terminate the calculations at iteration n if $\max_{x \in F} |p_M^{n+1}(x) - p_M^n(x)| < 10^{-11}$, as this proves to yield adequate accuracy.

We illustrate our results for an example in which $\xi = 2$, $\sigma^2 = 9$, $\alpha = 0.2$ with $M = 16$ and 32. The results for this case are typical, in the sense that the general solution properties shown are robust with respect to broad input parameter variations.

In this example we have $p^*(x) = 0$ for $x = 1.81$, while $p^*(x) = 16$ or 32 for $x = -15.87$ or $x = -33.55$, respectively. The policy iteration procedure yields: $x_0 = 1.15$, $x_{16} = -15.90$ for $M = 16$; and $x_0 = 1.15$, $x_{32} = -33.575$ for $M = 32$. The main effect of policy iteration is to shift x_0 down from $p^{*-1}(0) = 1.81$, reducing the magnitude of $p_M(x)$ for all x . Although $p_M^0(x)$ is not a good approximation to $p_M(x)$, the value functions for these two control laws are not very different. Figure 3 shows the worst-case difference between the value function V_0 corresponding to the truncated unconstrained optimal control and the optimal value function $V \doteq u_M$. This occurs for $M = 32$ near $x = 0$, in which case the two value functions differ by less than 0.8%. Thus, in this example (and many others as well) the controller can achieve surprisingly good performance using the very easily implemented control law $p_M^0(x)$.

Figure 4 shows the value functions u_{16} , u_{32} and u^* for $x \in [-25, 15]$. Detailed examination of results reveals that for $x \geq -15$, convergence to $u(x)$ is practically attained for $M = 16$: u_{16} and u_{32} agree to six digits and their seventh digits differ by at most three for all x in this region. As x decreases from -15 to -30 , $u_{16} - u_{32} > 0$

FIG. 3. Computer plots of u_{32} and V_0 .

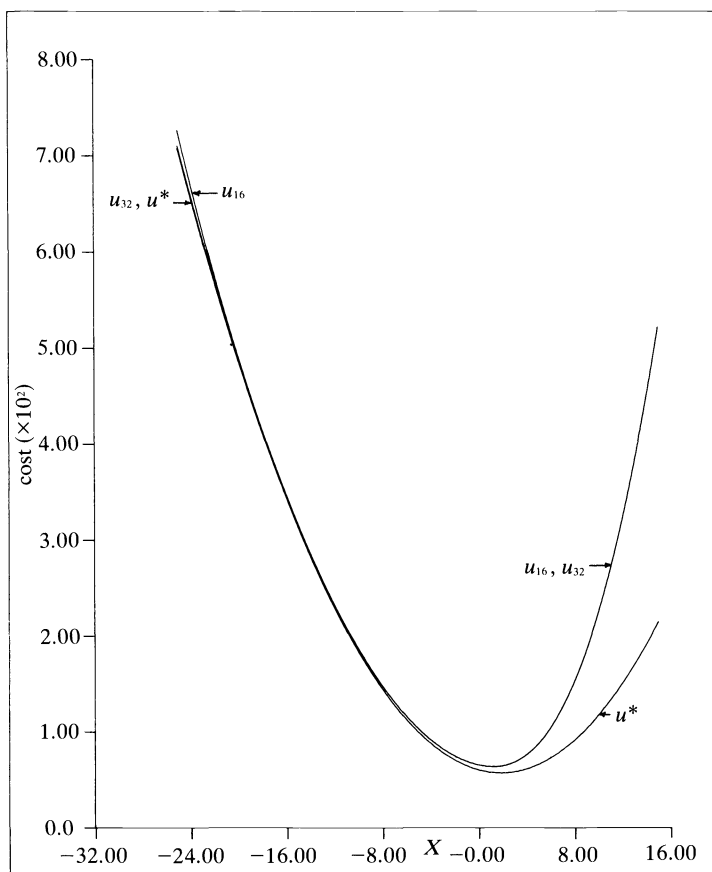
increases, while $u_{32} - u^* > 0$ is nearly a constant, decreasing very slowly. See Fig. 5 for a magnified view of the behavior of u_{16} , u_{32} and u^* for $x \in [-30, -25]$.

These results suggest that for moderate values of M ($M = 16$ in our example) the value function u is practically attained for $x > x_M$ by setting $u = u_M$. However, for $x < x_M$, $u - u_M$ is large, and we cannot achieve a good approximation to u in this region. As noted already, a direct solution to the $M = \infty$ problem does not appear to be feasible, and even finding u_M for large M is impractical because the step-length $h > 0$ needed in solving (33) would be so small that roundoff errors would invalidate the approximation of derivatives by finite differences (Smith [1969]).

In the Appendix we show that $u(x) - u^*(x) \rightarrow 0$ as $x \rightarrow -\infty$, but the numerical evidence for this asymptotic behavior is inconclusive. Given the convergence, however, the numerical evidence supports the conclusion that it is extremely slow, as it is in this case.

7. Extensions and concluding remarks. In this paper we have proved the existence of an optimal solution to a production planning problem with nonnegative production rates. We have characterized the value function $u(x)$ and have obtained computationally a very good approximation to it.

As an important aside, we point out that the introduction of a simple restriction on the production rates complicates enormously the stochastic production planning problem. This is not so in its deterministic counterpart.

FIG. 4. Computer plots of u_{16} , u_{32} and u^* .

Future extensions of our production planning problem should deal with constraints on inventory levels. These would complicate the problem substantially by requiring additional state variables.

Appendix: asymptotic behavior of $u(x)$ as $x \rightarrow \pm\infty$. From corollary 1, $u(x)$ satisfies

$$(A.1) \quad -\frac{1}{2}\sigma^2 u'' + \xi u' + \frac{1}{4}u'^2 + \alpha u = x^2, \quad x \leq \bar{x},$$

$$(A.2) \quad -\frac{1}{2}\sigma^2 u'' + \xi u' + \alpha u = x^2, \quad x \geq \bar{x},$$

$$(A.3) \quad u'(\bar{x}) = 0.$$

To study what happens as $x \rightarrow \pm\infty$, we can assume that \bar{x} is known.

For $x \rightarrow +\infty$, the task is easy since (A2) is a linear equation. Boundary conditions are (A3) and a quadratic growth as $x \rightarrow \infty$.

Case: $x \geq \bar{x}$.

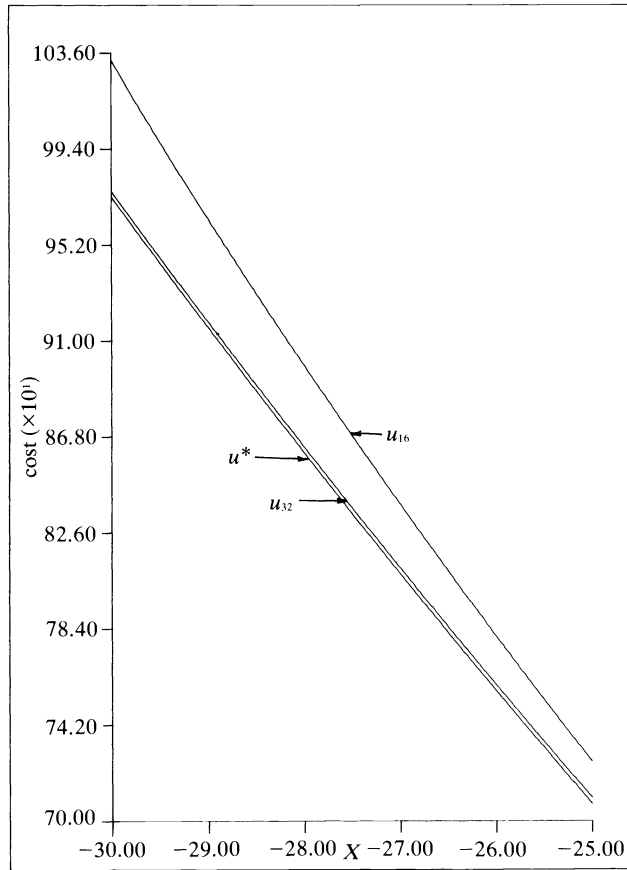
$$(A.4) \quad u(x) = \phi(x) + K e^{-\gamma x},$$

where $\phi(x)$ is given in (11),

$$\gamma = [-\xi + \sqrt{\xi^2 + 2\alpha\sigma^2}]/\sigma^2,$$

and K satisfies

$$(2/\alpha)\bar{x} - (2\xi/\alpha^2) - \gamma K e^{-\gamma\bar{x}} = 0.$$

FIG. 5. Magnified view of u_{16} , u_{32} and u^* .

It follows that

$$(A.5) \quad |u(x) - \phi(x)| \rightarrow 0 \quad \text{as } x \rightarrow +\infty.$$

We also note that

$$(A.6) \quad u(\bar{x}) = (1/\alpha)\bar{x}^2 - (2\xi/\alpha^2)\bar{x} + (1/\alpha^2)[2\xi^2/\alpha + \sigma^2] + (1/\gamma)[(2/\alpha)\bar{x} - 2\xi/\alpha^2].$$

Case: $x \leq \bar{x}$. To obtain the behavior of $u(x)$ as $x \rightarrow -\infty$, we first observe that the solution $u^*(x) = mx^2 + nx + q$ of the unconstrained problem specified in § 5 satisfies (A1); see also Sethi and Thompson [9]. Therefore, we have the following situation. u and u^* are solutions of the same nonlinear equation (A1) for $x \leq \bar{x}$, they have quadratic growth, and it is easily seen that $u(\bar{x}) \neq u^*(\bar{x})$. Setting

$$u - v = z,$$

we get

$$(A.7) \quad \begin{aligned} -\frac{1}{2}\sigma^2 z'' + \xi z' + \frac{1}{4}z'^2 + \frac{1}{2}z'(2mx + n) + \alpha z &= 0, \\ z(\bar{x}) = u(\bar{x}) - u^*(\bar{x}) &= \delta > 0. \end{aligned}$$

Note that m and n are specified in § 5. From (A7), it follows that

$$0 \leq z(x) \leq \delta.$$

More precisely, we have

$$(A.8) \quad 0 \leq z(x) \leq \zeta(x)$$

where ζ is the solution of

$$(A.9) \quad \begin{aligned} -\frac{1}{2}\sigma^2\zeta'' + \xi\zeta' + \frac{1}{2}\zeta'(2mx+n) + \alpha\zeta &= 0, \\ \zeta(\bar{x}) &= \delta. \end{aligned}$$

But

$$(A.10) \quad \zeta(x) = \delta\eta(x)$$

where,

$$(A.11) \quad \begin{aligned} -\frac{1}{2}\sigma^2\eta'' + \eta'\left(m + \frac{n}{2} + \xi\right) + \alpha\eta &= 0, \\ \eta(\bar{x}) &= 1. \end{aligned}$$

We shall now obtain the following estimate

$$(A.12) \quad \eta(x) \leq \frac{1}{(1 + (\bar{x} - x)^2)^s} = \theta(x)$$

where $s > 0$ is conveniently chosen.

We need only to show that it is possible to choose s such that

$$(A.13) \quad \begin{aligned} -\frac{\sigma^2}{2}\theta'' + \theta'\left(m + \frac{n}{2} + \xi\right) + \alpha\theta &\geq 0, \\ \theta(\bar{x}) &= \eta(\bar{x}) = 1. \end{aligned}$$

For this, we observe that

$$\theta' = \frac{-2s(x - \bar{x})}{(1 + (x - \bar{x})^2)^{s+1}}$$

and

$$\theta'' = \frac{-2s}{(1 + (x - \bar{x})^2)^{s+1}} + \frac{4s(s+1)(x - \bar{x})^2}{(1 + (x - \bar{x})^2)^{s+2}}.$$

Thus, we have

$$\begin{aligned} &-\frac{\sigma^2}{2}\theta'' + \theta'\left(m + \frac{n}{2} + \xi\right) + \alpha\theta \\ &= \frac{-2\sigma^2s(s+1)(x - \bar{x})^2}{(1 + (x - \bar{x})^2)^{s+2}} + \frac{\sigma^2s}{(1 + (x - \bar{x})^2)^{s+1}} \\ &\quad - \frac{2s(x - \bar{x})(m + n/2 + \xi)}{(1 + (x - \bar{x})^2)^{s+1}} + \frac{\alpha}{(1 + (x - \bar{x})^2)^s} \\ &\geq \frac{1}{(1 + (x - \bar{x})^2)^s} \left[\alpha - \frac{2s(x - \bar{x})(m + n/2 + \xi)}{1 + (x - \bar{x})^2} - 2\sigma^2s(s+1) \right]. \end{aligned}$$

But

$$\frac{2(x - \bar{x})(mx + n/2 + \xi)}{1 + (x - \bar{x})^2} \leq 2m + |m\bar{x} + n/2 + \xi|,$$

and, therefore, we can choose s such that $s < 1$ and

$$(A.14) \quad \alpha \geq 2s[2m + |m\bar{x} + n/2 + \xi| + 2\sigma^2]$$

to imply (A.13).

We have, therefore, proved the estimate

$$(A.15) \quad u^*(x) \leq u(x) \leq u^*(x) + \frac{(u(\bar{x}) - v(\bar{x}))}{(1 + (x - \bar{x})^2)^s}$$

where \bar{x} is the minimum of u . This completes the proof of Theorem 3.

REFERENCES

- [1] L. ARNOLD, *Stochastic Differential Equations*, Wiley, New York, 1974.
- [2] A. BENSOUSSAN, *Stochastic Control by Functional Analysis Methods*, North-Holland, New York, 1982.
- [3] E. V. DENARDO, *Contraction mappings in the theory underlying dynamic programming*, SIAM Rev., 9 (1967), pp. 165–177.
- [4] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.
- [5] A. FRIEDMAN, *Stochastic Differential Equations and Applications*, Vol. 1, Academic Press, New York, 1975.
- [6] C. W. GEAR, *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
- [7] H. J. KUSHNER, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.
- [8] O. A. LADYZHENSKAYA AND N. N. URALTSEVA, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.
- [9] S. P. SETHI AND G. L. THOMPSON, *Simple models in stochastic production planning*, in *Applied Stochastic Control in Econometrics and Management Science*, A Bensoussan, P. Kleindorfer, and C. Tapiero, eds., North-Holland, New York, 1981, pp. 295–304.
- [10] ———, *Optimal Control Theory: Applications to Management Science*, Martinus Nijhoff Pub., Boston, 1981.
- [11] G. D. SMITH, *Numerical Solution of Partial Differential Equations*, Oxford Univ. Press, London, 1969.
- [12] G. L. THOMPSON AND S. P. SETHI, *Turnpike horizons for production planning*, *Management Science*, 26 (1980), pp. 229–241.
- [13] J. A. E. VAN NUNEN, *Contracting Markov Decision Processes*, Mathematical Centre Tracts 71, Mathematisch Centrum, Amsterdam, 1976.

TWO-METRIC PROJECTION METHODS FOR CONSTRAINED OPTIMIZATION*

ELI M. GAFNI† AND DIMITRI P. BERTSEKAS‡

Abstract. This paper is concerned with the problem $\min \{f(x) | x \in X\}$ where X is a convex subset of a linear space H , and f is a smooth real-valued function on H . We propose the class of methods $x_{k+1} = P(x_k - \alpha_k g_k)$, where P denotes projection on X with respect to a Hilbert space norm $\|\cdot\|$, g_k denotes the Frechet derivative of f at x_k with respect to another Hilbert space norm $\|\cdot\|_k$ on H , and α_k is a positive scalar stepsize. We thus remove an important restriction in the original proposal of Goldstein [1] and Levitin and Poljak [2], where the norms $\|\cdot\|$ and $\|\cdot\|_k$ must be the same. It is therefore possible to match the norm $\|\cdot\|$ with the structure of X so that the projection operation is simplified while at the same time reserving the option to choose $\|\cdot\|_k$ on the basis of approximations to the Hessian of f so as to attain a typically superlinear rate of convergence. The resulting methods are particularly attractive for large-scale problems with specially structured constraint sets such as optimal control and nonlinear multi-commodity network flow problems. The latter class of problems is discussed in some detail.

Key words. constrained optimization, gradient projection, convergence analysis, multicommodity flow problems, large-scale optimization

1. Introduction. Projection methods stemming from the original proposal of Goldstein [1], and Levitin and Poljak [2] are often very useful for solving the problem

$$(1) \quad \begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in X \end{aligned}$$

where $f: H \rightarrow R$ and X is a convex subset of a linear space H . They take the form

$$(2) \quad x_{k+1} = P_k(x_k - \alpha_k g_k)$$

where α_k is a positive scalar stepsize, $P_k(\cdot)$ denotes projection on X with respect to some Hilbert space norm $\|\cdot\|_k$ on H and g_k denotes the Frechet derivative of f with respect to $\|\cdot\|_k$, i.e., g_k is the vector in H satisfying

$$(3) \quad f(x) = f(x_k) + \langle g_k, x - x_k \rangle_k + o(\|x - x_k\|_k),$$

where $\langle \cdot, \cdot \rangle_k$ denotes the inner product corresponding to $\|\cdot\|_k$.

As an example let $H = R^n$, and B_k be an $n \times n$ positive definite symmetric matrix. Consider the inner product and norm corresponding to B_k

$$(4) \quad \langle x, y \rangle_k = x' B_k y, \quad \|x\|_k = (\langle x, x \rangle_k)^{1/2} \quad \forall x, y \in H,$$

where all vectors above are considered to be column vectors and prime denotes transposition. With respect to this norm we have (cf. (3))

$$(5) \quad g_k = B_k^{-1} \nabla f(x_k),$$

* Received by the editors August 8, 1982, and in revised form July 27, 1983.

† This work was supported by the National Science Foundation under grant NSF/ECS 79-20834. Computer Science Department, University of California, Los Angeles, California 90024.

‡ Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

where $\nabla f(x_k)$ is the vector of first partial derivatives of f

$$(6) \quad \nabla f(x_k) = \begin{bmatrix} \frac{\partial f(x_k)}{\partial x^1} \\ \vdots \\ \frac{\partial f(x_k)}{\partial x^n} \end{bmatrix}.$$

When problem (1) is unconstrained ($X = H$), iteration (2) takes the familiar form

$$x_{k+1} = x_k - \alpha_k B_k^{-1} \nabla f(x_k).$$

Otherwise the vector

$$x_{k+1} = P_k(x_k - \alpha_k g_k)$$

is the solution of the problem

$$\begin{aligned} &\text{minimize } \|x - x_k + \alpha_k g_k\|_k^2 \\ &\text{subject to } x \in X. \end{aligned}$$

A straightforward computation using (4) and (5) shows that the problem above is equivalent to the problem

$$(7) \quad \begin{aligned} &\text{minimize } \nabla f(x_k)'(x - x_k) + \frac{1}{2\alpha_k} (x - x_k)' B_k (x - x_k) \\ &\text{subject to } x \in X. \end{aligned}$$

When X is a polyhedral set and B_k is a quasi-Newton approximation of the Hessian of f , the resulting method is closely related to recursive quadratic programming methods which currently enjoy a great deal of popularity (e.g., Garcia-Palomares [3], Gill et al. [4]).

It is generally recognized that in order for the methods above to be effective it is essential that the computational overhead for solving the quadratic programming problem (7) should not be excessive. For large-scale problems this overhead can be greatly reduced if the matrix B_k is chosen in a way that matches the structure of the constraint set. For example if X is the Cartesian product $\prod_{i=1}^m X_i$ of m simpler sets X_i , the matrix B_k can be chosen to be block diagonal with one block corresponding to each set X_i , in which case the projection problem (7) decomposes naturally. Unfortunately, such a choice of B_k precludes the possibility of superlinear convergence of the algorithm, which typically cannot be achieved unless B_k is chosen to be a suitable approximation of the Hessian matrix of f [3], [5].

The purpose of this paper is to propose projection methods of the form

$$(8) \quad x_{k+1} = P(x_k - \alpha_k g_k)$$

where the norms $\|\cdot\|$ and $\|\cdot\|_k$ corresponding to the projection and the differentiation operators respectively can be different. This allows the option to choose $\|\cdot\|$ to match the structure of X , thereby making the projection operation computationally efficient, while reserving the option to choose $\|\cdot\|_k$ on the basis of second derivatives of f thereby making the algorithm capable of superlinear convergence. When $H = R^n$, the projection norm $\|\cdot\|$ is the standard Euclidean norm

$$(9) \quad \|x\| = (x'x)^{1/2} = |x|,$$

and the derivative norm $\|\cdot\|_k$ is specified by an $n \times n$ positive definite symmetric matrix B_k

$$(10) \quad \|x\|_k = (x' B_k x)^{1/2},$$

the vector x_{k+1} of (8) is obtained by solving the quadratic programming subproblem

$$(11) \quad \begin{aligned} &\text{minimize } g'_k(x - x_k) + \frac{1}{2\alpha_k} |x - x_k|^2 \\ &\text{subject to } x \in X \end{aligned}$$

where

$$(12) \quad g_k = B_k^{-1} \nabla f(x_k).$$

The quadratic programming problem (11) may be very easy to solve if X has special structure. As an example consider the case of an orthant constraint

$$(13) \quad X = \{x | 0 \leq x^i, i = 1, \dots, n\}.$$

Then, the iteration takes the form

$$(14) \quad x_{k+1} = [x_k - \alpha_k B_k^{-1} \nabla f(x_k)]^+$$

where for any vector $v \in R^n$ with coordinates v^i , $i = 1, \dots, n$ we denote by v^+ the vector with coordinates

$$(v^i)^+ = \max \{0, v^i\}.$$

Iteration (14) was first proposed in Bertsekas [6], and served as the starting point for the present paper. It was originally developed for use in a practical application reported in [18]. The computational overhead involved in (14) is much smaller than the one involved in solving the corresponding quadratic program (7) particularly for problems of large dimension. Indeed large optimal control problems have been solved using (14) (see [6]) that, in our view, would be impossible to solve by setting up the corresponding quadratic programming (7) and using standard pivoting techniques. Similarly (14) holds an important advantage over active set methods [4] where only one constraint is allowed to enter the active set at each iteration. Such methods require at least as many iterations as the number of active constraints at the optimal solution which are not active at the starting vector, and are in our view a poor choice for problems of very large dimension.

An important point is that it is not true in general that for an arbitrary positive definite choice B_k , iteration (14) is a descent iteration (in the sense that if x_k is not a critical point, then for α_k sufficiently small we have $f(x_{k+1}) < f(x_k)$). Indeed this is the main difficulty in constructing two-metric extensions of the Goldstein–Levitin–Poljak method. It was shown, however, in [6] (see also [19]) that if B_k is chosen to be partially diagonal with respect to a suitable subset of coordinates, then (14) becomes a descent iteration. We give a nontrivial extension of this result in the next section (Proposition 1). The construction of the “scaled gradient” g_k satisfying the descent condition

$$(15) \quad \langle g_k, \nabla f(x_k) \rangle > 0$$

is based on a decomposition of the negative gradient into two orthogonal components by projection on an appropriate pair of cones that are dual to each other. One of the two components is then “scaled” by multiplication with a positive definite self-adjoint operator (which may incorporate second derivative information) and added to the first

component to yield g_k . The method of construction is such that g_k , in addition to (15), also satisfies

$$f[P(x_k - \alpha g_k)] < f(x_k)$$

for all α in an interval $(0, \bar{\alpha}_k]$, $\bar{\alpha}_k > 0$.

Section 3 describes the main algorithm and proves its convergence. While other stepsize rules are possible, we restrict attention to an Armijo-like stepsize rule for selecting α_k on the arc

$$\{z | z = P(x_k - \alpha g_k), \alpha > 0\}$$

which is patterned after similar rules proposed in Bertsekas [6], [7]. Variations of the basic algorithm are considered in § 5, while in § 4 we consider rate of convergence aspects of algorithm (8), (11), (12) as applied to finite dimensional problems. We show that the descent direction g_k can be constructed on the basis of second derivatives of f so that the method has a typically superlinear rate of convergence. Here we restrict attention to Newton-like versions of the algorithm. Quasi-Newton, and approximate Newton implementations based on successive overrelaxation or conjugate gradient methods are also possible. A superlinearly convergent conjugate gradient-based implementation is applied to a large-scale multicommodity flow problem in the last section of the paper.

While the algorithm is stated and analyzed in general terms, we pay special attention to the case where X is a finite dimensional polyhedral set with a decomposable structure since we believe that this is the case where the algorithm of this paper is most likely to find application.

2. The algorithmic map and its descent properties. Consider the problem

$$(16) \quad \begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in X \end{aligned}$$

where f is a real-valued function on a Hilbert space H , and X is a nonempty, closed, convex subset of H . The inner product and norm on H will be denoted by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ respectively. We say that two vectors $x, y \in H$ are orthogonal if $\langle x, y \rangle = 0$. For any $z \in H$ we denote by $P(z)$ the unique projection of z on X , i.e.,

$$(17) \quad P(z) = \arg \min \{\|x - z\| | x \in X\}.$$

We assume that f is continuously Frechet differentiable on H . The Frechet derivative at a vector $x \in H$ will be denoted by $\nabla f(x)$. It is the unique vector in H satisfying

$$f(z) = f(x) + \langle \nabla f(x), z - x \rangle + o(\|z - x\|)$$

where $o(\|z - x\|)/\|z - x\| \rightarrow 0$ as $z \rightarrow x$. We say that a vector $x^* \in X$ is *critical* with respect to problem (16) if

$$(18) \quad \langle \nabla f(x^*), x - x^* \rangle \geq 0 \quad \forall x \in X,$$

or equivalently, if $x^* = P[x^* - \nabla f(x^*)]$.

It will be convenient for our purposes to represent the set X as an intersection of half spaces

$$(19) \quad X = \{x | \langle a_i, x \rangle \leq b_i, \forall i \in I\},$$

where I is a, possibly infinite, index set and, for each $i \in I$, a_i is a nonzero vector in H and b_i is a scalar. For each closed convex set X there exists at least one such

representation. We will assume that the set I is nonempty—the case where I is empty corresponds to an unconstrained problem which is not the subject of this paper. *Our algorithm will be defined in terms of a specific collection $\{(a_i, b_i) | i \in I\}$ satisfying (19) which will be assumed given.* This is not an important restriction for many problems of interest including, of course, the case where X is a polyhedron in R^n .

We now describe the algorithmic mapping on which our method is based. For a given vector $x \in X$ we will define an arc of points $\{x(\alpha) | \alpha \geq 0\}$ which depends on an index set $I_x \subset I$ and an operator D_x which will be described further shortly. The index set I_x is required to satisfy

$$(20) \quad I_x \supset \{i \in I | \langle a_i, x \rangle \geq b_i - \varepsilon \|a_i\|\}$$

where ε is some positive scalar. Let C_x be the cone defined by

$$(21) \quad C_x = \{z | \langle a_i, z \rangle \leq 0, \forall i \in I_x\}$$

and C_x^+ be the dual cone of C_x

$$(22) \quad C_x^+ = \{z | \langle y, z \rangle \leq 0, \forall y \in C_x\}.$$

For orientation purposes we mention that if X is a polyhedral subset of R^n (or more generally if the index set I is finite), and ε is sufficiently small, then I_x can consist of the indexes of the active constraints at x , i.e., we may take $I_x = \{i | \langle a_i, x \rangle = b_i, i \in I\}$. In that case C_x is the cone of feasible directions at x , while C_x^+ is the cone generated by the vectors a_i corresponding to the active constraints at x . More generally C_x is a (possibly empty) subset of the set of feasible directions at x , and for any $\Delta x \in C_x$ with $\|\Delta x\| \leq \varepsilon$ the vector $x + \Delta x$ belongs to X .

Let d_x be the projection of $[-\nabla f(x)]$ on C_x , i.e.,

$$(23) \quad d_x = \arg \min \{\|z + \nabla f(x)\| | z \in C_x\}.$$

Define

$$(24) \quad d_x^+ = -[\nabla f(x) + d_x].$$

It can be easily seen that the vectors d_x and d_x^+ are orthogonal and that d_x^+ is the projection of $[-\nabla f(x)]$ on C_x^+ , i.e.,

$$(25) \quad d_x^+ = \arg \min \{\|z + \nabla f(x)\| | z \in C_x^+\}.$$

Note that if the norm $\|\cdot\|$ on H is such that projection on the set X is relatively simple, then typically the same is true for the projection (23), required to compute d_x and d_x^+ .

Let Γ_x be the subspace spanned by the elements of C_x which are orthogonal to d_x^+ , i.e.,

$$(26) \quad \Gamma_x = \text{span} \{C_x \cap \{z | \langle z, d_x^+ \rangle = 0\}\}.$$

Note that

$$(27) \quad d_x \in \Gamma_x$$

since d_x belongs to C_x and is orthogonal to d_x^+ . Let $D_x: \Gamma_x \rightarrow \Gamma_x$ be a positive definite self-adjoint operator mapping Γ_x into itself. Consider the projection \tilde{d}_x of $D_x d_x$ on the closed cone $C_x \cap \{z | \langle z, d_x^+ \rangle = 0\}$, i.e.

$$(28) \quad \tilde{d}_x = \arg \min \{\|z - D_x d_x\| | z \in C_x, \langle z, d_x^+ \rangle = 0\}.$$

Consider also the direction vector

$$(29) \quad g = -(d_x^+ + \tilde{d}_x).$$

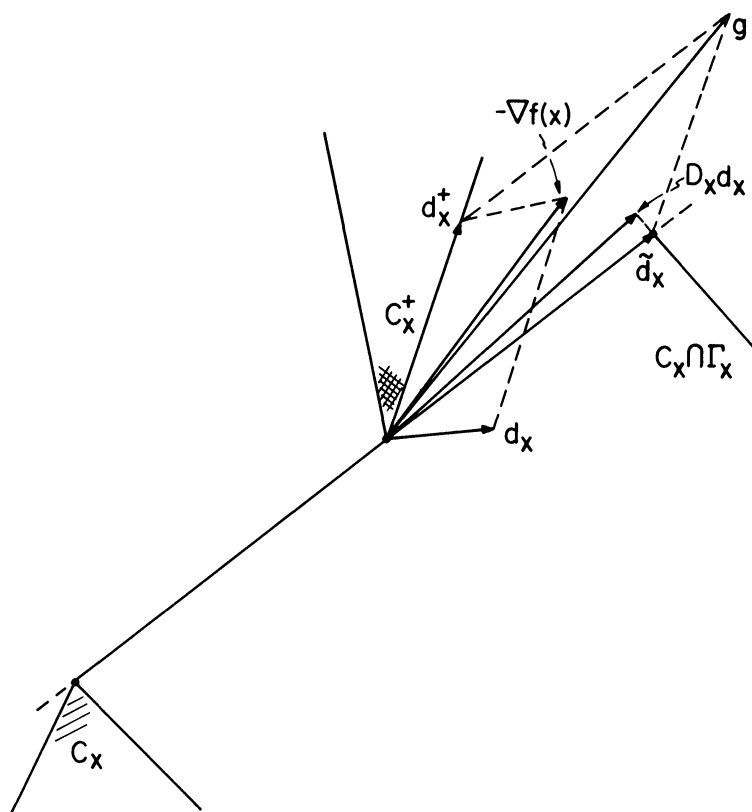


FIG. 2. Obtaining g for a case where C_x^+ lies on a two-dimensional manifold in R^3 .

where

$$\hat{I}_x = \left\{ i \mid i \in I_x \text{ and } \frac{\partial f(x)}{\partial x^i} > 0 \right\}.$$

If \hat{I}_x is empty then $\Gamma_x = R^n$ and we have $d_x = -\nabla f(x)$, $d_x^+ = 0$. In this case $g = -D_x d_x = D_x \nabla f(x)$ where D_x is any $n \times n$ positive definite symmetric matrix. If I_x is not empty, by rearranging indices if necessary assume that for some integer p with $0 \leq p \leq n-1$ we have $\hat{I}_x = \{p+1, \dots, n\}$. Partition $\nabla f(x)$ as

$$\nabla f(x) = \begin{bmatrix} \tilde{w} \\ \hat{w} \end{bmatrix}$$

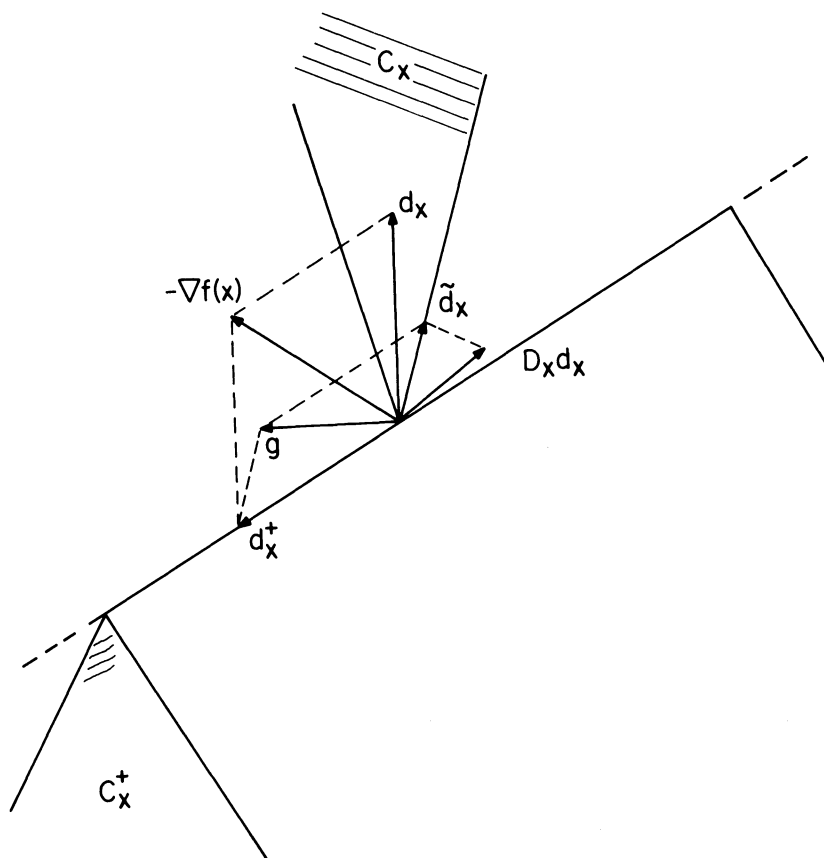
where $\tilde{w} \in R^p$ and $w \in R^{n-p}$. The vector g is given by

$$g = \begin{bmatrix} (D_x \tilde{w})^\# \\ \hat{w} \end{bmatrix}$$

where D_x is a $p \times p$ positive definite symmetric matrix, $(D_x \tilde{w})^\#$ denotes projection of $D_x \tilde{w}$ on C_x , i.e., $(D_x \tilde{w})^\#$ is obtained from $D_x \tilde{w}$ by setting to zero those coordinates of $D_x \tilde{w}$ which are negative and whose indices belong to I_x .

Example 2. Let $H = R^n$, and X be the unit simplex

$$(31) \quad X = \left\{ x \mid \sum_{i=1}^n x^i = 1, x^i \geq 0, i = 1, \dots, n \right\}.$$

FIG. 3. Obtaining g for a case where C_x lies on a two-dimensional manifold in R^3 .

Suppose the inner product on R^n is taken to be

$$(32) \quad \langle x, y \rangle = \sum_{i=1}^n s^i x^i y^i$$

where s^i , $i = 1, \dots, n$ are some positive scalars. Let \hat{I}_x be a set of indices including those indices i such that $0 \leq x^i \leq \varepsilon / \sqrt{s^i}$. Then the cone C_x can be taken to be

$$(33) \quad C_x = \left\{ z \left| \sum_{i=1}^n z^i = 0, z^i \geq 0, \forall i \in \hat{I}_x \right. \right\}.$$

The vector d_x is obtained as the solution of the projection problem

$$(34) \quad \begin{aligned} & \text{minimize } \frac{1}{2} \sum_{i=1}^n s^i \left[z^i + \frac{1}{s^i} \frac{\partial f(x)}{\partial x^i} \right]^2 \\ & \text{subject to } \sum_{i=1}^n z^i = 0, \quad z^i \geq 0, \quad i \in \hat{I}_x. \end{aligned}$$

The solution of this problem is very simple. By introducing a Lagrange multiplier λ for the equality constraint $\sum_{i=1}^n z^i = 0$, we obtain that λ is the solution of the piecewise

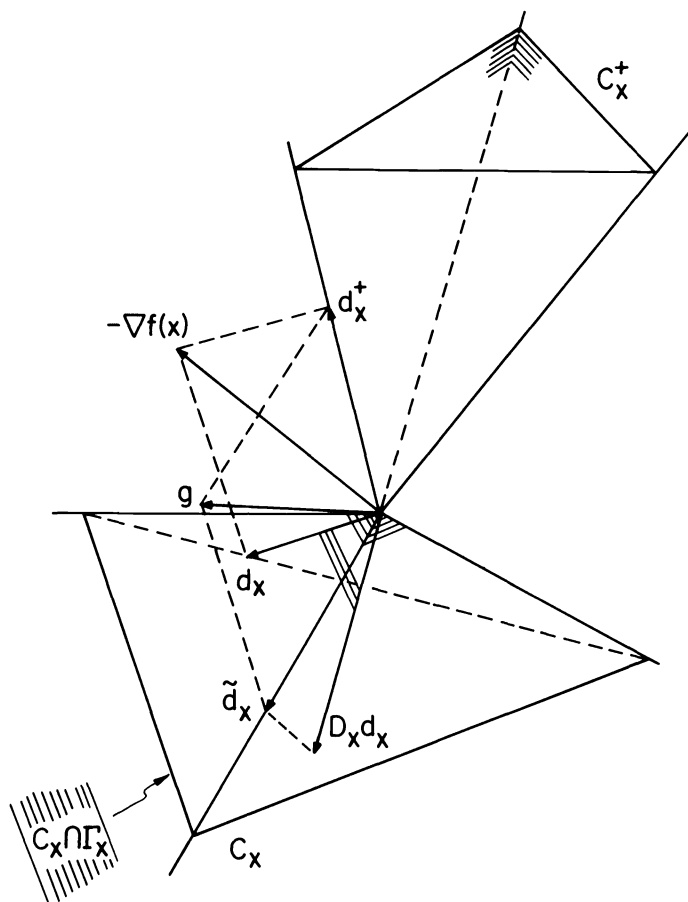


FIG. 4. Obtaining g for a case where both C_x^+ and C_x have nonempty interior in R^3 .

linear equation

$$(35) \quad \sum_{i \in \hat{I}_x} \frac{1}{s^i} \left[\lambda - \frac{\partial f(x)}{\partial x^i} \right]^+ + \sum_{i \notin \hat{I}_x} \frac{1}{s^i} \left[\lambda - \frac{\partial f(x)}{\partial x^i} \right] = 0.$$

This equation can be solved by the well-known method of sorting the breakpoints $\partial f(x)/\partial x^i$, $i \in \hat{I}_x$ in decreasing order, and testing the values of the left side at the breakpoints until two successive values bracket zero. Once λ is obtained, the coordinates of d_x are given by

$$(36) \quad d_x^i = \begin{cases} \frac{1}{s^i} \left[\lambda - \frac{\partial f(x)}{\partial x^i} \right]^+ & \text{if } i \in \hat{I}_x, \\ \frac{1}{s^i} \left[\lambda - \frac{\partial f(x)}{\partial x^i} \right] & \text{if } i \notin \hat{I}_x. \end{cases}$$

The vector d_x^+ is then obtained from the equation

$$d_x^+ = -[\nabla f(x) + d_x].$$

Let

$$(37) \quad \tilde{I}_x = \left\{ i \mid i \in \hat{I}_x \text{ and } \lambda < \frac{\partial f(x)}{\partial x^i} \right\}.$$

It is easily verified that the subspace Γ_x is given by

$$(38) \quad \Gamma_x = \left\{ z \left| \sum_{i=1}^n z^i = 0, z^i = 0, \forall i \in \tilde{I}_x \right. \right\}.$$

The vector \tilde{d}_x is obtained as the solution of the simple projection problem.

$$(39) \quad \begin{aligned} & \text{minimize } \frac{1}{2} \sum_{i=1}^n s^i [z^i - (D_x d_x)^i]^2 \\ & \text{subject to } \sum_{i=1}^n z^i = 0, z^i \geq 0 \quad \forall i \in \hat{I}_x, \quad z^j = 0 \quad \forall j \in \tilde{I}_x \end{aligned}$$

where $(D_x d_x)^i$ is the i th coordinate of the vector $D_x d_x$ obtained by multiplying d_x with an $n \times n$ symmetric matrix D_x which maps Γ_x into Γ_x and is positive definite on Γ_x . We will comment further on the choice of D_x in the last section of the paper. The vector g is given now by $g = -(\tilde{d}_x + d_x^+)$. Note that the solution of both projection problems (34) and (39), as well as the problem of projection on the simplex X of (31) is greatly simplified by the choice of the “diagonal” metric specified by (32).

Proposition 1 below is the main result regarding the algorithmic map specified by (20)–(24), (28)–(30). For its proof we will need the following lemma, the proof of which is given in Appendix A.

LEMMA 1. *Let Ω be a closed convex subset of a Hilbert space H , and let $P_\Omega(\cdot)$ denote projection on Ω . For every $x \in \Omega$ and $z \in H$:*

a) *The function $h: (0, \infty) \rightarrow \mathbb{R}$ defined by*

$$h(\alpha) = \frac{\|P_\Omega(x + \alpha z) - x\|}{\alpha} \quad \forall \alpha > 0$$

is monotonically nonincreasing.

b) *If y is any direction of recession of Ω (i.e., $(x + \alpha y) \in \Omega$ for all $\alpha \geq 0$), then*

$$(40) \quad \langle y, x + z \rangle \leq \langle y, P_\Omega(x + z) \rangle.$$

PROPOSITION 1. *For $x \in X$, let $\varepsilon > 0$ and I_x satisfy (20), and let $D_x: \Gamma_x \rightarrow \Gamma_x$ be a positive definite self-adjoint operator on the subspace Γ_x defined by (21)–(26). Consider the arc $\{x(\alpha) | \alpha \geq 0\}$ defined by (23), (24), (28)–(30).*

a) *If x is critical, then*

$$x(\alpha) = x \quad \forall \alpha \geq 0.$$

b) *If x is not critical, then*

$$(41) \quad \langle \nabla f(x), g \rangle > 0,$$

and

$$(42) \quad \langle \nabla f(x), x - x(\alpha) \rangle \geq \alpha \langle d_x, D_x d_x \rangle + \frac{1}{\alpha} \|x(\alpha) - (x + \alpha \tilde{d}_x)\|^2 > 0 \quad \forall \alpha \in \left(0, \frac{\varepsilon}{\|g\|}\right).$$

Furthermore there exists $\bar{\alpha} > 0$ such that

$$(43) \quad f(x) > f[x(\alpha)] \quad \forall \alpha \in (0, \bar{\alpha}].$$

Proof. a) It is easily seen that for every $z \in C_x$ we have

$$(44) \quad \left(x + \frac{\varepsilon}{\|z\|} z\right) \in X$$

in view of the definitions (19)–(21). Since x is critical, we have $\langle \nabla f(x), y - x \rangle \geq 0$ for all $y \in X$. Therefore using (44) we have

$$(45) \quad \langle \nabla f(x), z \rangle \geq 0 \quad \forall z \in C_x.$$

From the definitions of C_x^+ , d_x and d_x^+ (cf. (21)–(24)) and (45) it follows that

$$-\nabla f(x) \in C_x^+$$

and

$$d_x^+ = -\nabla f(x), \quad d_x = 0.$$

Using (28)–(30), we obtain $x(\alpha) = P[x - \alpha \nabla f(x)]$. Since x is critical, we have that $x = P[x - \alpha \nabla f(x)]$ for all $\alpha \geq 0$ and the conclusion follows.

b) We have by using the facts $\nabla f(x) = -(d_x + d_x^+)$ and $\langle \tilde{d}_x, d_x^+ \rangle = 0$

$$(46) \quad \langle \tilde{d}_x, \nabla f(x) \rangle = -\langle \tilde{d}_x, d_x + d_x^+ \rangle = -\langle \tilde{d}_x, d_x \rangle.$$

Now \tilde{d}_x is the projection of $D_x d_x$ on the cone $C_x \cap \{z | \langle z, d_x^+ \rangle = 0\}$, d_x belongs to this cone and therefore is a direction of recession. Using Lemma 1b), it follows that

$$(47) \quad \langle d_x, \tilde{d}_x \rangle \geq \langle d_x, D_x d_x \rangle.$$

Combining (46) and (47), we obtain

$$(48) \quad \langle \tilde{d}_x, \nabla f(x) \rangle \leq -\langle d_x, D_x d_x \rangle \leq 0$$

where the second inequality is strict if and only if $d_x \neq 0$. Also d_x^+ is the projection of $-\nabla f(x)$ on C_x^+ , so

$$(49) \quad \langle d_x^+, \nabla f(x) \rangle \leq 0$$

with strict inequality if and only if $d_x^+ \neq 0$. Combining (48) and (49) and using the fact $g = -(d_x^+ + \tilde{d}_x)$, we obtain

$$(50) \quad \langle g, \nabla f(x) \rangle \geq 0$$

with equality if and only if $d_x = 0$ and $d_x^+ = 0$, or, equivalently $\nabla f(x) = 0$. Since x is not critical, we must have $\nabla f(x) \neq 0$, so strict inequality holds in (50) and (41) is proved.

Take any $\alpha \in (0, \varepsilon / \|g\|)$. Since projection on a closed convex set is a nonexpansive operator (see e.g. [8] or use the Cauchy–Schwarz inequality to strengthen (B.16) in Appendix B), we have

$$(51) \quad \|x(\alpha) - x\| \leq \|x - \alpha g - x\| = \alpha \|g\| < \varepsilon.$$

Therefore we have

$$\langle a_i, x \rangle < b_i - \varepsilon \|a_i\| < b_i - \langle a_i, x(\alpha) - x \rangle \quad \forall i \in I_x$$

and as a result

$$\langle a_i, x(\alpha) \rangle < b_i \quad \forall i \in I_x.$$

It follows that $x(\alpha)$ is also the projection of the vector $x - \alpha g$ on the set $\Omega_x \supset X$ given by

$$\Omega_x = \{z | \langle a_i, z \rangle \leq b_i, i \in I_x\},$$

i.e.,

$$(52) \quad x(\alpha) = \arg \min \{\|z - (x - \alpha g)\| | z \in \Omega_x\}.$$

Now the vector d_x is easily seen to be a direction of recession of the set Ω_x , so by Lemma 1b) we have

$$\langle d_x, x(\alpha) \rangle \geq \langle d_x, x - \alpha g \rangle = \langle d_x, x + \alpha d_x^+ + \alpha \tilde{d}_x \rangle.$$

Since $\langle d_x, d_x^+ \rangle = 0$, the relation above is written by using also (47)

$$(53) \quad -\langle d_x, x - x(\alpha) \rangle \geq \alpha \langle d_x, D_x d_x \rangle.$$

In view of the fact $\tilde{d}_x \in C_x$ we have $(x + \alpha \tilde{d}_x) \in \Omega_x$, and since $x(\alpha)$ is the projection on Ω_x of $(x + \alpha d_x^+ + \alpha \tilde{d}_x)$ (cf. (52)), we have

$$\langle x + \alpha d_x^+ + \alpha \tilde{d}_x - x(\alpha), x + \alpha \tilde{d}_x - x(\alpha) \rangle \leq 0.$$

Equivalently, using the fact $\langle d_x^+, \tilde{d}_x \rangle = 0$,

$$(54) \quad -\langle d_x^+, x - x(\alpha) \rangle \geq \frac{\|x(\alpha) - (x + \alpha \tilde{d}_x)\|^2}{\alpha}.$$

By combining (53) and (54) and using the fact $\nabla f(x) = -(d_x + d_x^+)$, we obtain

$$(55) \quad \langle \nabla f(x), x - x(\alpha) \rangle \geq \alpha \langle d_x, D_x d_x \rangle + \frac{\|x(\alpha) - (x + \alpha \tilde{d}_x)\|^2}{\alpha}$$

which is the left inequality in (42). To show that the right side of (55) cannot be zero, note that if it were, then we would have both $d_x = 0$ (implying $\tilde{d}_x = 0$, $x(\alpha) = P(x - \alpha \nabla f(x))$) and $x(\alpha) = x + \alpha \tilde{d}_x$ (implying $P(x - \alpha \nabla f(x)) = x$). Since x is not critical, we arrive at a contradiction. Therefore the right inequality in (42) is also proved.

By using the mean value theorem, we have

$$(56) \quad f(x) - f[x(\alpha)] = \langle \nabla f(x), x - x(\alpha) \rangle + \langle \nabla f(\zeta_\alpha) - \nabla f(x), x - x(\alpha) \rangle$$

where ζ_α lies on the line segment joining x and $x(\alpha)$. Using (55) and (56), we obtain for all $\alpha \in (0, \varepsilon/\|g\|)$

$$(57) \quad \frac{1}{\alpha} \{f(x) - f[x(\alpha)]\} \geq \langle d_x, D_x d_x \rangle + \frac{\|x(\alpha) - (x + \alpha \tilde{d}_x)\|^2}{\alpha^2} + \left\langle \nabla f(\zeta_\alpha) - \nabla f(x), \frac{x - x(\alpha)}{\alpha} \right\rangle.$$

Using (51) and the Cauchy-Schwarz inequality, we see that

$$(58) \quad \left\langle \nabla f(\zeta_\alpha) - \nabla f(x), \frac{x - x(\alpha)}{\alpha} \right\rangle \geq -\|\nabla f(\zeta_\alpha) - \nabla f(x)\| \cdot \|g\|.$$

Since $\|\nabla f(\zeta_\alpha) - \nabla f(x)\| \rightarrow 0$ as $\alpha \rightarrow 0$, we see from (57) and (58) that if $d_x \neq 0$ then for all positive but sufficiently small α we have $f(x) > f[x(\alpha)]$. If $d_x = 0$ then $\tilde{d}_x = 0$ and using Lemma 1a)

$$(59) \quad \frac{\|x(\alpha) - (x + \alpha \tilde{d}_x)\|^2}{\alpha^2} = \frac{\|x(\alpha) - x\|^2}{\alpha^2} \geq \|x(1) - x\|^2 \quad \forall \alpha \in (0, 1].$$

From (57), (58) and (59) we see again that when $d_x = 0$, then for all positive but sufficiently small α we have $f(x) > f[x(\alpha)]$. Therefore, there exists $\bar{\alpha} > 0$ such that (43) holds in both cases where $d_x = 0$ and $d_x \neq 0$. Q.E.D.

3. Convergence analysis. The previous section has shown how a vector $x \in X$, a scalar $\varepsilon > 0$, an index set I_x satisfying

$$I_x \supset \{i \in I \mid \langle a_i, x \rangle \geq b_i - \varepsilon \|a_i\|\},$$

and a positive definite self-adjoint operator $D_x : \Gamma_x \rightarrow \Gamma_x$ where Γ_x is the subspace defined by (21)–(26), uniquely define an arc of points $x(\alpha) \in X$, $\alpha \geq 0$ where

$$x(\alpha) = P(x - \alpha g), \quad \alpha \geq 0$$

and g is defined via (23), (24), (28)–(30). Furthermore for each $x \in X$ which is not critical, Proposition 1b) shows that by choosing α sufficiently small, we can obtain a point of lower cost on this arc. Therefore any procedure that, for any given $x \in X$, chooses I_x , ε , and D_x satisfying the above requirements, coupled with a rule for selecting a point of lower cost on the corresponding arc $x(\alpha)$ leads to a descent algorithm. There is a large variety of possibilities along these lines but we will focus attention on the following broad class of methods:

We assume that we are given a continuous function $\varepsilon : X \rightarrow R$ such that

$$(60) \quad \varepsilon(x) \geq 0 \quad \forall x \in X,$$

$$(61) \quad \varepsilon(x) = 0 \Rightarrow x \text{ is critical}$$

(for example $\varepsilon(x) = \min \{\varepsilon, \|x - P[x - \nabla f(x)]\|\}$ where $\varepsilon > 0$ is a given constant). We are also given scalars $\beta \in (0, 1)$, $\sigma \in (0, 1/2)$, $\lambda_1 > 0$ and $\lambda_2 > 0$ with $\lambda_1 \leq \lambda_2$.

At the beginning of the k th iteration of the algorithm we have a vector $x_k \in X$. If x_k is critical, we set $x_{k+1} = x_k$. Else we obtain the next vector x_{k+1} as follows:

Step 1. Choose an index set $I_k \subset I$ satisfying

$$(62) \quad I_k \supset \{i \in I \mid \langle a_i, x_k \rangle \geq b_i - \varepsilon(x_k) \|a_i\|\},$$

and compute

$$(63) \quad d_k = \arg \min \{\|z + \nabla f(x_k)\| \mid z \in C_k\},$$

$$(64) \quad d_k^+ = -[\nabla f(x_k) + d_k]$$

where

$$(65) \quad C_k = \{z \mid \langle a_i, z \rangle \leq 0, i \in I_k\}.$$

Step 2. Choose a positive definite self-adjoint operator $D_k : \Gamma_k \rightarrow \Gamma_k$, where

$$(66) \quad \Gamma_k = \text{span} \{C_k \cap \{z \mid \langle z, d_k^+ \rangle = 0\}\},$$

and D_k satisfies

$$(67) \quad \|D_k\| \leq \lambda_2 \quad \text{and} \quad \lambda_1 \|z\|^2 \leq \langle z, D_k z \rangle \quad \forall z \in \Gamma_k.$$

Compute \tilde{d}_k given by

$$(68) \quad \tilde{d}_k = \arg \min \{\|z - D_k d_k\| \mid z \in C_k, \langle z, d_k^+ \rangle = 0\}.$$

Define

$$(69) \quad g_k = -(d_k^+ + \tilde{d}_k)$$

and

$$(70) \quad x_k(\alpha) = P(x_k - \alpha g_k) \quad \forall \alpha \geq 0.$$

Step 3. Set

$$(71) \quad x_{k+1} = x_k(\alpha_k)$$

where

$$(72) \quad \alpha_k = \beta^{m_k}$$

and m_k is the first nonnegative integer m satisfying

$$(73) \quad f(x_k) - f[x_k(\beta^m)] \geq \sigma \left\{ \beta^m \langle d_k, D_k d_k \rangle + \frac{\|x_k(\beta^m) - (x_k + \beta^m \tilde{d}_k)\|^2}{\beta^m} \right\}.$$

Proposition 1b) shows that x_{k+1} is well defined via the stepsize rule (71)–(73) in the sense that m_k is a (finite) integer and furthermore

$$f(x_k) > f(x_{k+1})$$

for all k for which x_k is not critical. The following proposition is our main convergence result.

PROPOSITION 2. *Every limit point of a sequence $\{x_k\}$ generated by the algorithm above is a critical point.*

Proof. Let $\{x_k\}_K$ be a subsequence of $\{x_k\}$ converging to a point \bar{x} which is not critical. We will arrive at a contradiction. Since $\{\alpha_k\}$ is bounded, we assume without loss of generality that

$$\lim_{\substack{k \rightarrow \infty \\ k \in K}} \alpha_k = \bar{\alpha}$$

where $\bar{\alpha} \in [0, 1]$. Since $\{f(x_k)\}$ decreases monotonically to $f(\bar{x})$, it follows from the form of the stepsize rule that

$$(74) \quad \lim_{\substack{k \rightarrow \infty \\ k \in K}} \alpha_k \langle d_k, D_k d_k \rangle = 0,$$

$$(75) \quad \lim_{\substack{k \rightarrow \infty \\ k \in K}} \frac{\|x_k(\alpha_k) - (x_k + \alpha_k \tilde{d}_k)\|^2}{\alpha_k} = 0.$$

We consider two cases:

Case 1 ($\bar{\alpha} > 0$). It follows from (74) and the fact $\langle d_k, D_k d_k \rangle \geq \lambda_1 \|d_k\|^2$ (cf. (67)) that $\lim_{k \rightarrow \infty, k \in K} d_k = 0$, and therefore also

$$\lim_{\substack{k \rightarrow \infty \\ k \in K}} \tilde{d}_k = 0, \quad \lim_{\substack{k \rightarrow \infty \\ k \in K}} d_k^+ = -\nabla f(\bar{x}).$$

By taking the limit as $k \rightarrow \infty$, $k \in K$, in the equation $x_k(\alpha_k) = P(x_k + \alpha_k d_k^+ + \alpha_k \tilde{d}_k)$, using the continuity of the P operator, we obtain

$$\lim_{\substack{k \rightarrow \infty \\ k \in K}} x_k(\alpha_k) = P[\bar{x} - \bar{\alpha} \nabla f(\bar{x})].$$

Therefore (75) yields

$$\bar{x} = P[\bar{x} - \bar{\alpha} \nabla f(\bar{x})].$$

Since $\bar{\alpha} > 0$ this implies that \bar{x} is critical, thereby contradicting our earlier assumption.

Case 2 ($\bar{\alpha} = 0$). It follows that for all $k \in K$ which are sufficiently large

$$(76) \quad f(x_k) - f\left[x_k\left(\frac{\alpha_k}{\beta}\right)\right] < \sigma \left\{ \frac{\alpha_k}{\beta} \langle d_k, D_k d_k \rangle + \frac{\|x_k(\alpha_k/\beta) - (x_k + (\alpha_k/\beta)\tilde{d}_k)\|^2}{\alpha_k/\beta} \right\},$$

i.e., the test (73) of the stepsize rule will be failed at least once for all $k \in K$ sufficiently large.

Since $g_k = -(d_k^+ + \tilde{d}_k)$, $\langle d_k^+, \tilde{d}_k \rangle = 0$, we have

$$(77) \quad \|g_k\|^2 = \|d_k^+\|^2 + \|\tilde{d}_k\|^2.$$

Since \tilde{d}_k is the projection of $D_k d_k$ on $C_x \cap \{z | \langle z, d_k^+ \rangle = 0\}$, we must have $\|\tilde{d}_k\| \leq \|D_k d_k\|$ and, using (67), $\|\tilde{d}_k\| \leq \lambda_2 \|d_k\|$. Therefore from (77) and the fact $\|d_k^+\| \leq \|\nabla f(x_k)\|$, $\|d_k\| \leq \|\nabla f(x_k)\|$ we obtain

$$\|g_k\|^2 \leq (1 + \lambda_2^2) \|\nabla f(x_k)\|^2.$$

It follows that

$$(78) \quad \limsup_{\substack{k \rightarrow \infty \\ k \in K}} \|g_k\| < \infty.$$

We also have

$$(79) \quad \lim_{\substack{k \rightarrow \infty \\ k \in K}} \varepsilon(x_k) = \varepsilon(\bar{x}) > 0.$$

It follows from (78), (79) and the fact $\bar{\alpha} = 0$ that for all $k \in K$ sufficiently large $\alpha_k/\beta \in (0, \varepsilon(x_k)/\|g_k\|)$ and therefore using Proposition 1b) (cf. (42)), we obtain

$$(80) \quad \left\langle \nabla f(x_k), x_k - x_k\left(\frac{\alpha_k}{\beta}\right) \right\rangle \geq \frac{\alpha_k}{\beta} \langle d_k, D_k d_k \rangle + \frac{\|x_k(\alpha_k/\beta) - (x_k + (\alpha_k/\beta)\tilde{d}_k)\|^2}{\alpha_k/\beta}.$$

Using the mean value theorem, we have

$$(81) \quad f(x_k) - f\left[x_k\left(\frac{\alpha_k}{\beta}\right)\right] = \left\langle \nabla f(x_k), x_k - x_k\left(\frac{\alpha_k}{\beta}\right) \right\rangle + \left\langle \nabla f(\zeta_k) - \nabla f(x_k), x_k - x_k\left(\frac{\alpha_k}{\beta}\right) \right\rangle$$

where ζ_k lies on the line segment connecting x_k and $x_k(\alpha_k/\beta)$. From (76), (80), and (81) we obtain for all $k \in K$ sufficiently large

$$(82) \quad (1 - \sigma) \left\{ \langle d_k, D_k d_k \rangle + \frac{\|x_k(\alpha_k/\beta) - (x_k + (\alpha_k/\beta)\tilde{d}_k)\|^2}{(\alpha_k/\beta)^2} \right\} \\ \leq \left\langle \nabla f(x_k) - \nabla f(\zeta_k), \frac{x_k - x_k(\alpha_k/\beta)}{\alpha_k/\beta} \right\rangle.$$

Since (cf. (51), (78)) we have

$$\limsup_{\substack{k \rightarrow \infty \\ k \in K}} \frac{\|x_k - x_k(\alpha_k/\beta)\|}{\alpha_k/\beta} \leq \limsup_{\substack{k \rightarrow \infty \\ k \in K}} \|g_k\| < \infty$$

and

$$\lim_{\substack{k \rightarrow \infty \\ k \in K}} \|\nabla f(x_k) - \nabla f(\zeta_k)\| = 0,$$

it follows that the right side of (82) tends to zero as $k \rightarrow \infty$, $k \in K$. Therefore so does the left side which implies that

$$(83) \quad \lim_{\substack{k \rightarrow \infty \\ k \in K}} d_k = 0, \quad \lim_{\substack{k \rightarrow \infty \\ k \in K}} \tilde{d}_k = 0$$

and

$$(84) \quad \lim_{\substack{k \rightarrow \infty \\ k \in K}} \frac{\|x_k(\alpha_k/\beta) - (x_k + (\alpha_k/\beta)\tilde{d}_k)\|^2}{(\alpha_k/\beta)^2} = 0.$$

Since it follows from (79) and (83) that there exists \bar{k} such that

$$x_k + \frac{\alpha_k}{\beta} \tilde{d}_k \in X \quad \forall k \geq \bar{k},$$

we obtain using Lemma 1a)

$$(85) \quad \frac{\|x_k(\alpha_k/\beta) - (x_k + (\alpha_k/\beta)\tilde{d}_k)\|^2}{(\alpha_k/\beta)^2} \geq \left\| P \left[\left(x_k + \frac{\alpha_k}{\beta} \tilde{d}_k \right) + d_k^+ \right] - \left(x_k + \frac{\alpha_k}{\beta} \tilde{d}_k \right) \right\|^2.$$

From (84) and (85) it follows that

$$\lim_{\substack{k \rightarrow \infty \\ k \in K}} \left\| P \left[\left(x_k + \frac{\alpha_k}{\beta} \tilde{d}_k \right) - (\nabla f(x_k) + d_k) \right] - \left(x_k + \frac{\alpha_k}{\beta} \tilde{d}_k \right) \right\|^2 = 0.$$

Using (83), we obtain

$$\|P[\bar{x} - \nabla f(\bar{x})] - \bar{x}\| = 0,$$

which contradicts the assumption that \bar{x} is not critical. Q.E.D.

We mention that some of the requirements on the sequences $\{\varepsilon(x_k)\}$ and $\{D_k\}$ can be relaxed without affecting the result of Proposition 2. In place of continuity of $\varepsilon(\cdot)$ and assumption (67) it is sufficient to require that if $\{x_k\}_K$ is a subsequence converging to a noncritical point \bar{x} , then

$$\liminf_{\substack{k \rightarrow \infty \\ k \in K}} \varepsilon(x_k) > 0,$$

$$\liminf_{\substack{k \rightarrow \infty \\ k \in K}} \inf \{ \langle z, D_k z \rangle \|z\| = 1, z \in \Gamma_k \} > 0$$

$$\limsup_{\substack{k \rightarrow \infty \\ k \in K}} \|D_k\| < \infty.$$

This can be verified by inspection of the proof of Proposition 2.

A practically important generalization of the algorithm results if we allow the norm on the Hilbert space H to change from one iteration to the next. By this we mean that at each iteration k a new inner product $\langle \cdot, \cdot \rangle_k$ and corresponding norm $\|\cdot\|_k$ on H are considered. The statement of the algorithm and corresponding assumptions must be modified as follows:

a) The gradient $\nabla f(x_k)$ will be with respect to the current inner product $\langle \cdot, \cdot \rangle_k$ (cf. (3)).

b) The projection defining d_k , d_k^+ , \tilde{d}_k and the arc $x_k(\cdot)$ should be with respect to the current norm $\|\cdot\|_k$.

c) The assumptions on I_k and D_k and the stepsize rule should be restated in terms of the current inner product and norm.

There is no difficulty in reworking the proof of Proposition 2 for this generalized version of the algorithm provided we assume that all the norms $\|\cdot\|_k$, $k=0, 1, \dots$ are “equivalent” to the original norm $\|\cdot\|$ on H in the sense that for some $m > 0$ and $M > 0$ we have

$$m\|z\| \leq \|z\|_k \leq M\|z\| \quad \forall z \in H, \quad k=0, 1, \dots$$

Naturally the norms $\|\cdot\|_k$ should be such that projection on X with respect to any one of them is relatively easy, for otherwise the purpose of the methodology of this paper is defeated. The motivation for considering a different inner product at each iteration stems from the fact that it is often desirable in nonlinear programming algorithms to introduce iteration-dependent scaling on the optimization variables. This is sometimes referred to as “preconditioning.” The use of the operator D_k fulfills that need to a great extent but while this operator scales the component d_x of the negative gradient, it does not affect at all the second component d_x^+ . The role of an iteration-dependent norm can be understood by considering situations where the index set I_k is so large that the cone C_k is empty. In this case $d_k^+ = -\nabla f(x_k)$, $\tilde{d}_k = 0$ and the k th iteration reduces to an iteration of the original Goldstein–Levitin–Poljak method, for which practical experience shows that simple, for example diagonal, scaling at each iteration can sometimes result in spectacular computational savings.

4. Rate of convergence. In this section we will analyze the rate of convergence of algorithm (62)–(73) for the case where X is polyhedral and H is finite dimensional. An important property of the Goldstein–Levitin–Poljak method (cf. [7]) is that if it generates a sequence $\{x_k\}$ converging to a strict local minimum \bar{x} satisfying certain sufficiency conditions (compare with [7]), then after some index \bar{k} the vectors x_k lie on the manifold of active constraints at \bar{x} , i.e., $x_k \in \bar{x} + N_{\bar{x}}$ where

$$(86) \quad N_{\bar{x}} = \{z | \langle a_i, z \rangle = 0, \forall i \in A_{\bar{x}}\}$$

and where

$$(87) \quad A_{\bar{x}} = \{i | i \in I, \langle a_i, \bar{x} \rangle = b_i\}.$$

Our algorithm preserves this important characteristic. Indeed, we will see that, under mild assumptions, our algorithm “identifies” the set of active constraints at the limit point in a finite number of iterations, and subsequently reduces to an unconstrained optimization method on this subspace. This brings to bear the rate of convergence results available from unconstrained optimization.

The rate of convergence analysis will be carried out under the following assumptions:

(A) H is finite dimensional, X is polyhedral, f is continuously Frechet differentiable, and ∇f is Lipschitz continuous on bounded sets, i.e., for every bounded set there exists $L > 0$ such that for every x and y in X we have

$$(88) \quad \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

(B) \bar{x} is a strict local minimum and there exists $\delta > 0$ such that

$$(89) \quad P(y) \in \bar{x} + N_{\bar{x}} \quad \forall y \text{ such that } \|\bar{x} - \nabla f(\bar{x}) - y\| \leq \delta.$$

(C) The function $\varepsilon(x)$ in the algorithm has the form

$$(90) \quad \varepsilon(x) = \min \{\varepsilon, \|x - P[x - \nabla f(x)]\|\},$$

where $\varepsilon > 0$ is a given scalar. Furthermore the set I_k in the algorithm is chosen to be (cf. (62))

$$(91) \quad I_k = \{i \in I \mid \langle a_i, x_k \rangle \geq b_i - \varepsilon(x_k) \|a_i\|\}.$$

The Lipschitz condition (88) is satisfied in particular if f is twice continuously differentiable. Condition (89) is a weakened version of an often employed regularity and strict complementarity assumption which requires that the set of vectors $\{a_i \mid i \in A_{\bar{x}}\}$ is linearly independent and all Lagrange multipliers corresponding to the active constraints are strictly positive. The form (90) for $\varepsilon(x)$ is required for technical purposes in our subsequent proof. The reader can verify that there are other forms of $\varepsilon(x)$ that are equally suitable. Finally the choice (91) for the set I_k is natural and is ordinarily the one that is best for algorithmic purposes.

The following proposition allows us to transfer rate of convergence results from unconstrained minimization to algorithm (62)–(73).

PROPOSITION 3. *Let \bar{x} be a limit point of the sequence $\{x_k\}$ generated by iteration (62)–(73), and let Assumptions (A)–(C) hold. Then*

$$(92) \quad \lim_{k \rightarrow \infty} x_k = \bar{x}$$

and there exists \bar{k} such that for all $k \geq \bar{k}$ we have

$$(93) \quad x_k \in \bar{x} + N_{\bar{x}},$$

$$(94) \quad \Gamma_k = \text{span} \{C_k \cap \{z \mid \langle z, d_k^+ \rangle = 0\}\} = N_{\bar{x}},$$

$$(95) \quad d_k = \arg \min \{\|\nabla f(x_k) + z\| \mid z \in N_{\bar{x}}\},$$

$$(96) \quad x_{k+1} = x_k + \alpha_k D_k d_k,$$

where $\alpha_k = \beta^{m_k}$ and m_k is the first nonnegative integer m for which

$$(97) \quad f(x_k) - f[x_k(\beta^m)] \geq \sigma \beta^m \langle d_k, D_k d_k \rangle.$$

The proof of Proposition 3 is given in Appendix B. From (96) and (97) we see that eventually the method reduces to an unconstrained minimization method on the manifold $\bar{x} + N_{\bar{x}}$. The proposition shows that if the matrix D_k is chosen so that for all k sufficiently large it is equal to the inverse Hessian of f restricted on the manifold $\bar{x} + N_{\bar{x}}$, then the method essentially reduces to the unconstrained Newton method and attains a superlinear rate of convergence.

5. Algorithmic variations. Many variations on iteration (62)–(73) are possible. One of them, changing the metric on the Hilbert space H from iteration to iteration, was discussed at the end of § 3. In this section we discuss other variations. These will include the use, in various cases, of a pseudometric on H instead of a metric, variations on the step size rules and finally variations on the various projections in (62)–(73). We will state the variations without a convergence proof. In each case, the reworking of the proofs of §§ 2–3 to show that the variation is valid, poses no difficulty.

Singular transformation of variables through a pseudometric. Here we address the case where X is not a solid body in H , i.e., for some linear manifold M we have $X \subset M \neq H$. In this case we observe that (42) is the only place where a metric as opposed to a pseudometric is needed. Noticing that if $X \subset M$, then all quantities in (42) belong to M , one can conclude that all that is necessary is to have a metric on M . This leads us to consider the use of pseudometric on H provided it induces a metric on M . Furthermore, we can change the pseudometric on H from iteration to

iteration, as we can change the metric, provided that the metrics induced on M are equivalent in the sense described in § 3. In some cases the introduction of a pseudometric serves to facilitate the projection further (see [17, Chap. 4]).

Stepsize rules. The Armijo-like rule (73) can be viewed as a combination of the Armijo rule used in unconstrained minimization [9], and an Armijo-like rule for constrained optimization proposed by Bertsekas in [7, cf. eq. (12)]. Corresponding to an alternate suggestion made there [7, cf. eq. (22)], we can replace (73) by

$$(98) \quad f(x_k) - f(x_k(\beta^m)) \geq \sigma \{ \beta^m \langle d_k, D_k d_k \rangle + \langle \nabla f(x_k), (x_k + \beta^m \tilde{d}_k) - x_k(\beta^m) \rangle \}.$$

Also, a variation of the Goldstein stepsize rule [9] can be employed, in which $\sigma < 0.5$ and α is chosen such that

$$(99) \quad \begin{aligned} & (1 - \sigma) \{ \alpha \langle d_k, D_k d_k \rangle + \langle \nabla f(x_k), (x_k + \alpha \tilde{d}_k) - x_k(\alpha) \rangle \} \\ & \geq f(x_k) - f(x_k(\alpha)) \\ & \geq \sigma \{ \alpha \langle d_k, D_k d_k \rangle + \langle \nabla f(x_k), (x_k + \alpha \tilde{d}_k) - x_k(\alpha) \rangle \}. \end{aligned}$$

The rule (99) is the counterpart of (98). The reader can easily construct the counterpart to (73).

Variations on the projections. There is one central observation in the paper, namely, the projections of $D_k d_k$ and d_k^+ on any closed convex set for which d_k is a direction of recession, result in descent directions. By employing different sets with this property, variations on the algorithm result since different directions may be obtained and different arcs may be searched.

The first variation is to replace C_k in (68) by $(\Omega_k - x_k)$, i.e.

$$(100) \quad \tilde{d}_k = \arg \min \{ \|z - D_k d_k\| \mid z \in \Omega_k - x_k, \langle z, d_k^+ \rangle = 0 \}$$

where

$$\Omega_k = \{z \mid \langle a_i, z \rangle \leq b_i, \forall i \in I_k\}.$$

Evidently

$$\Omega_k - x_k \supset C_k$$

and as a result d_k is a direction of recession of $\Omega_k - x_k$, which implies that \tilde{d}_k defined by (100) is a descent direction.

Interestingly, this variation gives rise to a variation in the stepsize search. Since the set $\{z \mid z \in C_k, \langle z, d_k^+ \rangle = 0\}$ is a cone, the vector \tilde{d}_k of (68) satisfies

$$\alpha \tilde{d}_k = \arg \min \{ \|\alpha D_k d_k - z\| \mid z \in C_k, \langle z, d_k^+ \rangle = 0 \}.$$

Thus, (70) can be interpreted as

$$x_k(\alpha) = P[x_k + \alpha d_k^+ + q_k(\alpha)]$$

where

$$q_k(\alpha) = \arg \min \{ \|\alpha D_k d_k - z\| \mid z \in C_k, \langle z, d_k^+ \rangle = 0 \}.$$

When C_k is replaced by $\Omega_k - x_k$, a new algorithm results by searching along the arc

$$x_k(\alpha) = P[x_k + \alpha d_k^+ + \tilde{q}_k(\alpha)]$$

where

$$\tilde{q}_k(\alpha) = \arg \min \{ \|\alpha D_k d_k - z\| \mid z \in \Omega_k - x_k, \langle z, d_k^+ \rangle = 0 \}.$$

Indeed, the particular algorithm suggested in [6] can be considered to be an implementation of the last variation for an orthant constraint.

6. Multicommodity network flow problems. In this last section we apply algorithm (62)–(73) to a classical nonlinear multicommodity network flow problem and present some computational results. In view of the typically very large number of variables and constraints of this problem, active set methods of the type presented in [4] are in our view entirely unsuitable.

We consider a network consisting of N nodes, $1, 2, \dots, N$, and a set of directed links denoted by \mathcal{L} . We assume that the network is connected in the sense that for any two nodes m, n , there is a directed path from m to n . We are given a set W of ordered node pairs referred to as origin-destination (or OD) pairs. For each OD pair $w \in W$, we are given a set of directed paths P_w that begin at the origin node and terminate at the destination node. For each $w \in W$ we are also given a positive scalar r_w referred to as the input of OD pair w . This input must be optimally divided among the paths in P_w so as to minimize a certain objective function.

For every path $p \in P_w$ corresponding to an OD pair $w \in W$ we denote by x^p the flow travelling on p . These flows must satisfy

$$(101) \quad \sum_{p \in P_w} x^p = r_w \quad \forall w \in W,$$

$$(102) \quad x^p \geq 0 \quad \forall p \in P_w, w \in W.$$

Equations (101), (102) define the constraint set of the optimization problem—a Cartesian product of simplices.

In Example 2 we discussed the application of our method to the case of a simplex constraint. It is not difficult to see that if we take a “diagonal” metric on the space, the multicommodity flow problem decomposes in the sense explained below.

Let x denote the vector of variables $x^p, p \in P_w, w \in W$, and let x^w denote the vector of variables $x^p, p \in P_w$. Let $C_x(x^w)$ and $\Gamma_x(x^w)$ denote the cone and subspace, respectively, in $R^{|w|}$, generated at x , when all variables aside from those in x^w are considered fixed and $\varepsilon = \varepsilon(x)$. Then

$$C_x = \prod_{w \in W} C_x(x^w), \quad \nabla f(x) = (\dots, \nabla_{x^w} f(x), \dots), \quad \Gamma_x = \prod_{w \in W} \Gamma_x(x^w).$$

Thus all projections decompose and therefore in many respects the multicommodity flow problem is not different from the problem with a single simplex constraint. The only points where the “interaction” among the simplices appears is in computing ε_k , and in computing $D_k d_k$.

To every set of path flows $\{x^p | p \in P_w, w \in W\}$ satisfying (101), (102) there corresponds a flow f^a for every link $a \in \mathcal{L}$. It is defined by the relation

$$(103) \quad f^a = \sum_{w \in W} \sum_{p \in P_w} 1_p(a) x^p \quad \forall a \in \mathcal{L}$$

where $1_p(a) = 1$ if the path p contains the link a and $1_p(a) = 0$ otherwise. If we denote by f the vector of link flows, we can write relation (103) as

$$(104) \quad f = Ex$$

where E is the arc-chain matrix of the network.

For each link $a \in \mathcal{L}$ we are given a convex, twice continuously differentiable scalar function $D_a(f^a)$ with strictly positive second derivative for all $f^a \geq 0$. The objective

function is given by

$$(105) \quad D(f) = \sum_{a \in \mathcal{L}} D_a(f^a).$$

By using (104), we can write the problem in terms of the path flow variables x^p as

$$\text{minimize } J(x) = D(Ex)$$

subject to:

$$\begin{aligned} \sum_{p \in P_w} x^p &= r_w \quad \forall w \in W, \\ x^p &\geq 0 \quad \forall p \in P_w, w \in W. \end{aligned}$$

In communication network applications the function D may express, for example, average delay per message [10], [11] or a flow control objective [12], while in transportation networks it may arise via a user or system optimization principle formulation [13], [14], [15]. We concentrate on the separable form of D given by (105), although what follows admits an extension to the nonseparable case.

A Newton-like method will be obtained if we chose $D_k d_k$ so that $x_k + D_k d_k$ is the minimum of the quadratic approximation to f on $x_k + \Gamma_k$. For this we must find $A\bar{v}$ where \bar{v} solves

$$(106) \quad \text{minimize}_v \langle \nabla J(x_k), Av \rangle + \frac{1}{2} \langle Av, \nabla^2 J(x_k) Av \rangle$$

and where A is a matrix such that its columns are linearly independent and span Γ_k .

The particular structure of the objective function (105) gives rise to a Hessian matrix which makes the solution of (106) relatively easy to obtain. Indeed, using (105) we can rewrite (106) as

$$(107) \quad \text{minimize}_v \langle E' \nabla D(f_k), Av \rangle + \frac{1}{2} \langle Av, E' \nabla^2 D(f_k) EA v \rangle,$$

where $f_k = Ex_k$ and prime denotes transposition. A key fact (described in detail in Bertsekas and Gafni [16]) is that problem (107), in light of $\nabla^2 D(f_k)$ being diagonal, can be solved by the Conjugate Gradient (C-G) method using graph type operations without explicitly storing the matrix

$$A' E' \nabla^2 D(f_k) EA.$$

Note that a solution to (107) exists since $E' \nabla D(f_k)$ is in the range of the nonnegative definite matrix $E' \nabla^2 D(f_k) E$.

Computational results. A version of the algorithm was run on an example of the multicommodity flow problem. The network is shown in Fig. 5. Each OD pair was restricted to use only two prespecified paths. This reduced the programming load significantly, yet captured the essence of the algorithm. It is conjectured that the results we obtained are representative of the behavior of the algorithm when applied to more complex multicommodity flow problems.

The algorithm was operated in three modes distinguished by the other rules according to which the C-G method was stopped. In the first mode (denoted by Newton) the C-G iteration was run to the exact solution of problem (107). In the second mode, (denoted by approximate Newton) the C-G iteration was run until its residual was reduced by a factor of $\frac{1}{8}$ over the starting residual (this factor was chosen on a heuristic basis). Finally, in the third mode the C-G method was allowed to perform only one

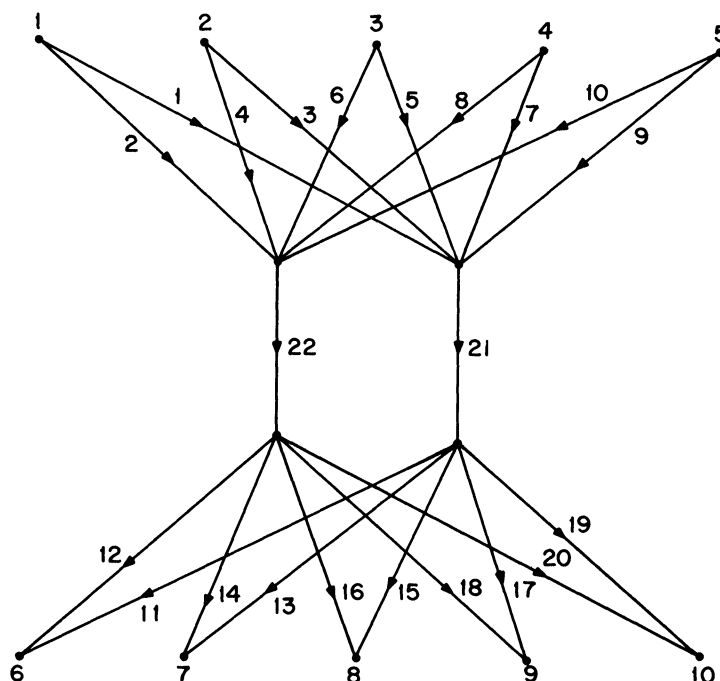


FIG. 5. The network; initially all flows traverse link 21.

step (denoted by 1-step—this results in a diagonally scaled version of the original Goldstein–Levitin–Poljak method). In all these modes, in addition to their particular stopping rule, the C-G method was stopped whenever for any OD pair w the flow on the path with the smallest partial derivative of cost became negative. Each time this happened, the last point in the sequence of points generated by the C-G method subiteration was connected by a line to the point preceding it. The point on the line at which the particular path flow became zero was taken as the result of the C-G iteration. We used different values ε_k for different OD pairs, according to a variation of (60) (with $\varepsilon = 0.2$).

We used two types of objective functions. The first is

$$D_a(f^a) = \frac{f^a}{C_a - f^a} \quad \forall a \in \mathcal{L}$$

where C_a is a given positive scalar expressing the “capacity” of link a . This function is typically used to express queueing delay in communication networks. The second type was taken to be quadratic. We used two sets of inputs, one to simulate heavy loading and one to simulate light loading. For each combination of cost function and input we present the results corresponding to the three versions in Table 1.

Our main observation from the results of Table 3 as well as additional experimentation with multicommodity flow problems is that in the early iterations the 1-step method makes almost as much progress as the other two more sophisticated methods but tends to slow down considerably after reaching the vicinity of the optimum. Also the approximate Newton method does almost as well as Newton’s method in terms of number of iterations. However the computational overhead per iteration for Newton’s method is considerably larger. This is reflected in the results of Table 3 which show

TABLE 1
Capacities.

$M =$	16.5	20	7.5	10	10
	15	5	9	7.5	7.5
	3	15	6	8	3
	10	6	10	10	14
	50	35	x	x	x

$C_a = m_{ij}, i = \frac{a}{5} + 1, j = a - 5(i - 1)$

TABLE 2
Low input. High input = low input \times 1.75.

origin \ destination	6	7	8	9	10
1	0.5	1	1.5	2	2.5
2	1	1	1	1	1
3	0.5	0.5	1.5	1.5	3.5
4	0.25	0.25	2	0.25	0.25
5	0.75	0.75	0.75	0	0

TABLE 3

	Initial value	Final value	No. of iterations	Total no. of C-G subiterations
Low load				
Nonquadratic objective	$1.600616 \cdot 10^6$			
Newton		8.743550	16	29
Approximate Newton		8.758665	16	16
1-step		8.758665	16	16
Quadratic objective	$1.866326 \cdot 10^1$			
Newton		7.255231	5	17
Approximate Newton		7.255231	7	13
1-step		7.255231	12	12
High load				
Nonquadratic objective	$9.759996 \cdot 10^6$			
Newton		$3.737092 \cdot 10^1$	14	117
Approximate Newton		$3.737745 \cdot 10^1$	15	30
1-step		$3.747400 \cdot 10^1$	15	15
Quadratic objective	$9.759996 \cdot 10^6$			
Newton		$1.521299 \cdot 10^1$	5	24
Approximate Newton		$1.521299 \cdot 10^1$	13	27
1-step		$1.521301 \cdot 10^1$	16	16

in three cases out of four a larger number of conjugate gradient subiterations for Newton's method. Throughout our computational experiments (see also [17]) the approximate Newton method based on conjugate gradient subiterations has performed very well and, together with its variations, is in our view the most powerful class of methods available at present for nonlinear multicommodity network flow problems.

Appendix A.

Proof of Lemma 1. a) Fix $x \in X$, $z \in H$ and $\gamma > 1$. Denote

$$(A.1) \quad a = x + z, \quad b = x + \gamma z.$$

Let \bar{a} and \bar{b} be the projections on X of a and b respectively. It will suffice to show that

$$(A.2) \quad \|\bar{b} - x\| \leq \gamma \|\bar{a} - x\|.$$

If $\bar{a} = x$ then clearly $\bar{b} = x$, so (A.2) holds. Also if $a \in X$ then $\bar{a} = a = x + z$ so (A.2) becomes $\|\bar{b} - x\| \leq \gamma \|z\| = \|b - x\|$ which again holds by the contraction property of the projection. Finally if $\bar{a} = \bar{b}$ then (A.2) also holds. Therefore it will suffice to show (A.2) in the case where $\bar{a} \neq \bar{b}$, $\bar{a} \neq x$, $\bar{b} \neq x$, $\bar{a} \notin X$, $\bar{b} \notin X$ shown in Fig. (A.1).

Let H_a and H_b be the two hyperplanes that are orthogonal to $(\bar{b} - \bar{a})$ and pass through \bar{a} and \bar{b} respectively. Since $\langle \bar{b} - \bar{a}, b - \bar{b} \rangle \geq 0$ and $\langle \bar{b} - \bar{a}, a - \bar{a} \rangle \leq 0$, we have that neither a nor b lie strictly between the two hyperplanes H_a and H_b . Furthermore x lies on the same side of H_a as a , and $x \notin H_a$. Denote the intersections of the line $\{x + \alpha(b - x) | \alpha \in \mathbb{R}\}$ with H_a and H_b by s_a and s_b respectively. Denote the intersection of the line $\{x + \alpha(\bar{a} - x) | \alpha \in \mathbb{R}\}$ with H_b by w . We have

$$(A.3) \quad \begin{aligned} \gamma &= \frac{\|b - x\|}{\|a - x\|} \geq \frac{\|s_b - x\|}{\|s_a - x\|} = \frac{\|w - x\|}{\|\bar{a} - x\|} = \frac{\|w - \bar{a}\| + \|\bar{a} - x\|}{\|\bar{a} - x\|} \\ &\geq \frac{\|\bar{b} - \bar{a}\| + \|\bar{a} - x\|}{\|\bar{a} - x\|} \geq \frac{\|\bar{b} - x\|}{\|\bar{a} - x\|} \end{aligned}$$

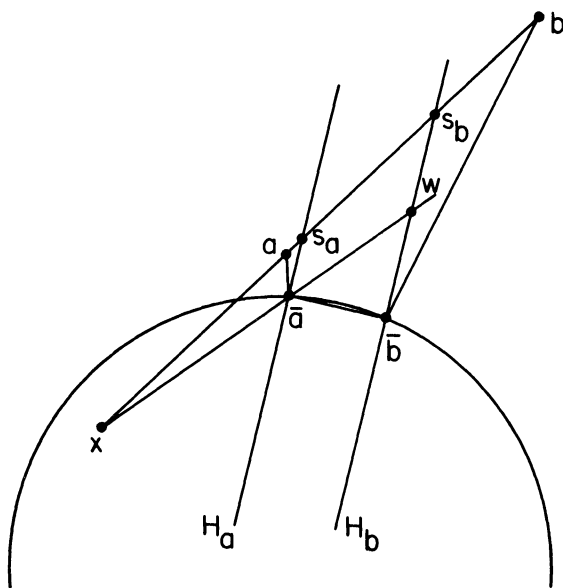


FIG. A.1

where the third equality is by similarity of triangles, the next to last inequality follows from the orthogonality relation $\langle w - \bar{b}, \bar{b} - \bar{a} \rangle = 0$, and the last inequality is obtained from the triangle inequality. From (A.3) we obtain (A.2) which was to be proved.

b) Since y is a direction of recession of Ω , we have

$$(A.4) \quad P_{\Omega}(x+z) + y \in \Omega.$$

Thus by definition of projection on a closed convex set

$$(A.5) \quad \langle (x+z) - P_{\Omega}(x+z), (P_{\Omega}(x+z) + y) - P_{\Omega}(x+z) \rangle \leq 0$$

or equivalently

$$\langle (x+z) - P_{\Omega}(x+z), y \rangle \leq 0,$$

and (40) follows. Q.E.D.

Appendix B. We develop the main arguments for the proof of Proposition 3 through a sequence of lemmas. In what follows we use the word “eventually” to mean “there exists \bar{k} such that for all $k \geq \bar{k}$,” where \bar{k} may be different for each case.

LEMMA B.1. *Under the conditions of Proposition 3, $\lim_{k \rightarrow \infty} x_k = \bar{x}$ and eventually*

$$(B.1) \quad I_k = A_{\bar{x}}.$$

Proof. By relation (73), since \bar{x} is a limit point and the algorithm decreases the value of the objective function at each iteration, we have

$$\lim_{k \rightarrow \infty} \|x_{k+1} - x_k\| = 0,$$

which implies, again by the descent property and the fact that \bar{x} is a strict local minimum

$$(B.2) \quad \lim_{k \rightarrow \infty} x_k = \bar{x}.$$

Therefore from (90)

$$(B.3) \quad \lim_{k \rightarrow \infty} \varepsilon(x_k) = \varepsilon(\bar{x}) = 0.$$

Since the set I is finite, it follows from (87), (91) and (B.3) that eventually

$$(B.4) \quad I_k \subset A_{\bar{x}}.$$

To show the reverse inclusion we must show that eventually

$$(B.5) \quad \langle a_i, x_k \rangle \geq b_i - \varepsilon(x_k) \|a_i\| \quad \forall i \in A_{\bar{x}}.$$

By the Cauchy-Schwarz inequality, (B.3) and (90) we have eventually

$$\varepsilon(x_k) \|a_i\| = \|x_k - P[x_k - \nabla f(x_k)]\| \cdot \|a_i\| \geq \langle P[x_k - \nabla f(x_k)] - x_k, a_i \rangle.$$

Therefore in order to show (B.5) it suffices to show that eventually

$$\langle a_i, P[x_k - \nabla f(x_k)] \rangle = b_i \quad \forall i \in A_{\bar{x}}$$

or equivalently

$$P[x_k - \nabla f(x_k)] \varepsilon \bar{x} + N_{\bar{x}}.$$

Since $x_k \rightarrow \bar{x}$ this follows from Assumption (B). Q.E.D.

LEMMA B.2. *Under the conditions of Proposition 3 for each $\bar{\alpha} \in (0, 1]$, eventually we have*

$$(B.6) \quad x_k(\alpha) \in \bar{x} + N_{\bar{x}} \quad \forall \alpha \in [\bar{\alpha}, 1].$$

Proof. From Lemma B.1 we have $x_k \rightarrow \bar{x}$ and eventually $C_k = \bar{C}$ where

$$(B.7) \quad \bar{C} = \{z | \langle z, a_i \rangle \leq 0, \forall i \in A_{\bar{x}}\}.$$

Since the projection of $-\nabla f(\bar{x})$ on \bar{C} is the zero vector and d_k is eventually the projection of $-\nabla f(x_k)$ on \bar{C} it follows that

$$(B.8) \quad \lim_{k \rightarrow \infty} d_k = 0.$$

Since \tilde{d}_k is the projection of $D_k d_k$ on a subset of C_k , and $\{\|D_k\|\}$ is bounded above (cf. (67), (68)), it follows that

$$(B.9) \quad \lim_{k \rightarrow \infty} \tilde{d}_k = 0.$$

Since $-\nabla f(x_k) = d_k^+ + d_k$ and $g_k = -(d_k^+ + \tilde{d}_k)$

$$(B.10) \quad \lim_{k \rightarrow \infty} g_k = \nabla f(\bar{x}).$$

A simple argument shows that Assumption (B) implies that for all $\alpha \in [0, 1]$

$$(B.11) \quad P(y) \in \bar{x} + N_{\bar{x}} \quad \forall y \text{ such that } \|\bar{x} - \alpha \nabla f(\bar{x}) - y\| \leq \alpha \delta.$$

For any $\bar{\alpha} \in (0, 1]$, equation (B.10) shows that we have eventually

$$\|\bar{x} - \alpha \nabla f(\bar{x}) - (x_k - \alpha g_k)\| \leq \alpha \delta \quad \forall \alpha \in [\bar{\alpha}, 1].$$

Therefore from (B.11) we have eventually

$$x_k(\alpha) = P(x_k - \alpha g_k) \in \bar{x} + N_{\bar{x}} \quad \forall \alpha \in [\bar{\alpha}, 1]. \quad \text{Q.E.D.}$$

LEMMA B.3. *Under the conditions of Proposition 3*

$$\liminf_{k \rightarrow \infty} \alpha_k > 0.$$

Proof. From Lemma B.1 we have eventually $I_k = A_{\bar{x}}$ and $x_k \rightarrow \bar{x}$, while from (B.8) we have $\|g_k\| \rightarrow \|\nabla f(\bar{x})\|$. Therefore from Proposition 1b) [cf. (42)] it follows that there exists $\hat{\alpha} > 0$ such that eventually

$$\langle \nabla f(x_k), x_k - x_k(\alpha) \rangle \geq \alpha \langle d_k, D_k d_k \rangle + \frac{1}{\alpha} \|x_k(\alpha) - (x_k + \alpha \tilde{d}_k)\|^2 \quad \forall \alpha \in (0, \hat{\alpha}).$$

Using this relation, we get that eventually

$$\begin{aligned} f(x_k) - f[x_k(\alpha)] &\geq \langle \nabla f(x_k), x_k - x_k(\alpha) \rangle - \frac{L}{2} \|x_k(\alpha) - x_k\|^2 \\ &\geq \alpha \langle d_k, D_k d_k \rangle + \frac{1}{\alpha} \|x_k(\alpha) - (x_k + \alpha \tilde{d}_k)\|^2 - \frac{L}{2} \|x_k(\alpha) - x_k\|^2 \\ &\geq \alpha \langle d_k, D_k d_k \rangle + \frac{1}{\alpha} \|x_k(\alpha) - (x_k + \alpha \tilde{d}_k)\|^2 \\ &\quad - L \|\alpha \tilde{d}_k\|^2 - L \|x_k(\alpha) - (x_k + \alpha \tilde{d}_k)\|^2 \\ &\geq \alpha (1 - \alpha L \lambda_2) \langle d_k, D_k d_k \rangle + \left(\frac{1}{\alpha} - L \right) \|x_k(\alpha) - (x_k + \alpha \tilde{d}_k)\|^2 \end{aligned}$$

where the third inequality follows from

$$\|x + y\|^2 \leq 2\|x\|^2 + 2\|y\|^2,$$

the last inequality follows from (67) and L is a Lipschitz constant that corresponds to any nonempty bounded neighborhood of \bar{x} . Taking any $\bar{\alpha} > 0$ satisfying

$$\bar{\alpha} \leq \hat{\alpha}, \quad 1 - \bar{\alpha}L\lambda_2 > \sigma, \quad \bar{\alpha} \left(\frac{1}{\bar{\alpha}} - L \right) > \sigma$$

we obtain, using (73) that

$$\liminf_{k \rightarrow \infty} \alpha_k > \bar{\alpha}$$

and the Lemma is proved. Q.E.D.

Proof of Proposition 3. The fact $\lim_{k \rightarrow \infty} x_k = \bar{x}$ is part of Lemma B.1, while (93) follows from Lemmas B.2 and B.3.

In order to show (94) we note that from Lemma B.1 and (B.8) we have eventually

$$(B.12) \quad C_k = \bar{C}, \quad C_k^+ = \bar{C}^+$$

and

$$(B.13) \quad \lim_{k \rightarrow \infty} d_k^+ = -\nabla f(\bar{x}).$$

Equation (B.13) implies that eventually assumption (B) holds with d_k^+ replacing $-\nabla f(\bar{x})$ and $\delta/2$ replacing δ . Therefore for all $i \in A_{\bar{x}}$ and $\rho_i > 0$ such that $\|\rho_i a_i\| < \delta/2$ we have

$$(B.14) \quad P(\bar{x} + d_k^+ \pm \rho_i a_i) \in \bar{x} + N_{\bar{x}},$$

$$(B.15) \quad P(\bar{x} + d_k^+) \in \bar{x} + N_{\bar{x}}.$$

For any $z_1, z_2 \in H$ we have from a general property of projection on X

$$\langle z_1 - P(z_1), P(z_2) - P(z_1) \rangle \leq 0,$$

$$\langle z_2 - P(z_2), P(z_1) - P(z_2) \rangle \leq 0.$$

By adding these two inequalities, we obtain

$$(B.16) \quad \|P(z_1) - P(z_2)\|^2 \leq \langle z_1 - z_2, P(z_1) - P(z_2) \rangle \quad \forall z_1, z_2 \in H.$$

By applying (B.16), we obtain

$$(B.17) \quad \|P(\bar{x} + d_k^+ \pm \rho_i a_i) - P(\bar{x} + d_k^+)\|^2 \leq \langle \pm \rho_i a_i, P(\bar{x} + d_k^+ \pm \rho_i a_i) - P(\bar{x} + d_k^+) \rangle.$$

Since $\langle a_i, z \rangle = 0$ for all $z \in N_{\bar{x}}$, $i \in A_{\bar{x}}$ it follows from (B.14), (B.15) that the right side of (B.17) is zero and therefore eventually

$$P(\bar{x} + d_k^+ \pm \rho_i a_i) = P(\bar{x} + d_k^+) \quad \forall i \in A_{\bar{x}}.$$

Since from (B.12) we have eventually $d_k^+ \in \bar{C}^+$, it follows that $P(\bar{x} + d_k^+) = \bar{x}$ and therefore also

$$P(\bar{x} + d_k^+ \pm \rho_i a_i) = \bar{x} \quad \forall i \in A_{\bar{x}}.$$

Hence eventually

$$d_k^+ \pm \rho_i a_i \in \bar{C}^+ \quad \forall i \in A_{\bar{x}}$$

which implies that

$$(B.18) \quad \langle d_k^+ \pm \rho_i a_i, y \rangle \leq 0 \quad \forall y \in \bar{C}, i \in A_{\bar{x}}.$$

Let

$$y \in \{z | z \in C_k, \langle z, d_k^+ \rangle = 0\}.$$

From (B.12) and (B.18) we have eventually

$$\langle a_i, y \rangle = 0 \quad \forall i \in A_{\bar{x}},$$

or equivalently $y \in N_{\bar{x}}$. Hence eventually

$$N_{\bar{x}} \supset \{z | z \in C_k, \langle z, d_k^+ \rangle = 0\}$$

and it follows that

$$\text{span } N_{\bar{x}} = N_{\bar{x}} \supset \text{span } \{C_k \cap \{z | \langle z, d_k^+ \rangle = 0\}\} = \Gamma_k.$$

To show the reverse inclusion, note that if $y \in N_{\bar{x}}$ then by Assumption (B) and (B.12) we have eventually

$$\langle y, d_k^+ \rangle = 0.$$

Since $N_{\bar{x}} \subset \bar{C}$ and eventually $C_k = \bar{C}$, it follows that eventually $y \in C_k \cap \{z | \langle z, d_k^+ \rangle = 0\}$ and a fortiori $y \in \text{span } \{C_k \cap \{z | \langle z, d_k^+ \rangle = 0\}\} = \Gamma_k$. Therefore eventually

$$N_{\bar{x}} \subset \Gamma_k$$

and the proof of (94) is complete.

Since d_k is the projection of $-\nabla f(x_k)$ on $C_k \cap \{z | \langle z, d_k^+ \rangle = 0\}$, equation (95) follows easily from (94).

Also from (93) and (94) we have eventually $x_k \in \bar{x} + N_{\bar{x}}$, $d_k \in N_{\bar{x}}$, $\tilde{d}_k \in N_{\bar{x}}$ and d_k^+ is orthogonal to $N_{\bar{x}}$, while by Lemma B.2 the vector x_{k+1} is the projection of $x_k + \alpha_k(\tilde{d}_k + d_k^+)$ on $\bar{x} + N_{\bar{x}}$. Therefore (96) and (97) follow. Q.E.D.

REFERENCES

- [1] A. A. GOLDSTEIN, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709–710.
- [2] E. S. LEVITIN AND B. T. POLJAK, *Constrained minimization problems*, USSR Comp. Math. Phys., 6 (1966), pp. 1–50.
- [3] U. M. GARCIA-PALOMARES, *Superlinearly convergent algorithms for linearly constrained optimization*, in Nonlinear Programming 2, O. L. Mangasarian, R. R. Meyer and S. M. Robinson, eds., Academic Press, New York, 1975, pp. 101–121.
- [4] P. E. GILL, W. MURRAY AND M. H. WRIGHT, *Practical Optimization*, Academic Press, New York, 1981.
- [5] J. C. DUNN, *Global and asymptotic rate of convergence estimates for a class of projected gradient processes*, this Journal, 18 (1981), pp. 659–674.
- [6] D. P. BERTSEKAS, *Projected Newton methods for optimization problems with simple constraints*, this Journal, 20 (1982), pp. 221–246.
- [7] ———, *On the Goldstein–Levitin–Poljak gradient projection method*, Proc. 1974 IEEE Conf. on Decision and Control, Phoenix, Az., pp. 47–52; IEEE Trans. Automat. Control, AC-20 (1976), pp. 174–184.
- [8] J.-J. MOREAU, *Convexity and duality*, in Functional Analysis and Optimization, E. R. Caianiello, ed., Academic Press, New York, 1966.
- [9] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.
- [10] R. G. GALLAGER, *A minimum delay routing algorithm using distributed computation*, IEEE Trans. Comm., COM-25 (1977), pp. 73–85.
- [11] D. P. BERTSEKAS, E. M. GAFNI AND R. G. GALLAGER, *Second derivative algorithms for minimum delay distributed routing in networks*, Report LIDS-R-1082, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, May 1979; IEEE Trans. Comm., COM-32 (1984), to appear.
- [12] R. G. GALLAGER AND S. J. GOLESTAANI, *Flow control and routing algorithms for data networks*, Proc. Fifth International Conference on Computer Communication (ICCC-80), Atlanta, GA, Nov. 1980, pp. 779–784.
- [13] D. P. BERTSEKAS AND E. M. GAFNI, *Projection methods for variational inequalities with application to the traffic assignment problem*, in Math. Prog. Study, D. C. Sorensen and R. J.-B. Wets, eds., North-Holland, Amsterdam, 1982, pp. 139–159.

- [14] S. DAFERMOS, *Traffic equilibrium and variational inequalities*, Transportation Sci., 14 (1980), pp. 42–54.
- [15] H. Z. AASHTIANI AND T. L. MAGNANTI, *Equilibria on a congested transportation network*, SIAM J. Alg. Disc. Meth., 2 (1981), pp. 213–226.
- [16] D. P. BERTSEKAS AND E. M. GAFNI, *Projected Newton methods and optimization of multicommodity flows*, LIDS Rep. P-1140, Massachusetts Institute of Technology, Cambridge, 1981, IEEE Trans. Automat. Control, AC-28 (1983), pp. 1090–1096.
- [17] E. M. GAFNI, *The integration of routing and flow-control for voice and data in a computer communication network*, PhD. dissertation, Dept. Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Aug. 1982.
- [18] D. P. BERTSEKAS, G. S. LAUER, N. R. SANDELL, JR AND T. A. POSBERGH, *Optimal short-time scheduling of large-scale power systems*, IEEE Trans. Automat Control, AC-28 (1983), pp. 1–11.
- [19] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

ON MINIMUM COST PER UNIT TIME CONTROL OF MARKOV CHAINS*

VIVEK S. BORKAR†

Abstract. The “minimum cost per unit time” control problem is studied for a class of Markov chains that, though important in applications, does not fit the conventional framework for this problem. Existence of optimal stationary strategies and necessary and sufficient conditions for optimality are established.

Key words. Optimal control, cost per unit time, stationary strategy, stable strategy, empirical measures, invariant probabilities

1. Introduction. This paper studies the minimum cost per unit time control problem for a class of Markov chains. This class is characterised by the fact that the chain can move from any given state to at most finitely many neighbouring states. Though this class arises often in practice, it fails to fall within the conventional framework for this problem, as will be explained later. Thus the traditional approach of treating this problem as a limiting case of the discounted cost problem fails in this case. (A concrete example of such a failure occurs in the control of two queues in tandem. See [6].) Here, a different and more direct method is employed to establish the existence of optimal stationary strategies.

2. Problem description. Suppose $\{X_n, n = 1, 2, \dots\}$ is a controlled Markov chain on the state space $S = \{1, 2, \dots\}$. Associated with it is a transition matrix P_u indexed by the “control vector” $u = [u_1, u_2, \dots]$, such that: For each $i, j \in S$, u_i belongs to a compact metric space $D(i)$ and the (i, j) th element of P_u is $p(i, j, u_i) \in [0, 1]$ with

$$\sum_{j \in S} p(i, j, u_i) = 1, \quad i \in S.$$

The functions $p(i, j, \cdot)$ are assumed to be continuous. Let $L = \prod_{i \in S} D(i)$ with the product topology suitably metrized. Throughout this paper, for any Polish space Z , $M(Z)$ will denote the Polish space of probability measures on Z with the metric topology of the Prokhorov metric [1]. Let $\bar{M}(L)$ denote the closed subset of $M(L)$ such that $\Phi \in \bar{M}(L)$ has the form

$$\Phi = \prod_{i \in S} \Phi(i)$$

where $\Phi(i)$ is a probability measure on $D(i)$ for each i .

A control strategy (CS) is a sequence $\{\xi_n\}$ of L -valued random variables such that for each n , the law of ξ_n is in $\bar{M}(L)$ and $\xi_n = [\xi_n(1), \xi_n(2), \dots]$ satisfies:

$$(*) \quad P(X_{n+1} = i / G_n) = p(X_n, i, \xi_n(X_n)), \quad i \in S,$$

where G_n is the smallest σ -field with respect to which $X_m, \xi_m, m \leq n$, are measurable.

Remarks. (i) More generally, we may consider $\{\xi_n\}$ as above without the restriction that the law of ξ_n should be in $\bar{M}(L)$ for each n . However, (*) shows that the transition mechanism of the Markov chain at time n depends only on $\xi_n(X_n)$. Hence the above restriction entails no loss of generality.

(ii) This notation is in the spirit of [2]. It differs from the more conventional notations, but seems better suited for the purposes of this paper.

* Received by the editors August 16, 1982, and in revised form November 21, 1983.

† Tata Institute of Fundamental Research, P.O. Box 1234, Bangalore 560 012, India.

The following subclasses of control strategies will be of interest:

(1) If $\{\xi_n\}$ are i.i.d. with common law $\Phi \in \bar{M}(L)$, call it a stationary randomized strategy (SRS) and denote it by $\gamma[\Phi]$.

(2) If Φ above is a point mass concentrated at $\xi \in L$, call it a stationary strategy (SS) and denote it by $\gamma\{\xi\}$.

Under either of these, the chain has stationary transition probabilities given respectively by $E_\Phi[p(i, j, \xi'(i))]$ and $p(i, j, \xi(i))$, $i, j \in S$. (Here, $\xi' = [\xi'(1), \xi'(2), \dots]$ is an L -valued random variable with law Φ , $E_\Phi[\cdot]$ being the expectation with respect to Φ .) Let $P[\Phi]$, $P\{\xi\} = P_\xi$ denote the corresponding transition probability matrices. We shall assume that the chain has a single communicating class under any SRS.

(iii) If the chain is positive recurrent under $\gamma[\Phi]$, call $\gamma[\Phi]$ a stable SRS (write SSRS).

(iv) If the chain is positive recurrent under $\gamma\{\xi\}$, call $\gamma\{\xi\}$ a stable SS (write SSS).

Under an SSRS $\gamma[\Phi]$ or an SSS $\gamma\{\xi\}$, the chain will have a unique invariant probability, denoted $\pi[\Phi]$, $\pi\{\xi\}$ resp.

Let $k: S \rightarrow [0, \infty)$ be a given cost function and let

$$\psi_{kn} = \left(\sum_{m=1}^n k(X_m) \right) / n,$$

$$\psi_{k\infty} = \limsup_{n \rightarrow \infty} \psi_{kn},$$

$$\bar{\psi}_{k\infty} = \liminf_{n \rightarrow \infty} \psi_{kn}.$$

The “minimum cost per unit time” control problem consists of finding a CS that a.s. minimizes $\psi_{k\infty}$. In general, such a CS need not exist [7]. If it does, call it an optimal CS. Under an SSRS $\gamma[\Phi]$ or SSS $\gamma\{\xi\}$, $\psi_{k\infty}$ a.s. equals the expectation of k w.r.t. $\pi[\Phi]$ ($\pi\{\xi\}$ resp.) denoted by $C_k[\Phi]$ ($C_k\{\xi\}$ resp.). (This expectation can be $+\infty$.) Let α be the infimum of $C_k\{\xi\}$ over all SSS and β the infimum of $C_k[\Phi]$ over all SSRS. We assume that at least one SSS exists and thus α, β are well-defined. Since each SSS is an SSRS, $\beta \leq \alpha$. Let

$$\eta = \liminf_{i \rightarrow \infty} k(i).$$

We shall assume that $\eta < \infty$. (The case $\eta = \infty$ has already been studied in [2].) If k is monotone increasing, $\eta > \beta$. With this in mind, we call k “almost monotone” when $\eta > \beta$.

We introduce another weaker concept of optimality for a SSS. We shall say that an SSS $\gamma\{\xi_0\}$ is weak sense optimal if $C_k\{\xi_0\} = \alpha$ and under any arbitrary CS,

$$\lim_{n \rightarrow \infty} P(\psi_{kn} \leq \alpha - \varepsilon) = 0 \quad \forall \varepsilon > 0.$$

This concept of optimality is the one that we shall be using in § 5.

Before stating our main assumptions, some further notation is necessary. For $i \in S$, let

$$\tau(i) = \min \{n > 1 | X_n = i\}.$$

Let the chain be governed by an SSRS $\gamma[\Phi]$ and let f be a bounded map $S \rightarrow R$. Define $a[\Phi] = [a[\Phi](1), a[\Phi](2), \dots]$, $b[\Phi] = [b[\Phi](1), b[\Phi](2), \dots]$ and $g_f[\Phi] = [g_f[\Phi](1),$

$g_f[\Phi](2), \dots]$ as follows:

$$\begin{aligned} a[\Phi](i) &= E[\tau(1)/X_1 = i] - 1, \\ b[\Phi](i) &= E[\tau(i)/X_1 = 1] - 1, \\ g_f[\Phi](i) &= E\left[\sum_{m=1}^{\tau(i)-1} f(X_m) \middle/ X_1 = u\right]. \end{aligned}$$

It is well known that $a[\Phi](i)$, $b[\Phi](i)$ and hence $g_f[\Phi](i)$ are finite. Let $C_f[\Phi]$ denote the expectation of f w.r.t. $\pi[\Phi]$. (This is consistent with our definition of $C_k[\Phi]$.) Define the column vector $V_f[\Phi] = [V_f[\Phi](1), V_f[\Phi](2), \dots]^T$ by $V_f[\Phi](i) = g_f[\Phi](i) - C_f[\Phi]a[\Phi](i)$, $i \in S$. If Φ is a point mass concentrated at $\xi \in L$, denote $a[\Phi]$, $b[\Phi]$, $g_f[\Phi]$, $C_f[\Phi]$, $V_f[\Phi]$ by $a\{\xi\}$, $b\{\xi\}$, $g_f\{\xi\}$, $C_f\{\xi\}$, $V_f\{\xi\}$ resp. These quantities can also be defined for arbitrary $f: S \rightarrow R$ as long as $C_{|f|}[\Phi]$ (or $C_{|f|}\{\xi\}$) is finite. (See Lemma 3.1 below.) We shall use the convention that $C_f[\Phi] = f(\infty)$ ($C_f\{\xi\} = f(\infty)$) when $\gamma[\Phi]$ ($\gamma\{\xi\}$) is not an SSRS (SSS) and $f(\infty) = \lim_{u \rightarrow \infty} f(u)$ exists and is finite.

Our main assumptions are as follows.

A1. For all $i \in S$, there exists a finite subset R_i of S such that $p(i, l, \cdot) = 0$ for all $l \notin R_i$.

A2. For each finite subset A of S and integer M , there exists a finite integer N such that the minimum path length from i to any state in A exceeds M under any SRS whenever $i \geq N$.

Remarks. (i) By A1, all conditional expectations of the form $E[f(X_{m+n})/X_m]$ or $E[f(X_{m+n})/G_m]$ for finite m, n , arbitrary $f: S \rightarrow R$, and arbitrary CS, can be written as finite sums of real random variables and thus make perfect sense even when $E[f(X_{m+n})]$ is either undefined or unbounded.

(ii) A standard assumption in the classical treatment of the minimum cost per unit time control problem is that for some $j \in S$,

$$\sup_{i \in S} E[\tau(j)/X_1 = i]$$

is uniformly bounded for all SSS. Many other assumptions arising in literature can be shown to be equivalent to this assumption (See [3], [7].) By A2,

$$\lim_{i \rightarrow \infty} E[\tau(j)/X_1 = i] = \infty,$$

violating the above condition. Thus this set-up does not fit the classical framework.

3. Preliminary results. Let $f: S \rightarrow R$. If $C_{|f|}[\Phi] < \infty$, we have the following two lemmas.

LEMMA 3.1. $|g_f[\Phi](i)| < \infty$ for $i \in S$ and $V_f[\Phi](1) = g_f[\Phi](1) - C_f[\Phi]a[\Phi](1) = 0$.

Proof. Fix $i \in S$. Let $\tau_1 = \min\{m | X_m = i\}$, $\tau_{2n} = \min\{m > \tau_{2n-1} | X_m = 1\}$, $\tau_{2n+1} = \min\{m > \tau_{2n} | X_m = i\}$, $n = 1, 2, \dots$. Then

$$(3.1) \quad 0 \leq \left[\left(\sum_{j=1}^n \sum_{m=\tau_{2j-1}}^{\tau_{2j}-1} |f(X_m)| \right) / n \right] / (\tau_{2n}/n) \leq \left(\sum_{m=1}^{\tau_{2n}} |f(X_m)| \right) / \tau_{2n}.$$

Letting $n \rightarrow \infty$,

$$0 \leq g_{|f|}[\Phi](i) / (a[\Phi](i) + b[\Phi](i)) \leq C_{|f|}[\Phi].$$

The first claim follows on noting that

$$|g_f[\Phi](i)| \leq g_{|f|}[\Phi](i).$$

For $i = 1$,

$$\lim_{n \rightarrow \infty} \left[\left(\sum_{j=1}^{n-1} \sum_{m=\tau_j}^{\tau_{j+1}-1} f(X_m) \right) / n \right] / (\tau_n/n) = \lim_{n \rightarrow \infty} \left(\sum_{m=1}^{\tau_n} f(X_m) \right) / \tau_n.$$

Hence $g_f[\Phi](1)/a[\Phi](1) = C_f[\Phi]$. Q.E.D.

Let 1_c , U , Q_f denote, respectively, the infinite column vector with all elements equal to 1, the infinite-dimensional identity matrix and the infinite column vector whose i th element is $f(i)$.

LEMMA 3.2. For each SSRS $\gamma[\Phi]$, $V_f[\Phi]$ satisfies

$$(3.2) \quad C_f[\Phi]1_c = (P[\Phi] - U)V_f[\Phi] + Q_f.$$

Moreover, any solution W of (3.2) satisfying $|W(i) - V_f[\Phi](i)| \leq K$ for some $K < \infty$, differs from $V_f[\Phi]$ by at most a scalar multiple of 1_c . (Note that (3.2) is unaltered if $V_f[\Phi]$ is changed by a scalar multiple of 1_c .)

Proof. Consider the chain controlled by $\gamma[\Phi]$. Since $V_f[\Phi](1) = 0$, we have

$$\begin{aligned} V_f[\Phi](i) &= E \left[\sum_{m=1}^{\tau(1)-1} (f(X_m) - C_f[\Phi]) / X_1 = i \right] \\ &= f(i) - C_f[\Phi] + E \left[\sum_{m=2}^{\tau(1)-1} (f(X_m) - C_f[\Phi]) / X_1 = i \right] \\ &= f(i) - C_f[\Phi] + (P[\Phi])_i V_f[\Phi] \end{aligned}$$

for $i \in S$, where $(P[\Phi])_i$ is the i th row of $P[\Phi]$. The first claim follows. Suppose $W = [W(1), W(2), \dots]^T$ satisfies (3.2) and $Y = W - V_f[\Phi]$ satisfies: $\{\gamma(i)\}$ are uniformly bounded. Let $P^n[\Phi]$, $n = 1, 2, \dots$, be the n -times matrix product of $P[\Phi]$ with itself and $P^\infty[\Phi]$ the matrix each row of which is $[\pi[\Phi](1), \pi[\Phi](2), \dots]$. Then $P[\Phi]Y = Y = P^n[\Phi]Y$. Since the chain is positive recurrent,

$$\frac{1}{n} \sum_{m=1}^n P^m[\Phi] \rightarrow P^\infty[\Phi]$$

termwise. By Scheffe's theorem [1], the rows of the left-hand side converge to the corresponding rows of the right-hand side in l_1 . Therefore,

$$Y = \frac{1}{n} \sum_{m=1}^n P^m[\Phi]Y \rightarrow P^\infty[\Phi]Y.$$

Thus

$$Y(i) = \sum_{j \in S} \pi[\Phi](j) Y(j) \quad \text{for all } i \in S.$$

The claim follows. Q.E.D.

The rest of this section is devoted to the discussion of certain spaces of probability measures and random variables taking values in the same. From now on, we assume without any loss of generality that $D(i)$'s are replicas of a fixed compact metric space D . (If not, take $D = L$ and replace $p(i, j, \cdot) : D(i) \rightarrow [0, 1]$ by $p(i, j, pr_i(\cdot)) : D \rightarrow [0, 1]$ where $pr_i : D \rightarrow D(i)$ is the projection onto the i th coordinate space.) Let $\bar{S} = S \cup \{\infty\}$ be the one point compactification of S suitably metrized.

LEMMA 3.3. Each $\nu \in M(\bar{S} \times D)$ has a decomposition

$$\nu(A) = \delta_\nu \nu'(A \cap (S \times D)) + (1 - \delta_\nu) \nu''(A \cap (\{\infty\} \times D)),$$

for A Borel in $\bar{S} \times D$, where $\nu' \in M(S \times D)$, $\nu'' \in M(\{\infty\} \times D)$ and $0 \leq \delta_\nu \leq 1$. If we use

the convention that $\nu' = \nu_0$ ($\nu'' = \nu_1$) when $\delta_\nu = 0$ ($\delta_\nu = 1$) for arbitrarily fixed $\nu_0 \in M(S \times D)$ ($\nu_1 \in M(\{\infty\} \times D)$), this decomposition is unique.

The proof is trivial. For $\nu \in M(\bar{S} \times D)$, let $m_\nu \in M(S)$ be defined by

$$m_\nu(A) = \nu'(A \times D), \quad A \text{ Borel in } S. \quad \text{Q.E.D.}$$

LEMMA 3.4. *There exists a measurable map $h_\nu: S \rightarrow M(D)$ such that for all bounded continuous $f: S \times D \rightarrow R$,*

$$\int_{S \times D} f d\nu' = \int_S \int_D f(w_1, w_2) dh_\nu(w_1) dm_\nu$$

where w_1, w_2 are dummy variables of integration for the outer and the inner integral respectively.

Proof. Take h_ν to be a version of the appropriate regular conditional distribution. Q.E.D.

Let Φ_ν denote the product measure $\prod_{i \in S} h_\nu(i)$ on L . For each $n = 1, 2, \dots$, define

$$\mu'_n(A \times B) = \left(\sum_{m=1}^n I\{X_m \in A, \xi_m(X_m) \in B\} \right) / n$$

for $A \times B \in \mathcal{a} = \{A' \times B' | A', B' \text{ Borel sets in } \bar{S}, D \text{ resp.}\}$. For each sample path μ'_n is a probability measure on the field \mathcal{a} . Since \mathcal{a} generates the product σ -field of $\bar{S} \times D$, μ'_n has a unique extension $\mu_n \in M(\bar{S} \times D)$. Then μ_n, m_{μ_n} are simply the empirical measures for processes $\{(X_m, \xi_m(X_m))\}$ ($\{X_m\}$ resp.) at time n .

LEMMA 3.5. *For each sample path, $\{\mu_n\}$ converges to a sample path-dependent compact set in $M(\bar{S} \times D)$.*

Proof. By Prokhorov's theorem [1], $M(\bar{S} \times D)$ is compact. The claim follows easily. Q.E.D.

Denote by \bar{C}_b , (C_b resp.) the Banach space of bounded continuous maps $\bar{S} \rightarrow R$ and the Banach space of bounded maps $f: S \rightarrow R$ such that $f(\infty) = \lim_{i \rightarrow \infty} f(i)$ exists, each endowed with the supremum norm. Let $r: \bar{C}_b \rightarrow C_b$ map $f \in \bar{C}_b$ to its restriction to S . Then r is an isometric isomorphism between \bar{C}_b, C_b .

We are now ready to state the main result of this section.

LEMMA 3.6. *Suppose the chain is governed by an arbitrary CS. Then there exists a null set N such that for all sample paths outside N and each limit point ν of $\{\mu_n\}$ in $M(\bar{S} \times D)$ for which $\delta_\nu > 0$, $m_\nu = \pi[\Phi_\nu]$.*

Proof. Let \bar{G} be a countable dense set in \bar{C}_b . Then $G = r(\bar{G})$ is countable dense in C_b . For each $f \in \bar{G}$, the martingale stability theorem [4, p. 53] yields

$$(3.3) \quad \lim_{n \rightarrow \infty} \left(\sum_{m=2}^n (f(X_m) - E[f(X_m) | \mathcal{G}_{m-1}]) \right) / n = 0 \quad \text{a.s.}$$

Let N be the null set outside which (3.3) holds for all $f \in \bar{G}$. Consider a fixed sample path outside N . Define $f': \bar{S} \times D \rightarrow R$, $f'': \bar{S} \times D \rightarrow R$ by

$$f'(i, u) = f(i), \quad i \in \bar{S}, \quad u \in D,$$

$$f''(i, u) = \sum_{j \in S} p(i, j, u) f(j), \quad i \in S, \quad u \in D,$$

$$f''(\infty, u) = f(\infty).$$

Using A1 and A2, one can easily verify that f'' is bounded continuous. Note that

$$\left(\sum_{m=1}^n f(X_m) \right) / n = \int f' d\mu_n,$$

$$\left(\sum_{m=1}^n E[f(X_{m+1}) / G_m] \right) / n = \int f'' d\mu_n.$$

By (3.3), we conclude that

$$\int f' d\nu = \int f'' d\nu$$

for each limit point ν of $\{\mu_n\}$. If $\delta_\nu > 0$, we have

$$\sum_{i \in S} m_\nu(i) f(i) = \sum_{i \in S} m_\nu(i) \sum_{j \in S} E_{\Phi_\nu}[p(i, j, u_i)] f(j)$$

for all $f \in G$ and hence for all $f \in C_b$. (Here, $\{u_i\}$ are dummy variables of integration and $E_{\Phi_\nu}[\cdot]$ the expectation w.r.t. Φ_ν). Hence for all $i \in S$,

$$m_\nu(i) = \sum_{j \in S} m_\nu(j) E_{\Phi_\nu}[p(j, i, u_j)],$$

implying $m_\nu = \pi[\Phi_\nu]$. Q.E.D.

4. Existence result for almost monotone cost functions. Throughout this section, we assume that k is almost monotone.

Let $\{\Phi^n\}$ be a sequence of probability measures in $\bar{M}(L)$ such that $C_k[\Phi^n] \downarrow \beta$. For each n , let q_n be the unique element of $M(\bar{S} \times D)$ such that $\delta_{q_n} = 1$, $m_{q_n} = \pi[\Phi^n]$ and $\Phi_{q_n} = \Phi^n$.

LEMMA 4.1. *For any limit point ν of $\{q_n\}$ in $M(\bar{S} \times D)$, $\delta_\nu = 1$ and $C_k[\Phi_\nu] = \beta$.*

Proof. Define $k_m: S \rightarrow R$, $m = 1, 2, \dots$, by

$$k_m(i) = \begin{cases} k(i) & \text{if } i \leq m, \\ w(i)[(k(i) - \eta) \vee 0] - (\eta - k(0)) \vee 0 + \eta & \text{if } i > m, \end{cases}$$

where $\{w(i)\}$ are positive numbers in $(0, 1]$ satisfying $w(i) \downarrow 0$ and $w(i)k(i) \rightarrow 0$ as $i \rightarrow \infty$. Note that $k_m(i) \rightarrow \eta$ as $i \rightarrow \infty$ for each m , implying $k_m \in C_b$. Also, $k_m \uparrow k$ as $m \rightarrow \infty$. Define $\bar{k}_m: \bar{S} \times D \rightarrow R$, $m = 1, 2, \dots$, by $\bar{k}_m(i, u) = r^{-1}(k_m)(i)$, $i \in \bar{S}$, $u \in D$. Then \bar{k}_m is bounded continuous. For each m , $\int \bar{k}_m dq_n \rightarrow \int \bar{k}_m d\nu$ along a subsequence. But

$$\int \bar{k}_m dq_n = \int k_m dm_{q_n} \leq \int k dm_{q_n} = \int k d\pi[\Phi^n] = C_k[\Phi^n] \rightarrow \beta.$$

Hence $\int \bar{k}_m d\nu \leq \beta$. Since $\bar{k}_m(\infty, u) = \eta$ for all $u \in D$,

$$(4.1) \quad \int \bar{k}_m d\nu = \delta_\nu \int k_m dm_\nu + (1 - \delta_\nu)\eta.$$

Since $\eta > \beta$ and $\int \bar{k}_m d\nu \leq \beta$, we must have $\delta_\nu > 0$ and $\int k_m dm_\nu \leq \beta$. Let $f \in \bar{G}$ and f' , f'' as in the proof of Lemma 3.6. Then for $n = 1, 2, \dots$,

$$\int f' dq_n = \int f'' dq_n$$

by the definition of q_n and hence

$$\int f' d\nu = \int f'' d\nu.$$

Since $\delta_\nu > 0$, argue as in Lemma 3.6 to conclude that $m_\nu = \pi[\Phi_\nu]$. Recall that $\int k_m dm_\nu \leq \beta$. By the monotone convergence theorem,

$$\int k_m dm_\nu \uparrow \int k dm_\nu = \int k d\pi[\Phi_\nu] = C_k[\Phi_\nu]$$

as $m \rightarrow \infty$. Hence $C_k[\Phi_\nu] \leq \beta$. But $C_k[\Phi_\nu] \geq \beta$ by the definition of β . Hence $C_k[\Phi_\nu] = \beta$. From (4.1),

$$\beta \geq \delta_\nu \int k_m dm_\nu + (1 - \delta_\nu)\eta.$$

Since $\eta > \beta$ and $\int k_m dm_\nu \uparrow \beta$, we must have $\delta_\nu = 1$. Q.E.D.

LEMMA 4.2. *An optimal SSRS exists.*

Proof. Pick $\{\Phi^n\}$ as before and let ν be as in the preceding lemma. Then $C_k[\Phi_\nu] = \beta$. Consider $\{X_n\}$ governed by an arbitrary CS $\{\xi_n\}$. Let N be the null set in Lemma 3.6. Consider a fixed sample path outside N . Define \bar{k}_m, k_m as in the preceding lemma. Let $\bar{k}: \bar{S} \times D \rightarrow R$ be defined by $\bar{k}(i, u) = k(i)$ for $i \in S, u \in D$ and $\bar{k}(\infty, u) = \eta$ for $u \in D$. Then $\bar{k} \geq k_m$ termwise for each m . Let $\{\eta_l\}$ be a subsequence of $\{\eta\}$ such that $\psi_{k_{\eta_l}} \rightarrow \psi'$ for some ψ' . Let ϕ be a limit point of $\{\mu_{\eta_l}\}$ in $M(\bar{S} \times D)$. Then

$$\psi' = \lim_{l \rightarrow \infty} \int \bar{k} d\mu_{\eta_l} \geq \lim_{l \rightarrow \infty} \int \bar{k}_m d\mu_{\eta_l} = \int \bar{k}_m d\phi = \delta_\phi \int k_m dm_\phi + (1 - \delta_\phi)\eta.$$

Letting $m \rightarrow \infty$ on the right-hand side, we get

$$\psi' \geq \delta_\phi \int k dm_\phi + (1 - \delta_\phi)\eta.$$

But $\eta > \beta$ and whenever $\delta_\phi > 0$,

$$\int k dm_\phi = \int k d\pi[\Phi_\phi] = C_k[\Phi_\phi] \geq \beta.$$

Hence outside the null set N ,

$$\psi_{k_{\infty}} \geq \bar{\psi}_{k_{\infty}} \geq \beta = C_k[\Phi_\nu].$$

Thus $\gamma[\Phi_\nu]$ is an optimal SSRS. Q.E.D.

THEOREM 4.1. *An optimal SSS exists.*

Proof. Let $\gamma[\Phi]$ be an optimal SSRS. By A1, for each $i \in S$,

$$\sum_{j \in S} p(i, j, u) V_k[\Phi](j), \quad u \in D,$$

is a finite sum. By the usual compactness-continuity arguments, it attains its minimum w.r.t. u at some $\xi(i) \in D$. Let

$$\xi = [\xi(1), \xi(2), \dots].$$

Let $l(i)$ denote the i th element of $(P[\Phi] - U)V_k[\Phi]$. Consider the chain governed by $\gamma\{\xi\}$. Then for $n = 1, 2, \dots$,

$$\beta = l(X_n) + k(X_n) \geq \sum_{j \in S} p(X_n, j, \xi(X_n)) V_k[\Phi](j) - V_k[\Phi](X_n) + k(X_n).$$

Thus

$$\begin{aligned} \beta \geq & \frac{1}{n} E[V_k[\Phi](X_{n+1})/X_n] + \left(\sum_{m=2}^n (E[V_k[\Phi](X_m)/X_{m-1}] - V_k[\Phi](X_m)) \right) / n \\ & - V_k[\Phi](X_1)/n + \psi_{k_n}. \end{aligned}$$

Taking expectations conditioned on X_1 on both sides,

$$\beta \geq E[V_k[\Phi](X_{n+1})/X_1]/n - V_k[\Phi](X_1)/n + E[\psi_{kn}/X_1].$$

Since $\eta > \beta$, there exists an $\varepsilon > 0$ such that $k(i) \geq \beta + \varepsilon$ from some i onwards. Hence from A2 and the definition of $V_k[\Phi]$, it is easily verified that $V_k[\Phi](i) \rightarrow \infty$ as $i \rightarrow \infty$. Thus the set

$$B = \{i \in S \mid V_k[\Phi](i) < 0\}$$

is at most finite and hence $V_k[\Phi](X_{n+1})I\{X_{n+1} \in B\}$ is bounded uniformly. Now,

$$\beta \geq E[V_k[\Phi](X_{n+1})I\{X_{n+1} \in B\}/X_1]/n - V_k[\Phi](X_1)/n + E[\psi_{kn}/X_1].$$

By Fatou's lemma, we have

$$\beta \geq \limsup_{n \rightarrow \infty} E[\psi_{kn}/X_1] \geq E\left[\liminf_{n \rightarrow \infty} \psi_{kn}/X_1\right].$$

If $\gamma\{\xi\}$ is not an SSS, then $\{X_n\}$ is either null recurrent or transient. In either case, the following holds for every finite $F \subset S$:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{m=1}^n I\{X_m \in F\} = 0 \quad \text{a.s.}$$

From this, it is easily seen that

$$\liminf_{n \rightarrow \infty} \psi_{kn} \geq n \quad \text{a.s.,}$$

contradicting $\eta > \beta$. Hence $\gamma\{\xi\}$ is an SSS. Thus

$$\begin{aligned} \beta &\geq E\left[\liminf_{n \rightarrow \infty} \psi_{kn}/X_1\right] \geq E\left[\liminf_{n \rightarrow \infty} \left(\left(\sum_{m=1}^n (k(X_m) \wedge K)\right)/n\right)/X_1\right] \\ &= C_{k \wedge K}\{\xi\} \end{aligned}$$

where $K < \infty$ is arbitrary. Letting $K \uparrow \infty$, the monotone convergence theorem yields $\beta \geq C_k\{\xi\} \geq \alpha$. Thus $\alpha = \beta = C_k\{\xi\}$ and hence ξ is an optimal SSS. Q.E.D.

THEOREM 4.2. *Let $\gamma\{\xi\}$ be an optimal SSS. Then*

$$\alpha 1_c = (P\{\xi\} - U)V_k\{\xi\} + Q_k = \min_u (P_u - U)V_k\{\xi\} + Q_k$$

where the minimum is termwise. Conversely, any SSS $\gamma\{\xi\}$ satisfying the above is optimal.

Proof. Consider an optimal SSS $\gamma\{\xi\}$. Then the first equality above is immediate. Suppose the second equality is false. By virtue of the first equality, there exists a $\xi_0 = [\xi_0(1), \xi_0(2), \dots] \in L$, distinct from ξ , such that

$$\alpha 1_c = (P_{\xi_0} - U)V_k\{\xi\} + Y + Q_k$$

where $Y = [y(1), y(2), \dots]^T$ satisfies

$$\sup_i y(i) > 0, \quad \inf_i y(i) \geq 0.$$

Proceeding as in the proof of the preceding theorem, we can show that $\gamma\{\xi_0\}$ is an SSS and $\alpha \geq C_k\{\xi_0\} + C_Y\{\xi_0\} > C_k(\xi_0)$, contradicting the definition of α . Hence the second equality must hold. Conversely, let the above equalities hold for some SSS $\gamma\{\xi\}$. Consider the chain governed by $\gamma\{\xi\}$. Then for $n = 1, 2, \dots$,

$$\alpha = E[V_k\{\xi\}(X_{n+1})/X_n] - V_k\{\xi\}(X_n) + k(X_n).$$

As in the proof of the preceding theorem,

$$\alpha \geq E[V_k\{\xi\}(X_{n+1})I\{V_k\{\xi\}(X_{n+1}) < 0\}/X_1]/n - V_k\{\xi\}(X_1)/n + E[\psi_{kn}/X_1].$$

Letting $n \rightarrow \infty$, proceed as in the proof of Theorem 4.1 to conclude that $\alpha \geq C_k\{\xi\} \geq \alpha$. Q.E.D.

Remarks. We have proved that $\bar{\psi}_{k\infty} \geq \alpha = C_k\{\xi\}$ a.s. This is a stronger result than the one originally sought, viz. $\psi_{k\infty} \geq \alpha = C_k\{\xi\}$ a.s.

5. Existence result for the stable case. By "stable case" we mean that all SS are SSS and the set $E = \{\pi\{\xi\} | \xi \in L\}$ is tight. In addition, we assume that k is bounded. Let s denote the map $L \rightarrow M(S)$ mapping ξ into $\pi\{\xi\}$. For each bounded $f: S \rightarrow R$, let $e_f: L \rightarrow R$ be the map that maps ξ into $C_f\{\xi\}$.

LEMMA 5.1. *The maps s, e_f defined above are continuous and E is compact.*

Proof. Let $\xi_n \rightarrow \xi_\infty$ in L . By hypothesis, $\pi\{\xi_n\}$, $n = 1, 2, \dots$, is tight. Let π_∞ be a limit point of $\{\pi\{\xi_n\}, n = 1, 2, \dots\}$ in $M(S)$. Without any loss of generality, assume that $\pi\{\xi_n\} \rightarrow \pi_\infty$. By Scheffe's Theorem [1], $\pi\{\xi_n\} \rightarrow \pi_\infty$ in total variation. Note that $P\{\xi_n\} \rightarrow P\{\xi_\infty\}$ termwise. Since $\pi\{\xi_n\} = \pi\{\xi_n\}P\{\xi_n\}$, $n = 1, 2, \dots$, we have,

$$\pi_\infty - \pi_\infty P\{\xi_\infty\} = (\pi_\infty - \pi\{\xi_n\}) + (\pi\{\xi_n\} - \pi_\infty)P\{\xi_n\} + \pi_\infty(P\{\xi_n\} - P\{\xi_\infty\}).$$

From the foregoing, it follows that each of the three terms on the right tend to zero as $k \rightarrow \infty$. Hence $\pi_\infty = \pi\{\xi_\infty\}$. By [5, problem 9, p. 172], it follows that the map s is continuous. Continuity of e_f for $f: S \rightarrow R$ bounded, is immediate. E is a continuous image of a compact set and hence compact. Q.E.D.

COROLLARY 5.1. *For any bounded $f: S \rightarrow R$, there exists a $\xi_f \in L$ such that*

$$C_f\{\xi_f\} = \min_{\xi \in L} C_f\{\xi\} \triangleq \alpha_f.$$

Let h be any function in C_b such that $|h|$ is bounded by one. By the preceding corollary, there exists a $\xi_0 \in L$ such that

$$C_h\{\xi_0\} = \min_{\xi} C_h(\xi).$$

Clearly,

$$C_h\{\xi_0\}1_c = (P\{\xi_0\} - U)V_h\{\xi_0\} + Q_h.$$

LEMMA 5.2. *Let $\xi \in L$ satisfy $\xi(j) = \xi_0(j)$ for all $j \in S$ except $j = 1$. Consider the chain governed by $\gamma\{\xi\}$. Then*

$$E \left[\sum_{m=2}^{\tau(1)} (V_h\{\xi_0\}(X_m) - E[V_h\{\xi_0\}(X_m)/X_{m-1}]) / X_1 = 1 \right] = 0.$$

Proof. Let

$$M_1 = 0,$$

$$M_n = \sum_{m=2}^n (V_h\{\xi_0\}(X_m) - E[V_h\{\xi_0\}(X_m)/X_{m-1}]), \quad n = 2, 3, \dots$$

Consider the chain starting at $X_1 = 1$ with probability one. $\{M_n\}$ is a martingale with respect to the natural filtration of $\{X_n\}$. By optional sampling theorem, $E[M_{\tau(1) \wedge n}] = 0$

for each n . Thus

$$\begin{aligned} E[M_{\tau(1)}] &= E[M_{\tau(1)} - M_{\tau(1) \wedge n}] \\ &= V_h\{\xi_0\}(1) - E[V_h\{\xi_0\}(X_{\tau(1) \wedge n})] \\ &\quad - E\left[\sum_{m=\tau(1) \wedge n+1}^{\tau(1)} (E[V_h\{\xi_0\}(X_m)/X_{m-1}] - V_h\{\xi_0\}(X_{m-1}))\right]. \end{aligned}$$

Note that

$$\begin{aligned} &\left| \sum_{m=\tau(1) \wedge n+1}^{\tau(1)} (E[V_h\{\xi_0\}(X_m)/X_{m-1}] - V_h\{\xi_0\}(X_{m-1})) \right| \\ &\leq 2(\tau(1) - \tau(1) \wedge n) \downarrow 0 \quad \text{a.s. as } n \uparrow \infty. \end{aligned}$$

Also

$$E[V_h\{\xi_0\}(X_{\tau(1) \wedge n})] = V_h\{\xi_0\}(1)P(\tau(1) \leq n) + E[V_h\{\xi_0\}(X_n)I\{\tau(1) > n\}].$$

Thus it suffices to show that

$$\lim_{n \rightarrow \infty} E[V_h\{\xi_0\}(X_n)I\{\tau(1) > n\}] = 0.$$

From our choice of ξ , it can be easily verified that for $i \neq 1$,

$$V_h\{\xi_0\}(i) = E\left[\sum_{m=1}^{\tau(1)-1} (h(X_m) - C_h\{\xi_0\})/X_1 = i\right]$$

where the expectation is under $\gamma\{\xi\}$. Thus

$$\begin{aligned} |E[V_h\{\xi_0\}(X_n)I\{\tau(1) > n\}]| &= \left| E\left[\sum_{m=n}^{\tau(1)-1} (h(X_m) - C_h\{\xi_0\})I\{\tau(1) > n\}\right] \right| \\ &\leq 2E[\tau(1) - n | I\{\tau(1) > n\}] \rightarrow 0 \quad \text{as } n \uparrow \infty. \end{aligned} \quad \text{Q.E.D.}$$

LEMMA 5.3. $C_h\{\xi_0\}1_c = \min_u [(P_u - U)V_h\{\xi_0\} + Q_h]$, where the minimum is termwise.

Proof. Suppose not. Then there exist $i_0 \in S$, $u_0 \in D$ and $\delta > 0$ such that

$$C_h\{\xi_0\} - \delta = \sum_{j \in S} p(i_0, j, u_0) V_h\{\xi_0\}(j) - V_h\{\xi_0\}(i_0) + h(i_0).$$

Suppose $i_0 = 1$. Let $\xi \in L$ be such that $\xi(i) = \xi_0(i)$ for $i \neq 1$ and $\xi(1) = u_0$. Consider the chain starting at $X_1 = 1$ and governed by $\gamma\{\xi\}$. Let $\tau_0 = 1$ and τ_i denote the i th return time to 1 for $i = 1, 2, \dots$. Routine arguments show that for $n = 1, 2, \dots$

$$\begin{aligned} C_h\{\xi_0\} - n\delta/\tau_n &= E[V_h\{\xi_0\}(X_{\tau_{n+1}})/X_{\tau_n}] / \tau_n - V_h\{\xi_0\}(1)/\tau_n \\ &\quad - \left[\left(\sum_{m=2}^{\tau_n} (V_h\{\xi_0\}(X_m) - E[V_h\{\xi_0\}(X_m)/X_{m-1}]) \right) / n \right] / [\tau_n/n] \\ &\quad + \left[\left(\sum_{m=1}^{\tau_n} h(X_m) \right) / \tau_n \right]. \end{aligned}$$

Letting $n \uparrow \infty$ and noting that

$$\sum_{m=\tau_n+1}^{\tau_{n+1}} (V_h\{\xi_0\}(X_m) - E[V_h\{\xi_0\}(X_m)/X_{m-1}]), \quad n = 0, 1, 2, \dots$$

are i.i.d. by virtue of the strong Markov property, we get

$$\begin{aligned} C_h\{\xi_0\} - \delta / E[\tau(1)] \\ = -E\left[\sum_{m=2}^{\tau(1)} (V_h\{\xi_0\}(X_m) - E[V_h\{\xi_0\}(X_m)/X_{m-1}])\right] / E[\tau(1)] + C_h\{\xi\} \\ = C_h\{\xi\}. \end{aligned}$$

(The last equality follows from the preceding lemma.) Thus $C_f\{\xi_0\} > C_f\{\xi\}$, a contradiction. The claim follows for the case $i_0 = 1$. Suppose $i_0 \neq 1$. Define

$$\begin{aligned} V'_h\{\xi_0\} &= [V'_h\{\xi_0\}(1), V'_h\{\xi_0\}(2), \dots] \text{ by} \\ V'_h\{\xi_0\}(i) &= E\left[\sum_{m=1}^{\tau(i_0)-1} (h(X_m) - C_h\{\xi_0\}) / X_1 = i\right] \end{aligned}$$

where the expectation is according to $P\{\xi_0\}$. As in Lemma 3.2, one can show that

$$C_h\{\xi_0\}1_c = (P\{\xi_0\} - U)V'_h\{\xi_0\} + Q_h.$$

By the foregoing argument,

$$C_h\{\xi_0\}1_c = \min_u (P_u - U)V'_h\{\xi_0\} + Q_h.$$

By the second half of Lemma 3.2, the claim will follow if we show that

$$|V'_h\{\xi_0\}(i) - V_h\{\xi_0\}(i)| \leq K$$

for some $K < \infty$. But

$$\begin{aligned} |V'_h\{\xi_0\}(i) - V_h\{\xi_0\}(i)| &= \left| E\left[\sum_{m=1}^{\tau(i_0)-1} (h(X_m) - C_h\{\xi_0\}) \right. \right. \\ &\quad \left. \left. - \sum_{m=1}^{\tau(1)-1} (h(X_m) - C_h\{\xi_0\}) / X_1 = i\right] \right| \\ &= \left| E\left[\sum_{m=\tau(1) \wedge \tau(i_0)+1}^{\tau(1) \vee \tau(i_0)-1} (h(X_n) - C_h\{\xi_0\}) / X_1 = i\right] \right| \\ &\leq 2(E[\tau(1)/X_1 = i_0] + E[\tau(i_0)/X_1 = 1]) < \infty, \end{aligned}$$

where all the expectations are according to $P\{\xi_0\}$. The claim follows. Q.E.D.

The converse of this lemma (i.e., every $\xi_0 \in L$ satisfying

$$C_h\{\xi_0\}1_c = \min_u (P_u - U)V_h\{\xi_0\} + Q_h$$

satisfies $C_h\{\xi_0\} = \min_{\xi} C_h\{\xi\}$) can be proved in a similar way.

Now consider the chain governed by an arbitrary CS. Define $M(S)$ -valued random variables ν_n , $n = 1, 2, \dots$, by

$$\nu_n(A) = \frac{1}{n} \sum_{m=1}^n I\{X_m \in A\} (= m_{\nu_n}(A)),$$

for A Borel in S . Without any loss of generality, we may assume that the chain starts at a given state $i_0 \in S$ with probability one.

LEMMA 5.4. For each $\epsilon, \delta > 0$, there exists a finite integer N such that $P(\nu_n(A(N)) > \epsilon) < \delta$ for all n , where $A(N)$ denotes the subset $\{N+1, N+2, \dots\}$ of S .

Proof. Suppose not. Then there exist $\varepsilon, \delta > 0$ such that for all $N = 1, 2, \dots$, $P(\nu_n(A(N)) > \varepsilon) \geq \delta$ for infinitely many n . Since E is tight, we can pick N large enough so that $\pi\{\xi\}(A(N)) < \varepsilon\delta/2$ for all $\xi \in L$. Let $f(i) = I\{i > N\}$, $i \in S$. Pick $\xi_0 \in L$ such that

$$C_f\{\xi_0\} = \max_{\xi} C_f\{\xi\}.$$

Then the following holds by virtue of Lemma 5.3.

$$C_f\{\xi_0\} = \max_u (P_u - U) V_f\{\xi_0\} + Q_f = \pi\{\xi_0\}(A(N)).$$

Routine arguments show that

$$\begin{aligned} \frac{\varepsilon\delta}{2} > C_f\{\xi_0\} &\geq \limsup_{n \rightarrow \infty} E \left[\frac{1}{n} \sum_{m=1}^n f(X_m) \right] \\ &\geq \limsup_{n \rightarrow \infty} \varepsilon E[I\{\nu_n(A(N)) > \varepsilon\}] \geq \varepsilon\delta, \end{aligned}$$

a contradiction. The claim follows. Q.E.D.

LEMMA 5.5. Let B_1 be a set of maps $S \rightarrow R$ such that $|f|, f \in B_1$, has a common uniform bound $K < \infty$. For each $f \in B_1$, let $\phi_f: E \rightarrow R$ be the map that maps $\mu \in E$ to $\int f d\mu \in R$. Then $\{\phi_f | f \in B_1\}$ is an equicontinuous family.

Proof. Let $\varepsilon > 0$. Pick $N < \infty$ large enough so that

$$\sum_{i > N} \mu(i) < \varepsilon/4K \quad \text{for } \mu \in E.$$

(Since E is tight, such an N exists.) Let $d(\cdot, \cdot)$ be the Prokhorov metric on $M(S)$. From its definition, it follows that $d(\mu_1, \mu_2) < \varepsilon'$ for some $\varepsilon' > 0$ small enough will imply $|\mu_1(i) - \mu_2(i)| < \varepsilon'$ for $i \leq N$. Take $\varepsilon' < \varepsilon/4NK$. Then

$$|\phi_f(\mu_1) - \phi_f(\mu_2)| < \varepsilon.$$

The claim follows. Q.E.D.

COROLLARY 5.2. The map $C_b \rightarrow R$ mapping $f \in C_b$ to

$$\alpha_f \triangleq \min_{\xi} C_f\{\xi\} \in R$$

is continuous.

Proof. Let $f_n \rightarrow f_\infty$ in C_b . Then $\{\Phi_{f_n}, n = 1, 2, \dots, \infty\}$ is an equicontinuous family by the above lemma. Let $\varepsilon > 0$. Pick $\delta > 0$ such that $d(\mu_1, \mu_2) < \delta$ implies

$$|\Phi_{f_i}(\mu_1) - \Phi_{f_i}(\mu_2)| < \varepsilon/3, \quad i = 1, 2, \dots, \infty,$$

for $\mu_1, \mu_2 \in E$. Cover E by finitely many open d -spheres of radius δ and centers (say) $\mu_1, \mu_2, \dots, \mu_m$. Since $\int f_n d\mu \rightarrow \int f_\infty d\mu$ for $\mu \in E$, we can find a finite integer N such that

$$\left| \int f_n d\mu_i - \int f_\infty d\mu_i \right| < \varepsilon/3 \quad \text{for } 1 \leq i \leq m, \quad n \geq N.$$

It follows that

$$|\Phi_{f_n}(\mu) - \Phi_{f_\infty}(\mu)| < \varepsilon, \quad \mu \in E, \quad n \geq N,$$

implying

$$\limsup_{n \rightarrow \infty} \sup_{\mu \in E} |\Phi_{f_n}(\mu) - \Phi_{f_\infty}(\mu)| = 0.$$

Then, as $n \rightarrow \infty$,

$$|\alpha_{f_n} - \alpha_{f_\infty}| \leq \sup_{\mu \in E} |\Phi_{f_n}(\mu) - \Phi_{f_\infty}(\mu)| \rightarrow 0. \quad \text{Q.E.D.}$$

THEOREM 5.1. *A weak sense optimal SSS exists. Moreover, an SSS $\gamma\{\xi\}$ is weak sense optimal if and only if the following hold:*

$$C_k\{\xi\} = \min_u (P_u - U) V_k\{\xi\} + Q_k,$$

where the minimum is termwise.

Proof. Define $k_N \in C_b$, $N = 1, 2, \dots$, by

$$k_N(i) = k(i)I\{i \leq N\} + KI\{i > N\}$$

where $K \in (0, \infty)$ satisfies $|k(i)| \leq K$ for all $i \in S$. Choose ξ_0, ξ_N , $N = 1, 2, \dots$ such that $C_k\{\xi_0\} = \alpha$ and $C_{k_N}\{\xi_N\} = \alpha_{k_N}$. Let $\varepsilon, \delta > 0$. Consider the chain governed by an arbitrary CS. Choose N large enough so that

$$|\alpha_k - \alpha_{k_N}| < \varepsilon/2$$

and

$$P(\nu_n(A(N)) \geq \varepsilon') < \delta \quad \text{for all } n,$$

where ε' is chosen so that $0 < \varepsilon' < \varepsilon/8K$. Clearly,

$$|\psi_{kn} - \psi_{k_N n}| < \varepsilon/4$$

on the set $\{\nu_n(A(N)) < \varepsilon'\}$. Thus

$$\begin{aligned} P(\psi_{kn} \leq \alpha_k - \varepsilon) &\leq P(\psi_{kn} \leq \alpha_{k_N} - \varepsilon/2) \\ &\leq P(\psi_{kn} \leq \alpha_{k_N} - \varepsilon/2, \nu_n(A(N)) < \varepsilon') + P(\psi_{kn} \leq \alpha_{k_N} - \varepsilon/2, \nu_n(A(N)) \geq \varepsilon') \\ &\leq P(\psi_{k_N n} \leq \alpha_{k_N} - \varepsilon/4) + \delta. \end{aligned}$$

Recalling the remark following Theorem 4.2, we have $\liminf_{n \rightarrow \infty} \psi_{k_N n} \geq \alpha_{k_N}$ a.s. for all N . Hence, letting $n \rightarrow \infty$ in the above, we get $\limsup_{n \rightarrow \infty} P(\psi_{kn} \leq \alpha_k - \varepsilon) \leq \delta$. Since $\delta > 0$ was arbitrary,

$$\limsup_{n \rightarrow \infty} P(\psi_{kn} \leq \alpha_k - \varepsilon) = 0.$$

Hence $\gamma\{\xi\}$ is weak sense optimal. The second claim can be proved by arguments analogous to those used to prove Lemmas 5.2 and 5.3. Q.E.D.

6. The case when at least one SS is not an SSS. In this section, we assume that there exists at least one SS, say $\gamma\{\xi_0\}$, which is not an SSS, and also that $k \in C_b$.

THEOREM 6.1. *An optimal SS exists.*

Proof. If $\eta = k(\infty) > \beta$, the claim follows from Theorem 4.1. If $n \leq \beta$, then $C_k\{\xi_0\} = \eta$ implies $\eta = \beta$ and

$$C_k\{\xi_0\} = \min_{\xi} C_k\{\xi\} = \min_{\Phi} C_k[\Phi].$$

For $\xi \in L$, define $\hat{\pi}\{\xi\} \in M(\bar{S})$ by

$$\hat{\pi}\{\xi\}(A) = \begin{cases} \pi\{\xi\}(A \cap S) & \text{when } \gamma\{\xi\} \text{ is an SSS,} \\ I\{\infty \in A\} & \text{otherwise,} \end{cases}$$

for A Borel in \bar{S} . For an SRS $\gamma[\Phi]$, we can define $\hat{\pi}[\Phi]$ in an analogous manner. Let $F = \{\hat{\pi}\{\xi\} | \xi \in L\}$. We claim that for any SRS $\gamma[\Phi]$, $\hat{\pi}[\Phi]$ is in the closed convex hull of F . This is obvious when $\gamma[\Phi]$ is not an SSRS, because $\hat{\pi}[\Phi] = \hat{\pi}\{\xi_0\}$ for this case. Let $\gamma[\Phi]$ be an SSRS. Suppose the claim is false. Identifying F with a subset of l_1

(identify $\nu \in M(\bar{S})$ with $[\nu(\infty), \nu(1), \nu(2), \dots] \in l_1$), an application of the Hahn–Banach theorem shows that there exists a bounded $f: \bar{S} \rightarrow R$ such that

$$(6.1) \quad \int f d\hat{\pi}[\Phi] < \min_{\xi} \int f d\hat{\pi}\{\xi\}.$$

Without any loss of generality, we can take f to be nonnegative. Pick $K < \infty$ such that $f(i) \leq K$ for $i \in S \cup \{\infty\}$. Define f_n , $n = 1, 2, \dots$, by

$$f_n(i) = \begin{cases} f(i), & i = 1, 2, \dots, n, \\ K & \text{otherwise.} \end{cases}$$

Then by virtue of Theorem 4.1,

$$\int f_n d\hat{\pi}[\Phi] \geq \min_{\xi} \int f_n d\hat{\pi}\{\xi\}.$$

Letting $n \rightarrow \infty$, the dominated convergence theorem yields,

$$\int f d\hat{\pi}[\Phi] \geq \lim_{n \rightarrow \infty} \downarrow \min_{\xi} \int f_n d\hat{\pi}\{\xi\}.$$

It is easily verified that the right-hand side exceeds $\min_{\xi} \int f d\hat{\pi}\{\xi\}$. This contradicts (6.1), proving our claim. It now follows from Lemma 3.6 that under any CS, there exists a null set outside which every limit point of $\{\nu_n\}$ lies in the closed convex null of F . Recalling the observations at the beginning of this proof, the result follows. Q.E.D.

7. Some open problems. In conclusion, we list some unsolved problems related to the present work.

(i) Is the assumption that E is tight really necessary in § 5? It seems quite likely that the situation “all SS are SSS and E is not tight” does not occur at all.

(ii) Can “weak sense optimality” in Theorem 5.1 be strengthened to “optimality”? In the light of Theorem 6.1, the following may be true under the set-up of § 5: Outside a null set, all limit points of $\{\nu_n\}$ in $M(\bar{S})$ belong to the closed convex hull of E .

(iii) How do the recurrence properties of the Markov chain governed by an SSS $\gamma\{\xi\}$ depend on ξ ? A conjecture in this direction is as follows: There exist disjoint open sets O_1 and O_2 in L such that $O_1 \cup O_2$ is dense in L and the chain is positive recurrent, transient or null recurrent according to whether $\xi \in O_1$, $\xi \in O_2$ or $\xi \in (O_1 \cup O_2)^c$, respectively.

(iv) It would be interesting to extend these results to continuous time Markov chains, as these will have important applications to the control of queueing networks (see, e.g., [2]).

REFERENCES

- [1] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.
- [2] V. BORKAR, *Controlled Markov chains and stochastic networks*, this Journal, 21 (1983), pp. 652–666.
- [3] A. FEDERGRÜN, A. HORDIJK AND H. C. TIJMS, *A note on simultaneous recurrence conditions on a set of denumerable stochastic matrices*, J. Appl. Prob., 15 (1978), pp. 842–847.
- [4] M. LOEVE, *Probability Theory II*, 4th ed., Springer, New York, 1978.
- [5] J. R. MUNKRES, *Topology*, Prentice–Hall, Englewood Cliffs, NJ, 1975.
- [6] Z. ROSBERG, P. P. VARAIYA AND J. WALRAND, *Optimal control of service in tandem queues*, IEEE Trans. Automat. Control, AC-27 [1982], pp. 600–610.
- [7] S. ROSS, *Applied Probability Models with Optimization Applications*, Holden–Day, San Francisco, 1970.